2014 Electronic Imaging

SCIENCE AND TECHNOLOGY

2-6 February 2014

Technical Summaries

www.electronicimaging.org

Location

Hilton San Francisco, Union Square
San Francisco, California, USA

Conferences and Courses

2–6 February 2014

2014 Symposium Chair

Sergio R. Goma
Qualcomm Inc.

2014 Symposium Co-chair

Sheila S. Hemami
Northeastern University

2014 Short Course Chair

Choon-Woo Kim
Inha University

2014 Short Course Co-chair

Majid Rabbani
Eastman Kodak Co.

Contents

SPIE is the international society for optics and photonics, a not-for-profit organization founded in 1955 to advanced light-based technologies. The Society serves nearly 225,000 constituents from approximately 150 countries, offering conferences, continuing education, books, journals, and a digital library in support of interdisciplinary information exchange, professional growth, and patent precedent. SPIE provided over $3.3 million in support of education and outreach programs in 2012.

Click on the Conference Title to be sent to that page

Conference 9011: Stereoscopic Displays and Applications XXV

Monday - Wednesday 3 –5 February 2014

Part of Proceedings of SPIE Vol. 9011 Stereoscopic Displays and Applications XXV

9011-1, Session 1

Stereoscopic cell visualization: from mesoscopic to molecular scale

Björn Sommer, Christian Bender, Tobias Hoppe, Chrstian Gamroth, Lukas Jelonek, Univ. Bielefeld (Germany)

CONTEXT

Stereoscopic vision is a substantial aspect of three-dimensional visualization approaches. Although most recent animation movies created for cinemas are shown in stereoscopic 3D (S3D), there are still many areas which do not take advantage of this technology. One of these areas is cell visualization. Despite the fact that many protein crystallographers have preferred working with stereoscopic devices for over a decade, it is quite astonishing that cell visualization seems to have ignored S3D completely, even though stereoscopic visualization of the cellular cosmos not accessible to the human eye bears high potential.

Furthermore, the scientific community often works with interactive visualization environments. These tools usually provide S3D for different hardware configurations, but the intensity of the stereoscopic effect can only be manually adjusted by using slider buttons. This technique is sufficient to explore a single instance of a molecule, but it is inconvenient when navigating through a large environment on multiple scales.

OBJECTIVE

In this work approaches will be discussed to apply S3D to 1) rendered cell animations and 2) interactive cell environments by using freely available open source tools.

A very important aspect in cell visualization is the bridging of scales. The mesoscopic level starts at a few thousands of nanometers – related to the cell and its components – whereas the molecular level goes down to a few Angstrom, where single atoms are visible. Therefore, both scales may differ by a factor of 100,000. This is especially a problem if the stereoscopic effect should be adjusted during an interactive navigation process.

METHOD

For the rendered animations it will be shown how to use Blender in combination with Schneider's Stereoscopic Camera plug-in. An exemplary short movie was created, starting in the blood vessels, proceeding with the inner cell components and finally showing the translation and transcription process based on protein/PDB models.

The interactive exploration environments are provided by the CELLmicrocosmos project. On the molecular level, the MembraneEditor is used to show a fixed projection plane S3D method. The mesoscopic level is represented by CellExplorer which is equipped with a dynamic projection plane S3D method.

RESULTS

The stereoscopic cell animations rendered with Blender were successfully shown on notebook monitors and power walls as well as on large cinema projection screens. The CELLmicrocosmos projects were optimized to provide adequate interactive cell environments which were successfully used during different university projects and presentations. Because the software developer is not able to define the relative position of the user to the point of interest, the fixed projection plane S3D method was used in combination with smaller membrane structures. But the dynamic projection plane is furthermore compatible with cell environments featuring large scale differences.

NOVELTY

Cell visualization is an emerging area in scientific communication. This work should encourage cytological researchers to take S3D technology into account for future projects. Moreover, the stereoscopic capabilities of the CELLmicrocosmos project are shown which have been developed over several years and which have never been discussed in our previous publications.

9011-2, Session 1

The impact of stereo 3D sports TV broadcasts on user's depth perception and spatial presence experience

Kathrin Weigelt, Josef Wiemeyer, Technische Univ. Darmstadt (Germany)

CONTEXT

This work examines the impact of content and presentation parameters in 2D versus 3D on depth perception and spatial presence, so that guidelines for stereoscopic content development can be provided.

OBJECTIVE

Sport content potentially relevant to 3D broadcasts shows a great variety (e.g. basic conditions). Therefore, it is important to know which sports are appropriate for 3D TV broadcasts. 3D experience is supported by enhanced depth perception and spatial presence. The term 'spatial presence' denotes the subjective experience of 'being there', i.e., having the impression to be present in the mediated world ('self location') and to be able to interact with it ('possible actions'). The process model of spatial presence (PMSP; [1]) contends that two types of factors contribute to the experience of spatial presence: media and person factors. In our work, we analyse media content and presentation parameters as important factors contributing to depth perception and spatial presence.

METHOD

Based on a two level approach, a small-N study was performed. The study applied qualitative and quantitative approaches to assess spatial perception in 2D and 3D. In total, 8 participants performed stereo vision and vision acuity tests. After a test session, stereo and non-stereo video stimuli were presented, and spatial presence was assessed with the Measurements, Effects, Conditions Spatial Presence Questionnaire [2], and the Spatial Presence Experience Scale [3]. Depth perception was rated using a 5-point Likert scale according to the recommendations of the International Telecommunication Union [4].

With a repeated measures design, the impact of dimension (2D vs. 3D), content (soccer vs. boxing), scene layouts (camera position, field of view) were considered and analysed using ANOVAs with repeated measures on all factors. In line with the behavioural tests, participants had to recall spatial positions and motions of athletes. Qualitative interviews were conducted.

RESULTS

The study confirmed that 3D has an effect on user's depth perception and spatial presence experience compared to 2D. Effects of the camera distance could be shown on depth perception ($F_{1,7} = 12.94$, $p < .01$, partial $\eta^2 = .65$), self-location ($F_{1,7} = 16.32$, $p < .01$, partial $\eta^2 = .70$) and possible actions ($F_{1,7} = 8.99$, $p < .05$, partial $\eta^2 = .56$) in soccer. Moreover, a significant effect on self-location was found for the field of view ($F_{1,7} = 8.18$, $p < .05$, partial $\eta^2 = .54$). Several significant interactions could be revealed, but there were no further significant results for boxing or the comparison of sport disciplines.

NOVELTY

A big-N study will be implemented soon as a follow-up study. We expect to reproduce the results of the small-N study [5]. The novel contribution of the big-N study is the additional analysis of the direction of motion. Therefore, BMX-Miniramp will replace boxing as a sport discipline.

Moreover, the behavioural tests will be enhanced by the use of a three level approach that combines recall and recognition elements. The results of both, the big-N and the small-N study will be taken into account and will be transformed into guidelines for sports broadcasters on how to capture 3D sports for a 3D content development.

9011-83, Session Key1

Preservation and exhibition of historical 3D movies (provisional) *(Keynote Presentation)*

Jeff Joseph, Univ. of California, Los Angeles (United States)

No Abstract Available

9011-3, Session 2

A novel stereoscopic display technique with improved spatial and temporal properties

Paul V. Johnson, Joohwan Kim, Martin S. Banks, Univ. of California, Berkeley (United States)

CONTEXT

Common stereoscopic 3D display techniques utilize either temporal interlacing or spatial interlacing to present different images to the two eyes, but both methods are prone to different types of perceptual artifacts that detract from viewing experience. We propose a novel technique that is a spatiotemporal hybrid of these two methods, and have confirmed using psychophysical experiments that it reduces the visibility of artifacts.

OBJECTIVE

Temporal interlacing alternates left- and right-eye views in time, and is prone to visible flicker, unsmooth motion appearance, and depth distortions. Depth distortion occurs due to the Mach-Dvorak effect, whereby a temporal delay to one eye's input causes a horizontally moving object to appear displaced in depth. Spatial interlacing, on the other hand, sends even pixel rows to one eye and odd pixel rows to the other eye, resulting in lower effective spatial resolution under certain viewing conditions. In our proposed hybrid protocol, the left- and right-eye views are interlaced spatially, but the rows corresponding to each eye alternate every frame. We predict that this novel "hybrid" display protocol will combine the better properties of temporal and spatial interlacing. In this study we compare spatial, temporal, and hybrid interlacing in terms of spatial resolution, flicker, motion artifacts, and depth distortion.

METHOD

Using a mirror stereoscope we simulated spatial, temporal, and hybrid interlacing on a single setup. We measured motion artifacts by presenting a moving stimulus on a dark background: we found the stimulus speed above which motion artifacts– edge banding, motion blur, or judder– were perceived. We determined the critical flicker frequency for each display protocol by presenting a simple bright stimulus on a dark background at several different frame rates and having the viewer report whether they perceived visible fluctuations in luminance. We measured depth distortion by presenting two stimuli moving horizontally in opposite directions and having users adjust the disparity until they appeared at the same depth. We determined effective spatial resolution with a "tumbling E" task where viewers indicate the orientation of a letter E presented stereoscopically with several different sizes. Several viewing distances were used so that we could better ascertain the limitations of the display regardless of individuals' visual acuity.

RESULTS

Depth distortion, motion artifacts, and flicker were significantly reduced with the hybrid protocol compared to temporal interlacing. The effective spatial resolution was improved compared to spatial interlacing. The results suggest that hybrid interlacing combines the best qualities of both spatial and temporal interlacing and thereby provides a more realistic perceptual experience.

NOVELTY

To the best of our knowledge, this hybrid interlacing technique has not been characterized. We demonstrate using psychophysical experiments that it has significant improvements over existing methods and should provide a better viewing experience.

9011-4, Session 2

Frameless multiview display modules employing flat-panel displays for a large-screen autostereoscopic display

Kenji Hirabayashi, Masayuki Tokoro, Yasuhiro Takaki, Tokyo Univ. of Agriculture and Technology (Japan)

A large-screen autostereoscopic display enables life-size realistic communication. Its screen size should be more than 100 in. to cover one wall of a room, and it should support multiple viewers. A multiple projection system consisting of hundreds of projectors has been developed to construct large-screen autostereoscopic displays. The multiple projection system requires a long projection length; thus, a large space is required and installation is time consuming. Moreover, a dark room environment is required. In this study, we propose the tiling of frameless multi-view display modules employing flat-panel displays. This system requires less space and cost, and does not require a dark room. Moreover, the tiled screen can be configured in various ways such as landscape, portrait, and curved screens.

The flat-panel multi-view display consists of a flat-panel display and a lenticular lens. When the flat-panel multi-view displays are tiled, bezels of the flat-panel displays generate horizontal and vertical black bars in the tiled screen. In this study, an imaging system is combined with a flat-panel multi-view display to construct a frameless multi-view display module. The proposed module consists of a flat-panel multi-view display, an imaging lens, an aperture, a screen lens, and a vertical diffuser. The imaging lens projects the screen of the multi-view display on the vertical diffuser. When the magnification of the imaging lens is larger than one, a frameless screen is achieved. The aperture is placed on the plane where the viewpoints of the multi-view display are generated to eliminate repeated viewpoints. The aperture also limits rays in the vertical direction to reduce image degradation caused by aberration. The combination of the imaging lens and the screen lens projects the viewpoints of the multi-view display in the observation space to generate viewpoints for observers. The vertical diffuser increases the vertical viewing zone.

When the multi-view display modules are tiled, the screen lens is properly shifted in each module to produce a common viewing area for all modules. The common viewing area can also be produced by shifting the aperture in each module or by rotating the modules.

A liquid-crystal display with a resolution of 3,840 ? 2,400 was combined with a slanted lenticular lens to construct the multi-view display. The imaging lens consisted of two Fresnel lenses to decrease aberration. A Fresnel lens was also used as the screen lens. A lenticular lens was used as the vertical diffuser. The screen size of the multi-view display module was 27.3 in. and the module depth was 1.5 m. The 3D resolution was 320 ? 200 and the number of viewpoints was 144. The viewpoints were generated with a horizontal width of 2.4 m at a distance of 5.1 m from the screen. The horizontal pitch of the viewpoints was 16 mm.

A prototype display system with a medium-sized screen was constructed using four modules. These modules were vertically aligned to obtain a screen size of 62.4 in. (589 mm ? 1,472 mm). The prototype system can display almost human-size objects.

9011-5, Session 2

Interpolating vertical parallax for an autostereoscopic 3D projector array

Andrew V. Jones, Koki Nagano, USC Institute for Creative Technologies (United States); Jing Liu, Univ. of California, Santa Cruz (United States); Jay Busch, Xueming Yu, Mark T. Bolas, Paul Debevec, USC Institute for Creative Technologies (United States)

CONTEXT:

We present a technique for achieving tracked vertical parallax for multiple users for a variety of autostereoscopic projector array setups including front- and rear- projection, and curved display surfaces. This “hybrid parallax” approach allows for immediate horizontal parallax as viewers move left and right, and tracked parallax as they move up and down, allowing cues such as 3D perspective and eye contact to be conveyed faithfully.

OBJECTIVE:

Projector arrays are well suited for 3D displays because of their ability to generate dense and steerable arrangements of pixels. We have developed a new autostereoscopic display utilizing a single dense row of 69 pico projectors. The projectors are focused on a 30x30cm vertically anisotropic screen that scatters the light from each lens into a vertical stripe while preserving horizontal angular variation. Each viewer’s eye observes the combined effect of image stripes from multiple projectors which combine to form a seamless 3D image. As every viewer sees a different 3D image, it is possible to customize each view with a different vertical perspective. Given a sparse set of tracked viewer positions, the challenge is to create a continuous estimate of viewer height and distance for all potential viewing angles to provide consistent vertical perspective to both tracked and untracked viewers.

METHOD:

Rendering to a dense projector display requires multiple-center of projection imagery, as adjacent projector pixels diverge to different viewer positions. If you assume constant viewer height and distance for each projector, viewers may see significant cross-talk and geometric distortion particularly when multiple viewers are in close proximity. We solve this problem with a custom GPU vertex shader projection that dynamically interpolates multiple viewer heights and distances within each projector frame. Thus, each projector’s image is rendered in a distorted manner representing multiple centers of projection, and might show an object from above on the left and from below on the right.

RESULTS:

We use a low-cost RGB depth sensor to simultaneously track multiple viewer head positions in 3D and interactively update the imagery sent to the array. Even though each user sees slices of multiple projectors, the perceived 3D image is consistent and smooth from any vantage point with reduced cross-talk. This rendering framework also frees us to explore different projector configurations including front and rear-mounted projector arrays and non-flat screens. Our rendering algorithm does not add significant overhead enabling realistic dynamic scenes. Our display produces full color autostereoscopic 3D imagery, with zero horizontal latency, and a wide 110º field of view which can accommodate numerous viewers.

NOVELTY:

While user tracking has long been used for single-user glasses displays, and single-user autosteroscopic display [Perlin et al. 2000] in order to update both horizontal and vertical parallax,

our system is the first autostereoscopic projector array to incorporate tracking for vertical parallax. Our method could be adapted to other projector arrays [Rodriguez et al. 2007, Kawakita et al 2012, Kovacs and Zilly 2012, Yoshida et al 2011]. Furthermore, our display is reproducible with off-the-shelf projectors, screen materials, graphics cards, and video splitters.

9011-6, Session 2

Vertical parallax added tabletop-type 360-degee three-dimensional display

Yasuhiro Takaki, Junya Nakamura, Tokyo Univ. of Agriculture and Technology (Japan)

We previously proposed a 360-degree three-dimensional (3D) table-screen display using a small array of high-speed projectors, which provides 3D images on a tabletop that can be viewed by multiple viewers from all directions. This system combined the multi-projector system and the high-speed projector system that were proposed prior to the development of this system. The proposed system requires fewer projectors than the former system and provides more views and colors than the latter system. However, the proposed system provides only horizontal parallax. Therefore, the 3D images become distorted depending on the vertical viewing positions. In this study, we present a technique to add vertical parallax to the proposed system.

The small array of the high-speed projector system comprises a small number of high-speed projectors and a rotating screen. To enable the use of multiple projectors, the lens shift technique is used to superimpose all images generated by the projectors onto the rotating screen. Because the rotating screen has a lens function, the image of a projection lens is generated in the space, which becomes a viewpoint. Because the lens axis of the screen lens is shifted from the rotation center, the rotation of the screen generates a number of viewpoints on a circle around the rotating screen. Using multiple projectors enables us to increase the number of colors and viewpoints and reduce the screen rotation speed. Because all projectors generate the viewpoints on circles at the same height, the system can provide only horizontal parallax. A vertically diffusing function was added to the rotating screen to increase the vertical viewing zone.

In this study, all projectors are located at different heights from the screen to provide the vertical parallax. Because the heights of the viewpoints produced by the screen lens depend on the height of the projector, different projectors generate the viewpoints at different heights. Therefore, multiple viewpoints are aligned in the vertical direction so that the vertical parallax is obtained. When parallax images are properly generated to correspond to the positions of the viewpoints, 3D images with both horizontal and vertical parallaxes are generated. The extent of vertical diffusion by the rotating screen should be as large as the vertical pitch of the viewpoints.

The proposed technique was experimentally verified. Three DMD projectors were used to generate three viewpoints in the vertical direction. The heights of the projectors were 780, 840, and 900 mm. The corresponding heights of the viewpoints were 821, 764, and 720 mm, since the focal length of the screen lens was 400 mm. The frame rate of the projectors was 22.222 kHz. The projectors employed RGB LEDs to generate color images in a time-sequential manner. Each projector generated 900 viewpoints around the screen. The frame rate of the 3D image generation was 24.7 Hz, and the rotation speed of the screen was 1,481 rpm. The diameter of the rotating screen was 300 mm. The generation of 3D images with horizontal and vertical parallaxes was verified.

9011-80, Session 2

A variable-collimation display system

Robert Batchko, Sam Robinson, Holochip Corp. (United States); Benito Graniela, Naval Air Warfare Ctr. Training Systems Div. (United States)

CONTEXT

Two important human depth cues are accommodation and vergence. Normally, the eyes accommodate and converge/diverge in tandem; changes in viewing distance cause the eyes to simultaneously adjust

focus and orientation. However, ambiguity of these cues is a well-known limitation in many stereoscopic display technologies. This limitation also arises in state-of-the-art full-flight simulator displays wherein the lack of accurate short-range accommodation and vergence cues hinders the training of key operations such as take-off and landing. Hence, the development of a wide field-of-view autostereoscopic display which provides accurate accommodation and vergence from 3m to infinity could play a major role in improving pilot safety and training, and impact numerous other display applications currently limited by a lack of accurate depth cues.

OBJECTIVE

In current full?flight simulators, the out-the-window (OTW) display employs collimated cross-cockpit ("fixed-collimation") technology. Fixed collimation displays allow the pilot and copilot to perceive identical OTW imagery without angular errors or distortions; however, accommodation and vergence cues are limited to only long-range distances (> 18m). While this approach works well for long-range imagery, the ambiguity of depth cues at shorter range hinders the pilot's ability to gauge distances in critical maneuvers such as take-off and landing, posing a serious limitation. Here, we present the first in a series of papers on the development of a novel, 3D flight-simulator variable-collimation display (VCD) technology by Holochip Corporation under NAVY SBIR Topic N121-041. The VCD utilizes an adaptive-lens based collimated display architecture and is designed for integration into rotary-wing and vertical take-off and landing full-flight simulators. By providing accurate depth cues for viewing distances ranging from 3m to infinity, the VCD stands to greatly improve pilot training and safety.

METHOD

Conventional fixed-collimation displays consist of a bank of display projectors, illuminating a large curved rear-projection (RP) screen. The projectors and RP screen are mounted above the cockpit of the flight simulator. A very large parabolic mirror is positioned outside of the front window of the cockpit, filling the field-of-view of the pilots. The RP screen is reimaged by the mirror such that the pilots' eyes are inside the exit pupil of the system. The image light rays are generally collimated, making the image appear at long-range from the pilots. Holochip's approach modifies the fixed-collimation display with adaptive lens technology. This enables control of the image location (hence, degree of collimation) such that the image can be rapidly moved from short to long distances from the pilots, thereby providing accommodation cues. Further, the exit pupil encompasses both of the pilot's eyes, thereby providing vergence cues.

RESULTS

Simulation and test results, including image-plane display range, response time, image quality, brightness and resolution will be presented.

NOVELTY

Holochip's VCD system is the first simulator display technology to provide accurate accommodation and vergence cues for image distances ranging from 3m to infinity. With these dynamic depth cues, next-generation full-flight simulators enabled with VCDs will offer more realistic imagery than previously available, expanding the scope of simulated environments and improving the efficiency and quality of pilot training.

9011-7, Session 3

Subjective evaluation of a 3D video conferencing system

Hadi Rizek, Acreo Swedish ICT AB (Sweden); Kjell E. Brunnström, Kun Wang, Acreo Swedish ICT AB (Sweden) and Mid Sweden Univ. (Sweden); Börje Andrén, Acreo Swedish ICT AB (Sweden); Mathias Johanson, Alkit Communications AB (Sweden)

CONTEXT

Although video conferencing and telemeetings have advanced considerably recently, it is still physical meetings that are the most dominant type of meetings. Furthermore, the more important the meeting, the more likely it is to be a physical meeting. One reason for this is the lack of immersive experience with current video conferencing systems. One way to enhance the experience would be to incorporate stereoscopic presentation and in this case autostereoscopic 3D, since glasses will not be an option for video conferencing. Alkit Communications, Sweden, has developed a 3D video conference system using multiview autosterescopic screens1,2.

OBJECTIVE

The aim of this work was to evaluate a 3D video conferencing system with a similar 2D system. An additional objective was to evaluate how the current ITU standards3,4 for subjective testing of video conferencing could be applied to 3D video conferencing.

METHOD

The test was divided into two different sessions. Each test was carried out for a pair of subjects at the same time. In the first session, traditional 2D conferencing was presented to the subjects to compare it with the other methods involved in the test. The setup of the test was 2D-LCD display and a HD camera, see Figure 1(a). In the second session, an autostereoscopic screen presented the views from an 8-cameras setup, see Figure 1(b). The autosteroscopic screen was an eight view 47" LCD display from Alioscopy. The test subjects performed audio-visual tasks such as name guessing and free conversions as suggested in the ITU recommendations. After each task the test subject had to answer a number of questions, which they rated on a five-graded category scale mostly with the labels: Excellent, Good, Fair, Poor and Bad. For example:

- How would you rate the reality of the virtual representation of the other person?
- How would you rate the overall audio-visual quality?
- In what grade did you experience that the other party was present in the same room?
- How would you rate the quality of depth perception?
- How would you judge the effort needed to interrupt the other party?

In order evaluate the added value of the 3D depth information, two depth based task were developed and included in the test, see Figure 2.

RESULTS

Twenty-six test subjects participated in the test; they were aged between 19-37 years, 10 of whom were female and 16 were male. The subjects had different nationalities and different backgrounds. The results of the five questions presented above are shown in Figure 3. It can be noted that the 2D system outperforms the 3D system in every aspect, except on the perception of depth.

The results of the depth based tasks are shown in Figure 4. It clearly shows a statistically significant advantage for the stereoscopic 3D presentation.

One the post-questionnaire questions was whether the test person thought that 3D could bring added value to video conferencing. Figure 5 shows that a majority thought so, despite the result that 2D was better in almost every aspect.

NOVELTY

The main novelty of this study is the subjective evaluation of the a multiview stereoscopic 3D video conference system. The study introduces new methods to evaluate the added value of 3D for depth based tasks. The results show that currently the presentation is not good enough to compete with a good 2D system, but the test persons saw the potential in the 3D technology for this application. The depth based tasks showed that although the quality of the 3D system was poor in resolution, the viewing distance quite far, there were a big advantage with the stereoscopic 3D to perform the depth based tasks.

9011-8, Session 3

Subjective quality assessment for stereoscopic video: a case study on robust watermarking

Rania ben Said, Mihai P. Mitrea, Afef Chammem, Télécom SudParis (France); Touradj Ebrahimi, Ecole Polytechnique Fédérale de Lausanne (Switzerland)

State-of-the-art:

Regardless the final targeted application (compression, watermarking, depth-based special effects), protocols for evaluating the visual differences between the original and the processed stereoscopic video are required. In this respect, solutions directly inherited from 2D video are always considered, despite their inner limitations. First, they do not take into consideration the stereoscopic-related peculiarities of the HVS (human visual system). Second, they completely ignore the cultural differences between the 2D and stereoscopic video consumption in our society.

Paper main contribution:

The present paper goes one step further; it reconsiders the ITU-R BT.2021 (Subjective methods for the assessment of stereoscopic 3DTV systems) standard and investigates with statistical error control three of its key aspects: the minimal number of observers, the inter-genre variability of the results and the number of quality levels which can actually be correlated with the human visual system.

Testing conditions:

Image quality, depth perception and visual comfort are evaluated in laboratory conditions, at the ARTEMIS department. The general viewing conditions are set so as to meet the requirements expressed in ITU-R BT 500-12.

Two corpora are processed: 3DLive (summing up about two hours of HD 3D TV content captured by French professionals) and MPEG 3D video reference corpus (17 minutes, provided by both academic/industry, and encoded at different resolutions, from 320x192 to 640x480).

The processed video sequences are obtained by watermarking these two corpora with three types of methods (spread spectrum, side information and hybrid).

A panel of 60 observers (32 males and 28 females) was considered. Randomly selected sub-panels of 15, 20, 30, 40 and 50 observers were subsequently defined. All the subjects are screened for visual acuity using Snellen chart and color vision using the Ishihara test, cf. ITU-R BT 500-12.

At the beginning of the first session, from 2 to 5 training presentations are introduced to stabilize the observers' opinion. Each observer evaluates 17 randomly chosen video excerpts of 40 seconds each. These excerpts represent the two corpora and all the possibilities investigated in the experiments (original/original, watermarked/original, original/watermarked, watermarked/watermarked).

A DSCQS (double stimulus continuous quality scale) method has been adopted. The scoring is achieved by an Android application, running on a tablet; each of the three evaluation criteria (image quality, depth perception and visual comfort) have associate a sliding control; when the observers scores, the application record a value between 0 and 100; this value is a posteriori linearly mapped to the targeted number of quality levels (in our studies we successively considered 2, 3 and 5).

Statistical analysis:

The statistical investigation considers MOS (mean opinion score) estimation with 95% confidence limits, outliers detection, inner stationarity tests and paired t-Test (applied with both first and second type statistical error control).

9011-9, Session 3

Measuring perceived depth in natural images and study of its relation with monocular and binocular depth cues

Pierre Lebreton, Alexander Raake, Technische Univ. Berlin (Germany); Marcus Barkowsky, Patrick Le Callet, L'Univ. Nantes Angers le Mans (France) and Univ. de Nantes (France) and Institut de Recherche en Communications et en Cybernétique de Nantes (France)

CONTEXT:

The perception of depth in images and video sequences is based on different depth cues. Studies have considered depth perception threshold as a function of viewing distance (Cutting & Vishton,1995). The combination of different monocular depth cues and their quantitative relation with binocular depth cues have been studied (Landy, 1995). But none of these attempt to provide a quantitative contribution of monocular depth cues compared to each other in the particular case of natural images.

OBJECTIVE:

The objective of the study is to provide a quantitative evaluation of the strength of different depth cues compared to each other in the process of the construction of the perception of depth. And study how results on pooling strategies from the literature using artificial stimuli can be applied to the proposed database.

METHOD:

To perform this study, 200 different images were carefully selected. These images were selected from open image database or shot for the purpose of the experiment. The selection was done to fulfill the required characteristics: images should cover different combination of amount of monocular and binocular depth cues. To ensure this property, approximately 600 images were rated by two expert observers on 8 scales: linear perspective, relative size, texture gradient, defocus blur, areal perspective, interposition, light and shades and binocular depth. Based on their scores, the selection was done to separate the increase of one monocular depth cue compare to the binocular depth while maintaining the other depth cues to a level as low as possible. 8 set of 25 images was then established. The 25 images compose a "matrix" of 5 by 5 images corresponding to 5 level of monocular depth cue and 5 level of binocular depth cue. The 8 sets of images investigating the relationship between linear perspective, relative size, texture gradient and defocus blur compared to binocular depth. Two set of 25 images were defined per depth cue. These images were evaluated in two subjective experiments. One was conducted to evaluate the "global" perceived depth and the other targeted the evaluation of the seven previously stated monocular depth cues.

RESULTS:

From the data of the experiment, it was found that binocular depth cues have a strong effect on the overall depth perception. This is consistent to previous studies from the literature. A classification of the images in two categories: low- and high-value of each monocular depth cue shows a significant effect on the global score (Mann-Whitney U test ($p < 0.01$)). Except for the ``areal perspective" depth cue, results shows that increasing the level of monocular depth cue increases the global perceived depth. From the seven different monocular depth cues, texture gradient and areal perspective have been found to have the more influence on the global depth scores, followed by the relative size, the linear perspective, the interposition, the defocus blur and the light and shades.

NOVELTY:

The first novelty is to provide a study of the quantitative weighting of the importance of different monocular depth cues compare to each other in the particular context of natural images. The results are limited by the number of image considered, but the design for the images selection process considering as much variation of combinations of monocular

and binocular depth should reduce the dependency to the particular instantiation of the problem. The second contribution of this work is the release of the database to the scientific community.

9011-10, Session 3

Subjective evaluation of two stereoscopic imaging systems exploiting visual attention to improve 3D quality of experience

Philippe Hanhart, Touradj Ebrahimi, Ecole Polytechnique Fédérale de Lausanne (Switzerland)

CONTEXT

Crosstalk and vergence-accommodation conflicts negatively impacts the quality of experience (QoE) provided by stereoscopic displays. However, exploiting visual attention and on the fly adaptation of the 3D rendering process can reduce these drawbacks.

OBJECTIVE

Crosstalk, which is due to imperfect separation between the left and right views, is one of the most annoying distortions in stereoscopic displays [1]. This distortion reduces image and depth quality. When watching a stereoscopic display, the eyes converge to the location of the virtual object while the accommodation is always set for the screen surface. It is believed that this unnatural decoupling of vergence and accommodation increases visual discomfort [2].

To improve the QoE provided by stereoscopic displays, researchers have proposed to exploit visual attention [3,4,5,6]. Gaze tracking is performed to determine the virtual object fixated by the user. Then, either a hardware solution is used to control the display [3,4], or a software solution is used to adapt the rendering of the 3D content [5,6].

In particular, to reduce the vergence-accommodation conflict, the 3D content should be adjusted such that the fixated object lies on the screen plane. This can be achieved by performing view synthesis to generate a new stereo pair or simply by shifting horizontally the left and right views of the original stereo pair. By bringing the fixated object on the screen plane, which corresponds to the zero disparity plane (ZDP), perceived crosstalk is also reduced.

In this paper, we propose and evaluate two different solutions that exploit visual attention to improve the 3D QoE on stereoscopic displays: one system using an eye-tracker to determine the actual gaze location and one system using a 3D visual attention model to predict the gaze location.

METHOD

Two systems were developed based on the actual and predicted gaze location using a Smart Eye Pro remote eye tracking system and a 3D visual attention model [7], respectively. The gaze position was filtered taking into account the time required to fuse two images and the assumption that near objects are more salient. The filtered gaze position was used in conjunction with the disparity map to determine the horizontal shift necessary to bring the fixated object on the ZDP.

The user preference between standard 3D mode and the two proposed systems was evaluated in terms of image quality, depth quality, and visual discomfort through a subjective evaluation. Six stereoscopic video sequences with various characteristics were used: one for training and five for testing. The pair comparison methodology was selected. A total of 18 subjects took part in the experiment.

RESULTS

Results show that exploiting visual attention increases image and depth quality and decreases visual discomfort, especially when using real time gaze location.

NOVELTY

This work presents both hardware and software solutions that exploit visual attention to improve 3D QoE for stereoscopic displays, including their subjective evaluation and comparison to commercially available 3D imaging system.

REFERENCES

[1] P.J.H. Seuntiëns, L.M.J. Meesters, and W.A. IJsselsteijn, “Perceptual attributes of crosstalk in 3D images”, Displays, 26 (4–5), pp. 177-183, October 2005.

[2] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks, “Vergence–accommodation conflicts hinder visual performance and cause visual fatigue,” Journal of Vision, 8 (3), March 2008.

[3] S. Shiwa, K. Omura, and F. Kishino, “Proposal for a 3-D display with accommodative compensation: 3DDAC,” Journal of the Society for Information Display, 4 (4), pp. 255–261, December 1996.

[4] K. Talmi and J. Liu, “Eye and gaze tracking for visually controlled interactive stereoscopic displays,” Signal Processing: Image Communication, 14 (10), pp. 799-810, August 1999.

[5] E. Peli, T. R. Hedges, J. Tang, and D. Landmann, “A Binocular Stereoscopic Display System with Coupled Convergence and Accommodation Demands,” SID Symposium Digest of Technical Papers, 32 (1), pp. 1296–1299, June 2001.

[6] R. Yang and Z. Zhang, “Eye gaze correction with stereovision for video-teleconferencing,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 26 (7), pp. 956-960, July 2004.

[7] Y. Zhang, G. Jiang, M. Yu, and K. Chen, “Stereoscopic visual attention model for 3D video,” Proceedings of the 16th international conference on Advances in Multimedia Modeling, pp. 314-324, 2010.

9011-11, Session 3

Subjective quality and depth assessment in stereoscopic viewing of volume-rendered medical images

Johanna Rousson, Barco N.V. (Belgium); Jeanne Couturou, Télécom Saint-Etienne (France); Arnout Vetsuypens, Barco N.V. (Belgium); Ljiljana Platisa, Asli E. Kumcu, Univ. Gent (Belgium); Tom R. Kimpe, Barco N.V. (Belgium); Wilfried Philips, Univ. Gent (Belgium)

CONTEXT: 3D imaging has a proven value in minimally invasive and robotic surgery, and could further improve the detection of tumors in diagnostic applications. The benefit of using 3D over 2D images is the added depth information. For instance, depth information allows surgeons to perform suturing with higher accuracy and time efficiency as they get a more realistic spatial orientation and a better understanding of distances, positions and connectivities. Knowing the relationship between perceived depth (DP) and perceived image quality (IQ) is crucial for designing objective IQ metrics enabling automated and efficient evaluation of the quality of new medical 3D displays at every step of their development.

OBJECTIVE: We aimed at assessing the effects of common alterations occurring during the image acquisition stage or at the display side on IQ and DP, separately, on a stereoscopic medical image, to find out the relationship between IQ and DP. Additionally, we intended to investigate to what extent the undistorted view of one stereoscopic medical image could compensate for losses in terms of IQ and DP due to alterations present in the other view.

METHOD: The study was designed in consultation with surgeons. We conducted subjective experiments following a double stimulus five-point quality scale methodology with eight non-expert human observers. The reference stimulus was a volume-rendered stereoscopic image comprising a region of interest (ROI) with realistic colors and a complex internal vascular and bone structure. The stimuli were impaired versions of the reference stimulus altered by applying either single- or dual-view Gaussian blur filtering, adding white Gaussian noise, manipulating luminance, brightness and contrast, and adding crosstalk.

The stimuli comprised 992x960 pixels and were displayed on a 24inch

full HD stereoscopic surgical display comprising a patterned retarder. When assessing DP, observers were guided to concentrate on the relative positions and connectivity of different objects. To evaluate IQ, the observers were asked to focus on the sharpness, the color and the recognition of the different elements, mainly contained in the ROI.

RESULTS: Medians and first and third quartiles computed for all types and levels of impairments suggested that DP was very robust to luminance, contrast, and brightness alterations. Additionally, DP appeared even insensitive to noise distortions until standard deviation ?=20 and crosstalk rates of 7%. In contrast, IQ seemed sensitive to all distortions and especially to brightness, blur, noise, and crosstalk impairments. Finally, for both DP and IQ, the Friedman test indicated that the quality scores for dual-view distortions were significantly worse than scores for single-view distortions for multiple levels of blur and crosstalk impairments. No differences were found for most levels of brightness, contrast and noise distortions. So, DP and IQ didn't react equivalently to identical impairments, and both depended whether dual- or single-view distortions were applied.

NOVELTY: To the authors' knowledge, this is the first report of a study devoted to exploring the relationship between IQ and DP for a wide variety of impairments applied to stereoscopic medical datasets, and for both dual- and single-view distortions.

9011-12, Session 4

Interlopers 3D: experiences designing a stereoscopic game

James Weaver, Durham Univ. (United Kingdom); Nicolas S. Holliman, The Univ. of York (United Kingdom)

Novelty:

We present experiences designing and implementing a novel stereoscopic game, Interlopers 3D. The game is loosely based on the arcade favourite, Invaders, but Interlopers 3D has been designed so that depth is now a fundamental part of the game play.

Background:

Previous work by ourselves and others has demonstrated that stereoscopic 3D games, when designed to use the binocular depth cue, can result in better player performance in achieving game goals. Nevertheless there remain many unknowns about how best to integrate the binocular depth cue into games so that it benefits player performance and adds to the enjoyment of the game. This presents game designers with an opportunity to explore new binocular game mechanics and techniques that have previously been unavailable in monocular gaming.

Aims:

To investigate the visual cues that are present in binocular and monocular vision, identifying which are relevant when gaming on a stereoscopic display. To design a game whose mechanics are so reliant on binocular cues that the game becomes impossible in monocular mode. To evaluate player performance and to investigate whether players enjoyed gaming in stereoscopic 3D.

Method:

A stereoscopic 3D game was designed based on the "shoot-em-up" genre of Invaders style games. In Interlopers 3D the players objective is to shoot down Interlopers, that are advancing in depth, before they reach their destination. Scoring highly requires players to make accurate depth judgements and target the closest Interlopers first. The game was implemented in OpenGL, GLUT and C++ using an above-below stereoscopic output format compatible with current 3DTV systems, the software design adopted a Model-View-Controller software architecture.

Results:

A group of twenty participants played a basic and advanced version of the game in both monocular 2D and stereoscopic 3D, the results show that in both the basic and advanced game participants achieved significantly higher scores when playing in stereoscopic 3D. The advanced game also showed that by disrupting the depth from motion cue the game became more significantly more difficult in monoscopic 2D. In self-report feedback a number of players noted the game was more stressful and frantic when played in monoscopic 2D mode, a number of players also commented that the game was enjoyable in stereoscopic 3D mode.

Conclusions:

We have designed, implemented and tested a novel stereoscopic 3D game in the style of the Invaders genre of games. Although the game was not impossible to play in monoscopic 2D, participants' results show that it put them at a significant disadvantage when compared to playing in stereoscopic 3D. In addition players reported that the stereoscopic 3D mode was less difficult to play and more enjoyable in comparison to the monoscopic 2D mode.

9011-13, Session 4

Architecture for high performance stereoscopic game rendering on Android

Julien C. Flack, Hugh Sanderson, Dynamic Digital Depth Australia Pty. (Australia)

CONTEXT:

Stereoscopic gaming is a popular source of content for consumer 3D display systems. Traditionally PCs or gaming consoles such as the PlayStation 3 have been used to generate and render stereoscopic graphics. Recently there has been a significant shift in the gaming industry towards casual gaming for mobile devices designed on low cost ARM processors running the Android Operating System. Critically such ARM cores are now being integrated directly into the next generation of 3D TVs opening the way for gaming without the need for an external games console. Although native stereo support has been a feature of some high profile games on more mature platforms like PC and PS3 there is a lack of 3D support for the emerging Android platform.

OBJECTIVE:

In this paper we describe a framework for enabling stereoscopic 3D gaming on Android for applications on mobile devices, set top boxes and TVs. A core component of the architecture is a 3D game driver, which is integrated into the Android OpenGL ES graphics stack to convert existing 2D graphics applications into stereoscopic 3D in real-time. Several key innovations were required to realize this framework within the constraints of typical embedded ARM cores.

METHOD:

A high performance stereo renderer was developed in order to meet requirements for maintaining high frame rates (fps) without significant impact on battery life. Techniques traditionally used on PCs involve rendering multiple views into either separate buffers, or into separate viewports. For many embedded GPUs, these operations are prohibitively time consuming. In this paper we describe an innovative rendering technique to separate the views in the depth domain and render directly into the display buffer.

A significant proportion of popular casual games are effectively designed in 2D and rendered on to a 2D plane. Although such games are designed for 2D, excellent stereoscopic 3D effects can be achieved by separating the scene into layers and re-positioning these layers in 3D space. This 2D to 3D conversion process is not trivial due to the difficulty in identifying which layer graphics calls are associated with. We have developed a unique domain specific language that tracks information about the state of various layers to successfully assign correct depths to objects in the scene.

RESULTS:

We demonstrate the advantages of the stereo renderer by characterising the performance in comparison to more traditional techniques, including depth based image rendering, both in terms of frame rates and impact on battery consumption. We illustrate the analysis and processing of 2D layered games by showing how to automatically decompose a game into

layers which are positioned in 3D space.

The overall architecture of the system is described highlighting cross-platform SoC support and associated embedded graphics cores. Support for a variety of 3D display modes is detailed, including advanced autostereoscopic displays, which require the integration of eye tracking into the rendering core.

NOVELTY:

The paper describes a novel architecture for supporting OpenGL ES based 3D rendering on Android systems including a high performance rendering core and a heuristics engine to convert 2D games into 3D in real-time.

9011-14, Session 4

Comprehensive evaluation of latest 2D/3D monitors and comparison to a custom built 3D mirror based display in laparoscopic surgery

Ulrich Leiner, Lothar Mühlbach, Detlef Ruschin, Wolfgang Reiner, Fraunhofer-Institut für Nachrichtentechnik Heinrich-Hertz-Institut (Germany); Hubertus Feussner, Dirk Wilhelm, Silvano Reiser, Nils Kohn, Technische Univ. München (Germany); Michael Witte, Fraunhofer HHI (Germany)

CONTEXT

3D endoscopes combined with 3D displays did not achieve a breakthrough in laparoscopic surgery up to now. Though typical surgical tasks require spatial orientation and positioning, contrary to intuition several studies failed to establish significant performance advantage of 3D over 2D imaging. Rather than that, visual alterations, such as eye strain, diplopia and blur have been attributed to the use of stereoscopic displays. Advancements in technology suggested a re-evaluation of such findings.

OBJECTIVE

A large study was designed and conducted to investigate the question whether surgeons can benefit from using state-of-the-art 3D visualization systems. We also intended to establish a gold standard for 3D-displays and developed a mirror system that simulates a 3D display that is perfect in terms of delay, stereo effect and resolution. Furthermore we tested, whether experienced surgeons will profit in equal measure or less from 3D viewing, questioning the “novice benefit more than experts”-hypothesis.

METHOD

We conducted a randomized study on 48 individuals with different experience levels in laparoscopic surgery in which three different 3D displays were compared to a 2D display by assessing multiple performance and mental workload parameters during a laparoscopic suturing task. The 3D displays were an off-the-shelf 3D HD-screen with polarizing glasses and a prototype autostereoscopic display with head-tracking developed by Fraunhofer HHI.

A custom built system consisted of a mirror arrangement blocking the direct line of sight towards the workspace. This setup was used as a reference 3D display, due to its lack of latency, resolution losses and depth artefacts.

We measured the task complete time and the precision of stitching. Instrument tracking provided information on the movement path lengths and velocities. The NASA task load index was used for assessing the mental work load. Subjective user satisfaction was measured through rating scales.

RESULTS

All performance parameters were superior for the 3D HD glasses-based display as compared to the 2D and the autostereoscopic display, but were even exceeded by the mirror display. Subjects performed the task 20% faster and with a higher precision, with the 3D HD display or the mirror display and the instrument paths were shortened. Work-load parameters did not show significant differences between displays. Subjects complained about impaired vision while using the autostereoscopic monitor, mainly due to eye-tracking errors, while no visual discomfort was reported on the 3D HD display and the mirror display. Test results did not depend on the training level of the subjects, experts and novices improved their performance comparably.

NOVELTY

We achieved new and more precise results compared to previous studies, e.g. van Beurden09 regarding gains of 3D systems for surgical interventions. The invention of the 3D mirror system allowed for a benchmark tool for 3D visualization systems thus stimulating further research and development.

M van Beurden et al; Stereoscopic displays in medical domains: a review of perception and performance effects, Human Vision and Electronic Imaging XIV. Edited by Rogowitz, Bernice E.; Pappas, Thrasyvoulos N. Proceedings of the SPIE, Volume 7240 (2009), pp. 72400A-72400A-15 (2009).

9011-15, Session 4

A stereoscopic system for viewing the temporal evolution of brain activity clusters in response to linguistic stimuli

Kyle R Almryde, University of Arizona (United States); Angus Forbes, Javier Villegas, The Univ. of Arizona (United States); Elena Plante, University of Arizona (United States)

CONTEXT: The availability of fMRI data continues to open new paths of research in diverse disciplines. In the Department of Speech, Language, and Hearing Sciences at the University of Arizona, researchers analyze sequences of fMRI data in order to find multiple clusters of activation, related to language, that change over time in response to linguistic stimuli. This complicated dataset, containing both spatial and temporal statistics, creates a challenge for visualization designers who aim to create a system to assist in analysis tasks and to promote discussion among experts.

OBJECTIVE: In this paper, we present a novel application for the stereoscopic visualization of the fMRI data of subjects exposed to unfamiliar spoken languages. An analysis technique based on Independent Component Analysis (ICA) is used to identify statistically significant clusters of brain activity at different testing sessions. Our system is designed to illustrate the temporal evolution of the brain activity through displaying these clusters as they change over time.

METHOD: The fMRI raw data is presented as a stereoscopic pair in an immersive environment utilizing passive stereo rendering. For each point of view, different slices are textured and placed in the 3D scene at different distances from the camera. The temporal evolution of brain activity is superimposed atop the 3D image using different color-mappings for each component. User-controlled parameters available in real-time to the viewer include: changing the orientation, updating transparency of the brain structure and/or ICA clusters, and selecting particular time points in order to highlight interesting clustering behaviors.

RESULTS: Our system has been used by researchers in the Department of Speech, Language, and Hearing Sciences. We found that researches were easily able to identify patterns of interest as they manipulate the time parameters to show the evolution of brain activity in the visualization. Moreover, it promoted discussion and hypothesis-forming by encouraging researches to share opinions and by allowing them to physically pointing to regions in virtual space. Our application fully benefits from the extra dimension. Similar to the way that spatial information allows discrimination of voices in the cocktail party effect, the stereoscopic setup allows discrimination of individual events aligned in semi-transparent layers of brain data.

NOVELTY: Although commercial system for stereoscopic visualization

of fMRI data exists, our system was especially designed to take advantage of the mapping of depth. This mapping allows the viewer to simultaneously observe the time evolution of different regions of brain activity. Temporal changes associated with brain activity are often presented as one-dimensional signals that have to be observed independently from the 3D data. Our system incorporates the temporal information into the stereoscopic 3D rendering, making it easier for domain experts to integrate all the analysis information. Another contribution of our system is the simultaneous representation of selected components (resulting from ICA) as different colors that can blend when they intersect, illustrating components that are affecting the same region at the same time.

9011-16, Session 5

Fusion of Kinect depth data with trifocal disparity estimation for near real-time high quality depth maps generation

Guillaume Boisson, Paul Kerbiriou, Valter Drazic, Olivier Bureller, Neus Sabater, Arno Schubert, Technicolor S.A. (France)

CONTEXT:

Generating high quality depth maps in real-time for video streams would be extremely valuable in Cinema and Television production or post-production. Indeed, not to mention stereoscopic movies, depth information would be helpful for rendering visual effects for regular releases.

OBJECTIVE:

We address the problem of generating depth maps for professional – e.g. Cinema – purposes. To this end a rig consisting of a professional camera (RED ONE) flanked with two satellite HD cameras (Iconix), and a Kinect device on the top of it has been designed. We aim to combine inter-view matching and depth sensing to circumvent their mutual weaknesses and provide dense (HD), high quality depth maps on the flow together with the central view.

METHOD:

In this work a new two-pass, disparity-based calibration framework has been designed. After calibrating and registering left, central and right cameras, the Kinect signal is calibrated with respect to inter-view disparities instead of ground-truth. This makes depth values re-projected on central camera perfectly consistent with disparity estimation, so the fusion processing is straightforward. For the sake of speed, the fusion algorithm is hierarchical. At each resolution level, the disparity is estimated around the disparity estimated at the previous scale, by minimizing a global energy criterion. The proposed energy function has two terms: a data term weighted by its confidence and a consistency term encoding the coherence between the estimated disparity and the Kinect signal. Finally, at each stage, multilateral filtering is performed on depth maps, considering a global confidence measure in order to reject outliers. This is especially helpful to remove temporal artifacts.

RESULTS:

Our fusion approach generates fast, high-quality depth maps, on the flow with central camera view. Our CUDA implementation generates quarter-pel accurate HD720p depth maps at 15fps. The output is almost flawless, suitable for virtual view synthesis and 3D reconstruction. Thanks to the combination of disparity estimation with Kinect signal, depth maps are reliable both in textured and uniform areas. Furthermore, thanks to left and right satellite cameras, edges are sharp. Last, the delivered depth maps are temporally stable thanks to our post-filtering based on estimation reliability.

NOVELTY:

Multimodal fusion for depth mapping is widely addressed in the literature. Several papers deal with merging disparity estimation together with the use of a depth sensor (either a structured light device, like the Kinect, or a Time-of-Flight camera). The originality of our work is three-fold: the setup, the proposed algorithm, and the performances. There is no similar setup in the literature, with a heterogeneous rig consisting of a professional camera flanked by two compact satellite cameras. Regarding the approach, previous works propose to fuse depth estimations and depth measures a posteriori. On the contrary, our approach estimates disparity constrained by Kinect depth hierarchically, which, to the best of our knowledge, has never been proposed previously. Our

disparity-based calibration framework for the Kinect is novel as well. Finally, we achieve near real-time performances for state-of-the-art depth map quality, which makes our approach suitable for a lot of applications.

9011-17, Session 5

Depth map post-processing for depth-image-based rendering: a user study

Matej Nezveda, Nicole Brosch, Technische Univ. Wien (Austria); Florian H. Seitner, emotion3D (Austria); Margrit Gelautz, Technische Univ. Wien (Austria)

CONTEXT

Depth-Image-Based Rendering (DIBR) enables the generation of novel views/3D-content from color images and corresponding depth maps or stereo-derived disparity maps. The quality of the novel views and their underlying disparity map is often measured by comparison with 2D quality metrics or ground truth, respectively. However, objective quality metrics are not able to sufficiently determine the subjective visual quality of 3D-content [1]. Thus, the question of the subjective effectiveness of existing disparity map post-processing techniques in the context of DIBR/3D-content arises.

OBJECTIVE

The proposed paper examines the quality of disparity map post-processing techniques in the context of DIBR/3D-content. We conduct a user study which addresses (1) the effects of disparity map post-processing on the quality of stereo pairs that contain a novel view and (2) the question whether objective quality metrics are suitable for evaluating them.

METHOD

In order to investigate the effects of disparity map post-processing on the quality of stereo pairs that contain a novel view, a paired comparison user study with 18 participants is performed. We use six stereo pairs to generate their respective disparity maps. These image pairs contain difficult scenes for disparity map generation (e.g., with fuzzy object borders or thin vertical structures) that cause errors in the corresponding disparity maps. Before generating the novel views, six post-processing filters are applied on these disparity maps. For the main study, the best performing filters of a preliminary study with eight observers were selected. The finally selected filters also include the bilateral filter [2], the guided image filter [3] and the weighted mode filter [4]. Both the unprocessed and the post-processed disparity maps are used to generate the novel views. The original left views and the generated novel views form the 42 stereoscopic images that are used in our evaluations.

We additionally compare the results from the subjective study to the results obtained by twelve objective quality metrics [5]. These metrics include the peak-signal-to-noise ratio and the structural similarity index, which are commonly used to evaluate the quality of stereoscopic content [6,7].

RESULTS

Our user study shows that post-processing disparity maps enhances the perceived quality of stereo pairs that contain a novel view. In particular, edge-preserving filters, i.e., the bilateral filter and the guided image filter, achieve average subjective quality scores of 100 percent and 71 percent, respectively. These scores are more than twice as high as the average score of their unprocessed counterparts (i.e., 26 percent). We further observe that the employed objective quality metrics are not suitable to measure the visual quality of stereo pairs that contain a novel view. The latter observation is in agreement with [1].

NOVELTY

Various approaches for disparity post-processing have been proposed (e.g., [2-4,8]). However, the evaluations of these post-processing techniques focus on comparisons of the processed disparity maps to disparity ground truth. Contrary to these quantitative evaluations, the proposed study additionally considers human perception by performing a subjective quality assessment. Our evaluations further indicate that the correlation between subjective and objective quality results is weak.

(Please see the additional file for supplementary images and the reference list.)

9011-18, Session 5

Local disparity remapping to enhance depth quality of stereoscopic 3D images using stereoacuity function

Hosik Sohn, Yong Ju Jung, Yong Man Ro, KAIST (Korea, Republic of)

CONTEXT

Due to accommodation-vergence conflict in stereoscopic three-dimensional (S3D) contents and human's ability in binocular fusion, current S3D displays can provide high viewing quality within a limited depth range only. In addition, viewing quality of S3D content, which is targeted for a particular viewing environment, cannot be always maintained as the perceived depth range of the content varies with different display sizes and viewing distances. Given the limitations of the S3D display devices and the diversity of viewing environments, improving the viewing quality of S3D content is of great importance for both content creation and post production.

OBJECTIVE

Disparity remapping has been suggested as one of possible ways to improve the viewing quality of stereoscopic 3D content [1-3]. By adjusting disparities of a scene, it is possible to enlarge the perceived depth range and also relative depths between neighboring objects/regions in the scene. The simplest approach to enhance the depth sensation may be to scale the disparity range of a scene. However, it is not always applicable since there may be limits for the allowable depth range in S3D displays [1]. As such, given a limited depth range, enhancing the perceived relative depth is very important to enhance depth quality of a scene.

In this paper, we propose a local disparity remapping method to enhance depth quality of stereoscopic 3D image. In order to enlarge relative depths between objects (or regions) while preserving the original depth structure of a scene, the proposed approach decomposes a disparity map into a coarse and detail layer, where the coarse layer contains the information of depth structure and the detail layer contains the details of the depth structure (i.e., disparity gradient), respectively. Then, we adaptively manipulate the detail layer in depth and image spaces while the coarse layer is unchanged. In this way, we can effectively adjust relative depth between objects while providing spatial adaptability in the image space.

METHOD

In order to generate the coarse layer, we apply Gaussian smoothing to the input disparity map. The detail layer is generated by the subtraction of the coarse layer from the original disparity map. Then, the detail layer is locally scaled to emphasize relative depth information. Specifically, the detail layer is divided into multiple blocks and a scale factor for each block is determined to ensure that the relative depth between neighboring objects is perceivable. Note that, in this paper, we exploit a stereoacuity function [4] to guide the local disparity remapping. The stereoacuity function describes the minimum amount of perceivable disparity difference given disparity magnitudes. Lastly, a new disparity map is reconstructed by the addition of the processed detail layer to the coarse layer.

RESULTS

In order to evaluate the performance of the proposed approach, a subjective assessment of depth quality was conducted. We compared the subjective assessment results for the original and the processed images. The results showed that the proposed local disparity remapping method was capable of improving the depth quality of stereoscopic 3D images. Statistical analysis also showed that the improvement in the depth quality was significant.

NOVELTY

In this paper, we propose a simple but effective disparity remapping method to enhance depth quality of stereoscopic 3D images. The proposed local disparity remapping method is based on disparity map decomposition. The decomposition of disparity map enables us to manipulate the relative depth information of objects or regions in a scene while preserving their original depth structure. In addition, block-based processing of the detail layer allows for the adaptability in both depth and image spaces. The advantage of the proposed method over previous approaches [1].[5] is that it does not require any identification of objects (such as visual saliency estimation and object segmentation) and optimization process to manipulate the relative depth information. Thus, the proposed approach is more applicable for a real-time processing in S3D display devices.

Reference

[1] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross, "Nonlinear disparity mapping for stereoscopic 3D," ACM Trans. Graphics, vol. 29, no. 4, pp. 1-10, 2010.

[2] C.-H. Chang, C.-K. Liang, Y.-Y. Chuang, "Content-aware display adaptation and interactive editing for stereoscopic images," IEEE Trans. Multimedia, vol. 13, no. 4, pp. 589-601, August 2011.

[3] N. Holliman, "Mapping Perceived Depth to Regions of Interest in Stereoscopic Images," in Proc. SPIE, Stereoscopic displays and virtual reality systems XI, vol. 5291, 2004.

[4] C. M. Zaroff, M. Knutelska, and T. E. Frumkes, "Variation in stereoacuity: normative description, fixation disparity, and the roles of aging and gender," Invest. Ophth. Vis. Sci., vol. 44, pp. 891-900, Feb. 2003.

[5] S.-W. Jung and S.-J. Ko, "Depth sensation enhancement using the just noticeable depth difference," IEE Trans. Image Processing, vol. 21, no. 8, pp.3624-3637, 2012.

9011-19, Session 5

Efficient quality enhancement of disparity maps based on alpha matting

Nicole Brosch, Matej Nezveda, Margrit Gelautz, Technische Univ. Wien (Austria); Florian H. Seitner, emotion3D (Austria)

CONTEXT

In novel view generation for stereoscopic display/3D-content the quality of an underlying disparity map contributes to the quality of a synthesized view. Misalignment of disparity and color edges, inaccuracies due to mixed pixels and mismatches may lead to visual artifacts and object deformations.

OBJECTIVE

We propose an efficient post-processing method for disparity map and image enhancement. This enhancement focuses on (1) reducing the influence of mismatches in disparity maps, (2) aligning disparity and image edges even in the presence of mixed pixels, (3) sharpening image edges and (4) providing alpha values at disparity edges.

METHOD

Given two input stereo views and a derived disparity map, the proposed algorithm performs three steps: Segmentation, alpha matting and color and disparity re-assignment.

First, we segment the disparity map into disparity-layers. The addressed

errors typically occur near the border of these segments (i.e., disparity edges).

Second, we improve and apply a fast matting algorithm [1]. We introduce a confidence term that is based on color similarity of corresponding pixels in the stereo views, closeness to disparity edges and disparity and color coherence. Thus, the influence of erroneous disparity assignments is reduced. The improved matting algorithm is used to compute alpha values and new foreground colors for each segment. This step can be implemented efficiently (i.e., four milliseconds per disparity-layer with 620x360 pixels).

Third, we re-assign colors and disparities for each segment in ascending order of disparity. Since alpha values indicate the degree of affiliation to a segment versus all other segments, they can act as costs for disparity re-assignment. Thus, pixels with large alphas are assigned to the disparity of the current segment and pixels with low alphas are discarded. If, after processing all segments, a pixel in the new disparity map is not assigned to a disparity, it is assigned to the disparity at which it has the largest alpha value. The new disparity map has a better alignment of disparity and color edges. In the new color image, the combination of the new foreground colors of each segment, colors edges at disparity edges are sharper.

We additionally propose a temporal-coherent extension for videos.

RESULT

We quantitatively compare post-processed disparity maps with ground truth disparity and previously proposed post-processing techniques (including [1-5]), and report their respective runtimes. Since the objective quality of a disparity map not necessarily reflects the visual quality of its synthesized views, we show synthesized views that were generated with and without post-processing. We demonstrate that our method can improve the quality of disparity maps and synthesized views for images that contain fuzzy borders (e.g., fur).

NOVELTY

The main contributions of this paper are: (1) developing an efficient disparity map post-processing method that (2) aligns disparity and color edges even in the presence of mixed pixels, (3) sharpens color edges at disparity borders and (4) provides alpha values at disparity borders. In contrast to previous publications (e.g., [1-5]), the proposed method addresses mixed pixels and does not introduce mixed disparities that can lead to object deformations in synthesized views.

9011-20, Session 6

Description of a 3D display with motion parallax and direct interaction

Mark Flynn, Jerry Tu, zSpace (United States)

CONTEXT

The context is the creation of a stereoscopic display with low ghosting, motion parallax and direct interaction with the 3D scene presented.

OBJECTIVE:

In this abstract we describe the architecture of a new type of 3D display. It is our strong belief that all of the following attributes are required to provide a compelling end user experience.

1. Implementation of Motion Parallax. This provides acceptable viewing comfort and enables a very powerful 3D cue.

2. Full Resolution.

3. Direct interaction of with the scene, using a tracked stylus.

3. Very low ghosting.

4. Very low latency between the various components.

6. Lightweight and unpowered eyewear.

7. No perceptible flicker.

Based on these requirements, we designed and built the zSpace stereoscopic system described below.

METHOD:

The zSpace system consists of the following components.

- A stereoscopic display based on a time-sequential polarization switch.

- A head tracking system and direct interaction stylus based on Infra-Red camera sensors.

RESULTS:

The display has the components shown in Fig. 1, and listed below.

1. A segmented edge-lit LED backlight (BLU).

2. A full HD 24”? diagonal 120 Hz TN LCD.

3. A TN polarization switch (PS).

The PS functions as shown in Fig. 2. It takes as input the linearly polarized light from the LCD, and converts it to either left-handed circularly polarized (CP) light (LCP), or right handed CP (RCP), depending on whether a right or left image is being displayed on the LCD. These two states are then analyzed with CP eyewear, so that the left eye only sees the left image, and the right eye only sees the right image.

Tracking

In order to produce the effect of motion parallax, the 3D application must render the left and right images exactly where the left and right eyes are located at any instant in time. This requires accurate tracking of the position the user’s eyes and the direction they are pointing to. The zSpace system uses sophisticate hardware platform and software algorithms to measure the position and orientation (POSE) of the user head with high accuracy and low latency.

The zSpace Tracking Hardware System consist of Infra-red (IR) illuminating LEDs, IR reflectors, and IR image sensors (cameras). These are shown in Fig. 3. The IR LED illuminates the 5 IR reflectors affixed onto the user’s passive polarized glasses. The reflections from the 5 reflectors are captured by the two IR camera sensors.

The IR images are then processed by tracking algorithms in software. A correspondence algorithm identifies each of the five reflectors, then triangulation calculations produce the absolute x, y, and z locations. From these coordinates, we compute the cross products and obtain orientation measurements. The position and orientation are computed for each pair of images from the high speed camera sensors and sent to the 3D application so that it can properly implement motion parallax.

The system also includes a stylus which contains two inertial sensors: accelerometer and rate gyroscope. The raw inertial sensor measurements are fused with optical tracking information (via the IR cameras) in a Kalman state estimation algorithm. The output is the position and orientation of the stylus. This 6 degrees-of-freedom state for the stylus is sent to the 3D application to implement direct interaction, allowing the user to directly move, rotate, scale or otherwise manipulate virtual objects in the stereoscopic volume.

NOVELTY:

While various components described herein have been discussed before, we believe that this is the first time that they have all been integrated into a single compact system.

9011-21, Session 6

LCD masks for spatial augmented reality

Quinn Y. J. Smithwick, Daniel Reetz, Lanny S. Smoot, Walt Disney Imagineering (United States)

CONTEXT:

One aim of Spatial Augmented Reality (SAR)[1] is to visually integrate synthetic objects into real-world spaces amongst physical objects, viewable by many observers without encumbrances of 3D glasses, head-mounted goggles or mobile screens. In common implementations -- using beam-combiners, scrim projection, or near-future, transparent self-emissive displays -- the synthetic object’s and real-world scene’s light combine additively. As a result, the synthetic object appears low-contrast

and semi-transparent against well-lit backgrounds, and it does not cast shadows. These limitations prevent the synthetic object appearing solid and visually integrated into the real-world space.

OBJECTIVE

We use a transparent LCD panel as a programmable dynamic mask. The LCD panel displaying the synthetic object's silhouette mask is co-located with the object's color image, both staying aligned for all points-of-view. The mask blocks the background providing occlusion, presents a black level for high-contrast images, blocks scene illumination thus casting true shadows, and prevents blow-by in projection scrim arrangements.

METHOD

We have three implementations of SAR with LCD masks: 1) beam-combiner with an LCD mask, 2) scrim projection with an LCD mask, and 3) a transparent OLED with an LCD mask.

1) A 23" Viewsonic monitor, modified by removing its backlight, acts as a transparent LCD, a second 23" television displays the synthetic color object, and a half-silvered mylar mirror (Hudson Mirror, LLC) acts as a beam-combiner. The transparent LCD is at the reflected image plane of the color display. (Fig.1)

2) A modified 27" Dell monitor, acting as a transparent LCD mask, is covered by scrim material (RoseBrand), and projected onto using a NEC LCD projector. (Fig.3)

3) A layered 2" transparent OLED display (4D Systems) and a 4" transparent LCD module from a mini projector (FAVI Entertainment) are stacked together. (Fig.6)

Using Quartz Composer running on a Mac-mini with a Matrox Dual-Head-2-Go display expander, animated color imagery and corresponding silhouettes are displayed on a single large window spanning the two displays.

RESULTS

When implemented using the LCD mask (see figures 2b,4b,7b), the synthetic objects are opaque and high contrast (including black tones), as compared to low-contrast, semi-transparent synthetic objects without the mask (see figures 2a,4a,7a). Motion parallax and mutual occlusion/disocclusion between the synthetic character and the real-world is apparent in figures 5a,b.

CONCLUSIONS

The LCD mask provides occlusion, high-contrast imagery, and shadows over a wide field-of-view for many viewers, without requiring a single viewpoint, headtracking, nor a scene model as in programmable projection lighting [2][3]. Mask and image are colocated, staying aligned for all points-of-view. Semi-transparency or color translucency can be achieved using grey or color mask silhouettes. Opaque dimensional synthetic content has been achieved by displaying the color image on a depth fused display. Commodity LCD panels are now large enough to accommodate average human-sized characters. (See figure 8 for a modified 80" Sharp Aquos television w/ scrim projection).

References:

[1] O Bimber, and R Raskar. 2005. Spatial Augmented Reality: Merging Real and Virtual Worlds. A K PETERS Limited (MA)

[2] O Bimber, et al. 2001. The virtual showcase. IEEE Computer Graphics and Applications, 48–55.

[3] O Bimber, and B Fröhlich. 2002. Occlusion shadows: Using projected light to generate realistic occlusion effects for view-dependent optical see-through display. In Proc. IEEE and ACM Int. Symposium on Mixed and Augmented Reality, 186.

9011-22, Session 6

Transparent stereoscopic display and application

Nicola Ranieri, Hagen Seifert, ETH Zurich (Switzerland); Markus Gross, ETH Zürich (Switzerland)

CONTEXT

Augmented reality has become important to our society as it can enrich the actual world with virtual information.

Transparent 3D screens offer one possibility to overlay rendered scenes with the environment, acting both as display and window.

OBJECTIVE

Upcoming technologies as transparent LCD and transparent OLED are candidates to be used in augmented reality applications. However, transparent LCD are not self-emissive, transmit only 33% of the light through the color filters and absorb another 50% in the polarizer. Hence, extremely bright background would be needed to account for the resulting transparency of at most 17%, which narrows the area of application. Transparent OLED on the other hand are self-emissive but could not achieve high transparency so far neither.

Promising alternatives are transparent back-projection foils, which attain a much higher transparency. Though, existing screen technologies like anisotropic or isotropic back-projection have various limitations, which have to be considered when using them in a 3D-capable system.

In a first part of the project we assess different back-projection technologies for the use in combination with passive and active shutter glasses. Based on the derived insights, we build a transparent stereoscopic 3D system for augmented reality.

Once a transparent 3D display has been built, care has to be taken to align virtual content with reality. Proper alignment can be achieved by tracking the eyes of a viewer and adapting the stereoscopic rendering accordingly. Unfortunately, low frequency and high latency of the tracker can inhibit the immersive experience of the user, imposing further challenges to the system.

These issues can be overcome by using a Kalman filter and prediction, which has been successfully done for human motion in previous work. In a second part of our project we assess the quality of similar methods, when being used to add motion parallax and correct perspective cues to the binocular cues of stereoscopic rendering.

METHOD

Advantages, drawbacks and limitations of transparent anisotropic and isotropic back-projection foils for passive and active stereo are evaluated. Based on the results, a system using an anisotropic back-projection foil is proposed and stereoscopic 3D is achieved through passive, polarized glasses. A Kalman filter and prediction are applied to a Kinect tracker, to decrease tracking latency and increase the update frequency. Viewer adaptive stereoscopic content is rendered from correct perspectives, further increasing realism.

RESULTS

We have designed and built a transparent, back-projected and stereoscopic display based on an analysis of existing back-projection technologies. Realism and immersion into the displayed scenes are improved by enhancing a Kinect tracker with proper filtering and prediction. Our passive system provides a lightweight and immersive user experience for augmented reality applications.

NOVELTY

We provide an evaluation of existing transparent back-projection foils for the use in glasses based stereoscopic 3D screens. A transparent stereoscopic display using passive glasses has been built. Algorithms how to use improved Kinect tracking to add motion parallax to a stereoscopic system are presented and verified in an augmented reality application.

9011-28, Session 6

A hand-held immaterial volumetric display

Antti Sand, Ismo K. Rakkolainen, Tampere Univ. of Technology (Finland)

CONTEXT

We have created a movable, limitedly volumetric “immaterial” display. It can be swept to create mid-air slices of volumetric objects, or augmented reality on top of real objects.

OBJECTIVE

Usually volumetric displays produce the images in a confined space, which does not allow touch. Various kinds of volume slicing displays [e.g., 1] enable volumetric slices on a diffuse plastic sheet in a very limited range. However, the solid sheet may be too cumbersome or it may harm delicate physical objects when used for augmented reality, as it requires proximity of the screen.

We have created a movable, limitedly volumetric “immaterial” display. It can e.g., be swept across mid-air to create slices of volumetric objects (e.g., MRI or CT scan datasets) or to create augmented reality on top of real objects.

METHOD

Our prototype is based on the patented FogScreen [2] technology. It is also the first mobile, hand-held fogscreen. The previous FogScreen installations have been fixed, large-scale setups. We constructed a proof-of-concept immaterial volumetric display. The fogscreen flow unit, pico projector, and a smartphone (for tracking and rendering) are all merged to a light-weight hand-held unit, and only the fog is generated in a separate container.

The smartphone has a camera (and potentially inertial sensors) for tracking. It also serves as an image source for the Microvision ShowWX+ pico projector. Initially we used AR markers for tracking. A small depth camera can additional¬ly be used for improving the tracking or for interaction on the display volume.

RESULTS

The device can render mid-air volumetric images and it can be used in any orientation, e.g., upside down, sideways, etc. Our construction is still very crude and big, but the device can be made smaller. Instead of a hand-held unit, an alternative construction would be to move the fog plane back and forth automatically or manually using a desktop FogScreen, thus creating a reach-through sliced display volume.

The viewer looks always roughly towards the projector, which produces high brightness due to the high anisotropy of Mie scattering of light from the fog [3] and high resolution, as the neighboring pixels do not blend together. The hotspot of the projector can be removed with user tracking and proper rendering [4].

NOVELTY

The movable, volumetric FogScreen is a new concept, and the first hand-held fogscreen. The proof-of-concept display provides an easy way to visualize volumetric objects in mid-air. It can also pass through objects without touching them, being thus suitable for e.g., augmented reality in proximity of real objects. Applications for it include mixed reality, mid-air 3D user interfaces [5, 6], visualizations in mid-air, etc.

REFERENCES

[1] Cassinelli A. and Ishikawa, M. 2009. “The Volume Slicing Display”. Siggraph Asia 2009 Emerging Technologies, p. 88.

[2] Fogio Inc. 2013. FogScreen. http://www.fogscreen.com/.

[3] Rakkolainen, I. 2008. “Measurements and Experiments of Immaterial Virtual Reality Display”. 2nd IEEE 3DTV Conf. 37-40.

[4] Palovuori, K. and Rakkolainen, I. 2013. ”Improved Virtual Reality for Mid-Air Projection Screen Technology”. 3rd Int. Symp. on Communicability, Comp. Graph. and Innovative Design for Interactive Sys. (CCGIDIS 2013). 25-33.

[5] DiVerdi, S., Rakkolainen, I., Höllerer, T. and Olwal, A. 2006. “A Novel Walk-through 3D Display”. SPIE Electronic Imaging, Stereoscopic Displays and Virtual Reality Systems XIII, Vol. 6055. 1-10.

[6] Rakkolainen, I., Höllerer, T., DiVerdi, S. and Olwal, A. 2009. Mid-air Display Experiments to Create Novel User Interfaces. Multimedia Tools and Applications, Vol. 44, Issue 3. Springer Netherlands, 2009. 389-405.

9011-24, Session 7

Perceived crosstalk assessment on patterned retarder 3D display

Bochao Zou, Yue Liu, Yongtian Wang, Beijing Institute of Technology (China); Yi Huang, Beijing Institute of Technology (China) and Beijing Engineering Research Ctr. for Mixed Reality and Novel Display Technology (China)

CONTEXT: Nowadays, almost all stereoscopic displays suffer from crosstalk, which is one of the most dominant degradation factors of image quality and visual comfort for 3D display devices. To deal with such problems, it is worthy to quantify the amount of perceived crosstalk.

OBJECTIVE: Crosstalk measurements are usually based on some certain test patterns, but scene content effects are ignored. To evaluate the perceived crosstalk level for various scenes, subjective test may bring a more correct evaluation. However, it is a time consuming approach and is unsuitable for real-time applications. Therefore, an objective metric that can reliably predict the perceived crosstalk is needed. A correct objective assessment of crosstalk for different scene contents would be beneficial to the development of crosstalk minimization and cancellation algorithms which could be used to bring a good quality of experience to viewers.

METHOD: A patterned retarder 3D display is used to present 3D images in our experiment. By considering the mechanism of this kind of devices, an appropriate simulation of crosstalk is realized. It can be seen from the literature that the structures of scenes have a significant impact on the perceived crosstalk, so we first extract the differences of the structural information between original and distorted image pairs through Structural SIMilarity (SSIM) algorithm, which could directly evaluate the structural changes between two complex-structured signals. Then the structural changes of left view and right view are computed respectively and combined to an overall distortion map. Under 3D viewing condition, because of the added value of depth, the crosstalk of pop-out objects may be more perceptible. To model this effect, the depth map of a stereo pair is generated and the depth information is filtered by the distortion map. Moreover, human attention is one of important factors for crosstalk assessment due to the fact that when viewing 3D contents, perceptual salient regions are highly likely to be a major contributor to determining the quality of experience of 3D contents. To take this into account, perceptual significant regions are extracted, and a spatial pooling technique is used to combine structural distortion map, depth map and visual salience map together to predict the perceived crosstalk more precisely. To verify the performance of the proposed crosstalk assessment metric, subjective experiments are conducted with 24 participants viewing and rating 60 simuli (5 scenes * 4 crosstalk levels * 3 camera distances). After an outliers removal and statistical process, the correlation with subjective test is examined using Pearson and Spearman rank-order correlation coefficient. Furthermore, the proposed method is also compared with two traditional 2D metrics, PSNR and SSIM.

RESULTS: After the above-mentioned processes, the evaluation results demonstrate that the proposed metric is highly correlated with the subjective score when compared with the existing approaches. Because the Pearson coefficient of the proposed metric is 90.3%, it is promising for objective evaluation of the perceived crosstalk.

NOVELTY: The novelty contributions are twofold. First, an appropriate simulation of crosstalk by considering the characteristics of patterned retarder 3D display is developed.

Second, an objective crosstalk metric based on visual attention model is introduced.

9011-25, Session 7

Subjective evaluation of an active crosstalk reduction system for mobile autostereoscopic displays

Alexandre Chappuis, Martin Rerábek, Philippe Hanhart, Touradj Ebrahimi, Ecole Polytechnique Fédérale de Lausanne (Switzerland)

CONTEXT

The quality of experience (QoE) provided by autostereoscopic 3D displays strongly depends on the user position. For an optimal image quality, the observer should be located at one of the relevant positions, called sweet spots, where artifacts reducing the QoE, such as crosstalk, are minimum.

OBJECTIVE

When a user watching an autostereoscopic display moves away from the sweet spot along the frontal axis, the right (left) image progressively leaks into the left (right) eye (crosstalk), until a position where the left (right) eye only perceives the right (left) image (reverse stereo). This effect decreases depth and image quality and increases visual discomfort, especially when reverse stereo occurs.

To improve the QoE provided by autostereoscopic displays, researchers have proposed to perform active crosstalk reduction based on the user position [1,2,3]. The user is tracked using face and eye detection. Outside of the sweet spots, crosstalk is canceled by simply switching back to 2D mode [2] or reduced by performing an intelligent assignment of the pixel values based on the visibility profiles of different views [3]. However, characterization of the display is necessary to determine the visibility profiles. In case of reverse stereo, the left and right images are swapped [3].

In this paper, we propose and evaluate a complete active crosstalk reduction system running on an HTC EVO 3D smartphone.

METHOD

To determine the crosstalk level at each position, display characterization was performed according to [4]. User tracking is performed using the frontal camera, face detection, and eye detection. Based on the user inter-pupillary distance and crosstalk profile, the system first helps the user finding the sweet spot using visual feedback. If the user moves away from the sweet spot and crosstalk is above a certain threshold, then active crosstalk compensation is performed. If crosstalk rises above 50%, then the images are swapped to correct reverse stereo. If the viewing angle is too large, a 2D image is displayed by using the same image for the left and right views.

The user preference between standard 2D and 3D modes and the proposed system was evaluated in terms of image quality, depth quality, and user friendliness through a subjective evaluation. Six stereoscopic images were used: one for training and five for testing. The pair comparison methodology was selected. A total of 18 subjects took part in the experiment. Normalized continuous quality scores (NCQS) were obtained from the preference matrices by using the Bradley-Terry-Luce model.

RESULTS

In terms of depth perception, the proposed system (NCQS=100) clearly outperforms the 3D (NCQS=40) and 2D (NCQS=0) modes. In terms of image quality, 2D mode (NCQS=100) was found to be the best, but the proposed system (NCQS=30) outperforms 3D mode (NCQS=0). Regarding user friendliness, the 3D mode and proposed system were found to be identical.

NOVELTY

This work presents a complete active crosstalk reduction system for mobile devices, including its subjective evaluation and comparison to commercially available 2D and 3D imaging systems.

REFERENCES

[1] A. Boev, M. Georgiev, A. Gotchev, and K. Egiazarian, “Optimized Single-viewer Mode of Multiview Autostereoscopic Display,” 16th European Signal Processing Conference (EUSIPCO), Lausanne, Switzerland, August 2008.

[2] A. Boev, M. Georgiev, A. Gotchev, N. Daskalov, and K. Egiazarian, “Optimized Visualization of Stereo Images on an OMPA Platform with Integrated Parallax Barrier Auto-stereoscopic Display,” 17th European Signal Processing Conference (EUSIPCO), Glasgow, Scotland, August 2009.

[3] J. Park, D. Nam, G. Sung, Y. Kim, D. Park, and C. Kim, “Active Crosstalk Reduction on Multi-View Displays Using Eye Detection,” SID Symposium Digest of Technical Papers, 42 (1), pp. 920–923, June 2011.

[4] M. Sykora, J. Schultz, and R. Brott, “Optical characterization of autostereoscopic 3D displays,” Proc. SPIE 7863, Stereoscopic Displays and Applications XXII, 78630V, February 2011.

9011-26, Session 7

Study of blur discrimination for 3D stereo viewing

Mahesh M. Subedar, Arizona State Univ. (United States) and Intel Corp. (United States); Lina J. Karam, Arizona State Univ. (United States)

Blur is an important attribute in the study and modeling of the human visual system (HVS). In this paper, we present the details of subjective tests performed to study blur discrimination using 3D test patterns. Specifically, we wanted to determine the effect of disparity on the blur discrimination, and how to incorporate disparity into blur discrimination models.

In blur discrimination experiments, subjects are presented with two stimuli, one with a reference blur level and other with a reference plus an additional blur (referred to as target-blur level), in a random order. Using a Two-Interval Forced Choice (2IFC) subjective test methodology, subjects are asked to judge which stimulus has larger blur. The threshold is measured as the additional blur required in order to discern the reference blur level form the target blur level. In this study, stereo 3D test pattern objects are rendered at different depths to measure the blur discrimination thresholds. A Gaussian low-pass filter is applied to generate test patterns with different blur levels.

The stimulus used for the experiments is a vertical edge formed by two rectangular boxes. In order to generate a blurred edge, a Gaussian low pass filter is applied across the central edge, along the horizontal direction. The stimulus was rendered on a Hyundai passive stereoscopic 3DTV display (Model# S465D). Five reference blur levels and three disparity values were considered, which correspond to a total of 15 experiments. The five reference blur levels considered are [0.0, 0.53, 1.06, 2.12, 5.30] arcmin, which correspond to the standard deviation of the Gaussian low-pass filter in arcmin unit. The three disparity values considered are [0.0, -21.22, 21.22] arcmin. The Michelson contrast ratio across the edge is fixed at 0.83, which is calculated as the ratio of luminance difference over the sum of luminance values, across the edge.

Six subjects participated in this subjective study. The subjects were tested for visual acuity, color vision and stereo vision. Each subject repeated 15 experiments four times, in four separate sessions. The plots of the threshold values vs. reference blur for six subjects, across five reference blur levels, and three disparities are presented. For each subject, the mean value of the four sessions is plotted, along with the average threshold values for all the subjects. It can be observed from the plots that the threshold values for the three disparity values follow a similar trend. A dip in the threshold value is seen around a reference blur level of 1.0 arcmin, which is similar to the 2D blur discrimination experiments found in the literature. A fit of Weber model to the subjective test data is also presented.

The subjective test results indicate that the blur discrimination thresholds remain almost constant, as we vary the disparity value, which corresponds to varying the object depth. These findings further indicate that one can apply 2D blur discrimination models for 3D stereo blur discrimination. We can conclude that models developed for 2D blur discrimination, such as the Weber model, can be applied to model the 3D blur discrimination.

9011-27, Session 7

The effect of stereoscopic acquisition parameters on both distortion and comfort

Robert H. Black, Georg F. Meyer, Sophie M. Wuerger, Univ. of Liverpool (United Kingdom)

CONTEXT: With ever-increasing demands for 3D quality it is important to generate faithful and comfortable stereoscopic content. It is vital to measure both comfort and distortion for typical stereoscopic content acquisition values (interaxial, FOV and convergence). OBJECTIVE: This project investigates perceived distortion and comfort as a function of acquisition parameters. Stereoscopic calculator applications are based on ray-tracing models, not perception data. We will compare these predictions with psychophysical data for distortion and comfort. METHOD: To measure distortions we used a two alternative forced choice (2AFC) task asking observers whether a hinge appeared greater than or less than a right angle. Comfort was assessed on the five point ITU-R BT.500-11 comfort scale. Participants were screened for stereoacuity using the Titmus Fly Test and completed the Simulator Sickness Questionnaire after the experiment. Two experiments were conducted, each with 40 observers who made 250 2AFC judgments of a textured hinge stimulus. In the first experiment, stimulus width was constant and stimulus depth was varied. In the second experiment, stimulus depth was constant and stimulus width was varied. The main factor manipulated was virtual camera separation (20, 40, 60, 80, 100mm) which linked interaxial separation to horizontal image translation (HIT). This kept the point of zero parallax at the hinge apex. To obtain a psychometric function for each camera separation, one of series of rendered hinge angles (50-130°; in 10° steps) was presented on a particular trial. After stimulus presentation, the observer was asked to respond whether the angle was greater or smaller than 90°. Subsequently, the observer rated the comfort of the stimuli on the five-point Likert scale. The corresponding disparities presented were in the range (30-165 arcmin). Screen width (50cm), viewer distance (60cm), and virtual target distance (60cm) were calibrated to display an orthoscopic image with a matched 45° horizontal field of view to the angle of view. Angles perceptually equivalent to right angles were then compared with the predictions using a ray-tracing model. RESULTS: We report two main findings: (1) In both experiments, there are systematic distortions in the percept of right angles as a function of camera separation. For the 60mm camera separation (approx. ortho-stereo) a right angle is perceived faithfully and no distortions occur. For the non-ortho camera separations (20, 40, 80, 100mm), we find a regression towards a right angle in the perceived angle, inconsistent with the prediction of the ray-tracing model. For example, when rendered with a 40 mm camera separation, a 70° angle is perceived as a right angle, while a ray tracing model predicts that a 60° angle should be perceived as a right angle. (2) For fixed width/varied depth stimuli, comfort ratings obtained after each trial show a strong negative correlation with the amount of disparity. For fixed depth/varied width stimuli, comfort ratings show a slight positive correlation with stimulus width, consistent with previous reports. NOVELTY: This work is novel because it explores the relationship between stereoscopic acquisition parameters of controlled stimuli for both distortion and comfort. Improving stereoscopic calculators with psychophysical distortion and comfort data can inform 3D content production.

Acknowledgements: EPSRC Grant No. 113300095 'Human factors in the design of stereoscopic 3D' Creative Industries KTN & Sony Computer Entertainment Europe

9011-301, Session Plen1

Using fMRI To Reverse Engineer the Human Visual System

Jack L. Gallant, Univ. of California, Berkeley (United States)

No Abstract Available

9011-29, Session 8

Fully Automatic 2D to 3D Conversion with aid of High-Level Image Features

Vikram V. Appia, Umit Batur, Texas Instruments Inc. (United States)

We present a fully automated 2D-to-3D conversion algorithm for consumer images, which combines low- and high-level image features in pseudo depth map generation for 3D view synthesis. Algorithms in prior art typically use low-level features and require human intervention to interpret high-level image features for depth map generation. Our method automatically assigns relative depth value to various regions in a given 2D scene using both low- and high-level features. The algorithm consists of four major parts: a) Training on ground truth depth images, b) Pseudo depth map generation, c) Enhancing depth map using high-level features and d) Depth Image based rendering.

We train our depth model using a database of ground truth color images with their associated depth maps. Our supervised learning model learns the relationship of low-level image features like color, texture, gradient and pixel co-ordinates with scene depth. For 2D-to-3D conversion of each test image, we create a unique scene model for the given image based on activity distribution within the image.

We then segment the test image into various regions using a mean shift clustering algorithm. The clustering algorithm uses color and edge information to segment the image. We use the generated scene model to assign relative depth values to each region in the image. Since the capture environment is unknown, using low-level features alone will create inaccuracies in depth maps. Such flawed depth creates various artifacts (unrealistic depth ordering, geometric distortion in faces, non-uniform depth to contiguous 3D objects, distorted foreground objects, etc.) in the rendered 3D image, causing an unpleasant viewing experience.

We use certain high-level feature detectors to enhance the depth map. We use feature detectors such as sky, foliage and orientation detector, as well as face detector, to enforce certain rules and overcome these imperfections. The orientation detection algorithm and the sky and foliage detectors are used to correct the relative ordering of depth values in the pseudo depth maps. Human faces in the image tend to cluster with the background due to variation in color of the clothing. We use the face detector to identify faces and ensure uniform depth on the human subjects in the images.

Finally, we use Depth Image Based Rendering algorithm to generate the stereo pair for rendered 3D images. Creating 3D image pair from a 2D image generates holes in the rendered views. The rendered views are generated either by replacing the missing pixels with simple techniques like inpainting, extrapolation or smoothing the depth maps to introduce distortions. Inpainting and extrapolation based approaches create various artifacts in highly textured regions, whereas smoothing based techniques introduce geometric distortions in the foreground objects. Such distortions can cause a very unpleasant viewing experience. We use the high-level feature detectors to identify the foreground objects and transfer the required distortion to a narrow band in the background. This leaves the foreground (object of interest such as faces and regions with high texture) distortion free to create a pleasing 3D view.

9011-30, Session 8

Stereoscopy for visual simulation of materials of complex appearance

Fernando E. da Graça, Alexis Paljic, Dominique Lafon-Pham, Mines ParisTech (France)

CONTEXT:

The context of this work is the use of Stereoscopy for visual simulation of materials of complex appearance.

The perception of materials in everyday life relies on different sensori-motor information, among which binocular vision and head motion have a strong role. There is a strong need for the computer graphics community to use virtual reality in the simulation of physically based material appearance, since it can provide these perception cues. However, for complex materials such as car paints, there are few works that study their display and their perception within stereoscopic systems. Through the use of metallic flakes within a substrate, car manufacturers create materials with goniochromatic effects that produce binocular differences. In this context we want to study if stereoscopic information is important for user perception of characteristics of such materials.

During our collaborations with car paint designers, they have identified that flake density has a strong role in the visual appearance and aesthetics of car paint. Works on perception of paints are mostly interested in perception of other specific characteristics: gloss [1, 2, 3], shape, slant [4]. To our knowledge, there are no studies on stereoscopic perception of flake density.

OBJECTIVE:

The present work studies the role of stereoscopy on perceived surface aspect. The objective is to investigate if, and how, the additional information conveyed by the binocular vision affects the observer judgment on the evaluation of flake density in a car paint simulation.

METHOD:

We have set up a heuristic flake model, on the basis of observation, with a Voronoi modelization of flakes. The model was implemented in our rendering engine using global illumination, ray tracing, with an off axis-frustum method for the calculation of stereo pairs.

We conducted a user study of flake density discrimination task to determine Perception Thresholds (JNDs) for flake density, and to compare Stereoscopic and Monoscopic Display (see pictures in appendix).

RESULTS:

Results show that the perception threshold is better with the stereoscopic display. Our results are discussed on the basis of two flake density metrics: percentage of the surface occupied by flakes, and number of flakes on the surface. Using morphology image processing techniques we found a specific scale of observation for which the changes in stimuli is important, this corresponds to flake size of approximately 404µm. This means that flakes of specific scale strongly contribute to density perception. This shows the combined role of flakes and background, but it is unclear if the stereoscopic effects are stemming only from flakes or also their surrounding background.

This sets up our next experiment of studying the stereo differences in perception without background information.

NOVELTY :

This work contributes to a better stereo visualization of goniochromatic materials within CG simulations. We provide perception thresholds and density metrics, which improve the understanding of limits of perception of flake density for stereoscopic and monoscopic display. To characterize our results, we also introduce a methodology to measure the density, and to identify the main visual characteristics of the plate.

REFERENCES:

[1] Obein, G., Knoblauch, K., Viénot, F.: Difference scaling of gloss: nonlinearity, binocularity, and constancy. Journal of vision 4 (2004) 711-720

[2] McCamy, C.S.: Observation and measurement of the appearance of metallic materials. part II. micro appearance. Color Research and Application 23 (1998) 362-373

[3] Sakano, Y., Ando, H.: Effects of head motion and stereo viewing on perceived glossiness. (2010) 1-15

[4] Knill D. C., Saunders J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? Vision Research, 43, 2539-2558

9011-31, Session 8

Stereoscopic depth perception in video see-through augmented reality within action space

Mikko Kytö, Aleksi Mäkinen, Timo Tossavainen, Pirkko T. Oittinen, Aalto Univ. School of Science and Technology (Finland)

CONTEXT:

Depth perception is an important component in many augmented reality (AR) applications. It is, however, affected by multiple error sources. Most studies on stereoscopic AR have focused on the personal space whereas we address the action space (at distances beyond 2 m; in this study 6-10 m) using a video see-through display (HWD). This is relevant for example in the navigation and architecture domains.

OBJECTIVE:

For design guideline purposes there is a considerable lack of quantitative knowledge of the visual capabilities facilitated by stereoscopic HWDs. To fill the gap two interrelated experiments were conducted: Experiment 1 had the goal of finding the effect of viewing through a HWD using real objects while Experiment 2 dealt with variation of the relative size of the augmentations in the monoscopic and binocular conditions.

METHOD:

In Experiment 1, the participants judged depths of physical objects in a matching task using the Howard-Dolman test. The order of viewing conditions (naked eyes and HWD) and initial positions of the rods were varied.

In Experiment 2, the participants judged the depth of an augmented object of interest (AOI) by comparing the disparity and size to auxiliary augmentations (AA). The task was to match the distance of a physical pointer to same distance with the AOI. The approach of using AAs has been recently introduced (Kytö et al. 2013). The AAs were added to the scene following literature-based spatial recommendations.

RESULTS:

The data from Experiment 1 indicated that the participants made more accurate depth judgments with HWD when the test was performed first with naked eyes. A hysteresis effect was observed with a bias of the judgments towards the starting position.

As for Experiment 2, binocular viewing improved the depth judgments of AOI over the distance range. The binocular disparity and relative size interacted additively; the most accurate results were obtained when the depth cues were combined. The results have similar characteristics with a previous study (Kytö et al. 2013), where the effects of disparity and relative size were studied in X-Ray visualization case at shorter distances.

Comparison of the two experiments showed that stereoscopic depth judgments were more accurate with physical objects (mean absolute error 1.13 arcmin) than with graphical objects (mean absolute error 3.77 arcmin).

NOVELTY:

The study fills the knowledge gap on exocentric depth perception in AR by quantitative insight of the effect of binocular disparity and relative size. It found that additional depth cues facilitate stereoscopic perception significantly. Relative size between the main and auxiliary augmentations turned out to be a successful facilitator. This can be traced to the fact that binocular disparity is accurate at short distances and the accuracy of

relative size remains constant at long distances. Overall, these results act as guidelines for depth cueing in stereoscopic AR applications.

References

Kytö, M., Mäkinen, A., Häkkinen, J., & Oittinen, P. (2013). Improving Relative Depth Judgments in Augmented Reality with Auxiliary Augmentations. ACM Transactions on Applied Perception, 10(1), 1–22.

9011-32, Session 8

A multilayer display augmented by alternating layers of lenticular sheets

Hironobu Gotoda, National Institute of Informatics (Japan)

A multilayer display is an auto-stereoscopic display constructed by stacking multiple layers of LC (liquid crystal) panels on top of a light source. It is capable of presenting the light field representing an arbitrary 3D scene to viewers situated within a designated viewing zone. Compared with other auto-stereoscopic displays, such as lenticular displays, the viewing zone provided by a multilayer display is rather narrow. However, the depth of field provided by a multilayer display is usually larger than that of lenticular displays. Furthermore, depth perception via accommodation is also possible with multilayer displays, which partially resolves the vergence-accommodation conflict.

One drawback of multilayer displays is the lack of capability representing the occlusion of objects in 3D scenes. To address this problem, research efforts were reported that incorporated directional lighting into the conventional multilayer displays. Here we propose another method to mitigate this problem. We consider an integration of multilayer and lenticular displays, where multiple LC panels sandwiched between pairs of lenticular sheets are stacked. By associating lenticular sheets with a LC panel, each pixel in the panel is turned into a “view-dependent pixel”, which is visible only from a particular direction. The entire set of pixels in the display is now partitioned into several disjoint classes, each of which corresponds to a very narrow viewing zone. Since pixels belonging to different classes are controlled independently, the visibility of objects should be resolved only within the narrow zone.

To obtain a suitable combination of lenticular sheets and LC panels, the optical properties of the layer’s sandwiched structure have been analyzed. We choose to use different sheets attached to the front-side (facing the viewers) and back-side (facing the back light) of a LC panel. In particular, the pitch of the front-side lenticular sheet is N times larger than that of the back-side sheet. The same relationship also applies to the focal lengths of the sheets. In any cases, the focal points lie in the center of the LC panel. The pixel pitch of LC panel is 1/M-th of the pitch of front-side lenticular sheet. These particular choices of parameters bring us the following features. When parallel rays of light arrive at a lens in the front-side sheet, they are transformed to two sets of parallel rays going out from the back-side sheet. The directions of incoming and outgoing rays are not, in general, coincident with each other. However, their differences can be made small when we take a large N. We have examined several combinations of (N, M) using a computer simulation model, and found a suitable one.

Finally, we have developed a prototype display consisting of two layers. Although the number of layers is small, we can still observe the differences between the conventional multilayer display, the lenticular display and the proposed display. In particular, objects’ occlusion is properly represented in the constructed prototype. The computer simulation model suggests that stacking larger number of layers is also possible. The construction of such displays is the target of our on-going work.

9011-81, Session Key2

Compressive displays: combining optical fabrication, computational processing, and perceptual tricks to build the displays of the future *(Keynote Presentation)*

Gordon Wetzstein, MIT Media Lab. (United States)

In this talk, we explore modern approaches to glasses-free 3D display using compressive light field displays. In contrast to conventional technology, compressive displays aim for a joint-design of optics, electronics, and computational processing that together exploit compressibility of the presented data. For instance, multiview images or light fields show the same 3D scene from different perspectives - all these images are very similar and therefore compressible. By combining displays that use multilayer architectures or directional backlighting combined with optimal light field factorizations, limitations of existing devices, for instance resolution, depth of field, and field of view, can be overcome. In addition to light field display, we will discuss approaches to compressive super-resolution image display and compressive high dynamic range display. As with compressive light field displays, these technologies rely on multiplexing image content in time such that the visual system of a human observer combines presented patterns into a consistent 3D, high-resolution, or high-contrast image.

With the invention of integral imaging and parallax barriers in the beginning of the 20th century, glasses-free 3D displays have become feasible. With rapid advances in optical fabrication, digital processing power, and computational perception, a new generation of display technology is emerging: compressive displays exploring the co-design of optical elements and computational processing while taking particular characteristics of the human visual system into account. We will review these techniques and also give an outlook on next-generation compressive light field camera technology.

9011-33, Session 9

(JEI invited) Acquisition of omnidirectional stereoscopic images and videos of dynamic scenes: a review *(Invited Paper)*

Luis E. Gurrieri, Eric Dubois, Univ. of Ottawa (Canada)

Different camera configurations to capture panoramic images and videos are commercially available today. However, capturing omnistereoscopic snapshots and videos of dynamic scenes is still an open problem. Several methods to produce stereoscopic panoramas have been proposed in the last decade, some of which were conceived in the realm of robot navigation and three-dimensional (3-D) structure acquisition. Even though some of these methods can estimate omnidirectional depth in real time, they were not conceived to render panoramic images for binocular human viewing. Alternatively, sequential acquisition methods, such as rotating image sensors, can produce remarkable stereoscopic panoramas, but they are unable to capture real-time events. Hence, there is a need for a panoramic camera to enable the consistent and correct stereoscopic rendering of the scene in every direction. Potential uses for a stereo panoramic camera with such characteristics are free-viewpoint 3-D TV and image-based stereoscopic telepresence, among others. A comparative study of the different cameras and methods to create stereoscopic panoramas of a scene, highlighting those that can be used for the real-time acquisition of imagery and video, is presented.

9011-35, Session 9

Depth consistency and vertical disparities in stereoscopic panoramas

Luis E. Gurrieri, Eric Dubois, Univ. of Ottawa (Canada)

CONTEXT: In recent years, the problem of acquiring omnidirectional stereoscopic imagery of dynamic scenes has gained commercial interest and, consequently, new techniques have been proposed to address this problem [1]. The goal of many of these novel panoramic methods is to provide practical solutions for acquiring real-time omnidirectional stereoscopic imagery suitable to stimulate binocular human stereopsis in any gazing direction [2][3]. In particular, methods based on the acquisition of partially overlapped stereoscopic snapshots of the scene are the most attractive for real-time omnistereoscopic capture [1]. However, there is a need to rigorously model these acquisition techniques in order to provide useful design constraints for the corresponding omnidirectional stereoscopic systems.

OBJECTIVE: Our main goal in this work is to propose an omnidirectional camera model, which is sufficiently flexible to describe a variety of omnistereoscopic camera configurations. We have developed a projective camera model suitable to describe a range of omnistereoscopic camera configurations and usable to determine constraints relevant to the design of omnistereoscopic acquisition systems. In addition, we applied our camera model to estimate the system constraints for the rendering approach based on mosaicking partially overlapped stereoscopic snapshots of the scene.

METHOD: First, we grouped the possible stereoscopic panoramic methods, suitable to produce horizontal stereo for human viewing in every azimuthal direction, into four camera configurations. Then, we propose an omnistereoscopic camera model based on projective geometry which is suitable for describing each of the four camera configurations. Finally, we applied this model to obtain expressions for the horizontal and vertical disparity errors encountered when creating a stereoscopic panorama by mosaicking partial stereoscopic snapshots of the scene.

RESULTS: We simulated the parameters of interest using the proposed geometric model combined with a ray tracing approach for each camera model. From these simulations, we extracted conclusions that can be used in the design of omnistereoscopic cameras for the acquisition of dynamic scenes. One important parameter used to contrast different camera configurations is the minimum distance to the scene to provide a continuous perception of depth in any gazing direction after mosaicking partial stereoscopic views. The other important contribution is to characterize the vertical disparities that cause ghosting at the stitching boundaries between mosaics. In the simulation, we studied the effect of the field-of-view of the lenses, and the pixel size and dimension of the sensor in the design of the system.

NOVELTY: The main contribution of this work is to provide a tractable method for analyzing multiple camera configurations intended for omnistereoscopic imaging. In addition, we estimated and compared the system constraints to attain a continuous depth perception in all azimuth directions. Also important for the rendering process, we characterized mathematically the vertical disparities that would affect the mosaicking process in each omnistereoscopic configuration. This work complements and extends our previous work in stereoscopic panoramas acquisition [1][2][3] by proposing a mathematical framework to contrast different omnistereoscopic image acquisition strategies.

REFERENCES:

[1] Gurrieri L. E., Dubois E., “Acquisition of omnidirectional stereoscopic images and videos of dynamic scenes: a review,” J. Electron. Imaging, vol. 22, no. 3, pp. 030902–030902, Jul. 2013. http://dx.doi.org/10.1117/1.JEI.22.3.030902

[2] Gurrieri L. E., Dubois E., “Efficient panoramic sampling of real-world environments for image-based stereoscopic telepresence,” Proc. SPIE. 8288, 82882D (2012). http://dx.doi.org/10.1117/12.908794

[3] Gurrieri L. E., Dubois E., “Stereoscopic cameras for the real-time acquisition of panoramic 3d images and videos,” Proc. SPIE. 8648, 86481W (2013). http://dx.doi.org/10.1117/12.2002129

9011-36, Session 9

Integration of multiple view plus depth data for free viewpoint 3D display

Kazuyoshi Suzuki, Nagoya Univ. (Japan); Yuko Yoshida, Tetsuya Kawamoto, Chukyo TV Broadcasting Corp. (Japan); Toshiaki Fujii, Kenji Mase, Nagoya Univ. (Japan)

CONTEXT:

In recent years, interactive communication systems are spreading and a study for developing a system that enables us to watch free viewpoint video of delivered live videos in real time has been popular.

With the spread of simple depth cameras that capture depth information of subjects in real time, the groundwork for reconstructing 3D graphical models of real three-dimensional objects in a short time has been developed rapidly.

OBJECTIVE:

Methods for generating free viewpoint video are roughly divided into two groups. One of them is Image Based Rendering (IBR) and the other is Model Based Rendering (MBR).?In both methods, the accuracies of depth information and camera parameters affect to the quality of generated free viewpoint images. The Marching Cubes method (MC method) is a candidate method for generating free viewpoint video by using MBR. The MC method is a simple and popular 3D visualization method for generating polygons by spanning triangular patches to given volume data. The accuracy of position of volume data of subjects affects to the quality of polygons of the subjects.

The resolution of real multiview videos and the corresponding depth information that are captured by cameras and depth cameras in the market are not so high. In addition, when integrating point cloud data that are captured by some cameras, consistency between them in terms of coordinate system or scale is necessarily insured.

For improving the subjective quality of polygons generated by the MC method, we propose a method for revising outlier in input point cloud data by using color information of input views and input depth information.

METHOD:

Entire circumferential point cloud models of subjects are constructed from multiple view plus depth data captured by multiple cameras and the camera parameters of them. Before that, view plus depth data at each viewpoint are converted to point cloud data that is a set of points with colors in three-dimensional space. The obtained point cloud data in all viewpoints are integrated into single point cloud data.

Next, input view at each viewpoint is segmented into superpixels by using the color information of the image and the corresponding depth data. All pixels in a same segment are considered to be on a same continuous plane. Points having outlier depth values are revised to values that are similar to those of the neighboring pixels. If the depth value of a pixel is lacked, a candidate depth value is interpolated from the values of neighboring pixels. In connected regions of point cloud data in different viewpoints, point data in the region are revised so as to vary smoothly by extracting segments that may be in the same surface of real object.

RESULTS:

We performed a subjective evaluation for comparing polygons generated by the pure MC method with those generated by the proposed method. We displayed free viewpoint images on a 2D display and freely specified virtual viewpoints by using a mouse. In addition, we generated stereoscopic images by capturing the constructed polygons at two virtual viewpoints corresponding to human eyes and display them on a 3D display. We confirmed that the images generated by the proposed method are superior to those generated by the pure MC method.

NOVELTY:

We proposed a method for revising integrated volume data for the purpose of improving the subjective quality of polygons generated by the Marching Cubes method.

9011-82, Session 9

Automatic detection of artifacts in converted S3D video

Dmitriy Vatolin, Lomonosov Moscow State Univ. (Russian Federation)

CONTEXT: In this work we address the problem of quality assessment of converted S3D video. There are specific types of artifacts which can be found in converted video and can't appear in captured content. One of the typical ways to create S3D video from 2D source is depth-image-based rendering; quality of the result depends both from depth maps and warping algorithms. Processing object boundaries is a principal concern, it may require hole filling and semitransparent-edges handling. Low-quality depth maps may cause a lot of problems from annoying jitter on object boundaries to totally unfeasible perception of scene.

OBJECTIVE: The objective of the performed work is development of the algorithms which will help to identify scenes with potential issues in converted S3D content.

METHOD: We propose three algorithms for detecting potential problems in converted S3D content. The first algorithm analyzes quality of edge processing and detects the boundaries with edge-sharpness mismatch between views. This artifact may appear when dealing with semitransparent edges or blurry depth map. After warping the boundary of the object becomes sharper in one view and blurrier in other view, what yields binocular rivalry. To identify this issue we estimate disparity map and extract object boundaries with noticeable depth difference. For these boundaries we analyze sharpness correspondence between RGB images, and if we have strong edge-sharpness mismatch, we signal potential issue. Special heuristics are applied to process scenes with complex background and large occlusion areas, because naive approach provides a lot of false positives.

Two other algorithms estimate quality of depth map, they help to detect scenes and objects with lack of depth volume. The algorithm for flat-scene detection estimates disparity map and then builds its linear approximation: it models the situation when linear gradient was used instead of real depth map for DIBR. Then we perform reconstruction of the right view from the left one on the basis of disparity-map linear approximation, and if it is matched perfectly we analyze the amount of feature points in the RGB image. If the image is feature-rich, we assume that depth map lacks details and signal the scene as potentially problem. The algorithm for flat-object detection exploits similar idea: we take the closest-to-viewer object and if it has uniform depth and is large enough with a lot of details in RGB image, the scene is also considered to be bad. For disparity estimation we use block-based algorithm with post-processing for adjusting boundaries and masking outliers in unreliable areas.

RESULTS: To evaluate the algorithms we performed automatic analysis of 10 Blu-ray 3D releases with converted S3D content including "Clash of the Titans", "The Avengers" and "The Chronicles of Narnia: The Voyage of the Dawn Treader". As the result of analysis we got tens of example scenes with described problems what demonstrates possible utility of proposed algorithm for QA purposes during production. The example scenes are provided in the attached file.

CONCLUSION: We described algorithms for detection of artifacts which are specific for converted S3D stereo. The proposed algorithms can be used for improving automated quality assessment at production stage. Directions of future work are decreasing false alarm rate of the current algorithms and development of new algorithms to extend the scope of processed-scene types and artifact types. The usage of proposed algorithms should be supplemented with analysis of common for S3D video issues such as depth-budget distribution, stereo-window handling, etc.

9011-37, Session 10

Disparity modifications and the emotional effects of stereoscopic images

Takashi Kawai, Daiki Atsuta, Yuya Tomiyama, Sanghyun Kim, Waseda Univ. (Japan); Hiroyuki Morikawa, Aoyama Gakuin Univ. (Japan) and Waseda Univ. (Japan); Reiko Mitsuya, Waseda Univ. (Japan); Jukka P. Häkkinen, Univ. of Helsinki (Finland)

CONTEXT

This paper describes a study focusing on the emotional effects of stereoscopic (3D) images. The authors have performed the disparity analysis of emotional scenes in well-known 3D movies using the center and the range of representable 3D space as the indexes. From the results of the analysis, the range of 3D space increased, and the trends of disparity modifications differed depending on the type of emotions.

OBJECTIVE

Above mentioned results suggested the possibility of a dramatic association between disparity modifications and emotional representations. In this study, the authors carried out an experiment to conduct the emotional effects of disparity modifications for four types of emotions.

METHODS

As the experimental stimulus, the International Affective Picture System (IAPS), a large set of normative emotional pictures, was chosen. IAPS has been evaluated by valence and arousal dimensions, and the images used in the experiment were extracted from IAPS by collating the evaluated values with Russell's circumplex model of emotion.

In concretely, three images judged to induce each of four types of emotions, happy, surprise, sad and fear, were selected. Since IAPS was a set of 2D pictures, the selected images were converted to 3D based on the monocular cues. In addition to the 2D and 3D condition, Emotional 3D (E3) condition applied above mentioned disparity modifications were prepared, and the three experimental conditions were compared in terms of valence and arousal.

The experimental stimulus was randomly presented for five seconds three times. Participants were asked to evaluate each stimulus using Self-Assessment Manikin (SAM). The stimulus was presented by using a 24-inch 3D display (Hyundai IT) with the polarizing filter glasses, and the viewing distance was about 90 cm. The participants were 20 university students with normal binocular vision.

RESULTS

From the results of two way factorial analysis of variance to examine the effects of the emotional types and the experimental conditions, a significant main effect of the emotional types was found in valence. On the other hand, significant main effects and interaction were found in arousal.

DISCUSSIONS

From the results of valence, it was thought that the image content affected more than the experimental conditions since a significant difference was found between the emotional types. However, in the results of arousal, it was confirmed that the emotions, happy, surprise and fear, were increased in the order from 2D, 3D to E3, significantly. Therefore, it was suggested that there was a possibility to increase arousal by the disparity modifications based on the results of the disparity analysis of emotional scenes.

9011-38, Session 10

Improving Perception of Binocular Stereo Motion on 3D Display Devices

Petr Kellnhofer, Tobias Ritschel, Karol Myszkowski, Hans-Peter Seidel, Max-Planck-Institut für Informatik (Germany)

Stereoscopic displays that reproduce binocular disparity are widespread in cinemas, home theaters and even mobile devices. Humans are familiar with this important depth cue and its presence is the key to a natural viewing experience.

However the presentation of disparity in current displays is severely limited. In contrast to real life, disparity is reduced to a limited range to avoid conflicts with accommodation to the screen plane. Current 3D displays, do not reproduce the full lightfield, and instead, present two 2D images, multiplexed to each eye. The final perception suffers from this issue combined with other luminance-related limitations like hold-type blur, limited dynamic range and spatial resolution. In this paper we discuss the influence of display limitations on stereo perception.

The motion can also get into conflict with the multiplexing protocol used to present images to both eyes. We investigate two very common practical scenarios of such interaction.

The first scenario is the most basic and cheap 3D equipment available: a conventional display observed using anaglyph glasses. The effect of using different color filters for each eye in combination with the different visual processing speed (Pulfrich effect) of each color -- respectively, eye -- is analyzed. We observe false depth percepts from disparity generated for moving objects as a result. We provide measurements of the effect magnitude with various filters, luminance intensities, disparity magnitudes and motion speeds and suggest correction. A simple web browser-based tool is provided to calibrate specific type of glasses and display.

Second, the sequential presentation of the left and right frame requires either a high refresh rate display and active shutter glasses or two projectors with passive polarized glasses. The first setup is common for 3D TVs and gaming, the second for 3D cinemas. Sequential presentation is done at frequencies beyond the flickering perception threshold. However, the delay between eyes can still be perceived for moving objects. If both frames are captured at the same simulation time, e.g. for simultaneous presentation, we perceive apparent motion in depth. Time offset-capturing is a straight forward cure for simple situation when each frame is displayed once. However, movie and game frame rates are not high enough to allow for such presentation without flickering. Then, frame repetition is used. We observe that false disparity depends also on the content and eye pursuit motion. We use a simple model of dynamic attention saliency to estimate the eye pursuit and optimized the capturing process accordingly. For applications where capturing cannot be optimized in presentation time we propose remedy based on image space warping. The output of our method is a dynamic stereo 3D content optimized for sequential presentation with possibly time-varying presence of frame repetition that reduces the depth-from-disparity-distortion.

Motion can additionally interact with processing applied to the stereo content before its presentation, e.g. viewing comfort driven disparity compression. We describe how disparity manipulating algorithms can affect perceived motion in depth and suggest a method for detection and correction of such distortion.

9011-39, Session 10

Measurement of perceived stereoscopic sensation through disparity metrics and compositions

Satoshi Toyosawa, Tokuyama Univ. (Japan); Takashi Kawai, Waseda Univ. (Japan)

CONTEXT:

Disparity is widely used as a principle measure evaluating discomfort, influence to artifacts, or movie production styles. Still, no agreement is made on what disparity based features are most effective describing stereoscopic sensation perceived from images in general.

OBJECTIVE:

The current study examines correlations between mean scores of subjective depth sensation and a various disparity based features. The features include conventional statistics such as average and standard deviation, percentiles from disparity histogram, contrasts, and width: They are extracted from disparity maps of each images presented. The contrast is defined as the difference between disparities from anterior and posterior regions, and represents overall anteroposterior range. The width is similar to the contrast but is computed from the posterior disparity of the back end object and anterior disparity of the front object.

It is expected that the level of sensation would differ depending on how pixels with different disparities are distributed. We therefore examine effects of the disparity distributions to the correlations. Here, the distribution is categorized as two classes based on a number of peaks in disparity histogram: single-peak and multiple-peaks distributions.

METHOD:

Fourteen images from two commercial 3D movies were presented to 15 subjects through an anaglyph stereoscope. Each subject answered perceived sense of depth magnitude and depth range in the 7 points Likert scale for each image. The correlation coefficients between the mean scores and each disparity feature were calculated.

RESULTS:

The result showed that the following features correlated significantly to the perceived depth magnitude: minimum (crossed), the contrast (95 percentile minus 5 percentile), and the width for the single-peak distribution, and maximum and 95 percentile (uncrossed) for the multiple-peaks distribution. The following features correlated significantly to the perceived depth range for the multiple-peaks distribution: the width. No significant feature was found for the single-peak distribution for the perceived depth range.

The result suggests that the measures for the perceived depth sensation should be adaptively chosen depending on the disparity distribution. The difference could come from viewers' different style in observing 3D contents as previous studies have suggested. In order to evaluate depth magnitude, the viewers tend to focus more on both the anterior regions and anteroposterior range when one foreground object is dominant. On the other hand, they concentrate their attention to the posterior regions when multiple objects exist. The viewers do not make judgment on anteroposterior space for one dominant object, but they rely on the range between the nearest and furthest objects for multiple objects, rather than simply checking the disparities at the nearest or furthest regions.

NOVELTY:

The previous studies have employed different disparity features to evaluate effects of binocular stereoscopy, but rational for their use has not clearly explained. The work by Sohn et al. (2012) has shown the relation between object thickness (similar to our width) and visual discomfort, but it did not discuss other possible features. The present study provides a guideline for what features to be used for depth sensation measurement.

9011-40, Session 10

Stereo and motion cues effect on depth judgment of volumetric data

Isaac Cho, Zachary J. Wartell, Wenwen Dou, Xiaoyu Wang, William Ribarsky, The Univ. of North Carolina at Charlotte (United States)

CONTEXT: Displays supporting stereoscopic and head-coupled motion parallax can enhance human perception of 3D surfaces and 3D networks but less for so volumetric data. Volumetric data is characterized by a heavy presence of transparency, occlusion and highly ambiguous spatial structure. There are many different rendering and visualization algorithms and interactive techniques that enhance perception of volume data and these techniques' effectiveness have been evaluated. However, how VR display technologies affect perception of volume data is less well studied.

OBJECTIVE: Our goal of this paper is finding effects of VR technologies (stereoscopic and head-coupled display) for a person's correct perception of depth ordering of volumetric objects. We examine the effect of stereoscopy and structure-from-motion on depth discrimination

and depth ordering tasks for a volumetric dataset.

METHOD: Our two experiments examine the effect of stereoscopy and/or head-tracking on the perception of volumetric data. Experiment 1, ”depth discrimination”, has four display conditions stereoscopy (2) x structure-from-motion (2), and Experiment 2, “depth ordering”, has six, stereoscopy (2) x structure-from-motion (3). We use a within-subject design with repeated measure. Each subject is randomly assigned a sequence of display conditions using Latin squares. The measures in our experiment are response answering time and error rate.

RESULTS: Results of a 2-way repeated measures ANOVA show that there were no interaction effects of stereo and motion on error rate in both experiments. However, we found that stereoscopy by itself improves depth perception in both depth ordering and depth discrimination tasks and head-tracking by itself helps only in the depth ordering task. In addition, 1-way repeated measures ANOVA shows the stereo plus motion condition is the most effective to reduce error rate in both experiments.

NOVELTY: In a prior shorter presentation of this study (), we presented the results of a simpler 2-way ANOVA analysis (display condition x subject pool: 4 x 2) which found an interaction effect of subject pool which partly correlated with gaming experience. Here we present a more complete and proper 3-way repeated-measures analysis, subject pool x stereo condition x motion condition (2 x 2 x 3), which shows no interaction with subject pool and a significant effect on error rate over all participants. We present a more elaborate comparison to previous studies with non-volumetric datasets. Those studies found motion alone showed greater advantage than stereo alone. In contrast, for our volume datasets the motion alone conditions did not demonstrate significant improvement over the stereo alone conditions. We compare our results to previous work by Ware et al.’s (http://dl.acm.org/citation.cfm?id=234975, http://dl.acm.org/citation.cfm?id=1279642) on 3D networks and discuss possible reasons for and implications of the different outcomes.

9011-51, Session PTues

Practical resolution requirements of measurement instruments for precise characterization of auto-stereoscopic 3D displays

Pierre M. Boher, Thierry Leroux, Véronique Collomb-Patton, Thibault Bignon, ELDIM (France)

Many papers have been devoted to the optical characterization of 3D displays these last years. Each class of 3D display provides stereoscopy in the eyes of observer with different means and so the requirements in terms of characterization are different. ELDIM has proposed in 2009 a new Fourier optics viewing angle measurement system with ultra-high angular resolution for the characterization of auto-stereoscopic 3D displays (1). Spectral polarization analysis with viewing angle instrument has also been proposed for passive glass 3D displays in 2010 (2). Temporal analysis of active glass 3D displays has been presented in 2011 (3).

In the present paper we are interested again in auto-stereoscopic 3D displays that are certainly the more demanding for precise optical characterization because of their very strong angular variation in the light emission. Different recently published papers present results obtained with standard instruments (like imaging luminance meters) that cannot be representative of what will be seen by the observer simply because these instruments are not optically equivalent to the human eye.

In the present paper, we are interested by this aspect of the optical measurement instruments that has not really being explorer presently and in particular by the impact of the optical resolution on the stereoscopic vision and the comparison to the human eye properties. Using viewing angle and imaging measurements made on various auto-stereoscopic displays with different size and working distances, we emphasize the effect of the resolution and show that results and in particular the 3D crosstalk can be strongly influenced by these parameters. It is true for viewing angle measurements at one location on the display if the angular resolution is higher than the resolution of the observer eye at the working distance. It is also true for imaging measurements made from one location for the entire display surface if the entrance iris of the instrument is larger than the human eye iris. Specific requirements for the optical measurement instruments with regards to display size and optimum working distance are derived and illustrated by measurement examples made in real conditions.

(1) “A new to characterize autostereoscopic 3D displays using Fourier optics instrument”, Boher, P., Leroux, T., Bignon, T., Collomb-Patton, V., , SPIE 7237, 37 (2009)

(2) ”Multispectral polarization viewing angle analysis of circular polarized stereoscopic 3D displays”, Boher, P., Leroux, T., Bignon, T., Collomb-Patton, V., SPIE 7524, 26 (2010)

(3) “Optical Characterization of Shutter Glasses Stereoscopic 3D displays”, Boher, P., Leroux, Collomb-Patton, T., Bignon, T., SPIE, 7863, 786312-1 (2011)

9011-52, Session PTues

Stereoscopic model for depth-fused 3D (DFD) display

Hirotsugu Yamamoto, Hiroshi Sonobe, Atsuhiro Tsunakawa, Junnosuke Kawakami, Shiro Suyama, Univ. of Tokushima (Japan)

CONTEXT: There are many reports on 3D perception based on Depth-Fused 3-D (DFD), including depth modulation with luminance ratio and protruding depth perception. However, no theoretical model has been presented to explain the depth perception characteristics.

OBJECTIVE: The purpose of this paper is to construct a theoretical stereoscopic model for DFD display that explains the continuous depth modulation and protruding depth perception.

METHOD: The model is composed of four steps: (0) preparation of DFD images, (1) geometrical calculation of viewed images, (2) human visual function for detecting intensity changes, and (3) stereoscopic depth perception. (0) Two types of displayed images for DFD display are prepared: the former pairs are for conventional DFD, where a fused image is located between the layered images; the latter pairs are for protruding DFD, where a fused image is located closer than the foreground image or further than the background image. A former pair is composed of a square in the background layer and a square with a reduced size in the foreground layer. The size is determined by the gap between the two layers and the viewing distance in order to match the apparent sizes of the images. The locations of the images are adjusted so that the apparent images overlap at the middle pint of the both eyes. A latter pair is composed of three squares in a layer and two rectangles in the other layer. The apparent images of the central square and the two rectangles are seamlessly connected when viewed at the middle point of both eye positions. Note that the former pair and the latter pair give different intensity distributions around the edges of the viewed images. (1) Viewed images at both eye positions are simulated based on geometrical optics. (2) In order to detect intensity changes, we have utilized Laplacian operation on a Gaussian blurred image, which was presented by Marr [David Marr, Vision (W. H. Freeman and Company, 1982)]. (3) Stereoscopic depths are calculated by matching the Laplacian operated images. Theoretical curves are compared to experimental results on conventional DFD and protruding DFD. Experiments on the conventional DFD are conducted by matching the distance of a reference image to the perceived distance of the fused image of two-layered images. Experiments on the protruding DFD are conducted in a caliper method [K. Sadakuni, et al., IEICE Trans. on Electronics, E95.C (2012) 1707, http://dx.doi.org/10.1587/transele.E95.C.1707].

RESULTS: Theoretical curves are calculated with programs based on OpenCV. By adjusting standard deviation of the Gaussian blur function, theoretical curves agreed with the experimental results. It is revealed that our stereoscopic model explains both conventional and protruding DFDs.

NOVELTY: DFD was originally invented by Suyama, who is the last author of this paper [S. Suyama, et al., Vision Research, 44 (2004) 785. http://dx.doi.org/10.1016/j.visres.2003.10.023]. Although there are many reports on experimental results of perceived depths and developments of hardware setups, there is no paper that provides theoretical framework on DFD perception. This is the first paper on stereoscopic model on DFD.

9011-53, Session PTues

Parallax multi-viewer auto-stereoscopic three-dimensional display

Lingdao Sha, Dan Schonfeld, Qun Li, Univ. of Illinois at Chicago (United States)

CONTEXT:

Currently, most products of either one-view or multi-view auto-stereoscopic (AS) three-dimensional (3D) display impose strict constraints on the viewing position and direction of the observers, which greatly reduces viewing comfort and flexibility. A breakthrough has been made by Perlin et al. who propose a real-time AS display system and algorithm with dynamic barrier that supports moving parallax of single mobile viewer. However, the application of the algorithm is quite narrow due to the limitation of single user.

OBJECTIVE:

We aim to extend Perlin's algorithm to handle multiple viewer case. To be specific, we focus our research on a display adjustment algorithm that enables high-quality AS display delivered to multiple viewers at various locations and with various orientations relative to the display.

METHOD:

We propose a parallax multi-viewer AS 3D display adjustment algorithm that combines both spatial multiplexing and temporal multiplexing to channel desired stereo pair to corresponding viewers according to their locations. The implementation of the algorithm requires an advanced parallax barrier that dynamically changes its strip width and directions in response to viewers' motion. To accommodate more users with a wider range of movement, the stripe width should be large enough. At the same time, however, the stripe should be unnoticeable so as to achieve high image resolution. To accomplish this, we rapidly animate them in a direction parallel to the viewing direction.

Assume the number of observers is M. At one time instant, the left eye of an observer sees 1/(2M) of the screen that delivers the left-eye image while the right eye seems nothing. At the next time instant, with the stripe flipped properly and screen display changed simultaneously, the right eye of the observer would see 1/(2M) of the screen that delivers the right-eye image. If the flipping is done fast enough, that is, M times faster than the time threshold before the eyes can notice the change, each eye would perceive its corresponding image in the stereo pair resulting in stereo perception. For simplicity, we show one-observer case to illustrate the dynamic shutter and screen adjustment scheme (see algorithm table in additional file).

RESULTS:

To visually demonstrate the effectiveness of the proposed algorithm without implementation of the dynamic barrier, we simulate AS stereo perception by red and cyan anaglyph stereo perception whose stereo pair can be easily channeled with red and cyan glasses. For simplicity, we assume two viewers at two different locations. At each time instant, we show the adjusted display with spatial multiplexing based on current user locations. We repeat the procedure multiple times with varying viewer locations to simulate time multiplexing. Preliminary testing results show that users are able to perceive the corresponding undistorted stereo scene correctly with the adjusted display.

NOVELTY:

To our best knowledge, this is the first time in literature that a novel system and display adjustment algorithm is proposed to enable high-quality AS display for multiple mobile viewers. The proposed method relies on spatio-temporal parallax barrier to channel desired stereo pair to corresponding viewers according to their locations.

9011-54, Session PTues

Floating volumetric display using an imaging element that consists of a 90º prism sheet and a linear Fresnel lens

Yuki Maeda, Daisuke Miyazaki, Osaka City Univ. (Japan); Satoshi Maekawa, Univ. of Hyogo (Japan); Takaaki Mukai, Osaka City Univ. (Japan)

CONTEXT

We have developed floating volumetric display systems [1,2] as shown in Fig. 1. In these studies, the size of a three-dimensional (3D) image was enlarged by widening a scanning range and an aperture of an optical scanner. A dihedral corner reflector array (DCRA) [3] was used as a distortion-free imaging element. The structure of the DCRA is two-dimensional array of square holes of approximately 100 [?m] on a side whose sidewalls are planar mirrors. Therefore, it is difficult to make a large DCRA without substantial cost. To display a larger image, an imaging element that has a large aperture was required.

OBJECTIVE

The aim of this study is to display a large floating 3D image by use of a novel imaging element that has a large aperture. We propose to use an imaging element consisting of a 90º prism sheet and a linear Fresnel lens.

METHOD

Figure 2 shows the basis of image forming by the proposed imaging element. The 90º prism sheet works as an one-dimensional array of dihedral corner reflectors, so that a diffuse ray in a transverse direction, which is shown as x axis in Fig. 2, converges at a point as a result of twice total internal reflection. On the other hand, a diffuse ray in a longitudinal direction, which is shown as z axis in Fig. 2, does not converge by the 90º prism sheet. To converge the diffuse ray in a longitudinal direction, a linear Fresnel lens is laid on the 90º prism sheet. A formed image by the proposed imaging element does not distort in the transverse direction because the ray in the transverse direction converges only by planar reflective surface.

RESULTS

Figure 3 shows a floating two-dimensional (2D) image formed by the proposed imaging element. A white bar was put between the 90º prism sheet and the linear Fresnel lens as a mark of a level plane of the imaging element. From motion parallax shown in Fig. 3(a) and 3(b), we confirmed that the 2D image of a blue flower was a floating image. The aperture area of the proposed imaging element was 20 [cm] ? 15 [cm], which was approximately twice larger than the DCRA used in our previous study [2]. A further large imaging element is able to make at low cost because of its simple structure. The image was formed at approximately 10 [cm] in front of the surface of the imaging element.

We are now improving the imaging element by designing a linear Fresnel lens to widen the aperture. And also, we are developing a volumetric 3D display including the proposed imaging element in place of the DCRA. We will use a rotational optical scanning method that is constructed in our previous study [2].

NOVELTY

The proposed imaging element can display a larger floating image at lower cost than the DCRA without distortion in the transverse direction. Including the proposed imaging element in our volumetric display system can display a large floating 3D image.

9011-56, Session PTues

A rendering approach for stereoscopic web pages

Jianlong Zhang, Wenmin Wang, Ronggang Wang, Peking Univ. Shenzhen Graduate School (China); Qinshui Chen, Shenzhen Graduate School of Peking University (China)

CONTEXT: Web technology provides a relatively easy way to generate contents for us to recognize the world, and with the development of stereoscopic display technology, the stereoscopic devices will become much more popular. The combination of web technology and stereoscopic display technology will bring revolutionary visual effect. OBJECTIVE: We think, the Stereoscopic 3D (S3D) web pages, in which text, image and video may have different depth, can be displayed on stereoscopic display devices. So that the approach about how to render two view S3D web pages should be provided: first, algorithm should be developed in order to display stereoscopic elements like text, rectangles by using 2D graphic library; second, a method should be presented to render stereoscopic web page based on current framework of the browser; third, a solution should be invented to fix some rendering problems.

METHOD: 2D elements with depth information should be projected in screen to create left view frame and right view frame. The paper deduces a simple matrix that helps transform the 2D elements into S3D ones. We create two frame buffers for left view frame and right view frame, apply the matrix to the boxes to render the two frames, and fuse them into a stereoscopic frame on the particular display at last. For the problem of box with large depth covering the box with small one, the paper presents the solution that the boxes which have same depth information are layered to the same layer and layers are rendered from back to front.

RESULTS: Webkit as the tool of our experiment is an open-source engine to allow browsers to render web page. Through adding new CSS label like 'depth', we assigned different depth to the boxes by writing HTML web pages with different value of depth label. In the result, the matrix deduced is successfully applied to rendering S3D elements and we implement a browser to support rendering S3D web page in the way. It demonstrates that the stereoscopic web page can bring very exiting visual effect.

NOVELTY: There exists a lot of research and applications about web-based 3D virtual technology, but has no applications about rendering stereoscopic web pages. Based on the S3D web pages, many stereoscopic games or other stereoscopic applications can be developed. The matrix deduced is fast to implement S3D rendering by current 2D graphic library, and the method presented is easy and successful to render S3D web pages in the current framework of the browser.

9011-57, Session PTues

The rendering context for stereoscopic 3D web

Qinshui Chen, Wenmin Wang, Ronggang Wang, Peking Univ. Shenzhen Graduate School (China)

CONTEXT: 3D technologies on the Web has been studied for many years, with popular solutions like X3D, O3D, 3DMLW and X3DOM, but these technologies are basically monoscopic 3D – 3D graphics painted on a 2D screen. With the stereoscopic technology gradually maturing, we are researching to integrate the binocular 3D technology into the Web, creating a stereoscopic 3D browser that will provide users with a brand new experience of human-computer interaction.

OBJECTIVE: In our project, we seek to apply stereoscopic 3D technology into web pages for a switch-over from current 2D web mode to 3D. Different elements of a web page can be assigned with different depth values; for each page our browser will create two offset images separately to the left and right eye of the viewer, which will be combined on the binocular 3D screen to generate the illusion of depth. From the user's viewpoint, some elements of the page appear to be sinking into the screen, some sticking on it, some even popping out of it. We also want to provide backward compatibility to current web standards, which means existing web pages can be correctly displayed in our browser without changing anything.

METHOD: To achieve binocular stereoscopic effects of a web page on a 3D screen, we need to extend current web standards to allow depth description and 3D interaction. New HTML tags are added for web site authors to specify different contents for the left-eye and right-eye frames; new CSS properties are added to assign depth values for elements of HTML pages; global objects and APIs in JavaScript are also extended, allowing users to query 3D screen information, and to interact with 3D pages. Implementing our system from scratch is too big a project, so we take the advantage of WebKit, which is open source and is one of the most popular browser rendering engines.

RESULTS: The flow of WebKit rending a web page includes the following steps: document parsing, DOM tree constructing, render tree constructing, layout and painting. To support new HTML tags and CSS properties, we have modified the document parsing and DOM tree constructing procedures. In addition to the original frame buffer in the painting procedure, we add another one so that it now has two frames for each eye. Moreover, as the depth of a box can change its final rendering position and size, we also do some minor changes to WebKit's rendering process. Finally, when two frames are combined together on the 3D screen, different elements of the displayed page can be seen floating in the air.

NOVELTY: The key aspect of our contribution is that we introduced stereoscopic 3D into the Web. We also built the 3D display ability into the DOM tree and rendering flow of the web browser, so rendering a 3D page is very natural, efficient and convenient. One more important feature is that our model is compatible with the W3C web standards.

9011-58, Session PTues

The development and future of 3D render technology for digital cinema

Darren Ma, Leonis Cinema (Beijing) Tech. Co., Ltd. (China)

Light efficiency is very important in 3D digital cinema render, but in fact the light efficiency of current 3D system (active, color filter and polarization modulator) is too low to satisfy the requirement of keeping enough color render and image contrast, and image resolution. A light recycle or light doubler implementation is proposed in this article, for both single projection system and dual projection system. All theatrical and tested data (light efficiency, contrast, pixel overlap, and color shift) of two type of light doubler are provide and analyzed. The improvement of light efficiency is between 50-80% over normal polarization modulator. We also analysis the disadvantage and advantage of several implementation ways of light doubler, such as embedded and independent light doubler, symtrical and non-symtrical light doubler. We also discuss the 3D technology for laser projection, such as polarization keep technology, light double for laser projector, and color filter technology of laser projector. And we think that color filter is best technology for laser projection based on analysis of active, color filter and polarization 3D technology.

9011-59, Session PTues

The design and implementation of stereoscopic 3D scalable vector graphics based on WebKit

Zhongxin Liu, Wenmin Wang, Ronggang Wang, Peking Univ. Shenzhen Graduate School (China)

CONTEXT: Scalable Vector Graphics (SVG), which can be embedded in HTML5, provides the ability to describe 2D-graphics and graphical applications, rather than 3D-graphics. We extended SVG to support Stereoscopic 3D graphics (S3D), which has not been done in previous works.

OBJECTIVE: We modified the SVG module of WebKit, which is a widely used open source rendering engine, to support stereoscopic 3D graphics, and transplanted it into S3D display. The modified WebKit is

able to convert both the 2D SVG shapes and Bezier curves into S3D mode and display them on S3D monitors, with a feeling of depth and thickness.

METHOD: We define the outside of a three dimensional figure as the front face and the inside as the back face. Without loss of generality, the original 2D SVG shapes are assumed to be the front face of 3D figures. First, the left graphic and right graphic are generated according to the original shape. Resolution of the screen is got from certain API. The distance between eyes is set to be 6.5 cm, and users' viewpoint is assigned to be at a distance from the center of screen. Secondly, a vanishing point can be assigned, which is assumed to be a point far enough from the center of screen by default. In order to convert an SVG shape into 3D mode, we calculate the corresponding shapes on the back face according to the front face shapes. Shapes seem to converge to the vanishing point with a ratio determined by distance from vanishing point to the front face as well as a thickness assigned. The shapes on back face of left and right view are calculated separately, to obtain a proper disparity and a feeling of thickness. In this way, back face shapes seem to be farther than front face shapes. Next, it is known that Bezier curves are obtained from certain control points with interpolation. Converting Bezier curves to S3D mode is equivalent to determining these control points on the back face in a similar way described above. Curves on the back face can be generated from these new points with interpolation.

RESULTS: Basic SVG shapes, such as circle, rectangle, polygon, line and ellipse, as well as Bezier curves, are converted into 3D graphics by our modified rendering engine and shown up on stereoscopic 3D display devices, in the case that they are contained in webpages. With a certain thickness, these graphics seem to be real objects rather than simple lines or curves. Meanwhile, as they are shown in stereoscopic mode, graphics with different depths are presented in different layers.

NOVELTY: We put forward idea of supporting stereoscopic 3D SVG, which displays graphics more vividly than previous work. Furthermore, with the knowledge of binocular stereo vision, we gave an approach to displaying SVG graphics on stereoscopic 3D screens, modified the SVG module of WebKit to provide 2D to S3D conversion of SVG and applied the modified rendering engine to stereoscopic display.

9011-60, Session PTues

Joint estimation of high resolution images and depth maps from light field cameras

Kazuki Ohashi, Keita Takahashi, Toshiaki Fujii, Nagoya Univ. (Japan)

Light field cameras (plenoptic cameras) are attracting much attention as tools for acquiring 3D information of a scene through a single camera. Several kinds of light field cameras have been developed and some of them are commercially available. Applications of light field cameras include digital refocusing, 3D reconstruction, and free-viewpoint image synthesis.

The main drawback of typical lenlets-based light field cameras is the limited resolution. This limitation comes from the structure of the camera where a microlens array is inserted between the sensor and the main lens. The microlens array projects 4D light field on a single 2D image sensor at the sacrifice of the resolution. The number of microlens corresponds to the angular resolution, and the number of pixels behind each microlens corresponds to the position resolution. Thus, the angular resolution and the position resolution trade-off under the fixed resolution of the image sensor. This fundamental trade-off remains after the raw light field image is converted to a set of sub-aperture images. These sub-aperture images can be regarded as a set of rectangular images whose viewpoints are arranged on a 2D surface corresponding to the aperture of the light field camera, and are used as basic units for further processing.

The purpose of our study is to estimate image and depth information with a higher resolution from low resolution sub-aperture images obtained by a light field camera. Specifically, we adopt super-resolution reconstruction to obtain a high resolution image from many sub-aperture images. In this reconstruction, these sub-aperture images should be registered as accurately as possible. This registration is equivalent to depth estimation, and thus, the depth information is also desired to be with a high resolution as much as possible. Meanwhile, depth information is originally estimated from the sub-aperture images by stereo matching, and the initial depth resolution is limited by the resolution of the sub-aperture images. Therefore, increasing the image resolution and increasing the depth resolution are two sides of the same coin. This intuition motivated us to develop a joint estimation method in which super-resolution and depth refinement are performed alternatively.

We evaluated our method by using a commercially available Lytro light field camera whose resolution is 3280 by 3280 pixels. From the raw light field image we obtained sub-aperture images with 330 by 380 pixels from 50 by 50 viewpoints. Using the proposed method, we increased by three times horizontally and vertically the resolution of the sub-aperture image and the depth map. Compared to the case without depth refinement, we can obtain not only the more accurate depth map but also the clearer image.

9011-61, Session PTues

Discontinuity preserving depth estimation using distance transform

Woo-Seok Jang, Yo-Sung Ho, Gwangju Institute of Science and Technology (Korea, Republic of)

CONTEXT

Recently, image generation methods for arbitrary view positions have been vital due to the development of multi-view display devices and three-dimensional (3D) contents. Accurate depth information is necessary for efficient image generation.

OBJECTIVE

The objective of this paper is to generate an accurate depth map using a stereo image. Over the past several decades, a variety of stereo-image-based depth estimation methods have been developed to obtain high-quality depth maps. However, accurate measurement of depth information from natural scene still remains problematic due to difficult correspondence matching in some regions. Especially, in case of the depth discontinuous region, i.e., the edge region, unclear color values exist, which lead to ineffective of corresponding matching. Thus, in this work, we propose a discontinuity preserving depth estimation method to solve the problems. Distance transform (DT) calculate the distance to the closest edge for each pixel of an input image. By controlling the color weighting term using DT values of left and right images, we carry out better correspondence matching in edge regions.

METHOD

The proposed method is initially motivated by hierarchical belief propagation. Due to the hierarchical structure, the previous work computes depth information accurately in the textureless region. Based on this work, the proposed method adds an algorithm to improve depth qualities in the edge region using distance transform. The algorithm is implemented as follows: 1) Distance transform (DT) is performed on the left and right images. Prior to the distance transform, the canny edge operator is used for extraction of color edge map from the image. Note that isolated edges are ignored by applying a median filter to input data prior to edge detection. 2) DT-based weighting function is computed. 3) Color weighting function is calculated. These weighting functions control the matching cost for better depth estimation in the edge region. 4) Block-based stereo depth estimation is carried out based on such weighting functions. Finally, we pursue improvement of depth information using post-filtering.

RESULTS

In order to evaluate the performance of the proposed method, we tested with stereo data sets with different image size. We compare our proposed method with other methods which have good performance. We objectively evaluate results by measuring the percentages of

bad matching pixels. The results indicate that the proposed method outperforms other comparative methods. Furthermore, the visual comparison of the experimental results demonstrates that the proposed stereo-image-based depth estimation method improve the quality in edge regions.

NOVELTY

This paper presents a new depth estimation method exploiting edge information. In general, pixel intensities around object edges are not clear due to mixed values located between the object and background. This causes problems when identifying discontinuous depth in object borders. In order to handle this, the proposed method generates depth information based on distance transform. This increases the depth quality at edges. Using distance transform produces better performance compared to the previous methods.

9011-62, Session PTues

Integrating multiview camera arrays into adaptive multiview laser scan displays

Leif Arne Rønningen, Norwegian Univ. of Science and Technology (Norway)

The scientific topic is about how to integrate a multi-view sparse aperture camera array into a multi-view laser scan based display array, and the extremely fast image processing needed. The technique is used to establish a physical unit that in turn can be a building block for seamless, arbitrary collaboration spaces for future telepresence.

Integrating the camera- and display arrays, is a major challenge, the plan is to apply arrays of GRIN lenses. However, the first experiments are carried out by simulating a sparse aperture camera using a high-quality camera with a large lens and sparse apertures in front, and simulating parallel hardware solutions in Matlab. Since laser scan is used for the display (65 x 65 mm), camera lenses have to be distributed on the boarders. Optimal distributions have been found, maximizing the contrast and resolution of images. Separate image capture for luminance and chrominance and following processing of sequences of images enhance contrast, resolution, color-quality and signal/noise ratios, especially at low-light conditions. The camera lenses are nearly invisible, and the views can cover any position and pose in the space. With superb video quality the collaboration space may generate data rate transients of up to 30 Terabits/second. This is reduced by adaptive views and intelligent dropping of content. Lighting of the scene, with all surfaces being collaborative surfaces is provided by display lasers not generating visible image views. The image processing involves segmentation of objects, object recognition and tracking, restoration (e.g., Wiener filtering), image stitching, registration and combining. Communication between building blocks is solved partly by implementing intelligent content drop functions, direct integrated circuit communication, and externally by PCIe communication.

The concepts proposed enhance the present 3D (two views) to near-natural multi-view/ continuous-view, and provided that the collaboration space- and network delays are reduced correspondingly, real telepresence feel can be obtained. The time delay from an optical event occurs in a collaboration space until the vision data is presented on a screen in another collaboration space can be less than 10 milliseconds plus propagation delay. This is important to many applications (such as telesurgery, musical conduction, games, VR, off-shore integrated operations, sports and scenography). The concepts open up new business opportunities and usages, enhanced services can be part of most offices and homes in say ten years.

GRIN lens arrays will be produced exposing diffusion-driven photopolymers to a laser, and this enables low-cost individual characterization of each lens, a prerequisite for the concept. Low latency and parallel heterogeneous processing, rather than minimizing the bit-rate, is emphasized.

The new concepts enable near-natural multi-view telepresence for the first time, and the total perceived quality are expected to outperform existing solutions. Examples of state-of-the-art solutions are telepresence collaborative spaces with a few 2D screens and 200 milliseconds end-to-end delay, and static, non-seamless, 2D OLED micro-displays combining display and image sensor functionality on a single CMOS IC. Near-natural quality- and seamless collaboration spaces of virtually any shape and size providing views from most angles, can be built by combining building blocks and interconnect directly or via collaboration space nodes.

9011-63, Session PTues

View synthesis from wide-baseline views using occlusion aware estimation of large disparities

Ahmed S. Elliethy, Hussein A. Aly, Military Technical College (Egypt); Gaurav Sharma, Univ. of Rochester (United States)

CONTEXT:

Viewer navigation in image-based virtual environments is gaining attention in many applications such as rendering of images for multi-view 3D displays [1], free viewpoint video, and multi-view compression. However, this navigation which requires a continuous generation of a virtual view from a set of wide-baseline (sparse) scene images is challenging, especially when there is occlusions due to multi-depth.

OBJECTIVE:

We present a new method to get an accurate novel synthetic view from four input views with large disparity among them and with possible occlusions. An example is depicted in figure 2, where given the capture image I1, I2, I3, I4 from different views and it is required to synthesize a novel view J between I2 and I3. This is different of the previous methods which either does not provide satisfactory results when dealing with large disparities, such as [3] or can deal with large disparities but without sufficient occlusion handling, such as [2].

METHOD:

The method has 3 stages:

Stage 1 “occlusion estimation at the interpolated view”: this is done by estimating the disparity fields from I1 towards I2 using forward Gabor transform based disparity estimation to be able to deal with large disparities, and then refining it using anisotropic regularization to enforce coherence; similar process is done from I4 towards I3. Both estimated disparity fields are used to estimate a visibility map which is used to guide stages 2 and 3. The idea behind estimating the visibility map is to get pixel locations in the interpolated view which have the number of accumulated visits of the forward compensated disparity vectors originating from a far view (I1 or I4) less than a predefined threshold and these pixels are identified as invisible from that view.

Stage 2 “estimate the backward disparity field” at the interpolated view using the proposed bi-lateral and uni-Lateral backward Gabor transform (figure 4), then refine it to obtain a coarse estimate of the interpolated view. This coarse estimate is used to guide the anisotropic regularization to obtain the final accurate disparity field.

Stage 3 “synthesize the interpolated view” using the obtained disparity field as well as the four input images with the aid of the proposed enhanced visibility map.

RESULTS:

We present results of our method for 3 test sequences: Flower-Garden , Lab [4], Ballroom [5], and a synthetic sequence. We simulated the sparse views by selecting images from the test sequences that are temporally far from each other. Figure 4 shows that our method reconstructs the novel view with high quality, specifically the occluded regions, quantitative results are presented in table 1 compared to the ground truth images using 2 metrics PSNR and SSIM [6].

We present experiments for a different range of disparities of our method compared to other methods in table 2. We found that for small disparities our method performs slightly better than method that don’t

take into consideration a good initial estimate of the disparity map in minimizing the objective function as in [3]. However, as the disparities get larger, our method performs much better as it make use of such a good estimate, this accounted for the proposed backward Gabor filter disparity estimation along with the enhanced visibility map.

NOVELTY:

1 We introduced enhanced visibility map estimation at the interpolated view to correctly guide the interpolation process to select the appropriate image pair used for the pre-pixel interpolation.

2 We introduced uni-lateral and bi-lateral Gabor filters that search based on backward disparity estimation.

References

[1]C. Fehn “Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3DTV”, SPIE USA, Vol. 5291, pp. 93-104, 2004.

[2] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, M. Tanimoto, “View generation with 3D warping using depth information for FTV”, Signal Processing: Image Communication, Vol. 24, pp. 65-72, January 2009.

[3] A.Radhakrishna, “SLIC Superpixels Compared to State-of-the-art Superpixel Methods”, IEEE Transactions, on Pattern Analysis and Machine Intelligence, Vol.34, pp. 2274-2282, May 2012.

9011-64, Session PTues

Superpixel-based 3D warping using view plus depth data from multiple viewpoints

Tomoyuki Tezuka, Keita Takahashi, Toshiaki Fujii, Nagoya Univ. (Japan)

Free viewpoint video is a technology that enables users to watch 3D objects from any viewpoint they want, and it has attracted much research interest recently. One of the common representations is view plus depth format, where a 2D color image and a depth map are provided from the same viewpoint. Such depth data can be obtained directly by using 3D cameras such as Kinect sensors, or by applying stereo matching to two or more viewpoints’ images. The focus of our research is how to visualize view plus depth data.

More specially, we want to see the object that is represented as a color image and depth map from different freely-chosen viewpoints. Such visualization is achieved by technologies that are referred to as 3D warping or depth image based rendering (DIBR) [1]. However, the images generated by existing methods have many holes that are very annoying due to occlusions and the limited sampling density. In [2], view plus depth data from the left and right cameras were successfully merged to generate virtual images from in-between viewpoints. However, their method was designed for view interpolation between the left and right cameras, but not for large viewpoint changes. In our work, we want to achieve larger viewpoint changes even from single set of view plus depth data.

3D points reconstructed from view plus depth data are isolated, i.e. not connected with each other. The gaps between these points appear as holes in the virtual viewpoint image. If the 3D points are rendered as properly connected ones, those holes can be inpainted. However, connecting all of the neighboring pixels is insufficient, because occlusion boundaries are kept vacant. To determine whether two pixels should be connected or not, we apply superpixel representation to the input image. Generally, a superpixel is a group of pixels that are gathered based on some distance metric. Typically, the distance is measured by color difference or position difference.?Among the various methods, we adopt SLIC [3] method, which is fast, memory efficient, and accurate, and we extended the original method to consider depth information. 3D points that fall into the same superpixel are connected, and the surfaces composed of those 3D points are inpainted using color interpolation. This inpainting drastically reduces holes in the virtual camera image even if the viewpoint moves largely from the real camera viewpoint. Meanwhile, 3D points of different superpixels are kept unconnected. Therefore, occlusion boundaries between foreground and background objects are kept vacant.

Our method performs reasonably even with single set of view plus depth data, but missing data from that viewpoint cannot completely interpolate. Therefore, we are investigating how to combine view plus depth data from multiple viewpoints to improve the quality of 3D visualization. We will report in this paper the latest experimental results using publicly available eight viewpoints data.

9011-65, Session PTues

Stereoscopic augmented reality with pseudo-realistic global illumination effects

Francois de Sorbier, Hideo Saito, Keio Univ. (Japan)

In this paper, we introduce an augmented reality framework for a credible integration of virtual objects in a real scene without markers. By credible, we mean that virtual objects are included with consideration of the surrounding environment and that we can render realistic global illumination effects like shadows. We also consider the rendering of this augmented scene from multiple slightly different viewpoints and describe how we can generate the images that can be required by an auto-stereoscopic display for instance.

The use of global illumination effects in augmented reality often needs the knowledge of the environment’s geometry which is difficult to obtain with a single color camera. As an example, common approaches usually ask for manually description of the surfaces, does not consider occlusions, requires fiducial markers or uses non real-time analysis methods that are not suitable in case of interactive augmented reality.

In our research, we take advantage of a Kinect-like depth camera for capturing the geometry of a scene with the corresponding color in real-time. However, such cameras suffer of noise, and many data are also often missing due to the specularity of the surfaces or the occlusions. We tackle this problem by performing a marker-free tracking of the viewpoint based on a real-time registration of the geometry. This allows us to improve the quality of the captured data over the time by integrating the previously captured data into the current view controlled by several criterions for managing sudden changes in the scene.

For global illumination effects like reflections and refractions, we need to access the environment surrounding the target scene. For this purpose, we take advantage of an environment map that is defined for each virtual object. For each captured frame, we use the estimated pose of the camera to update the environment map with the corresponding color and depth information. From the captured views, we also estimated the position of the light sources based on a segmentation of the saturated areas and considering their distance from the target scene.

Combining information from the environment map, the light sources and the camera tracking, we can display virtual objects with global illumination effects such as diffuse and mirror reflections, refraction and shadows in real time (about 20 frames per second) thanks to the GPU.

Since we can get accurate depth information based on the accumulation of data over the time, we can easily perform a real-time reconstruction of the scene. We can then render the scene from multiple virtual viewpoints and generate the different inputs images of an auto-stereoscopic display. We also demonstrate that our approach can be used with an HMD coupled with the depth camera.

9011-66, Session PTues

Development of free-viewpoint image synthesis system using time varying projection and spacetime stereo

Tatsuro Mori, Keita Takahashi, Toshiaki Fujii, Nagoya Univ. (Japan)

Stereoscopic image synthesis for 3D display using stereo or multi-view is researched actively and virtual view generation techniques have attracted much attention. To generate virtual view from multiple cameras, 3D model of objects is necessary to be reconstructed using multi-view images. As one of methods to realize that, on-the-fly estimation of view-dependent depth map has been researched. This method seems to be suitable for real-time synthesis because of unnecessary offline processing. Taguchi et al.[1] presented real-time free-viewpoint image synthesis system to estimate view-dependent depth map by implementing that estimation on a GPU using GPGPU(General-Purpose computation on GPUs) techniques. Now then, when using only passive sensor like a camera as input, the synthesis against textureless objects tends to be generally unstable and the problem leads to low quality virtual view. Global regularization methods for improving that are often employed but they require much computation time cost, so are unsuitable for real-time processing.

On the other hand, depth estimation methods introducing active sensor like a projector to processing have been traditionally researched. They can construct easily identifiable features by projecting texture into the scene and minimize the difficulty involved in determining correspondence against textureless objects. As one of them, Davis et al.[2] presented spacetime stereo reconstruction combining use of stereo camera and projector. They performed spatiotemporal matching under illumination varying overtime conditions such as projecting texture, and showed possibility to make depth estimation stable with small size of matching window against textureless object by introducing temporal matching under such conditions in addition to spatial one. And we also take notice of the fact that in this method although relatively simple estimation for matching is adopted, the depth is estimated well as a result. So, it is possible to adapt this method to real-time processing.

In this paper, we propose real-time free-viewpoint image synthesis system introducing spacetime stereo with projector. In the first step, active sensor projects random pattern texture into the scene. And when estimating view dependent depth map, we adopt spatiotemporal window matching making use of spatial and temporal features. This matching makes recovery of depth information against textureless areas robust and improves estimation accuracy using even small spatial window size although normally such size of window performs poorly for matching especially in regions with little texture. On the other hand, because of texture projection into the scene, a problem appears that cameras cannot acquire original texture of the scene. To solve this, projector sends a blank pattern every some frames, which is used to extract original texture of the scene, and we synthesize free-viewpoint image using that texture. We adopt these methods, and quality of virtual view is expected to increase relative to Taguchi et al.[1] while keeping real-time processing.

As a future work, we will develop this proposed synthesis system to GPU implementation appropriately for real-time estimation with multiple camera array (Currently, we are considering 16 camera array system). After that, we will relate our method with conventional ones and discuss the availability of proposed one.

References

[1]Yuichi Taguchi, Keita Takahashi, and Takashi Naemura,"Real-Time All-in-Focus Video-based Rendering Using a Network Camera Array",3D-TV-CON'08, pp. 241-244 ,May 28-30,2008.

[2]J.Davis, D.Nehab, R.Ramamoorthi, and S.Rusinkiewics,"Spacetime Stereo: A Unifying Framework for Depth from Triangulation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.27,No.2,pp.296~302,February 2005.

9011-67, Session PTues

General stereoscopic distortion rectification due to arbitrary viewer motion in binocular stereoscopic display

Qun Li, Dan Schonfeld, Univ. of Illinois at Chicago (United States)

CONTEXT: In binocular stereoscopic display, stereoscopic distortions due to viewer motion, such as depth distortion, shear distortion, and rotation distortion, result in misperception of the stereo content and reduce visual comfort dramatically. In the past, perceived depth distortion has been thoroughly addressed and depth adjustment techniques are now commonly available in most three-dimensional (3D) visual display products. Shear distortion has been discussed only in the context of multi-view display, where a different viewpoint of the stereo content is displayed according to the viewer position to accommodate motion parallax. However, few investigators have recognized the impact of rotation distortion as a consequence of viewer motion. As a result, no technique is available to address stereoscopic distortions due to general viewer motion.

OBJECTIVE: We discuss the mathematics and methodology to understand and rectify perceived stereoscopic distortions in binocular stereoscopic displays due to general user movement relative to the screen. In addition to depth distortions, shear and rotation distortions are also addressed at the same time. The paper aims at preserving an undistorted 3D perception of a stereo scene from a fixed perspective irrespective of viewing position.

METHOD: We propose a unified system and method that rectifies stereoscopic distortion due to general affine viewer motion and delivers a fixed perspective of the 3D scene irrespective of viewer motion. The proposed method encompasses perceived depth adjustment by disparity scaling as a special case. The system assumes eye tracking of the viewer and pixel-wisely adjusts the display location of the stereo video frame based on tracked viewer eye location. Specifically, for each point in the 3D scene, its coordinates in the viewer's left eye frame and right eye frame should both remain the same regardless of user motion in order to preserve stereo perception from a fixed perspective. Based on such geometry, the adjusted screen display location of each pixel for the stereo pair can be derived. Due to the limitation of the screen size, not all pixels are displayed after adjustment, hence linear interpolation is performed to smooth the stereo video frames.

RESULTS: For demonstration purpose, we implement our method on controlling perceived depth in binocular stereoscopic display of red and cyan anaglyph 3D. Anaglyph glasses are required in order to perceive the 3D scene. The user first perceives the designed perspective of the 3D scene at the reference position with 400mm distance to the monitor screen, centered both vertically and horizontally. The user then moves to 6 different positions with various distances and angles relative to the screen. At each position, the unadjusted image is first displayed followed by the adjusted image. At all positions, the tested users report to perceive a much more consistent scene with the adjusted display as that perceived at the reference position and at the same time, experience improved visual comfort.

NOVELTY: In the past, perceived depth distortion, shear distortion and barely rotation distortion have been discussed within the context of multi-view display to accommodate motion parallax. We first address such distortions with a different but also reasonable goal, that is, to maintain a fixed perspective of the stereo scene. In addition, we propose a unified solution that simultaneously rectifies the above stereoscopic distortions resulted from arbitrary viewer motion.

9011-68, Session PTues

Wide-field-of-view image pickup system for multiview volumetric 3D displays using multiple RGB-D cameras

Yuan Luo, Hideki Kakeya, Univ. of Tsukuba (Japan)

This paper proposes a real-time and wide-field-of-view image pickup system for coarse integral volumetric imaging (CIVI), which is a 3D display solution including multiview technology and volumetric imaging technology. This system is to apply CIVI display for live action videos generated by the real-time 3D reconstruction. Considering the system cost, RGB-D camera is chosen to be the solution to get depth data and color data from the real world. In order to have a complete surface of the objects and a wide field of views, multi RGB-D cameras from different directions should be working together at the same time. In this paper, the authors attain this goal by using two Kinect sensors.

One of the most important steps in this system is how to combine the reconstructed scene from different Kinects. Considering the daily use or the exhibition for the system, manual calibration in a traditional way is not fit for our purpose, for a fast and accurate calibration should be used in an automatic way. In this paper, we calculate the rotation and transformation matrix very quickly by using the checkerboard and the depth data.

The Proposed system works as follows. Firstly, a checkerboard should be put in front of the two Kinects and the system will calculate the conversion matrix automatically. Secondly, 3D point cloud data will obtained by each Kinect sensor and they are all converted into the same coordinate system. Thirdly, the multiview images are made by perspective transformation from different viewpoints. Finally the image for each viewpoint is divided depending on the depth of each pixel to realize volumetric image from each viewpoint. The prototype system based on the proposed method is constructed and it shows a better result than using only one Kinect. Many shadows behind the objects can be covered from another Kinect and a wider viewing angle can be shown in our CIVI displays. It is confirmed that the whole proposed system can achieve real-time 3D television system that can show not only the multiview volumetric 3D image but also a wide-field-of-view image on the real-time basis.

9011-69, Session PTues

Joint upsampling and noise reduction for real-time depth map enhancement

Kazuki Matsumoto, Chiyoung Song, Francois de Sorbier, Hideo Saito, Keio Univ. (Japan)

Recent advances in both computer vision technology and available computing power of ordinary systems brought an easier access to videos with high frame rate capturing not only the color information, but also the 3-dimensional physical geometry of the real world. These videos are getting more popularity among some of the on-going computer vision research fields including reconstructions and tracking. Depth cameras capture relatively reliable RGB-D video for casual applications; however, the raw depth information produced by depth camera suffers from too much noise to be used for more sophisticated projects, and therefore requires some degree of post-processing to increase the accuracy of the data. Moreover, the resolution of the depth map needs to be reduced for remotely operating applications requiring communications between host and client in order to reduce the bandwidth congestion. A prime example of such usage is the streaming of free-viewpoint video in realtime. While such downsampling, or lossy-compression is necessary to achieve the runtime efficiency, doing so inevitably reduces the amount of information available for the given video frame, as well as the trivial disagreement of RGB and depth resolutions. This unavoidable trade-off therefore calls for an efficient up-sampling algorithm for the sparse depth information to restore the dense and complex 3-dimensional geometry as much as possible, while having a minimum effect on overall runtime of the system.

We propose a full pipeline that upsamples the depth map produced by depth cameras, while also enhancing the quality of it by reducing the amount of structural noise which raw depth map from depth camera is heavily contaminated with. We first downsample the raw depth map D(u) where u is a pixel in an image plane, u = (u; v) ? R2 to obtain D0(u0) by subsampling D(u) around each neighborhood and averaging. Occluded pixels, or the pixels affected by the structural noise from D(u) are discart for averaging in order to reduce the effect of structural noises that appear as holes in the depth map, which results from random occlusions of pattern projected by Kinect. We then have D0(u0) go through a frame accumulation buffer for noise reduction. The buffer takes a weighted average of D0(u0) and previously-stored data G(u0) in the buffer, and stores the result back into G(u0) for future use. Then, the intensity image I(u) captured by Kinect at the same time is segmented to a number of clusters that groups pixels that have relatively small Cartesian distance with each other, with similar intensities. Under an assumption that the depth edges often coincide with sharp intensity changes, the segmentation of I(u) is used to partition D0(u0) to corresponding clusters whose depth values are expected to represent a planar structure. Therefore by fitting plane formula to each of depth cluster DCk, we gain a mathematical foundation to base the upsampling of D0(u0). However, our assumption of discontinuity similarities between depth and intensity images is not concrete, implying that some DCk might not depict a planar structure. To accommodate such clusters, a plane sanity check is performed while attempting the planar fitting. If a cluster is thought to be non-planar, we instead use upsampled G(u0), denoted as LIG(u) to the resolution of I(u) with simple linear interpolation.

As a result of our method, we can downsamples and upsamples the depth map captured by Microsoft Kinect by piecewise planar fitting approach based on higher-resolution RGB image stream, with dramatically reduced amount of noise while maintaining real-time requirements for aforementioned video streaming usage case, byusing GPGPU acceleration via nVidia CUDA architecture.

In the attached PDF file, the effect of the upsampling of depth image for generating free viewpoint image is demonstrated. Without our method, the number of the pixel of the input depth data is not sufficient, so a lot of holes are observed as shown in Fig.1 (c) and Fig. 2 (c). Using our method, we can upsample the depth data, so that we can successfully generate the free viewpoint images with smaller holes as shown in Fig.1 (d) and Fig. 2 (d).

9011-70, Session PTues

Experimental investigation of discomfort combination: towards visual discomfort prediction for stereoscopic videos

Seong-Il Lee, Yong Ju Jung, Hosik Sohn, Yong Man Ro, KAIST (Korea, Republic of)

CONTEXT

Recently, thanks to the unique viewing experience provided by stereoscopic three-dimensional (3D) displays, stereoscopic 3D services have surged. However, the issues of image safety (i.e., visual discomfort) associated with the viewing of stereoscopic images still remain as a possible obstacle to the widespread deployment of 3D video services. Thus, the development of an automatic prediction algorithm for visual discomfort is of a great importance to address the viewing safety issues in the viewing of stereoscopic images.

OBJECTIVE

In literature, disparity and motion characteristics have been known as major factors of visual discomfort caused by stereoscopic content. In particular, disparity characteristic of stereoscopic image content is known to affect the severity of accommodation-vergence conflict (AV conflict) [1], which is deemed one of fundamental discomfort factors for current stereoscopic displays. In general, AV conflict may worsen with larger screen disparities [2] (hereafter, we call the visual discomfort associated

with disparity characteristic as “disparity-induced discomfort”). In addition to disparity characteristic, visual discomfort may be affected by motion characteristic of stereoscopic video content. For instance, fast in-depth motions in stereoscopic content can lead to excessive changing demands on the AV linkage, which may induce visual discomfort [3] (hereafter, we call the visual discomfort associated with motion characteristic as “motion-induced discomfort”).

Inspired by the relationship among visual comfort, disparity, and motion characteristics, some previous studies have attempted to predict overall level of visual discomfort by analyzing disparity and motion characteristics of stereoscopic videos [3-5]. Although a few objective assessment algorithms for visual discomfort of stereoscopic videos have been proposed in previous studies [3-5], most of them have focused on individual predictions of disparity- and motion-induced discomforts that estimate the levels of visual discomfort associated with each of causes (i.e., disparity and motion characteristics in stereoscopic content). It should be pointed out that previous methods [3],[5] have used a simple weighted summation for the combination of the individual prediction values that are individual estimates of the levels of disparity- and motion-induced discomforts based on an assumption: the influence of disparity characteristics on the overall level of visual discomfort does not vary with that of motion characteristics and vice versa. However, rather surprisingly, there is no experimental evidence to support this assumption related with the subjective sensation of visual discomfort.

The current study investigates the way how to combine the individual prediction values of disparity- and motion-induced discomforts into an overall prediction value of visual discomfort for developing an algorithm that accurately predicts the level of visual discomfort for stereoscopic videos.

METHOD

This paper aimed at exploring how to combine individual prediction values for disparity- and motion-induced discomforts to more accurately estimate the perceived overall discomfort. For this investigation, we first reviewed and defined four possible combination methods: the weighted summation, multiplication, Minkowski summation, and max combination. To compare the prediction performance of the different combination methods, subjective assessment experiments were performed using a simple-scene dataset with varying disparity and motion characteristics, and then the prediction performance of each combination method was evaluated by measuring the similarity between subjective discomfort values and prediction values obtained by using each combination method. Furthermore, we tried to examine whether the results obtained by the simple-scene dataset were also valid for more complex and realistic scenes.

RESULTS

From the experimental results, we observed that the overall level of visual discomfort could be dominantly influenced by the most significant one between prediction values of disparity- and motion-induced discomforts. These results imply that the combination of disparity- and motion-induced discomforts follows the winner-takes-all (WTA) mechanism often observed in the human visual perception [6]. Also, consistent results were obtained for both synthetic simple-scenes and more realistic, natural scenes with various disparity and motion characteristics.

NOVELTY

To enhance the performance of a prediction algorithm for visual discomfort, it should be necessary to investigate not only individual prediction methods based on disparity and motion characteristics but also an effective combination method for the individual prediction values obtained using the individual prediction methods. To the best of our knowledge, in spite of its importance, there have been no attempts to study how to combine the individual prediction values of disparity- and motion-induced discomforts for the development of a better prediction algorithm of visual discomfort.

9011-71, Session PTues

Stereoscopic visual fatigue assessment and modeling

Danli Wang, Tingting Wang, Yue Gong, Institute of Software (China)

CONTEXT: Although three-dimensional (3D) displays have been applied in many areas, however, viewers often experience visual discomfort and fatigue during the viewing process, which may spoil viewer’s enthusiasm and seems to be one of the critical factors that impede the development of 3D applications. Visual fatigue is measured in either subjective or objective methods. Objective measures are more preferred for their capability to quantify the degree of human visual fatigue without being affected by individual variation [1, 2]. However, little research has been conducted on the integration of objective indicators, or the sensibility of each objective indicator in reflecting subjective fatigue.

OBJECTIVE: The main purpose of this paper is to explore a method to evaluate visual fatigue more objectively. This is to find several objective indicators in terms of the feasibility in measuring and the sensibility to reflect subjective fatigue, and analyze the effect of time on them. Ultimately, a model of the relation between the objective indicators and subjective fatigue is expected.

METHOD: Subjective and objective measurements are combined in the method. That is after each viewing session, viewers rate their visual fatigue with subjective scores (SS) according to a five-grading scale, followed by tests of the punctum maximum accommodation (PMA) and visual reaction time (VRT). Throughout the entire viewing process, an infrared camera enhanced with an infrared light is set up in straight front of the viewer for recording his/her eye movements. Later, pupil diameter (PD) and percentage of eyelid closure over the pupil over time (PERCLOS) are extracted from the videos processed by the algorithm. Based on the method, an experiment with 8 subjects was conducted to assess visual fatigue induced by 3D images on polarized 3D display. The experiment consisted of 10 sessions (5min per session), each containing the same 75 images displayed randomly. To eliminate the effect of individual variation and intrinsic measuring error, the amount of changes (each value minus pre-test value) rather than absolute values of these measures were analyzed in normalized unit.

RESULTS: Results of linear fitting showed that PMA, VRT and PERCLOS increased obviously along time axis and were the most sensitive to the amount of changes of SS, while PD showed slight decrease and the fluctuates became unsteady rather than maintaining natural amplitude. A model was derived from multiple regression (R2=0.989, SS’ is the predicted visual fatigue). Note that the model is preliminary and needs further verification in iterative experiments with various stimuli and large number of subjects. Even some better objective indicators may substitute for the ones in use.

NOVELTY: The proposed method in evaluating visual fatigue combined subjective assessment and objective measures. The sensibility and the amount of changes with time series of several objective indicators were observed. Finally, a predictive model of subjective visual fatigue was proposed, where the three objective indicators used as the independent variables were the most effective and could be measured simply.

9011-72, Session PTues

Visual discomfort under various brightness conditions using eye movements in watching stereoscopic 3D video

Sang-Hyun Cho, Hang-Bong Kang, The Catholic Univ. of Korea (Korea, Republic of)

CONTEXT

As stereoscopic 3D contents are popular, visual discomfort is one of major concerns in stereoscopic 3D contents industry.

OBJECTIVE:

We analyze visual discomfort which is considered illumination conditions using eye-movements and subjective test.

METHOD:

Our test consisted of 4 stages of illumination conditions: stable global; stable local; change global; change local. Each stage is separable as two sub-stages, bright and dark. The Illumination is changed from bright to dark or inverse direction in the middle of viewing time.

Our test procedure is as follows. First, the subjects get rested until visual discomfort of subject was perfectly removed. Then, the subjects answered the following eight questions (dizzy, double vision, stimulated, blurred vision, dry, tired, headache and light-headed) to check their subjective pre-stimulus discomfort. Each question was answered using a 5-point scale where 1 and 5 represented “not at all” and “yes, very much,” respectively. Next, the participant watched the 10-minute test stereoscopic 3D video clips in random order. Since the participant was wearing polarized glasses equipped with an eye-tracking device, we were able to measure eye-blinking, saccadic number and pupil size. In addition, we also measured the discomfort response at each 1-minute interval with a handheld slider. We use this discomfort scale to obtain a visual discomfort score at each 1-minute interval ranging in value from 1 (very comfortable) to 5 (extremely uncomfortable) according to ITU recommendations. After watching the video clip, the subject answered the eight questions above again to measure the subjective post-stimulus discomfort.

RESULTS:

The experimental results are as follows:

Stable local bright(sb:0.3377,bk: 0.7182,sc : 0.5710,ps:0.8414)

Stable global bright(sb:0.2386,bk: 0.7061,sc: 0.5758,ps:0.8092)

Stable local dark(sb:0.0958, bk: 0.5490, sc: 0.6290, ps:0.6796)

Stable global dark(sb:0.0417, bk: 0.4937, sc: 0.6303, ps:0.6743)

Change global from bright to dark(sb:0.2105,bk:0.6901,sc: 0.5842,ps:0.8903)

Change local from bright to dark (sb:0.1981,bk:0.6891,sc: 0.5914,ps:0.9083)

Change local from dark to bright(sb:0.1458,bk:0.6140,sc: 0.6093,ps:0.6924)

Change global from dark to bright(sb:0.1333,bk:0.5920,sc: 0.6246,ps:0.6443)

where sb is subjective, bk is blinking rate, sc is saccadic number and ps is pupil size score.

Note that all measurement values were normalized. Since viewers may concentrate their attention due to large contrast of brightness, viewers felt more visual discomfort in local case than in global case. In addition, in case of change illumination, viewers felt more visual discomfort when brightness changes from dark to bright than from bright to dark. Eye-blinking rates have positive correlation with Mean Opinion Score (MOS) of subjective test. Conversely, saccadic number and pupil size have negative correlation.

NOVELTY:

Visual discomfort is caused by various factors when watching stereoscopic 3D contents [1]. In particular, illumination change is known as one of major factors affecting visual discomfort [2]. However, most of previous researches about visual discomfort dealt with binocular disparity with accommodation and vergence linkage except illumination changes.

REFERENCES:

[1] M. Lambooij et al., “Visual discomfort and visual fatigue of stereoscopic displays: a review,” J. Imag. Sci. Technol. 53(3), 1–14 (2009).

[2] E. Simonson et al., “Effects of Illumination Level on Visual Performance and Fatigue,” J. Opt. Soc. Am. 38, 384-397 (1948).

9011-73, Session PTues

On the comparison of visual discomfort generated by 3D and 2D content based on eye-tracking features

Iana Iatsun, Mohamed-Chaker Larabi, XLIM-SIC (France)

During last years, there has been a noticeable expansion of 3D video. Its effect is comparable to moving from white and black to color television. It is expected to add a third dimension to TV broadcasting services. Among the advantages of 3D are the strong involvement of viewers and the increased feeling of presence. However, one of the main disadvantages is the possibility of affecting human health more than during watching conventional 2D content.

Human are able to perceive world in 3D thanks to two separated eyes gazing in the same point in the space. Like this the principal of creating 3D video is based on showing slightly different images for each of eye. The difference between natural viewing and the presentation of the stereo-images can influence on human’s comfort and satisfaction. Few reasons were revealed as causes of visual discomfort during watching stereo content. Theyareextendedlevelofdisparity,accommodation/vergencemismatchandunnaturalblur(crosstalk).1 These factors induce different strengths of visual discomfort, and in order to evaluate its effect various measurement of visual fatigue have been proposed. Emoto et al. proposed to evaluate visual fatigue by visual functions as viewer break point, recovery point, accommodation step response and evoked cortical potentials. Their results show that latency of evoked cortical potentials and range of relative vergence can be good indexes for visual fatigue evaluation.2 Blinking like essential function of visual processing, also was suggested as visual fatigue estimator in the work of Kim et al. It was given that eye blinking rate increased for higher visual fatigue.3 Part of the research was dedicated to find the link between visual fatigue and video characteristics. Thus, the investigation of Yano et al. showed that high level of discomfort was connected with large parallax and large amount of motion in stereoscopic sequences.4

In order to find the influence of watching stereo video on accumulation of visual fatigue, we conducted experiments using 2-dimensional and 3-dimensional video with the aim of comparing results. We investigated the effect of visual fatigue on eye-movements and eye-blinking in 2D and 3D experiment. Eye-movement was selected as a characteristic of the state of visual motor system, specially the ciliary muscle. Eye blinking as essential function of the eye was investigated also. The main function of eye blinking is to remove irritants from the surface of the cornea and conjunctiva. In such a way an increase of eye irritation is undoubtedly related to an increasing sensation of visual fatigue and changes the eye blinking behavior.

Following the goal of our experiment, we reproduced the same conditions for watching 2D and 3D video content. The atmosphere was close to the natural free-watching movie at home or in a cinema theatre. The room, where experiment was conducted, had a diffuse lightening with luminance around 67 lx in the middle of the screen and was noise-isolated. The same stimuli were used in 2D and 3D experiments. This was achieved by changing the display mode. The video sequences are parts of popular movies with different content (documentary, adventure, cartoon) in a random order for each test-sequence. The monitor Hyundai TriDef S465D with 60 Hz progressive scanning was used to display stimuli in stereoscopic and flat mode. To record participant’s eyes during watching eye-tracker TX-120 with capability to work with 3D glasses and large freedom of head movement was chosen. The distance between eye-tracker and display was 63 cm, which was calculated according to requirements of eye-tracker’ manufacturer. The eyes of subjects were located no more than 70 cm far from eye-tracker, in such a way that the distance between participant and monitor was about 2H of screen. As eye-tracker was set up in the same manner for all subjects, their position with different height was adjusted with a special chair, that allows to change it’s level and fix angle of rotation.

Before beginning of experiment subjects were asked about eye-diseases and injuries experienced during life. All participants passed Freiburg vision test to check their vision acuity, Ishihara test for detecting

colorblindness and randot stereotest to verify stereoacuity. Explanation of symptoms and perception of visual fatigue was given before the test, after that every individual rated his psychological sense of visual fatigue. Observers performed six sequences of 10 min watching 2D or 3D video separated by questionnaire of visual fatigue and simulator sickness. There was a long time gap between experiment for stereo and flat video. The order of sequences for each participant was different. Eye-movements and pupil diameter were measured simultaneously by eye-tracker TX-120 during watching video, calibration of device was done before each test-sequence to adjust position of observer after movements which were necessary to fill the questionnaire.

Results of the experiment showed that watching stereo video induced stronger feeling of visual fatigue than conventional 2D video. Subjective method showed that most of participants has an increased feeling of visual fatigue already after 20 min for stereo video, although, for 2D only after 1H of watching, visual fatigue reached slight level. Statistical analysis ANOVA was used in order to find the relationship between parameters of visual characteristics and video. It was found that frequency of saccade and blinking duration were affected by visual fatigue.

REFERENCES

1. M. Lambooij, M. Fortuin, I. Heynderickx, and W. IJsselsteijn, “Visual discomfort and visual fatigue of stereoscopic displays: a review,” Journal of Imaging Science and Technology 53(3), pp. 1–14, 2009.

2. M. Emoto, T. Niida, and F. Okano, “Repeated vergence adaptation causes the decline of visual functions in watching stereoscopic television,” Journal of Display Technology 1(2), pp. 328–340, 2005.

3. D. Kim, S. Choi, S. Park, and K. Sohn, “Stereoscopic visual fatigue measurement based on fusional response curve and eye-blinks,” in 17th International Conference on Digital Signal Processing, pp. 1–6, IEEE, 2011.

4. S. Yano, M. Emoto, and T. Mitsuhashi, “Two factors in visual fatigue caused by stereoscopic hdtv images,” Displays 25(4), pp. 141–150, 2004.

9011-74, Session PTues

Perception and annoyance of crosstalk in stereoscopic 3D projector systems

Kun Wang, Acreo Swedish ICT AB (Sweden) and Mid Sweden Univ. (Sweden); Börje Andrén, Mahir Hussain, Acreo Swedish ICT AB (Sweden); Kjell E. Brunnström, Acreo Swedish ICT AB (Sweden) and Mid Sweden Univ. (Sweden); Jesper Osterman, LC-Tec Displays AB (Sweden)

CONTEXT

Crosstalk i.e. light leakage between the left and right view, is the cause of a major perceptual problem in the 3D display system [1] shown itself mostly as ghosting. The crosstalk can be due to many factors depending on the 3D display technologies [2].

OBJECTIVE

The aim of this work is to investigate how much the physical variation of crosstalk that is perceived as visible artifacts and the annoyance of it in movie type contents from end-users’ Quality of Experience (QoE) point of view. Two types of 3D projector systems (one system using active shutter glasses, and the other system using passive polarized glasses) were compared in the experiment.

METHOD

A 3D projector system consisting of a stereoscopic projector (produced by DepthQ using time-multiplexing technology which alternating left and right view sequentially), an 80 inch silver screen, active shutter glasses or passive polarization glasses combined with a polarization modulator (produced by LC-Tec), was evaluated. The study comprised an objective measurement of crosstalk in the 3D projector system and a subjective experiment of the users’ experience of the visible distortions such as ghosting. The crosstalk from the system itself was measured first, and then additional simulated crosstalk was added into the 3D videos in order to find out, when the human observers’ starts to perceive the visible distortions and how the annoyance level could vary based on different amounts of crosstalk. Seven stereoscopic cinema contents were selected and processed in five crosstalk levels (0%, 2%, 7%, 12 %, and 20%) for the subjective experiment. The amount of crosstalk in the video was simulated according to the measured luminance characteristics of the projector system.

RESULTS

The objective measured crosstalk from the projector system itself was about 0.3% for the system using active shutter glasses and 2% for the system using passive polarized glasses (polarization modulator contributed less than 1%, the rest was due to other components in the system e.g. silver screen). A total of 26 naïve observers participated in the experiment. Figure 1 shows a linear relationship between the amount of overall crosstalk (sum of system crosstalk and simulated crosstalk) and users’ Mean Opinion Scores (MOS) of perceptual crosstalk experience. From that, we can see, it is necessary to control the overall crosstalk below 10% in order to keep observers not annoyed (MOS > 3.5). We call this the acceptance level. The level when the user starts to perceive the crosstalk distortion is about MOS=4.5 and this corresponds to about 3% crosstalk. The perception of crosstalk distortions also has variations due to different video contents. As Figure 2 shows that for source 3 and 6 people had difficulties to distinguish different levels of crosstalk distortion.

NOVELTY

Xing et al. [3] studied on the perceived crosstalk with a 3D projector system using two projectors based on the MPEG multiview materials [4]. Their study focused on the relationship among crosstalk level, camera baseline and scene contents. Our study focused on the relationship between, the perceptual threshold and the annoyance level on one side, and the variation of crosstalk levels, eye-glass technologies and scene contents on the other side. The major novelties are: a). found an acceptable crosstalk threshold for 3D movie contents, and a linear relationship between amount crosstalk and users’ annoyance level; b). compared the crosstalk performance of active and passive system solutions; c). the method to simulate crosstalk took projector system luminance characteristics into account, and considered both system crosstalk and simulated crosstalk into assessment.

9011-75, Session PTues

Eliciting steady state visual evoked potentials by means of stereoscopic displays

Enrico Calore, Davide Gadia, Daniele L. Marini, Univ. degli Studi di Milano (Italy)

CONTEXT:

Brain-Computer Interfaces (BCIs) are systems that provide users communication and control capabilities by recording and analyzing their brain activities [1].

BCIs could be implemented using various modalities. One of the most used for its robustness is based on the Steady State Visual Evoked Potential (SSVEP) detection: when a visual stimulus flickering at a constant frequency is gazed by an user, it is possible to detect in their Electroencephalograms (EEGs) a continuous brain response at the corresponding frequency.

Using SSVEP, BCIs could be implemented showing to the user targets flickering at different frequencies and detecting which is gazed by the observer, after linking to each of them a command.

Recently, BCIs are gaining attention in different fields, like gaming and Virtual Reality (VR). In particular, SSVEP-based BCIs have been successfully used, integrating flickering targets, in VR environments [2,3].

As stereoscopic displays are becoming widely available and often adopted for VR applications, it is interesting to evaluate their use to present SSVEP eliciting stimuli.

OBJECTIVE:

In this work we evaluate the use of stereoscopic displays for the presentation of SSVEP eliciting stimuli, comparing their effectiveness with the results presented for standard monoscopic displays. Moreover, we investigate the new possibilities offered by these devices for applications in the fields of BCIs and of vision research.

METHOD:

By means of a graphics API, we created an experimental scene to present flickering stimuli on an active stereoscopic display, obtaining reliable control of the targets' frequency independently for the two stereo views.

Using a 4 electrodes portable EEG acquisition device, we acquired and analyzed SSVEP responses from a group of subjects. During the experiments, we considered different frequencies and stereoscopic disparities in the flickering targets.

RESULTS:

From the preliminary results, we got evidence that stereoscopic displays represent valid devices for the presentation of SSVEP stimuli, since we obtained results comparable to those presented in the literature [5]. In particular, the shuttering glasses frequency does not influence the reliability of the presented stimuli, and different stereoscopic disparities do not influence the SSVEP response.

More interestingly, the use of different flickering frequencies for the two stereo views of the same target proved to elicit non-linear interactions, clearly visible in the EEG signal, between the stimulation frequencies. This suggests interesting applications for SSVEP-based BCIs in stereoscopic VR environments able to overcome some limitations imposed by the refresh frequency of standard displays [5].

Further studies are ongoing to assess if the presence of these non-linear interactions, which arise during stereo fusion, could be also used as a test for stereoscopic disparity evaluation and adjustment.

NOVELTY:

As far as we know, although similar interactions were reported in previous works using different stimulation techniques [4,5], this is the first work which exploits stereoscopic displays to elicit non-linear interactions between multiple SSVEP frequencies and to analyze the impact of stereoscopic displays on the stimuli presentation performance to elicit SSVEP responses.

9011-76, Session PTues

A new multimodal, interactive way of subjective scoring of 3D video quality of experience

Taewan Kim, Kwang-Hyun Lee, Sanghoon Lee, Yonsei Univ. (Korea, Republic of); Alan C. Bovik, The Univ. of Texas at Austin (United States)

CONTEXT

People now routinely experience a wide variety of 3D visual contents, such as 3D cinema, 3D TV and 3D games. It is highly desirable to be able to deploy reliable methodologies for obtaining the subjective experiences of large numbers of human viewers. The single stimulus continuous quality evaluation (SSCQE) protocol has been commonly used to assess such attributes as 3D presence, quality, fatigue and naturalness on stereoscopic video. Nevertheless, SSCQE migrated from tools developed for subjective assessment of 2D contents, and it suffers limitations when applied to 3D quality of experience (QoE) assessment tasks in regards to stably capturing relevant 3D quality attributes.

OBJECTIVE

When viewing 3D, subjects become deeply absorbed in the immersive video while experience and tend to become detached from the assessment process ('immersion problem'), leading to problems with the accuracy and latency of the responses. When performing subjective test with a continuous assessment method, losses of concentration may occur over long assessment periods, degrading the synchronization between the test sequence and the rating result. When scores are simply recorded at the termination of the 3D viewing, significant visual fatigue that may be felt can alter the scores that are recorded, relative to what was felt during the play of the video.

METHOD

We propose a new methodology that we call Multimodal Interactive Continuous Scoring of Quality (MICSQ) for subjective 3D QoE assessment. The main difference from conventional methodology is the interaction process between the subject(s) and the assessment environment. Conventional interfaces for 3D SSCQE are `one-way', since the subject decides and records scores on the same display as the video. However, MICSQ breaks this process into two separate interactions. First, the device interaction process is conducted via a direct interaction between the server and the assessment tool (tablet) in real-time. Second, the human interaction process using the assessment tool enables the subjects to avoid such problems as concentration loss and other effects from the immersion problem by the introduction of multimodal cues. Specifically, the tablet delivers audio and haptic cues to each subject throughout the task. For example, the tablet may prevent loss of concentration by supplying periodic audio and/or tactile reminders. Moreover, by audibly announcing the score at periodic intervals, the subjects remain aware of the scores they are delivering, even if they are fully absorbed in watching the video.

RESULTS

While much works remains to be done regarding determining the degree to which multimodal cues can enhance subject performance in visual tasks, the available evidence regarding neuroplastic enhancement by multimodal activity suggests that such approaches may prove to be highly effective. As a result, recorded human responses to 3D visualizations obtained via MICSQ correlate highly with measurements of spatial and temporal activity in 3D video content. We have also found that 3D QoE assessment results obtained using MICSQ are more reliable over a wide dynamic range of content than the conventional SSCQE protocol in terms of confidence interval and the range of subjective scores.

NOVELTY

The main contributions of this study are to develop a new reliable subjective assessment methodology for 3D content by way of demonstrating and validating it. The important advantage of the proposed methodology lies in its flexibility to enhance subjective assessment tasks in diverse viewing environments by decoupling the device - human interaction, by introducing multimodal (aural and tactile) cues, by allowing assessment tasks to occur under diverse illumination (even darkness), and the ease of hand-held portability over wireless and real-time recording. Using the new methodology, it is even possible for many users to perform the subjective assessment simultaneously. Moreover, introducing separation between the viewing device and the human interactions can enhance subjects' adaptability when participating in 3D perceptual and cognitive visual tasks. MICSQ will be introduced freely on the android market after this paper is published.

9011-77, Session PTues

Effect of local crosstalk on depth perception

Hiroshi Watanabe, Hiroyasu Ujike, National Institute of Advanced Industrial Science and Technology (Japan); John Penczek, Paul A. Boynton, National Institute of Standards and Technology (United States)

Context

Crosstalk is a major undesirable factor for the quality of stereoscopic 3D displays that utilize horizontal disparity. Ideally, two completely separate images should be provided, one for each eye. However, in producing and presenting stereoscopic images, crosstalk often arises, causing unintended depth-related effects. This can affect the 3D viewing experience, and distorted depth perception can lead to inadequate

observer response when interaction is required between observer and image.

Objective

This study investigates the effects of local crosstalk on depth perception. To do this, we measured depth threshold of five locations on 3D displays with glasses, and also optically measured crosstalk at each of those locations.

Methods

Observers

Twenty adults, aged 20 to 30 years (6 females and 14 males; 22.9 ± 2.6 years), were recruited from local Osaka residents. They were naïve about the experiment's purpose, and had normal, or corrected-to-normal, visual acuity, and no history of optic nerve disease.

Stimulus and Apparatus

Visual stimulus was two horizontally disparate dots, each of which has horizontal disparity of -6.4 to +6.4 arc min, with nine steps, relative to the screen position. Inter-dot distance was 2.6 degrees. A dot pair was presented at either of the five screen locations; the pair's center are at the screen center or at 270 pixels from the center towards the right, left, top or bottom. The 3D displays, for use with polarization glasses, were from two different manufacturers, and were the same display size (23 inches) and screen resolution (1920 ? 1080 dots).

Procedure

Observers wearing polarization glasses sat 57cm from a display and judged the depth order of two dots (whether the left or right one was nearer) using a PC keyboard (2-alternative forced choice task). They observed the stimulus dots through an aperture for limiting the vantage point of the observers. They responded 16 times per disparate dot pair. Order of disparity and location were randomized. All subjects performed the task for both types of 3D display on separate days, the number of trials thus totaling 1440 (9 disparities ? 16 repetitions ? 5 locations ? 2 displays).

Results:

We calculated the rate that observers judged the dot on the right to be nearer in 16 trials for each display, each screen location and each disparity. We then plotted the rate as a function of the left-right dot disparity, and fitted a psychometric function (cumulative distributed function) to the plot. The curve slope at the response probability of 50% was used to gauge depth order judgment sensitivity. Results show the sensitivity became lower at the location where crosstalk is higher for displays. Therefore, the distribution of the sensitivity on the display surface depends on the crosstalk distribution and its uniformity of the displays.

Conclusion

Distribution of depth sensitivity clearly reflects the local crosstalk of 3D displays with glasses. We speculate the distribution of depth sensitivity could be an index of the crosstalk uniformity of 3D displays.

9011-78, Session PTues

Investigation of retinal images and accommodation responses for super multi-view display

Junya Nakamura, Yasuhiro Takaki, Tokyo Univ. of Agriculture and Technology (Japan)

Super multi-view (SMV) displays have been developed to solve the vergence-accommodation conflict that causes visual fatigue. Viewpoints are generated with an interval smaller than the pupil diameter to allow eyes to focus on three-dimensional (3D) images. In our previous study, the accommodation responses were measured for 3D images produced by the recently developed reduced-view SMV display. We found that the SMV display condition evoked the accommodation responses, and we consider the variation range of the evoked accommodation responses as the depth of field (DOF) of eyes, i.e., the SMV condition increased the DOF of eyes. We used two test images; 'Maltese Cross' containing sharp edges, and 'Lenna' containing both sharp and unsharp edges, because the SMV display provides only horizontal parallax so that the retinal images blur differently in the horizontal and vertical directions. In this study, the retinal images are investigated to show the relationships between the accommodation responses and the retinal images.

The retinal images were captured using a cooled CCD camera by placing it at the viewpoints of the reduced-view SMV display, instead of using human eyes. The pitch of viewpoints of the display was 2.6 mm, and the pitch was made twice and thrice by displaying the same parallax images to adjacent two or three viewpoints. The focus of the camera was set at the four distances, (1) the nearest distance where eyes focused (the near limit of DOF), (2) the furthest distance where eyes focused (the far limit of DOF), (3) the distance of the display screen, and (4) the distance of the test images. The results obtained for the pitch of the viewpoints of 2.6 mm were explained below:

The retinal images obtained for 'Maltese Cross' contained horizontally separated two images when the camera was focused on the screen. When the camera was focused on the test image, the retinal images were not separated. However, the retinal images contained blur, and the vertical blur was larger than the horizontal one, because the SMV display controlled the ray directions only in the horizontal direction. When the camera was focused on the near and far limits of DOF, the vertical blur decreased to be comparable to the horizontal one.

The DOF range obtained for 'Lenna' was located further from viewers compared with that obtained for 'Maltese Cross.' When the camera was focused on the screen, the retinal images looked blurred images rather than twin images because 'Lenna' contained more unsharp edges than sharp ones. For the other three conditions, the same tendencies were obtained for the retinal images as obtained for 'Maltese Cross'. The retinal images obtained at the near and far limits of the DOF had better image quality than those obtained for the other two conditions.

From the results obtained for the two test images, the retinal images blurred equally in both the horizontal and vertical directions at the near and far limits of the DOF of eyes. The existence of the vertical blur might limit the increase of the DOF ranges of eyes. Therefore, the full-parallax SMV displays, which control rays both in the horizontal and vertical directions, could more increase the DOF ranges of eyes.

9011-300, Session Plen2

Integrated Imaging: Creating Images from the Tight Integration of Algorithms, Computation, and Sensors

Charles A. Bouman, Purdue Univ. (United States)

No Abstract Available

9011-41, Session 11

Compression for full-parallax light field displays

Danillo B. Graziosi, Zahir Y. Alpaslan, Ostendo Technologies, Inc. (United States)

Light field displays provide a more natural viewing experience by eliminating the need for glasses. Most commercially available light field displays use horizontal parallax-only input images to reduce the bandwidth demands of the display. Recent research suggests that such displays are not capable of eliminating discomfort associated with the vergence-accommodation conflict. For an eye strain free and natural 3D visualization full parallax light field displays are necessary. A full parallax display that eliminates vergence accommodation conflict and also displays an HD quality 3D image requires addressing of approximately

five orders of magnitude higher number of pixels compared to an HDTV. This paper introduces a compression method for full-parallax light field displays that exploits the spatio angular data redundancy in the light field to make real time display implementation possible.

Full-parallax light field images exhibit high spatio angular redundancy that can be exploited in compression. Many techniques based on standard video compression algorithms have been proposed. However, they are impractical in bandwidth, processing or power requirements for a real time implementation to support full parallax light field displays. Moreover, many algorithms are designed for horizontal parallax-only systems. We developed a novel compression algorithm for full-parallax light field displays that is able to cope with the large amount of light field image data given the strict timing requirements for real-time implementation.

Our compression method exploits the spatio angular redundancy of a full-parallax light field to reduce the bandwidth of the light field image, while simultaneously reducing the total computational load with minimal perceptual degradation. The method also increases processing speed and potentially provides power savings to the system.

We used well-known images such as the Stanford dragon, which was synthesized from given geometrical models. Our compression algorithm is able to reconstruct the texture of the entire light field with minimal perceptual quality degradation, achieving bandwidth reduction from 2 to 4 orders of magnitude. Subjective results showed that the introduced artifacts are imperceptible, and the system can successfully reproduce the 3D light field, providing natural binocular and motion parallax.

We describe a compression algorithm that is suitable for supporting full-parallax light field displays. Our unique approach makes it possible to compress the light field data during acquisition resulting in reduced processing requirements and potential power savings.

9011-42, Session 11

Non-linear rendering for autostereoscopic displays

Christian Riechert, Marcus Müller, Peter P. Kauff, Fraunhofer-Institut für Nachrichtentechnik Heinrich-Hertz-Institut (Germany)

CONTEXT:

Autostereoscopic devices suffer from a, compared to glasses-based 3D displays, limited depth range. This problem is further increased because S3D content in general does not use all available depth layers, but rather only some foreground layers, background layers and little in between. This often leads to a degraded depth perception on autostereoscopic displays. Especially foreground objects appear rather flat.

OBJECTIVE:

The limited depth range of autostereoscopic devices makes it necessary to remap depth layers during rendering in a non-linear way to make full use of all available depth layers. This way the overall depth impression is not significantly changed but fine-grained depth information in foreground objects will be preserved even on displays with rather few distinct depth layers. In this paper we propose a new non-linear depth rendering algorithm to improve depth impression of stereoscopic (or trifocal) content on autostereoscopic devices.

METHOD:

The proposed non-linear rendering method is based on existing multiview-video (MVD) rendering techniques. As input, S3D content (or even better trifocal content) is expected. In a first step depth maps for each camera are estimated using real-time capable high quality disparity estimation and post-processing. The calculated depth maps allow us to render virtual camera images to arbitrary positions between the original cameras. This is commonly done using linear depth image based rendering techniques (DIBR). We propose a new non-linear rendering method which allows, in addition to render virtual camera images, to remap depth to make better use of all available distinct depth layers. In contrast to other techniques like video-plus-depth (V+D) rendering or saliency based warping the renderer uses all two (or three) available input camera views to generate intermediate views. This can reduce visible artifacts significantly. The target non-linear rendering function can either be specified explicitly as samples of a strictly increasing function or calculated automatically from the source depth maps to optimally remap depth layers.

RESULTS:

Non-linear rendering for autostereoscopic displays allows you to optimally use the whole available depth range of a given autostereoscopic device. Experiments were done with publically available stereoscopic and multiview test sequences and commercially available 3D Blu-ray content. The results are promising and the general depth impression of many tested scenes could be improved in comparison to linear rendering. Especially, the typical flatness of foreground objects and the billboard effect could be reduced and objects tend to have more 3D volume when displayed on autostereoscopic displays.

NOVELTY:

Non-linear rendering techniques for 3D repurposing have been proposed by several authors. Popular methods are saliency based warping and non-linear video-plus-depth rendering. Both have the disadvantage of only using one source view and thus not using all available occlusion information. Our proposed method uses all available information from all source views and thus can handle disocclusion in the rendered image better.

9011-43, Session 11

Enhancing multi-view autostereoscopic displays by viewing distance control (VDC)

Silvio Jurk, Thomas Ebner, Fraunhofer-Institut für Nachrichtentechnik Heinrich-Hertz-Institut (Germany); Bernd Duckstein, Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute (Germany); Sylvain Renault, René de la Barré, Fraunhofer-Institut für Nachrichtentechnik Heinrich-Hertz-Institut (Germany)

CONTEXT

Multi-view autostereoscopic displays are supporting sufficient stereo quality only within a limited range around the nominal viewing distance. As a major drawback in usability, the manufacturer so far assigns this distance.

OBJECTIVE

Conventional multi-view displays spatially interlace various views of a 3D scene and form appropriate viewing channels. If the person maintains the nominal distance two slightly divergent views are projected to his eyes, both covering the entire screen. With increasing deviations from the nominal viewing distance the stereo image quality decreases.

Therefore, a novel approach is required that corrects the incorrect view assignments depending on the distance of the viewer. In particular, the authors suggest a continuous variation of multiple views. Because of the discrete arrangement of pixels at the display panel, a simple rearrangement of the view distribution becomes impracticable.

METHOD

The authors describe a software-based solution that enables continuous view adaptation based on the calculation of intermediate views and a column-by-column rendering method. Their algorithm controls each individual sub-pixel and generates a new interleaving pattern from selected views. In addition, color-coded test content is used to verify its efficiency. The resulting relative disparity can thereby be interpreted as hue angle within the HSV color space and can be further used as a measure for the visible depth.

RESULTS

This novel technology helps shifting the physically determined nominal distance to a user-defined distance thereby supporting stereopsis. The recent viewing positions can fall below or above the nominal distance of

the original setup. The authors successfully tested lenticular-based multi-view displays covering 8 to 28 views. Herein, the viewing distance could be corrected with factor of ±2.5. The approach is also extremely useful in extending the viewing range of multi-view displays in the single-user mode.

The crossing of light rays and the correction of the view assignments give rise to a floating perspective. A lateral movement of the viewer will not result in detectable viewing offsets. Therefore, the correction algorithm even enables persons with largely divergent interocular distance to experience 3D scenes without any limitation.

NOVELTY

This novel distance adaption system allows using multi-view autostereoscopic displays at variable viewing distances—independent from the fixed display architecture. It can be integrated in the control systems of a variety of lenticular-based multi-view displays including legacy versions.

9011-44, Session 12

Vision-based calibration of parallax barrier displays

Nicola Ranieri, Markus Gross, ETH Zurich (Switzerland)

CONTEXT

Parallax barrier displays deploy an image layer with two scrambled views and a barrier layer to multiplex the two views to distinct eye positions. Such a setup requires precise alignment or calibration of image and barrier layer.

OBJECTIVE

Since the introduction of dynamic parallax barriers, there are known ways to compute the barrier pattern for given eye positions. These methods, as the one introduced by Perlin et al. [Perlin2000], are usually iterative and hence, latency in barrier updates grows with increasing barrier resolution.

Furthermore, the computation of the barrier pattern relies on known display geometry. The higher the display resolution, the more severe become imperfections and misalignment caused by the manufacturing process. This can lead to undesired Moiré patterns, cross-talk and other artifacts.

In this work, we investigate computer vision based calibration of parallax barrier displays. Our method does not require prior knowledge about the display geometry and scales well with resolution. Based on the calibration results, each barrier position can be computed independently in parallel on a GPU.

METHOD

A specific eye position together with the barrier layer can be seen as a virtual camera, where the eye is the center of projection and the barrier layer the virtual image plane. Two eye positions thus define a virtual stereo camera pair with shared image plane. The image layer of the display induces a homography, relating pixels of one virtual camera to pixels of the other camera. Rays from eye positions through such pixel pairs intersect on the image layer [Hartley2000, page 325ff].

Thus, if one reference camera center is known with the homography relating the barrier layer to the image space, we can compute the relation for any other eye position. The underlying math and relations are illustrated in the supplementary material and will be described in full detail in the paper. Furthermore, we describe how standard camera calibration methods can be used to find all required information.

If the homography for an eye pair is known, the position of the i-th barrier can be computed iteratively by xi=H*x(i-1) (or xi=(H^i)*x0 with x0 as initial start point), similar to [Perlin2000]. We use the eigendecomposition of H=V*D*(V^(-1)) to compute H^i=V*(D^i)*(V^(-1)) without any iteration, as D is a diagonal matrix. Using V, D and V^(-1) in a GPU shader, each barrier position can be computed in parallel which scales well with barrier resolution.

RESULTS

We present a computer vision based calibration method for parallax barrier displays with sub-pixel accuracy. Without prior knowledge about the display geometry, we are able to relate the barrier layer to the image layer for arbitrary viewer positions. Our GPU implementation is stable and general and can be used to reduce latency and increase refresh rate of existing and upcoming barrier methods. Proof-of-concept for our method is shown on a dynamic parallax barrier prototype.

NOVELTY

The presented methods are novel to our knowledge. We provide insights, how calibration can be used to relate image layer and barrier layer for arbitrary and dynamic viewer positions. Our approach, to compute the barrier pattern without the use of iterations or recursions allows for hardware-accelerated implementations.

REFERENCES

[Perlin2000] An Autostereoscopic Display, K. Perlin et al., SIGGRAPH 2000 Conference Proceedings, http://www.mrl.nyu.edu/publications/autostereo/autostereo.pdf

[Hartley2000] Multiple View Geometry, R. Hartley and A. Zisserman, Cambridge University Press 2000, ISBN 978-0-521-54051-3

9011-45, Session 12

Viewing zone of autostereoscopic display with directional backlight using convex lens array

Shuta Ishizuka, Takuya Mukai, Hideki Kakeya, Univ. of Tsukuba (Japan)

When directional backlight to each eye alternates synchronously with the alternation of left-eye and right-eye images on?the display panel, the viewer can see a stereoscopic image without wearing special goggles. One way to realize directional backlight is to place a convex lens array in front of dot matrix light sources to generate collimated light. To implement this method, however, defocusing and field curvature of the lens should be taken into account. In this paper the viewing zone of autostereoscopic display with directional backlight using convex lens array is analyzed based on optical simulations. When the distance between the lens array and the light source plane is equal to the focal distance of the elemental lens, the viewing zone free from crosstalk becomes deeper in the central direction, while the viewing angle is not wide because of the field curvature of the lens. The viewing angle can be widened when the distance between the lens array and the backlight is shorter, while the viewing zone becomes shallower. This is due to the tradeoff between the effect of field curvature and defocusing of the lens, which should be balanced to meet with the requirement of practical stereoscopic vision. Use of lens made of material with large refraction index can widen the viewing angle according to the effect of reducing field curvature.

9011-46, Session 12

Time-division multiplexing parallax barrier based on primary colors

Qu Zhang, Hideki Kakeya, Univ. of Tsukuba (Japan)

So far, several methods have been proposed to deal with the resolution issue of conventional 2-view parallax barrier while the viewing zone issue yet remains. Since parallax barrier systems provide sweet spots only at certain positions, it takes extra efforts to realize a continuous viewing zone even with precise head tracking involved.

We consider multi-view as a practical solution for the viewing zone issue. For instance, by showing a “view1-view1-view2-view2” pattern on a 4-view system instead of “view1-view2-view3-view4”, a wider

range of viewing zone can be achieved for both view1 and view2. With head tracking brought in, it is possible to carry out a continuous viewing zone in the horizontal direction. To realize 4-view autostereoscopy with full display resolution per view, we introduce quadruple time-division multiplexing parallax barrier in our previous work. Though it should have required 240 Hz displays to present flickerless images, we succeed in reducing flickers under 120 Hz by using 1-pixel barrier while an annoying color breaking issue remains unsolved. Color breaking can be weakened by applying time-division multiplexing anaglyph, where a barrier pattern like “green-black-magenta-black” is created. However, crosstalk turns heavy as a trade-off.

In this research, we propose a new type of time-division multiplexing parallax barrier that is formed of a “red-green-blue-black” pattern. Correspondingly, an image pattern like “1R/4G/3B-2R/1G/4B-3R/2G/1B-4R/3G/2B” is used, where “1R/4G/3B” means that the red component of view1, the green component of view4 and the blue component of view3 are shown for this pixel. In one frame, each viewpoint will achieve 3/4 of the full resolution and 1/3 of the full color. By shifting the image pattern and barrier pattern by one phase every frame, four full-resolution and full-color viewpoints can be created every 4 frames. It is plain to see this type of pattern works similarly to the ones in the previous work do. However, since it shows least blank stripes when compared to the others, an improvement in viewing experience can be expected. Furthermore, changing the order of the element pixels in the barrier pattern may also make a difference. For example, “red-blue-green-black” is considered to have less color breaking than “red-green-blue-black” because the two darker components “blue” and “black” will not show in two adjacent frames.

We carried out a psychophysical experiment to figure out how this new method differs from the older ones and whether the order of the element pixels on the barrier pattern makes a difference. Not only are some positive results successfully achieved as expected, an acceptable performance in crosstalk is also shown. We consider this time-division multiplexing parallax barrier based on primary colors as a nice balanced one where crosstalk and color breaking are weakened to a low level at the same time.

9011-47, Session 12

Multi-user autostereoscopic display based on direction-controlled illumination using a slanted cylindrical lens array

Daisuke Miyazaki, Yui Hashimoto, Takahiro Toyota, Kenta Okoda, Osaka City Univ. (Japan); Tetsuro Okuyama, Toshikazu Ohtsuki, Akio Nishimura, Hiroyuki Yoshida, Panasonic Corp. (Japan)

CONTEXT:

The development of practical auto-stereoscopic display is expected to relax the restrictions for viewing position and the number of viewers, which are problems in many conventional auto-stereoscopic displays.

OBJECTIVE:

The purpose of this project is to develop an auto-stereoscopic display technique, which satisfies the following conditions.

1) Arbitrary viewing position within certain limited region.

2) Available for multiple viewers.

3) High resolution and large image size comparable to ordinary display devices for television.

4) Suppression of nonuniformity of luminance distribution.

5) Compact system configuration.

METHOD:

In the proposed system, an image display unit is illuminated with a direction-controlled illumination unit, which consists of a spatially modulated parallel light source and a steering optical system. The steering optical system is constructed with a slanted cylindrical array and vertical diffusers. This configuration led to the horizontally continuous change of light direction.

The direction-controlled illumination unit can control output position and horizontal angle of vertically diffused light. An LCD is illuminated by light from the direction-controlled illumination unit, which is controlled to shine the whole area of the LCD, and to form narrow exit pupil in front of the LCD. A viewer can watch the image only when an eye is located at the exit pupil.

Proposed system achieves auto-stereoscopic view by alternately switching the position of an exit pupil at viewer’s both eyes, and displaying parallax images are on the image display unit in synchronization with switching of the exit pupil. In addition, more than one people can view an auto-stereoscopic image simultaneously by forming exit pupils individually for viewers.

One of important problem of this method is nonuniformity of luminance distribution because of the difference of light intensity depending on the diffusing direction. We developed a method for suppression of the luminance nonuniformity by spatially modulating luminance distribution of the illumination unit.

We constructed an experimental system to verify the proposed method. We used a LCD projector and Fresnel lenses for the direction-controlled illumination unit, and a 32 inch full-HD LCD for image display.

RESULTS:

We confirmed that auto-stereoscopic images were generated by the experimental system. Crosstalk, which is amount of leak of light from other exit pupil, was enough small to view a stereoscopic image. We confirmed that multiple viewers watched a stereoscopic image simultaneously.Nonuniformity of luminance distribution was improved by the proposed method based on spatial modulation of illumination.

NOVELTY:

Surman et al. proposed a multi-user auto-stereoscopic display similar to our method. This conventional display system includes a steering optical system using complicated discrete optical elements, which causes luminance non-uniformity. We use simple cylindrical lenses, which have continuous structure.

The differences from our previous work are as follows.

1) Proposal of suppressing method for nonuniform luminance distribution based on spatial modulation of illumination.

2) Change of illumination unit of experimental system from the combination of LED

9011-48, Session 13

Accommodation measurements of integral photography

Hitoshi Hiura, Tomoyuki Mishina, Jun Arai, Yuichi Iwadate, NHK Science & Technical Research Labs. (Japan)

Integral photography (IP) reproduces the light rays from objects, in principle, the accommodation is considered to be consistent with the position of the reproduced 3D objects. Therefore, the visual fatigue that arises from accommodation-convergence conflict in binocular stereoscopic methods is not considered to occur. Nevertheless, there is no report that compares the accommodation responses in viewing the IP image and the binocular stereo image. We therefore measured the accommodation response during the viewing of 3D images produced by IP and by the binocular stereoscopic method.

A high-resolution display is required to measure the accommodation response; therefore, we developed a 3D display device that comprises a high-resolution liquid crystal display (LCD) and a high-density lens array. The LCD has a resolution of 468dpi and a diagonal size of 4.8 inches. The high-density lens array comprises 107x69 micro lenses that have a focal length of 3mm and diameter of 1mm arranged in a honeycomb pattern. The gap between the display and the lens array matches the 3-mm focal length of the lens array. In the case of the IP and binocular stereoscopic

targets, we presented the 3D target in front of and behind the 3D display. In the case of the real targets, we presented the 2D target on the 3D display at the same depth positions of the 3D targets reconstructed by IP and the binocular stereoscopic method.

In our experiments, we measured the accommodation response for three display conditions (IP, binocular stereoscopic and real object) and two viewing conditions (binocular and monocular). The visual target was a "Maltese cross", which subtended 1.9°x1.9°. The 3D display was positioned 60cm away from the observer. The target was presented at eight distances relative to the 3D display: 15cm, 10cm, and 5cm in front of the display, on the screen, and 5cm, 10cm, 15cm and 30cm behind the screen. The accommodation response was measured with a WAM-5500 optometric system (Grand Seiko Co., Ltd.), which allows the observer to view the target with both eyes.

As a result, we obtained the accommodation responses in viewing the IP images, the binocular stereoscopic images and the real objects. In the binocular viewing condition, the accommodation responses to the real objects varied with the target distance. Whereas the accommodation response of IP images showed weaker response, however, it approximately also varied along with the responses of the real objects. In the monocular viewing condition, the accommodation responses varied with the target distance for the IP and real object display conditions. The accommodation responses to the binocular stereoscopic images were stable at the display position. Therefore, IP has an ability to induce the accommodation to the depth positions of the 3D target.

9011-49, Session 13

Optimized design of directional backlight system for time-multiplexed auto-stereoscopic display based on VHOE

Yong Seok Hwang, Byeong Mok Kim, Eun-Soo Kim, Kwangwoon Univ. (Korea, Republic of)

Recently, the time-multiplexed autostereoscopic 3D displays with dual directional backlight system was produced showing a two-view, full resolution display by optically forming distinct viewing regions for left and right eyes. However, the various novel backlight systems with LGP and a film have been designed and proposed for the time-multiplexed autostereoscopic 3D displays. For example by Sasagawa et al., a double-sided prism sheet with two fast switching light sources, and similarly by 3M there is double-sided prism sheet. However, the unavoidable misalignment of displacement between lenticular lenses and prisms causes the boundary angle of separation shifted from the normal viewing, and scattered normal rays from small peaked ridge of prism structure. The adopted backlight system by Sasagawa was similar with the dual directional backlight system which was proposed by 3M. The discrimination between two backlight systems is based on optical characteristics of disposed material between light guide and the liquid crystal display panel. The role of the disposed film is to form the viewing zone utilizing directional beam from light guide plate (LGP). To form separated optical viewing zones at interocular distance without crosstalk by the disposed film, the precise design for structure of LGP is required. The structure of LGP in the 3M backlight system is designed well for the double sided prism film. However with LGP in the 3M backlight system, we could not induce optimized optical characteristics such as optical power efficiency and formation of output beam which is produced by angular and lateral distribution of rays, and reduction of the crosstalk for VHOE-based approach. Thus, the newly proposal for the structure of LGP at the VHOE based backlight system is necessary.

In this paper, a VHOE based-light guide plate (LGP) with directional and quasi-collimated beam for time multiplexed auto-stereoscopic display based on VHOE is proposed. Proposed LGP was designed, simulated through theoretical analysis at wave based ray viewpoint by MATLAB and LightTools. To design proposed LGP for VHOE-based auto-stereoscopic 3D display, three targets should be satisfied such as uniform distribution of intensity, uniform angular distribution and control of output angle from LGP without backshift. Thus, we established some adequate optical parameters which give main effect to three targets for the proposed VHOE-based backlight system and designed the optimized structures of LGP considering practical producible circumstance through optical measurement such as angular distribution and plane intensity distribution using optic detector in LightTools and intensity distribution using wave based ray tracing equations. According to simulation results which were carried out for three targets and analysis, it can be expected to achieve more accurate optical realization of proposed LGP optimized to VHOE layer. The uniform prism array structure and the viariable angle and width prism array are compared on the bottom of LGP. In this paper, at the uniform prism array structure, when the east and west angle of prism is near to 3 degree, the uniform intensity distribution and the directivity of the rays coming from LGP are produced showing the best results. However, the divergent angle is 15 degree. To get more collimate beam, the variable prism structure is proposed to adopt in the time-multiplexed backlight system based on VHOE. The width of prism is determined as 80um and the east and west angle of prism are varied reverse-symmetrically.

9011-50, Session 13

Analysis of multiple recording methods for full resolution multi-view auto-stereoscopic 3D display system incorporating VHOE

Yong Seok Hwang, Kyu Ha Cho, Eun-Soo Kim, Kwangwoon Univ. (Korea, Republic of)

Conventional multi-view 3D-display without glasses, such as parallax barrier, lenticular lens sheet type, has problem that the resolution is decreased according to the number of view and narrow viewing zone is built. To overcome the above problems in this paper, we analyze and establish multiple recording methods of photopolymer for a full-resolution multi-view auto-stereoscopic 3D display system incorporating VHOE (Volume Holographic Optical Element) and face-tracking. It is presented that VHOE using this proposed recording method can be easily applied to time-multiplexed multi-view display system. First, we found multiple recording conditions of VHOE for 6-multi-view display through analysis of optical properties. A VHOE is optically made by angle-multiplexed recording process to get 3-viewing zone. First, exposure time scheduling for multiple recording is adopted from the multiple recording process for hologram memory. The 10 and 20 multi-recordings on a photopolymer plate are chosen. The accumulated strength is calculated with the uniform time-interval of the recordings. And then, optimized exposure-time scheduling scheme for 3 viewing zones is calculated and is established through iteration measurement with induced equation from as spline curve-fitting method. To get 6-multi-view zone, two VHOE with displaying 3-veiwing zone are stacked reversely. To get minimum color dispersions, optical measurements for angular selectivity and wavelength selectivity are performed comparing with simulations based on Kogelnik's theory. Here, angular selectivity is about 2 degree. The wavelength selectivity is about 10nm. And then, the optimized conditions of the parameters such as thickness, refractive index modulation, diffusion rate, viscosity of the material are established through the time-variable simulation of FDTD for the diffusion theory of VHOE material. The optimized index modulations among parameters are analyzed and calculated for the multiplexed recording according to exposure time scheduling. It is assumed that the each recording include the dark response while the other recording is performed. For feasibility of the analyzed recording conditions theses recorded VHOEs combined with LCD panel and VHOE-based 6-view stereoscopic display system behind collimated white bulb light system is implemented incorporating face tracking technique. Two incident beams on VHOE layers in the opposite directions with symmetric angles which function as reference beams are sequentially synchronized with the respective view stereo images displayed on the LCD panel. It is verified that optically optimized recording conditions of VHOE are feasible for full resolution multi-view auto-stereoscopic display system incorporating VHOE and face-tracking.

Conference 9012: The Engineering Reality of Virtual Reality 2014

Monday - Tuesday 3 –4 February 2014

Part of Proceedings of SPIE Vol. 9012 The Engineering Reality of Virtual Reality 2014

9012-1, Session 1

Interactive projection for aerial dance using depth sensing camera

Tammuz Dubnov, Univ. of California, Berkeley (United States); Zachary Seldess, Shlomo Dubnov, Univ. of California, San Diego (United States)

This paper describes an interactive performance system for Floor and Aerial Dance performance that controls visual and sonic aspects of the presentation via a depth sensing camera (MS Kinect). The proliferation of cheap and high quality Infra-Red (IR) and color (RGB) depth cameras (RGB/IR-D) opened novel possibilities for Virtual Reality in general and for the arts in particular. Besides its original use for gaming, Kinect cameras have been used for various applications such as telepresence, security, as a scanning device, and even for high quality tracking of surgical instruments, to mention a few. What makes our project unique and different from these engineering tasks is the scale, both in time and space, and the open nature of the system design that needs to support a large range of sensing scenarios and offer a palette of creative mappings from human actions to animated projection, lighting and sound, based on the actors positions in space, their movement and gestures. Maybe the most interesting aspect of this project is the challenge of coming up with a new language for stage, lighting and sound design that utilizes technological opportunities by transforming the stage into an augmented reality reactive space in order to produce novel means for artistic expression. Accordingly, our paper will be structured as follows:

After brief review of RGB/IR-D technology, we will specify the tracking requirements for a combined floor and aerial dance. Since this type of art combines dance movement with circus-like aerial acrobatics on apparatuses such as the silks, lyra and more, the system is required to detect, measure and track free movement in space in terms of an overall location, body orientation and more detailed limbs positions and gestures. In the paper we will distinguish between skeleton tracking and general purpose 3 and 6 degree of freedom (6-DOF) tracking in space (on the ground and in the air) using IR markers. Segmentation of an actor body in space can be achieved based on depth information, which in turn allows projection of separate imagery on the performer and the backdrop. Such projection can serve different artistic roles, from augmented costume design through body mapping, to virtualization and abstraction of the dancer body immersed in a digital narrative. Other design considerations include establishing causality relations between the human actors and the virtual objects. This transforms the role of what traditionally would be considered as inanimate props into active players entailed with autonomous behavior and even a character. These and other situations demonstrate how traditional concepts of stage design and dramaturgy have to be reconsidered in augmented reality performance situations. It terms of existing genres, the work combines aspects of scripted and choreographed mastery with a thrill of a gamer experience that often interferes with the former. To address these concerns the last portion of the paper will be devoted to considering the work in terms of cybernetic, information and system art theories, which, to quote Burnam, “goes beyond a concern with staged environments and happenings; … [where] there are no contrived confines such as the theater proscenium or picture frame. Conceptual focus rather than material limits define the system.”

Initial results and videos of rehearsals with the system can be found at http://shlomodubnov.wikidot.com/kinect-aerial-dance or https://www.facebook.com/ZuzorKinect

9012-3, Session 1

3D whiteboard: collaborative sketching with self-tracked smartphones

James Lue, Jürgen P. Schulze, Univ. of California, San Diego (United States)

Engineers and scientists often use whiteboards to discuss spatial phenomena or data. This is straightforward in 2D scenarios, but harder in 3D, because one has to mentally project 3D shapes to 2D, which not everyone is good at without training. In those scenarios it would be good to have a 3D whiteboard. Our research project investigates the feasibility of 3D sketching with Android-based smartphones with built-in cameras.

Collaborative virtual environments built with traditional VR technology have been around for many, but we are not aware of any competing attempt at 3D sketching with mobile phones.

Our pose estimation system is based on AR Toolkit-style fiduciary markers in the working environment. There have to be enough markers so that there is always at least one marker the smart phone's camera can see. If no markers are seen, we use the gyroscope to update just the orientation. Prior to using the sketching application, the marker arrangement has to be calibrated with a custom Android app running on one of the smart phones. The calibration result will be sent to the other smart phone, so that both will use the same coordinate space.

To draw collaboratively in 3D, both phones have to run the 3D sketching app. We developed using a Nexus One and a Galaxy S II, but the app runs better on faster phones. It uses OpenCV to calculate the phone's pose, and OpenSceneGraph for rendering on the screen. The phones communicate with each other directly over wi-fi, using a custom communication protocol. Image processing with OpenCV is computationally expensive, hence we reduce the size of the camera image to 320x240 pixels, which allows us to achieve reasonably interactive frame rates, but reduces accuracy. With the Galaxy S II we get about ten pose updates per second, on the Nexus One about five.

During sketching, each user can draw independently. Draw commands are shared with the other phone as they happen, so that the 3D scenes stay synchronized at all times. Each user can see what the other user draws by pointing their phone's camera in the direction of that user's virtual sketch, and will see the sketch as if the phone's display was a window into the shared virtual world.

Our initial experiments with our system show that it is possible to create an application for 3D sketching solely based on smart phones, without additional servers. However, the user experience can still be improved with more accurate pose estimation and higher frame rates – currently, the frame rates are too low for productive use. We expect that newer phones will automatically solve this problem. Besides using new hardware, in the future we want to support more types of geometry. Our application would also benefit from a marker-less pose estimation approach, which would dramatically improve its practicability because markers and calibration would no longer be needed, but this approach is currently computationally too expensive for mobile phones.

9012-4, Session 1

Scalable metadata environments (MDE): artistically-impelled immersive environments for large-scale data exploration

Ruth G. West, Univ. of North Texas (United States); Todd Margolis, Andrew Prudhomme, Jürgen P. Schulze, Univ. of California, San Diego (United States); John P. Lewis, Weta Digital Ltd. (New Zealand) and Victoria Univ. of Wellington (New Zealand); Joachim H. Gossmann, Rajvikram Singh, Univ. of California, San Diego (United States); Iman Mostafavi, Limbic Software, Inc. (United States)

Scalable Metadata Environments (MDEs) are a new artistic approach for designing immersive environments for large scale data exploration in which users interact with data by forming multiscale patterns that they alternatively disrupt and reform. Developed and prototyped as part of an art-science research collaboration, we define an MDE as a 4D virtual environment structured by quantitative and qualitative metadata describing multidimensional data collections. Entire data sets (e.g.10s of millions of records) can be visualized and sonified at multiple scales and at different levels of detail so they can be explored interactively in real-time within MDEs. They are designed to reflect similarities and differences in the underlying data or metadata such that patterns can be visually/aurally sorted in an exploratory fashion by an observer who is not familiar with the details of the mapping from data to visual, auditory or dynamic attributes. While many approaches for visual and auditory data mining exist, MDEs are distinct in that they utilize qualitative and quantitative data and metadata to construct multiple interrelated conceptual coordinate systems. These "regions" function as conceptual lattices for scalable auditory and visual representations within virtual environments computationally driven by multi-GPU CUDA-enabled fluid dyamics systems. Additionally, MDEs present data sets of many millions of records and provide non-distortion-based detail in context views of one record in the context of 10s of millions of records with drill down/ drill-through capability. We employ the notions of "context" and "pattern" to frame and facilitate exploration of massive multidimensional data when one may not know what one is looking for. These concepts make data accessible to a broad audience spanning researchers, citizen scientists, educators and the general public. MDEs attempt to reach beyond the concept of "scale" to that of integration of multiple modalities and resolutions with context as an animating force that reveals patterns and interconnections within highly abstract data. They explore the potential for aesthetic /artistically-impelled techniques to support data exploration and hypothesis-generation in ways that transcend disciplinary boundaries and domain expertise while simultaneously interrogating the cultural context of the underlying science and technologies. Purposefully positioned on the edge between art and science, MDEs resist definition as either.

9012-5, Session 1

Virtual art revisited

Silvia P. Ruzanka, Rensselaer Polytechnic Institute (United States)

Art and VR Revisited

Virtual reality art at the turn of the millennium saw an explosion of creative exploration around this nascent technology. Though VR art has much in common with media art in general, the affordances of the technology gave rise to unique experiences, discourses, and artistic investigations. Women artists were at the forefront of the medium, shaping its aesthetic and technical development, and VR fostered a range of artistic concerns and experimentation that was largely distinct from closely related forms such as digital games.

Today, a new wave of consumer technologies including 3D TV's, gestural and motion tracking interfaces, and head-mount displays as viable, low-cost gaming peripherals drives a resurgence in interest in VR for interactive art and entertainment. Designers, game developers, and artists working with these technologies are in many cases discovering them anew. This paper explores ways of reconnecting this current moment in VR with its past. Can the artistic investigations begun in previous waves of VR be continued? How do the similarities and differences in contexts, communities, technologies, and discourses affect the development of the medium?

9012-6, Session 2

Navigating large-scale virtual environments within immersive displays: interfaces and cybersickness

Daniel R. Mestre, Aix-Marseille Univ. (France)

Today, virtual reality has a number of valuable applications, in particular in the fields of engineering and ergonomics, enabling the user to acquire spatial knowledge about a virtual environment through active exploration. However, when a 1:1 sensorimotor scale between the user and the VE is required and when the size of the VE is larger than the physical displacement space offered to the user, natural locomotion is no longer possible; locomotion interfaces have to be designed. A serious negative aspect of these interfaces in the fact that a significant number of users tend, during or after exposure to immersive virtual reality, to exhibit a symptom called cybersickness. They become dizzy, nauseous and/or disoriented. Cybersickness appears to be an extreme form of motion sickness and to be linked to the fact that different sensorial modalities deliver conflicting stimulation. The user experiences self-motion through visual imagery (especially within cave-like setups, in which the entire visual field is stimulated). On the other side, s/he experiences stationarity (no vestibular stimulation). In this matter, it has been often suggested that fast rotational visual motion was an aggravating factor. In this paper, we will report the results of a recent experimental study, in which we manipulated a number of factors, while measuring cybersickness symptoms, using the standard “simulator sickness questionnaire” (SSQ, Kennedy et al., 1993). Immersed and standing inside a 4-surface cave system, the participants simply had to reach the exit of a one-way maze. To do so, they were given three locomotion interfaces, across three different test sessions (presented in counterbalanced order across 12 volunteer participants). The first interface was an Xbox ® controller, on which two minijoysticks independently controlled forward motion and horizontal rotation. The other two interfaces consisted in custom adaptations of a “redirected walking” algorithm: users had to physically turn their body toward the direction they wanted to go and a slow counter rotation was then applied to the virtual environment, such that they were redirected to face the cave’s front wall. The general hypothesis was that this would prevent fast visual rotations, hence potentially reducing cybersickness. In the second interface, users had to walk in place, and passive captors on their knees were used to calculate locomotion (forward) speed. In the third interface, the Xbox controller was used for forward speed along with the “redirected” algorithm. The comparison between these last two interfaces was meant to measure if active stepping would reduce cybersickness. Results from this study show that 1) redirected walking is easily accepted and intuitively used by all participants; 2) that walking in place is reported as tiring and results in greater speed variability than the other two interfaces; 3) that, on average, all interfaces result in self-reports of cybersickness, even if exposure time was quite short, always under 10 minutes. Finally, it appears that users having no experience playing video games exhibit higher scores of cybersickness than “gamers”, suggesting that, besides interface properties, individuals factors come into play.

9012-7, Session 2

CaveCAD: a tool for architectural design in immersive virtual environments

Lelin Zhang, Cathleen E. Hughes, Eve Edelstein, Jürgen P. Schulze, Eduardo Macagno, Univ. of California, San Diego (United States)

We developed CaveCAD to be a computer-aided design tool for architects to design directly within 3D immersive virtual reality environments, such as CAVEs. All existing commercial software in this domain, such as Autodesk 3dsMax, Maya, AutoCAD, Trimble SketchUp or Blender was designed to run on standard desktop PCs, operated with mouse and keyboard. Approaches exist (e.g., TechViz or Conduit) to enable those software tools to render in VR systems, but these solutions normally only allow viewing, but do not allow controlling the modeling application from the VR system. Another distinctive feature CaveCAD provides is an intuitive approach to the design process by using specialized widgets for certain tasks, which allow using the software without intensive training and to build blueprints from scratch in a quick and convenient way. For example, latitude and longitude for the model are selected with a 3D globe. In order to reduce the amount of buttons in the GUI and make it more intuitively usable, we created a spherical menu without pull down menus or buttons. Each category of functions is visually represented within a spherical widget. Some less often used functionality is provided by more traditional pull-down menus.

Geometry in CaveCAD is organized in similar formats to those of existing commercial packages: models can be created, modified, colored or textured in either “group mode” with combinations of basic shapes, or in “geometry mode” with access to vertices, edges or surfaces.

CaveCAD uses the concept of a “virtual geometry manipulator”, which allows the user to manipulate objects outside of easy reach of their hands. It uses ideas from the World-in-Miniature approach, but applies them only to the selected shape, and does not use it to move the camera, but to interact with the object. When this manipulator is enabled, the user manipulates a piece of basic geometry (such as a cube) by editing a wireframe proxy of this geometry. This wireframe proxy is located within easy reach of one’s hands, but the affected object can be at a different scale and location in the scene. The scale can be changed in power of 10 increments by a button on the input device.

To the best of our knowledge, especially after the 2013 IEEE 3DUI competition, CaveCAD is unique in allowing 3D modeling and the creation of architectural models on a basic geometry level directly in a virtual environment. The most closely related project is probably Sixense’s MakeVR, but that is aimed more towards the modeling hobbyist rather than architect -- however, it is a much more complete package than CaveCAD.

Other typical 3D modeling features CaveCAD supports include save and load functions for 3D models, a variety of textures to apply to geometry, a choice of optional panoramic backgrounds, snap-to-grid and undo-redo functions, a coordinate plane display, an animation system for the creation of camera paths, real-time shadows based on location, date and time, and a library of pre-defined 3D objects, such as trees.

9012-8, Session 2

A hardware and software architecture to deal with multimodal and collaborative interactions in multiuser virtual reality environments

Pierre Martin, Anthony Tseu, Nicolas Férey, Lab. d’Informatique pour la Mécanique et les Sciences de l’Ingénieur (France); Damien Touraine, LIMSI CNRS (France); Patrick Bourdot, Lab. d’Informatique pour la Mécanique et les Sciences de l’Ingénieur (France)

Most advanced immersive devices provide collaborative environment within several users have their distinct head-tracked stereoscopic point of view. Combining with common used interactive features such as voice and gesture recognition, 3D mouse, haptic feedback, and spatialized audio rendering, these environments should faithfully reproduce a real context. However, even if many studies have been carried out on multimodal systems, we are far to definitively solve the issue of multimodal fusion, which consists in merging multimodal events coming from users and devices, into interpretable commands performed by the application. Multimodality and collaboration was often studied separately, despite of the fact that these two aspects share interesting similarities. We discuss how we address this problem, thought the design and implementation of a supervisor that is able to deal with both

multimodal fusion and collaborative aspects. The aim of this supervisor is to ensure the merge of user's input from virtual reality devices in order to control immersive multi-user applications. We deal with this problem according to a practical point of view, because the main requirements of this supervisor was defined according to a industrial task proposed by our automotive partner, that as to be performed with multimodal and collaborative interactions in a co-located multi-user environment. In this this task, two co-located workers of a virtual assembly chain has to cooperate to insert a seat into the bodywork of a car, using haptic devices to feel collision and to manipulate objects, combining speech recognition and two hands gesture recognition as multimodal instructions. Besides the architectural aspect of this supervisor, we described how we ensure the modularity of our solution that could apply on different virtual reality platforms, interactive contexts and virtual contents. A virtual context observer included in this supervisor in was especially designed to be independent to the content of the virtual scene of targeted application, and is use to report high-level interactive and collaborative events. This context observer allows the supervisor to merge these interactive and collaborative events, but is also used to deal with new issues coming from our observation of two co-located users in an immersive device performing this assembly task. We highlight the fact that when speech recognition features are provided to the two users, it is required to automatically detect according to the interactive context, whether the vocal instructions must be translated into commands that have to be performed by the machine, or whether they take a part of the natural communication necessary for collaboration. Information coming from this context observer that indicates a user is looking at its collaborator, is important to detect if the user is talking to its partner. Moreover, as the users are physically co-localised and head-tracking is used to provide high fidelity stereoscopic rendering, and natural walking navigation in the virtual scene, we have to deals with collision and screen occlusion between the co-located users in the physical work space. Working area and focus of each user, computed and reported by the context observer is necessary to prevent or

avoid these situations.

9012-9, Session 2

Z-depth integration: a new technique for manipulating z-depth properties in composited scenes

Kayla Steckel, David M. Whittinghill, Purdue Univ. (United States)

This paper presents a new technique in the production pipeline of asset creation for virtual environments called Z-Depth Integration (ZeDI). ZeDI is intended to reduce the time required to place elements at the appropriate z-depth within a scene. Though ZeDI is intended for use primarily in two-dimensional scene composition, depth-dependent "flat" animated objects are often critical elements of augmented and virtual reality applications (AR/VR). ZeDI is derived from "deep image compositing", a capacity implemented within the OpenEXR file format. Many scenes are composed of several overlapping elements composited together to produce a single unified image. In order to trick the human eye into perceiving scene elements as being in front of or behind one another, the developer must manually manipulate which pixels of an element are visible in relation to other objects embedded within the environment's image sequence. ZeDI improves on this process by providing a means for interacting with programmatically extracted z-depth data from a virtual environment scene. By streamlining the process of defining objects' depth characteristics, it is expected that the time and energy required for developers to create compelling AR/VR scenes will be reduced. In the proof of concept presented in this manuscript, ZeDI is implemented for pre-rendered virtual scene construction via an AfterEffects software plug-in. Future implementations will be augmented and expanded to allow ZeDI's functionality to run adaptively and at interactive frame rates.

9012-10, Session Panel

When VR really hits the streets

Jacquelyn Ford Morie, All These Worlds, LLC (United States)

Back in the early 1990s, the hype about virtual reality was that it would change life as we knew it, opening new doors to experience and enabling us to visit places that existed only in the imagination. Unfortunately it never lived up to that dream, partly because it was still an emerging technology (e.g. hard to manage and create for) and it required expensive equipment to enable a fully immersive experience.

Today, the world has changed. Inexpensive modeling and animation programs are easily accessible; user-generated content is de facto for many; platform code is ubiquitous. But perhaps the most significant change is the new generation of inventors and entrepreneurs who don't believe that things need to be expensive to work, and have found ways to make sophisticated equipment not only cheaper but even better then the original counterparts. Enter Build-it-yourself 3D printers, the Maker Fair mentality and the Oculus Rift. Couple this with a generation of people who grew up on interactive games and believe they can create their own, and you have the beginning of VR2.0 – VR accessible to 100s of thousands more people than could have been dreamed of 30 years ago. Looking at the Oculus Rift as a case in point, it's Kickstarter campaign aimed to raise $250k, and went on to raise almost 2.5 million. Selling at $300 each, 15,000 developer kits for the wide field of view HMD were sold. Think of that: in the whole history of VR there were never that many HMDs in existence at one time. The game was irrevocably changed with this small group of people believing they could make something better.

9012-11, Session 3

Automatic exposure for panoramic systems in uncontrolled lighting conditions: a football stadium case study

Vamsidhar Reddy Gaddam, Univ. I Oslo (Norway) and Simula Research Lab. (Norway); Marius Tennoe, Espen Helgedagsrud, Mikkel Nass, Henrik K. Alstad, Univ. I Oslo (Norway); Haakon K. Stensland, Carsten Griwodz, Pål Halvorsen, Univ. I Oslo (Norway) and Simula Research Lab. (Norway)

One of the most common ways of capturing wide field-of-view scenes like arena sports is by capturing panoramic videos. Using an array of multiple cameras with limited overlapping in the corresponding images, one can achieve a good panorama. Using the panorama, several immersive display options can be explored. There is a two fold synchronization problem associated to such a system. One is the temporal synchronization and the other is the automatic exposure synchronization. The former can be handled by using a common triggering solution to control the shutter of the camera, the latter is not straight forward when the light conditions are uncontrolled like in the case of an open football stadium.

In this paper we present the challenges and approaches for creating a completely automatic real-time panoramic capture system. The straight forward challenge in building such a system is that there is not one common area that is visible to all the cameras that can be used for metering. One approach we followed is to use the green of the field to measure the light. This approach allowed us to achieve visually pleasing results generally, but when the field is lit with direct sunlight, the measure is inaccurate due to the directional reflectance of grass. A second approach was devised where the overlapping areas between adjacent cameras are exploited, thus creating pairs of perfectly matched video streams. But there still existed some disparity among the pairs. We finally developed an approach where the time between two temporal frames is exploited to communicate the exposures among the cameras and thus achieving a perfectly synchronized array. An in-depth analysis of the driving system and experimental results are presented in this paper.

9012-12, Session 3

Museum as spacecraft: a building in virtual space

Julieta C. Aguilera, Adler Planetarium & Astronomy Museum (United States) and Plymouth Univ. (United Kingdom)

This paper presents immersion and interaction projects that engage visitors of the museum in order to build a mental model of the building as the body in relation to different scales in the Universe.Aspects of embodiment and visualization are presented as they correspond to current trans-disciplinary developments.

9012-14, Session 3

Enhancing audiovisual integration in virtual environments with libCollider

Joachim H. Gossmann, Eric Hamdan, Qualcomm Institute (United States)

There is a persistent demand for an audio component to the CalVR visualization framework that is under development at the Immersive Visualization Laboratory at the Qualcomm Institute, San Diego,

On the other hand, the integration of auditory elements into VR displays and data display environments is characterized by difficulties of interdisciplinary communication.

In this paper, we describe some of the typical problems as well as our proposed solution: libCollider, a client library for SuperCollider's 'scsynth' sound synthesis engine that provides C++ application developers with direct access to scsynth's sophisticated capabilities for real-time audio synthesis and rendering through an API with abstracted bi-directional network support.

9012-25, Session 3

Recent improvements in SPE3D: a VR-based surgery planning environment

Marcin Witkowski, Robert Sitnik, Warsaw Univ. of Technology (Poland); Nico Verdonschot, Univ. Twente (Netherlands) and Radboud Univ. Nijmegen Medical Ctr. (Netherlands)

TLEMsafe project [1] aims to create a patient-specific surgical navigation system, based on innovative ICT tools, for training, pre-operative planning and execution of complex musculo-skeletal (M-S) surgery. It aims to help the surgeon to safely reach the optimal functional result for the patient, and it will be a user-friendly training for surgeons. SPE3D (pronounced speed) is a surgery planning environment developed within TLEMsafe project. It enables the operator to virtually plan a surgical procedure on the customized musculo-skeletal model of the patient's lower limbs, send the modified model to the bio-mechanical analysis module, and export the scenario's parameters to the surgical navigation system. An earlier version of the application was described in [2].

Personalized M-S models originate from the AnyBody Modeling System 5.0 (AnyBody Technology A/S, Denmark) and are created based on medical imaging as well as functional trials. For the surgical intervention planning the M-S model is registered with magnetic resonance imaging (MRI) data that has been merged from five MRI section scans with a custom software tool from Materialise N.V., Belgium.

The SPE3D application provides the operator with a combined 2D/3D view of the model. The 3D model containing bones and muscles is visualized together with spatially aligned MRI planes (sagittal, coronal and transverse). Moreover, an arbitrary MRI cross-section can be rendered on the 3D M-S model. The position and orientation of the arbitrary cross-section may be modified by the operator in real time (a dedicated GPU shader has been created to implement this feature). Additionally, sagittal, coronal and transverse MRI planes may be simultaneously shown in classic 2D side panels that surgeons are familiar with.

Apart from the feature of thorough model inspection the interface provides the operator with virtual tools with which he or she can perform model modification as:

- cutting bones with an arbitrary plane or MRI cross-section,

- manipulating bones and their fragments,

- removal of unwanted bone fragments,

- repositioning muscle insertion points, modifying muscle force, removing muscles,

- placing implants from an implant library.

Having prepared a surgery scenario the operator may export its parameters to the AnyBody Modeling System for bio-mechanical analysis of functional outcome. If the results are satisfactory the scenario data may be exported to the surgical navigation system (BrainLab GmbH, Germany) to be used during the actual surgery.

The SPE3D environment has been developed with Virtools 5 (Dassault Systemes, France) environment which allows creation of 3D compositions and programming of human-model interactions. The interface supports stereoscopic viewing with nVidia 3D Vision (nVidia, USA) and the M-S model may be inspected/manipulated in a natural way with use of two Phantom Omni (Sensable, USA) haptic devices. Alternatively, the application may be controlled in a traditional way with use of a standard computer keyboard, mouse and 2D display (requires no specialized hardware) or with a touch screen (e.g. in an operating room).

The interface may be utilized in two main fields. Experienced surgeons may use it to simulate their operative plans and prepare input data for a surgical navigation system. It may also be used by student or novice surgeons who can use it for training.

SPE3D has been developed in the TLEMsafe project [1] funded by the European Commission FP7 program.

[1] TLEMsafe project website www.tlemsafe.eu

[2] M. Witkowski, J. Lenar, R. Sitnik, N. Verdonschot, A virtual reality interface for pre-planning of surgical operations based on a customized model of the patient, Proc. SPIE 8289, 2012.

9012-15, Session 4

Game engines and immersive displays

Benjamin Chang, Marc Destefano, Rensselaer Polytechnic Institute (United States)

While virtual reality and digital games share many core technologies, the programming environments, toolkits, and workflows for developing games and VR environments are often distinct. VR toolkits designed for applications in visualization and simulation often have a different feature set or design philosophy than game engines, while popular game engines often lack support for VR hardware. Extending a game engine to support systems such as the CAVE gives developers a unified development environment and the ability to easily port projects, but involves challenges beyond just adding stereo 3D visuals.

In this paper we outline the issues involved in adapting a game engine for use with an immersive display system including stereoscopy, tracking, and clustering, and present example implementation details using Unity3D. We discuss application development and workflow approaches including camera management, rendering synchronization, GUI design, and issues specific to Unity3D, and present examples of projects created for a multi-wall, clustered, stereoscopic display.

9012-16, Session 4

Gestural interfaces for immersive environments

Todd Margolis, Univ. of California, San Diego (United States)

Modern computer vision has taken a revolutionary leap forward making consumer hardware and software available for everyday use. With this new affordance in user interfaces, how can gestures be used to improve interaction for large scale immersive display environments. Through the investigation of full body tracking and hand tracking, this paper will discuss various modalities of gesture recognition and compare their usability to other forms of interactivity.

9012-17, Session 4

5555555 55555 555: A (constrained) narrative on the z-axis

Elif E. Ayiter, Sabanci Univ. (Turkey)

This paper discusses the creation of typographic systems as artworks in three dimensional online builder's worlds, also known as the metaverse. A series of such typographic installations, that have been created by the author, will be presented as case studies that will attempt to delineate the playful/aleatoric approach that has been taken whilst constructing textually based art ecologies that can be traversed and experienced through the agency of avatars. This approach has largely been based upon David Small's findings (1996) that note that when it comes to screen based 3D there are significant shortcoming when it comes to legibility. These are due to shifting view-points which result in a continuous perspective distortion that is unsuited for reading since the positioning of the reader in relation to the angle of the text remains awkward at best. While certain word shapes may still be recognizable in less than ideal circumstances, in general there are few viewpoints from which a text will hold its legibility. What differentiates these viewpoints from their Real Life counterparts is that these perspectives are displayed upon a flat surface, i.e., the screen. In a screen based 3D environments each new angle will result in a differently shaped letter and at extreme angles this shape can even be reduced to a line.

The author's focus therefore is upon how such screen-based virtual 3D spaces may utilize text within a context that departs from the primary attribute with which writing has inherently been associated – namely readability. In such environments readability may be displaced through the usage of text as a playful device, displayed as artefacts that are riddle-like configurations, or constructs that are meant to be understood through means other than straightforward reading; bringing about states of heightened engagement and 'play' through their manipulation or indeed simply by being immersed inside them. The awareness that legibility might not be an appropriate goal, when it comes to creating typographic artwork in a metaverse, lead to a search for textual sources that would address a need for play, for personal readings and interpretations; in other words, text that is meant to be 'felt' as an artwork, rather than to be 'read' as informational content. This brought to the fore asemic/aleatoric writing, and generative text as suitable sources for providing the material for a non-traditional usage of type.

Set against a general overview of the author's three dimensional typographic practice, three installations will be discussed in detail: "Asemia," a work which utilizes asemic (wordless, semantically open) texts; "The Tower of Heteronyms" that was built as a homage to the Portuguese poet Fernando Pessoa, whose poetry provides the source material; and "RGB," a space in which deconstructed, generative text and primary colors were used to create a aleatoric, non-linear narrative playground for avatars.

9012-18, Session 4

Embodiments, visualizations, and immersion with enactive affective systems

Diana G. Domingues, LART (Brazil); Cristiano J. Miosso, Suélia S. Rodrigues Fleury Rosa, Carla Silva Rocha Aguiar, Univ. de Brasilia (Brazil); Tiago F. Lucena, Univ. de Brasilia (Brazil) and LART (Brazil); Mateus Miranda, Adson F. Rocha, Univ. de Brasilia (Brazil); Ramesh Raskar, Massachusetts Institute of Technology (United States)

In the beginning, virtual reality systems for immersive experience used different types of proprioceptive devices and movement trackers to map a participant's displacements or gestures and provide a feedback for navigation and positioning. This principle constitutes the basic idea of kinesthesia in the domain of aesthetics, which was largely explored in many applications in virtual reality.

Our proposal in bioart and biomedical engineering for affective aesthetics focuses on the expanded sensorium and investigates problems regarding enactive systems. These systems enhance the sensorial experiences and amplify kinesthesia by adding the sensations that are formed in response to the physical world, which aesthetically constitutes the principle of synaesthesia. This approach relates the Cinetic Art of the 80s and the experiences with disruptive technologies for perception and sensorial measurement, offering a compelling experience in data landscapes. Simulators technologically add to the displacements the effects of collisions, vibrations, trepidations etc., when inhabiting synthetic spaces. Also, the historical intuitive interfaces, such as shutter glasses, data gloves, trackers, emitters, force biofeedback devices, joysticks, among others, are empowered by sensors that capture physiological signals in order to enhance the immersion experience and make it more intense. The embodied experience then corresponds to a Spinozan body with the affective ability to communicate with the environment through the several measured signals. This body sends and receives data, as it exchanges energies and signals with the environment, in an enactive condition described by Varela.

In this paper, we also present enactive systems inside the CAVE, configuring compelling experiences in data landscapes and human affective narratives. The interaction occurs through the acquisition, data visualization and analysis of several synchronized physiological signals, to which the landscapes respond and provide immediate feedback, according to the detected participants' actions and the intertwined responses of the environment.

The signals we use to analyze the human states include the electrocardiography (ECG) signal, the respiratory flow, the galvanic skin response (GSR) signal, plantar pressures, the pulse signal and others. Each signal is collected by using a specifically designed dedicated electronic board, with reduced dimensions, so it does not interfere with normal movements, according to the principles of transparent technologies. Also, the electronic boards are implemented in a modular approach, so they are independent, and can be used in many different desired combinations, and at the same time provide synchronization between the collected data.

In order to collect the signals while the participants live the immersive experience, we developed special support biomaterials, such as an insole made out of Brazilian latex. This is a biocompatible material that helps fixing the electrodes in the participants' skin, without displacements during the experience and without risks of allergic or other undesired reactions. This is especially important for easily positioning the plantar pressure sensors, through an easy-to-put insole with the electrodes already fixed.

In this context, the human actions and interactions generate living maps, body narratives through enactions in an ouroboric perception by Domingues, corresponding to Gibson's concept of ecological perception. In this work, our Laboratory of Art and TechnoScience (LART) at the University of Brasília at Gama, in Brazil, intends to collaborate with discussions on embodiments, data visualizations and immersions with enactive affective systems, which echo the enacionist theory of

autopoiesis and to the naturalization of the technologies, by Couchot (2012).

9012-19, Session 4

The visceral umbilicus: entwining body image and body schema in VR

Diane Gromala, Bryn Ludlow, Simon Fraser Univ. (Canada)

The immersive VR system: Oculus Rift HMD, sensors, biofeedback. Experimental conditions are outlined: immersive VR vs. stereoscopic Kinect.

9012-13, Session PTues

Retinal projection type super multi-view head-mounted display

Yutaka Ito, Hideya Takahashi, Seigo Nakata, Osaka City Univ. (Japan); Kenji Yamada, Osaka Univ. (Japan)

CONTEXT: Mixed Reality (MR) technology has recently been proposed as an approach for practical use of Virtual Reality (VR) technology. See-through head-mounted displays (HMD) provide an effective capability for MR. By using a see-through HMD, an observer can see both real world and virtual world at the same time. Some conventional see-through HMDs have been developed. They can display two-dimensional virtual information like a document. However, when two-dimensional virtual information is displayed by a HMD, an observer is hard to understand that it is related with a real-world object. What an observer sees needs to be augmented by three-dimensional (3D) virtual information image in accordance with the real object.

OBJECTIVE: We propose a retinal projection type super multi-view HMD. The smooth motion parallax provided by the super multi-view technique enables a precise superposition of 3D images on real objects. Moreover, if a viewer focuses his or her eyes on the displayed 3D image, the stimulus of accommodation of the human eye is reproduced naturally by the super multi-view technique. Therefore, although proposed HMD is a monocular HMD, it provides observers with natural 3D images. In addition, proposed retinal projection type super multi-view HMD use the principle of the Maxwellian view. In the Maxwellian view, parallel rays are converged directly at the center of the pupil, and projected onto the retina directly. Thus, proposed HMD can provide an extreme long focal depth image, and a provided 3D image by proposed HMD is clear and high contrast.

METHOD: Proposed retinal projection type super multi-view HMD use the principle of the Maxwellian view. In the Maxwellian view, parallel rays are converged directly at the center of the pupil. In the proposed HMD, the super multi-view condition is realized by converging projected some parallax images at the different points on the pupil. The proposed HMD works as a light field head-mounted display. The proposed HMD consists of an image projection optical system and a holographic optical element (HOE). The HOE is used as a combiner that superimposes the virtual image on the real scene to realize a function of the see-through. The HOE also works as a condenser lens to implement the Maxwellian view. The image projection optical system projects some parallax images onto the HOE, and projected images converged on the pupil.

RESULTS: In order to verify the effectiveness of the proposed HMD, we constructed the prototype HMD. In the prototype HMD, the number of parallax images is two, thus the number of convergence points on the pupil is two. The distance between two convergence points is 2 mm. The angle of view is about 40 degrees horizontally and vertically. And the observed virtual image is monochrome green because a green LED is used as an illumination light source. We displayed 3D images at the distance from 40 cm to 210 cm in front of the pupil, and confirmed the accommodation.

NOVELTY: The retinal projection type super multi-view HMD is proposed. It can provide a see-through 3D image by a monocular HMD.

9012-20, Session PTues

Comparing two input devices for virtual walkthroughs using a head mounted display (HMD)

Beatriz Sousa Santos, Paulo M. Dias, Paulo J. Santos, Carlos M. Ferreira, Univ. de Aveiro (Portugal)

1. Introduction

Virtual Reality (VR) platforms using Head Mounted Displays (HMDs) and orientation sensors are important for applications such as training and simulation and commonly used for virtual walkthroughs. These HMDs may have significant advantages when compared to other displays since they may allow users to always see the virtual world as they turn their heads offering a 360º field of regard. However, the type of input devices that can be used along with HMDs is limited as users cannot see the physical world while using them, which makes the choice of input devices more demanding.

In the scope of previous studies we observed and interviewed users, and logged data characterizing their performance, while doing virtual walkthroughs in a Virtual Environment (VE) using a setup including a HMD and a two button mouse as input for direction (forward and backward). Throughout those studies, participants recurrently suggested that we should use instead a gamepad with a joystick as in their opinion it would be more usable.

The users' suggestion to replace the mouse by a gamepad joystick seemed promising in terms of usability for the following reasons:

1 - a more intuitive mapping can be used with the joystick than with the mouse;

2 - gamepads are meant to be held with both hands while playing (in a similar way that the mouse had been used by many users in our studies), while mice were designed to be used on a working surface;

3 - many young users are very experienced in using gamepads.

Thus, it seemed that better results concerning performance, satisfaction and comfort might be possible with the gamepad joystick as compared to the mouse.

This made us consider performing a new user study to tackle the following question: Will users' performance, satisfaction and comfort while using the HMD in a virtual walkthrough improve with a joystick as input device as compared to a mouse?

This paper reports the methods and the results obtained in such a user study performed with the collaboration of 45 young participants.

2. Methods

We had the collaboration of 45 high school volunteer participants (16 girls and 29 boys aged between 15 and 18 years) recruited in the scope of technology demonstrations for young people. Due to practical constraints of the demonstration, a between-groups experimental design had to be used and thus, to assess potential individual differences between the two groups, all the participants also performed the walkthrough in the same Virtual Environment (VE) using a desktop setup.

2.1 Setups

The HMD was a i-glasses SVGA Pro, with stereoscopic capability with an orientation sensor (tracker) InterTrax 2 from InterSense with three degrees of freedom (DOF) (roll, pitch and yaw). Interaction was performed using either a standard two button mouse or a mini joystick of a gamepad (Fig. 1).

A combination of physical rotation (via head-tracking) and virtual translation (via input device) was used. Gaze-directed steering was chosen. The input device in both cases allowed only moving forward or backward.

The mappings used for the two input devices were the following:

1 - Mouse: push left button – move forward; push right button – move backwards;

2 - Gamepad: tilt joystick forward – move forward; tilt joystick backwards – move backwards;

The desktop used had a 19" wide screen monitor with a resolution of 800x600 pixels at a viewing distance of 25"; interaction was performed using a mouse and keyboard.

2.2 Virtual Environment and tasks

The used VE was a maze where users had to find objects designed to incorporate several tasks that users perform spontaneously, not needing directions from the experimenter; to be more attractive to participants it was presented to the users as a game.

2.3 Experimental design

This study was designed to test the following null hypothesis: users would have similar performance whether using a mouse or a joystick while navigating our VE using a HMD.

Input device and corresponding mapping (mouse vs. joystick) was the independent variable and performance, satisfaction and comfort were the dependent variables. Performance was measured through the number of caught objects, velocity, walked distance and number of collisions with the walls. Satisfaction and comfort (regarding nausea, vertigo and disorientation) were assessed through a post-test questionnaire in order to avoid eliciting an increase in simulator sickness felt by users. Data characterizing profile and preference were also collected through the same questionnaire. After completing the experiment, participants were invited to express their opinions and suggestions.

2.4 Procedure

Each participant played the game for five minutes on the HMD set-up using one of the input devices and also the desktop. The order in which the two set-ups were used was varied among users.

3. Results and discussion

The 45 high school volunteer students were divided in two groups: 22 used the mouse and 23 used the joystick while using the HMD. Performance data were analysed using Exploratory Data Analysis (EDA) and non-parametric tests; all performed using STATISTICA. These tests were chosen due to the relatively low number of participants and lack of normality of the obtained data.

The main finding was that no statistically significant difference was observed between the performance of the two groups that have used different input devices (mouse and joystick) with the HMD, contrary to what we expected. The obtained results also show that participants generally had better performances with the desktop, confirming results we had obtained in previous studies.

Our results contradict what many participants in previous studies had suggested, and seemed reasonable. This may illustrate the know phenomenon that what users prefer or believe would result in a more usable user interface may not work out better and reinforces the need to perform user studies in order to obtain guidelines that can be applied in the development of Virtual Environments.

The final version of this paper will present a more comprehensive description of the methods, detailed results, and discussion of possible causes of the findings.

9012-21, Session PTues

Automatic generation of Chinese character using features fusion from calligraphy and font

Cao Shi, Jianguo Xiao, Peking Univ. (China); Canhui Xu, Wenhua Jia, Peking Univ (China)

Chinese character deeply influences people's daily life in Southeast Asia, which is an indispensable tool to transmit information and manifest handwriting art. It is significant to explore glyph, which influences human cognition and vision, utilizing advanced modern digital algorithms. Recently, much research has been made on automatic imitation of Chinese calligraphy. According to generation mechanism, algorithms on automatic generation of Chinese character can be generally divided into two categories: strokes and radicals reused algorithm (SRRA) and stroke order drawing algorithm (SODA). SODA imitates handwriting process depending on stroke order. The other category of methodology to automatically generate Chinese character is SRRA, which arranges sample strokes and radicals to compose characters.

However, few consider making well use of font to synthesize calligraphy. This paper proposes a novel strategy to generate Chinese character through features fusion from local contour feature of calligraphy and elaborate structural feature of font.

A spatial statistic based contour feature representation is proposed to achieve extraction of local contour feature from Chinese calligraphy character. The contour is extracted from Chinese calligraphy character utilizing Canny edge detector. Each character contour image is divided into blocks, and then each block is filtered by a series of line detectors with one pixel width and various slopes. The Gaussian distribution is used to estimate the statistical distribution of a particular slope from blocks, which are at the same position in all collected calligraphy character images from a particular person. The parameters of Gaussian function of all slopes at all positions, including Mean and Standard Deviation, in this paper, are employed as the extracted local contour feature (for the particular person).

A features fusion strategy is designed to automatically generate new hybrid character, making well use of contour feature of calligraphy and structural feature of font. In the features fusion strategy, at a particular position in a font image, dilation operator and erosion operator are employed to inject the extracted contour feature from Chinese calligraphy into font at the particular position. The slope of dilation operator is controlled by the parameters of Gaussian distribution at the particular position. The whole process of features fusion executes dilation and erosion operations iteratively at all positions in the font image, which are similar to "pad" and "cut" in a sculpture progress. The extents of "pad" and "cut" can be controlled by the sizes of dilation operator and erosion operator.

Experimental results demonstrate that the generated new hybrid character hold both contour feature of calligraphy and structural feature of font. Especially, two kinds of Chinese calligraphy skills called "Fei Bai" and "Zhang Mo" are imitated in the hybrid character. "Fei Bai" depicts a phenomenon that part of a stroke fade out due to the fast movement of hair brush or the lack of ink, and "Zhang Mo" describes a condition that hair brush holds so much ink that strokes overlap.

9012-22, Session PTues

Platform for setting up interactive virtual environments

Danilo D. Souza, Paulo M. Dias, Beatriz Sousa Santos, Daniel Santos, Instituto de Engenharia Electrónica e Telemática de Aveiro (Portugal)

The demand for virtual worlds to solve different constraints or to enable the exploration of physically impossible tasks has been growing to the date, even more with the advance of the technology associated to the creation of such worlds. However, setting up an immersive virtual environments is still a complex task that demands a lot of time and resources (either financial or human) and requiring expertise in programming languages. This is a barrier for many possible users, that are unable to set up a virtual environment and in most case reduce the interaction to viewpoint control without possibility to interact with additional material.

In the context of a project involving several industries, we include, beside a simpler configuration, the possibility to interact in the virtual environment with content associate to the models (for example consulting documentation or viewing a movie related to a particular

equipment), The main idea is to provide useful tools to help evaluate and train in a virtual factory before or during the setup of the line.

A platform was created by us to ease the configuration of virtual environments - it's called pSIVE, or "Platform for Setting up Interactive Virtual Environments" which contains only free open-source technologies and aims to allow non-experts to set up interactive and immersive virtual environments by providing a simplified configuration graphic interface, flexible hardware handling, interaction techniques and the possibility to be easily extended. Within the context of its creation, one could easily build and deploy an interactive environment to visit and train on even nonexistent facilities allowing interaction and manipulation of relevant information. But possible users of pSIVE are not limited to the group of companies previously mentioned; they can be present in many different fields, from academic to industrial environments.

By the date, pSIVE is already functional, allowing the integration of different kinds of hardware (input and output), multiple interaction and navigation styles, semi-automated layout creation and embedding content to be interacted with, such as videos, documents and pictures. All controlled by a user friendly interface that orchestrates the configuration of the environment.

In this article we present a brief overview of the existing tools and frameworks that handle virtual environments creation and why we chose some of those to be part of pSIVE, followed by a presentation of its architectural aspects and details on the usage. Then we show some preliminary results and discussions from a real case scenario, where the platform was used to provide a virtual copy of a production line with information associated to different elements, also to act as an evaluation environment to assess different styles of selecting tri-dimensional elements and allowing users to try the platform and leave their comments and opinions. Finally, some conclusions are presented.

9012-24, Session PTues

MARVEL agents: mobile augmented reality for validating embodied learning

Quincy Conley, Robert K. Atkinson, Arizona State Univ. (United States)

The purpose of this proposed submission is to share an educational application of augmented reality (AR) that could potentially provide engaging, collaborative and embodied learning experiences for understanding concepts related to statistics education. The hope for this session is to share and discuss our progress on a study we are conducting to develop new empirical strategies for leveraging mobile and AR technologies for learning statistical reasoning skills such as calculating probability and sampling.

Statistics reasoning impacts virtually every aspect of our daily lives. However, it seems that across all levels of education (elementary, middle school, high school, or college), teachers have to assume no knowledge of statistics, thus, starting from scratch each time they set out to instruct students. This issue could be the result of how teaching these essential statistical concepts are currently taught. Statistical reasoning concepts are typically taught in two parts: (1) students are given the theoretical perspective in the form of a lecture or a reading assignment, and (2) separate from the instruction, they are given homework to practice applying the theory. What is problematic with this common but passive method of instruction is that it requires the students to understand the theory disconnected from the context in which they will practice in most cases. To progress beyond this passive instructional practice, more reliance needs to be on active techniques such as collaboration, embodiment, and situated learning theories.

The participants targeted for our study consist of approximately 180 undergraduate students from a major university located in the southwest United States. The participants were randomly assigned to a one of three levels of the independent variables (high AR experience permits the participants to move around the space, low restricts the participants to complete the tasks without moving around the space, and the control group will complete the same tasks without the use of AR). The learning task for this study required students in pairs to calculate the pollution level of a body of water by calculating proportion of fish infected by pollutants. By analyzing either 3-D or paper-based images of fish of varying color, size and species, students will informally estimate sample size, statistical proportions, and identify factors affecting sampling variability. Additionally, the learning task covered topics such as populations, samples distributions, and what it means for a sample to be representative of the population. During the AR experience scenarios, students used tablets wirelessly connected to the Internet to view different areas of a bay, which will project various 3-D, interactive images of fish. Using a 3 x 1 research design, we compared and analyzed the performance of students given three learning conditions; a high AR experience (with movement throughout a physical space), a low AR experience (without movement throughout a physical space), and business as usual (2-D images and text). As a starting point for further research, the intent of this session is to receive feedback on our research and to also stimulate further discussions on how to design and examine AR learning experiences that can foster the acquisition of knowledge and skills among a young generation of learners.

9012-26, Session PTues

Ergonomic approaches to designing educational materials for immersive multi-projection system

Takashi Shibata, Tokyo Univ. of Social Welfare (Japan); JaeLin Lee, Tetsuri Inoue, Kanagawa Institute of Technology (Japan)

Rapid advances in computer and display technologies have made it possible to present a virtual reality (VR) environment in which a user experiences a feeling of immersion. Virtual environments are applied to various fields including simulations, training, entertainment and telecommunications. To use such virtual environments effectively, research should be carried out into how the users see, perceive, and react to the virtual environments in view of human factors. In the field of VR technology, we have a great interest in application of VR technology to education because immersive feeling could be useful for experiential learning. In this study, we created the VR content of sea fish for science education, and then conducted an experiment to examine how observers perceived object's size and depth within their grasps.

The purpose of the VR content was to learn the type of sea fish and a variety of fish habitats. The VR content was created for a multi-projection system that had three rear-projection screens (front, right, and left) arranged in a 2.5m cube, and one screen (floor) projected by a combination of a front projector and a mirror. Stereoscopic images were presented by two projectors with polarized filters. There are two reasons why we chose the multi-projection system for presenting the educational VR content. First, the system can provide actual-size objects (e.g., Kenyon et al., Presence, 2007). It is thought to be important in education like learning the type of fish. The observer can compare the size of object with his/her body. Second, the system can produce stereo images located close to the observer so that the observer can reach the 3D image. Conventional 3D displays generally cannot show stereo images with a high value of disparity enough to close to the observer for several reasons.

In order to obtain the basic findings for effective use of the educational VR content, we conducted an experiment that examined the perceived size and depth of an object. The object was several kinds of fish extracted from the VR content. The background scene was the bottom of sea. The object was presented at five different distances from a subject: 47cm, 54cm, 63cm, 96cm, and 130cm. These distances correspond to theoretical positions calculated by binocular disparity: six, five, four, two, and one degrees, respectively. The distance from the subject to the front screen was 200cm. The sizes of the object were 14cm, 27cm, and 46cm. Two experimental conditions were established, one with and one without body action. In the condition of body action, observers

were able to estimate object's size and depth by moving their own arms. Nine subjects participated in the experiment. The results showed that both size and depth were more accurate in the body-action condition than in the non-body-action condition. It suggests that the presenting images within observer's grasp could be useful for education in virtual environment.

We are also running an experiment to evaluate visual fatigue when viewing stereo image with a large disparity in virtual environment. Our pilot test showed a possibility to be different from the case of viewing conventional 3D displays (e.g., Shibata et al., Journal of Vision, 2011). The detail will be discussed in the presentation.

Conference 9013: 3D Image Processing, Measurement (3DIPM), and Applications 2014

Wednesday 5 –5 February 2014 • Part of Proceedings of SPIE Vol. 9013 Three-Dimensional Image Processing, Measurement (3DIPM), and Applications 2014

9013-12, Session PWed

Human machine interface by using stereo-based depth extraction

Chao Kang Liao, Chi Hao Wu, Hsueh Yi Lin, Tung Yang Lin, Ting Ting Chang, Chin Chun Hsiao, Po Kuan Huang, Imec Taiwan Co. (Taiwan)

Hand gesture detection and body posture recognition have become an extensive investigation topic of the human machine interface (HMI). In order to capture a reliable body posture, user images associating with the depth map is commonly used for simultaneously three-dimensional information acquisition. Depth map capture methods have been proposed for HMI system purpose. They can be considered into three major ways, including the time-of-flight camera, structural light depth measurement (such as Microsoft Kinect), and the stereo matching method as the proposed method here. The former two methods generate a quality depth map with an active infrared light source which produces a temporal-coded pattern or a spatial-coded. Therefore, a considered problem of these methods as compared to the stereo matching method would be the interference of environment light. In addition, these methods require to equip their specific image sensors to capture the pattern from the light source. On cost and size perspective, the stereo matching is in a better position on commodity product.

In this article, a real-time stereo-based HMI system is present to generate on-the-fly user skeleton for HMI purpose. This HMI system consists a stereo camera, a FPGA-based depth generator, and a PC with a skeleton extraction middleware. A lattice dual sensor camera is employed to be the stereo image source. A rectification module followed by a local-based stereo matching algorithm are adopted. A stereo matching algorithm was built to process up to 128 pixel disparities. Sub-pixel estimation was also built to achieve 1/8 pixel resolution that drives the output up to 1024-disparity levels. The generated depth image is then processed by an image enhancement engine to improve the object edge and the textureless region. Those are the two major common challenges for the stereo-matching technology. This low-latency stereo matching system is built on a FPGA system to obtain the real-time depth map. The depth map stream is then transferred to a PC via a USB3.0 interface at a resolution of 1280x720 and a frame rate up to 60FPS for skeleton extraction. Since the image size and frame rate of this depth extraction system is in a considerable condition, the depth map stream is adoptable for many depth applications. In order to test the performance of the stereo-matching system, a middleware with USB video class (UVC) host was built to be the interface of OpenNI data source and to generate the skeleton data for application software. We conduct two applications to examine the accuracy of 3D body gesture recognition. The first experiment is the recognition accuracy of single user's body gestures. The examined body gesture recognition software constructs the body skeleton and provides 3D gesture library including hand circling, punching, swing, twisting sideward, slight leaning, and defense. The accuracy of body gestures can exceed 92% consistently when using a regular stereo camcorders. The second experiment is the multiuser skeleton recognition. The purpose of this experiment is to test whether the depth stream accuracy can allow the body gesture software to construct the skeleton for each user and recognize their gestures. During this experiment, two to four users stand closely and do body gestures independently. The result shows that multiuser scenario does not impact the accuracy of body gesture and skeleton construction. The design challenge of an efficient hardware architecture which meets the real-time requirement of high resolution 3D gesture games and high tolerance of lighting condition, variance of lens distortion and stereo image misalignment were addressed. An accuracy of less than 1/32 pixel error was measured after the on-the-fly rectification engine leading to a disparity resolution up to 1/8 pixels in the output.

9013-14, Session PWed

Tabu search for human pose recognition

William J. Dyce, Univ. Montpellier 2 (France); Nancy Rodriguez, Lab. d'Informatique de Robotique et de Microelectronique de Montpellier (France); Benoit Lange, Univ. Pierre et Marie Curie (France); Sebastien Andary, Antoine Seilles, NaturalPad (France)

Computer vision techniques are widely used to extract meaningful human poses from a stream of images.

Over the last few years Microsoft's Kinect depth sensor and accompanying software libraries (Kinect for Windows, NITE, ...) have lowered the barrier of entry to human movement recognition. Our interest is in using the Kinect for movement analysis, in line with the work of to Oulasvirta et al. and Stone et al. We are interested in particular in applying this movement analysis in a therapeutic context, for automatic post-stroke mobility evaluation for example.

Microsoft's "Kinect for Windows" library estimates the user's pose in the form of tree graph of labeled nodes representing joints, and presents this to the application programmer via a software interface. Primesense, the company responsible for developing the Kinect sensor, has made available similar pose-tracking middleware. Both these libraries are closed-source and implementation details are scant, but it is known that Microsoft Research reformulated the problem of body part recognition as a classification problem and used extensive machine learning to train their classifiers.

There are many problems with this approach which hamper its use for movement analysis, as Oulasvirta et al. noted in the context of their work on body movements. To begin with poses not present in the training set, for instance lying down or facing away from the camera, are not detected. Furthermore since the algorithm simply connects the patches together to form a "skeleton", it is possible for the resulting skeleton to defy the constraints of the human body, for instance by putting an arm through its torso. There is also no "bone" rotation information: bone roll values are needed, for example, to measure "dorsiflexion" during a post-stroke Fugl-Meyer et al. assessment of sensorimotor recovery.

It is with these limitations in mind that we present the beginnings of an alternative solution to the real-time human pose estimation, or "user skeleton extraction", problem.

For our implementation we used the open source OpenNI library to stream depth-labeled pixels from the infrared sensor. OpenNI depth images use lower values to represent closer areas and higher values to represent more distant ones, though the 0 value (black) is used for "dead" areas which are either too near, too far or subject to UV interference (from the sun or other Kinect sensors).

Our method uses the entire set of depth-labeled pixels provided by the Kinect sensor. An initial preprocessing phase separates the user's silhouette from the background. Next we approximate the appendage positions by calculating 5 paths from the centre of the silhouette with geodesically distant endpoints. Finally we remove as many path nodes as is possible while remaining inside the user silhouette.

We scan the histogram's categories from nearest (darkest) to furthest (lightest) looking for one containing at least T depth pixels. We then move backwards to the first category containing fewer than TxA and forwards to the first category containing fewer than TxB elements. Everything outside of this depth interval is set to 0 ("dead") on the map and ignored.

The cvBlob library contains an implementation of the Senior et al. blob detection algorithm. We use cvBlob to identify the pixel-area occupied by the user (assumed to be the largest blob) and filter out the rest. From here on our method uses the depth-labeled silhouette of the user.

We consider each of the remaining depth-pixels to be a node connected

to its 0-8 neighbours in the depth image. Two pixels adjacent are considered to be connected unless the difference in value (depth) between the two is less than a given epsilon. Thus if an arm is front

of the body, for instance, paths will not cross it. The difference in depth is also take into account when calculating the geodesic length of a path. It should be noted that for our prototype we downsample the original image to 1/8th its original size to speed up calculations: the performance increase more than makes up for the slight decrease in accuracy.

Tierny et al. identify an initial set of prominent features based on the geodesic distances of each vertex from the extremities of the mesh diameter. Their algorithm makes no assumptions about the topology of the model, but since we are dealing with humans we can generally assume that there will be 5 appendages. As such we search not for the two most geodesically distant points, but rather for the 5 points most geodesically distant both from the blob centroid C and from each other.

These 5 optimal “appendage paths” are approximated sequentially using a modified version of Dijkstra’s algorithm, rooted at the blob centroid. We perform a best-first graph exploration in order to find the path P for which F(p) = length(p)/tabu(p) is maximal.

“Tabu search” is a local optimisation technique proposed by Glover et al. which avoids local optima by placing a tabou on previous results. Here we optimise appendage length by applying a “tabu” value to the neighbourhood of each appendage end-point that we discover. This adds an artificial penalty cost to future paths which enter the same geodesic region, so ensures that the 5 appendages diverge.

Since the end goal is an articulated skeleton we need to remove as many nodes as possible from the appendage paths we discovered during the previous step so as to arrive at a small number of long, straight segments; However the resulting ``bone” segments should not leave the user’s silhouette.

Starting at the tip of the appendage path we explore backwards along its path towards the silhouette centroid, using Bresenham’s line algorithm to perform ``ray casts”. In so doing we can calculate the number of pixels D outside of the silhouette that the candidate bone will cross.

For B the length of the candidate bone and K a constant, a new bone is started when BxD > K and the process is repeated until we arrive back at the silhouette centroid.

In this paper we provided a short summary of our technique for extracting the topology of a user from a single frame of depth-data in order to infer their current pose. This is done by finding 5 appendage paths with end-points that are geodesically distant both from the center of the user’s silhouette and from each other.

Our application was developed in C++ using the OpenNI, OpenCV and cvBlob libraries: it is not yet optimised, but is already capable of processing several frames per second on consumer hardware. We hope with more work to attain real-time speeds.

Aside from speed optimisations, future work will look at ways to improve on the na”{i}ve foreground segmentation method, to place bone joints closer to the center of the blob using sphere-packing in a similar manner to Baran et al. and to fit the final graph to an anatomical model of a human being for use in natural interaction applications. We hope for this to adapt the work of Kar et al.

9013-16, Session PWed

A multiple wavelength unwrapping algorithm for digital fringe profilometry based on spatial shift estimation

Pu Cao, Jiangtao Xi, Yanguang Yu, Qinghua Guo, Univ. of Wollongong (Australia)

This paper studies the unwrapping problem associated with the spatial shift estimation based approach for 3D profile measurement proposed in [1], and presents a multiple wavelength unwrapping algorithm to solve the problem.

In recent years, optical noncontact three-dimension profile measurement has attracted increasing research efforts due to its distinct advantages over contact methods. Among other approaches, the fringe pattern profilometry based on digital fringe projection has been proven to be one of the most promising techniques due to the advantages of simple system structure, flexible fringe pattern generation and high accuracy. Several fringe pattern profilometry approaches have been developed during the past decades. The most widely used methods are on the basis of phase difference estimation. In these approaches, the projected fringe patterns are sinusoidal or periodic, and the deformed one reflected from the object surface is considered as the result of phase modulation of the original fringe pattern. Hence the detection of phase maps from original and deformed fringe patterns enables the retrieval of the 3D shape of object surface. Although phase based approaches have been considered as the most popular, they suffers from a number of weaknesses. A major restriction is that fringe patterns must be either sinusoidal or ideal periodic. However such a requirement is hard to meet in practice due to some factors, such as the nonlinear distortion on inherent to digital video projections. In order to solve the problem, Hu et al [1] proposed a spatial shift estimation based profilometry approach. Compared to the conventional phase difference estimation profilometry approaches [2,3], the spatial shift estimation tries to obtain the 3D profile by evaluating the spatial shift between the two fringe patterns, one projected on the reference plane and the other on the target object. With the spatial shift estimation technique, the projected fringe patterns are no more required to be sinusoidal, and accurate reconstruction can still be obtained when there exist nonlinear distortions on the fringe patterns.

Although the spatial shift estimation approach is suitable for any fringe pattern, use of a periodic fringe patterns is still necessary in order to have enough measurement resolution. In the spatial shift estimation-based approaches, spatial shift between corresponding pixels on the two fringe patterns is arbitrary, it can only be detected without ambiguity within the range of [0, ?], where ? is the wavelength, or the spatial width of the individual fringe, i.e., number of pixels per fringe stripe. Obviously, shift unwrapping is also required in order to correctly restore the 3D shape of the object surface. However, spatial shift unwrapping for complex object using spatial shift estimation-based fringe pattern profilometry is still an outstanding issue, which motivated the work presented in this paper.

Since the spatial shift unwrapping problem exists in spatial shift estimation approach which is similar to the phase unwrapping problem in phase difference estimation based fringe pattern profilometry, a review of the phase unwrapping problem is given in this paper. Phase unwrapping problem is a major problem associated with phase difference estimation-based fringe pattern profilometry approaches. This problem arises because the phase difference can only be detected within the main value range of $[-\pi, \pi]$, but the true phase difference can be arbitrary. In order to retrieve the actual surface shape of the object, phase unwrapping must be carried out to obtain the actual phase maps. To solve the unwrapping problem in phase difference estimation approach, Zhang et al [4] introduced a multiple wavelength phase unwrapping algorithm. In his method, an image which only has a single fringe covers the whole measurement area is first projected, and then a series of fringe images with a wavelength decreased by a factor from its previous wavelength is used. Since the first image only contains one fringe, the unwrapping step is not required. Then the phase of second image can be unwrapped by referring the first image. After the phase of second image is obtained, it can be used to correct the third image. In general, the phase of each wavelength is unwrapped by referring to the unwrapped longer wavelength phase pixel by pixel.

Based on Zhang’s method, we introduce a multiple wavelength unwrapping algorithm for spatial shift estimation approach. In our method, a series of fringe images with a wavelength decreased by a factor from its previous wavelength is also applied. The spatial shift of each wavelength is then unwrapped by referring to the unwrapped longer wavelength spatial shift pixel by pixel. Since the longest wavelength covers the whole measurement area, no spatial shift unwrapping step is necessary. The proposed method solves the spatial shift unwrapping problem in spatial shift estimation, hence it can enable the measurement of complex objects with significant step height or multiple separate objects using spatial shift estimation approach.

This paper is organized as follows. In this paper we firstly give a brief introduction on the conventional phase difference estimation based fringe pattern profilometry and the spatial shift estimation based technique, including their principles, system structures and relevant algorithms. Then we indicate that the unwrapping problem exists in spatial shift estimation approach which is similar to the phase unwrapping problem in phase difference estimation based fringe pattern profilometry. The paper then gives a review of the multiple wavelength phase unwrapping algorithm introduced by Zhang, based on which we introduce a multiple wavelength unwrapping algorithm for spatial shift estimation approach. Finally experimental results are given to demonstrate the proposed method can be used to measure complex objects with significant step height or multiple separate objects using spatial shift estimation approach.

References:

[1] Y. Hu, J. Xi, E. Li, J. Chicharo, and Z. Yang, “Three-dimensional profilometry based on shift estimation of projected fringe patterns,” Applied Optics, vol. 45, no. 4, pp. 678–687, February 2006.

[2] X. Su and W. Chen, “Fourier transform profilometry: a review,” Optics and Lasers in Engineering, vol. 35, pp. 263-284, 2001.

[3] M. Halioua and H. C. Liu, “Optical three-dimensional sensing by phase measuring profilometry,” Optics and Lasers in Engineering, vol. 11, pp.185-215, 1989.

[4] S. Zhang, “Phase unwrapping error reduction framework for a multiple-wavelength phase-shifting algorithm”, Optical Engineering, vol. 48, pp.105601, 2009.

9013-17, Session PWed

Experimental demonstration of parallel phase-shifting digital holography under weak light condition

Lin Miao, Kobe Univ. (Japan); Tatsuki Tahara, Kansai Univ. (Japan); Peng Xia, Yasunori Ito, Yasuhiro Awatsuji, Kyoto Institute of Technology (Japan); Osamu Matoba, Kobe Univ. (Japan)

Digital holography is a technique to record the information of three-dimensional (3D) objects as a digital hologram by using optical interferometer with a digital image sensor. For the reconstruction of 3D objects, numerical wave propagation is used and the quantitative evaluation is possible. It is not required to realize the same speed of the recording, but GPGPU can be used for fast reconstruction. There are several methods in digital holography to improve the reconstructed image quality both in the recording and the reconstruction processes. Here, we focus on the recording methods.

One of the most successful methods in digital holography is phase-shifting digital holography. In phase-shifting digital holography, at least two holograms are required to extract the complex amplitude distribution of the object wave. Therefore, it cannot be applied to recording a fast 3D event such as a movable object or a deformable object because the sequential recording of holograms with appropriate phase retardation is required. Parallel phase-shifting digital holography (PPSDH) solves this problem. In PPSDH, an array pattern consists of at least two amounts of phase retardation are described in the spatial distribution of the reference light. The detected hologram is called as multiplexed hologram that includes all phase retardation information. After divided into at least two subholograms according to the phase retardation amount, interpolation is applied to obtain the values in the blanked pixels. After that, the conventional phase-shifting method is applied to obtain the complex amplitude of the object wave.

One of the attractive features of PPSDH to other digital holography methods is fast recording of the multiplexed hologram. So far, 3D recording with 180,000 frames per second was demonstrated [1]. The reconstruction can be implemented by calculating numerical wave propagation of the extracted complex amplitude of the object wave in a computer. When fast dynamic events 3D objects are measured by using PPSDH, observed light energy becomes low. Therefore, minimum optical energy of the multiplexed hologram is required to design the measurement parameters such as illuminated intensity of light to the object and observation time of the multiplexed hologram. We have been evaluated numerically minimum optical energy of the multiplexed hologram in PPSDH by using the photon counting method [2]. When average numbers of photons is 35,565,588 where the laser output energy is about 10 pJ, reconstructed image with 256 x 256 pixels can be seen clearly.

In this paper, we demonstrate the experiments to show the effectiveness of the numerical evaluation. In the experiment, four-step PPSDH is used. A Nd:YAG laser operated at a wavelength of 532 nm is used as a coherent source. An image sensor with a polarization mask is used to capture the multiplexed hologram with the number of pixels of 1024?1024 and intensity resolution of 12 bits. The exposure time is 0.5 ms. Laser output power can be changed by a neutral density (ND) filter to realize weak light condition. Under the detection of weak light of the object wave, it is better to set the intensity of reference wave being larger than that of the object wave. We set the power ratio of the reference to the object waves being 31. Two transparent films are used as objects located at 460 mm and 530 mm, respectively. The power meter is used to measure the optical power of the multiplexed hologram. The detected optical power is changed from 50 pJ to 10 nJ.

We calculate the reconstruction error by subtracting the reconstructed amplitude distribution when the detected power is 10 nJ as an ideal case from that with small detected power. In the reconstructed images, there are speckle noises. Therefore, we applied a low-pass filter of 31 x 31 pixels to reduce the effect of the speckle noise. With the increase of laser output power, the reconstruction error decreases monotonically. For the 5% reconstruction error, the required optical energy is 2 nJ. We compare the experimental and numerical results. In the numerical results, the required energy with the same conditions in the experiment is 0.42 nJ for 5% reconstruction error. The optical energy required in the experiments is about 5 times larger than that of numerical simulation result. If the sensitivity of image sensor with polarization mask is set to be 1/5, the experimental results are in good agreement with the numerical results. Sensitivity of 1/5 in the image sensor is reasonable because 1/2 reduction is caused in the polarization mask due to the circular polarization and quantum efficiency of the image sensor is about 0.4. Therefore, 2 nJ is required for the successful reconstruction in PPSDH when the pixel number is 1024 x 1024. This value is useful for setting the measurement parameters of illuminated intensity to the object and the exposure time of the multiplexed hologram.

References:

[1] T. Kakue, R. Yonesaka, T. Tahara, Y. Awatsuji, K. Nishio, S. Ura, T. Kubota, and O. Matoba, “High-speed phase imaging by parallel phase-shifting digital holography,” Opt. Lett., Vol. 36, pp.4131-4133 (2011).

[2] L. Miao, K. Nitta, O. Matoba, and Y. Awatsuji, “Assessment of weak light condition in four-step parallel phase-shifting digital holography,” Appl. Opt. Vol. 52, pp. A131-A135 (2013).

9013-18, Session PWed

Global color calibration for 3D images through polynomial fitting

Pierre Yver, Sébastien Kramm, Abdelaziz Bensrhair, Institut National des Sciences Appliquées de Rouen (France)

Introduction

In the field of consumer or industrial 3D viewing, several aspects can influence the user experience and can lead to negative perception.

The geometric aspects are the most evident but the color consistency between the two images is also very important. Even minor differences of colors of the same area can lead to visual fatigue. As it can not be guaranteed that the cameras have exactly the same chromatic settings, it is thus important to minimize color differences by post processing the image pair.

This can be achieved manually for some applications but several situations may need a real time color correction, for example medical applications (endoscopy) or real-time broadcasting of sports events.

With a standardized color chart one can then manually adjust the camera settings in order to have exactly the same chromatic response. However, this manual approach has several defaults. Moreover, cameras parameters may be subject to drifting.

In previous works, the authors either used a color chart [Ilie et al.,2005] or computed a matching using local descriptors [Tehrani et al., 2010; Wang et al., 2011; Hasan et al., 2011]. This sparse matching is then used to compute a color model around these points.

Proposed method

In this paper, we propose a novel auto adaptive method that can correct the color differences between two aligned images by using the disparity map that give the relation between the position of a same 3D point in the two images. We consider here one of the images as the reference image, and correct the colors of the second one so that it matches the first image.

Besides the specific application of 3D video production, this technique can be useful to other fields such as computer stereovision, image stitching or Free Viewpoint Video systems.

The disparity map can be computed using classical techniques, most often using the processing power of modern GPU, for example [Mei et al., 2011].

It gives for every pixel, the position of the corresponding pixel in the other image. The map does not need to be complete: with many methods, homogeneous areas or region boundaries are difficult to match, thus leaving holes in the disparity map. Filtering techniques have been proposed to improve those maps and fill these and smooth the variations in order to have many accurate points.

The main advantage of using a dense map instead of only sparse features as in previous works is that we are able to build a model using the whole range of pixel values. In sparse approaches, some color values will likely be missing in the computed correspondences, thus the correction will fail for these colors.

Using the disparity map, we extract for each color plane the corresponding pixel value in the two images and build a set of data points.

For color correction, we follow here the global approach described in previous color-correction works and remap all pixel values from one image to a new value given by the correction curve. However, we do not limit ourselves to the classical Gamma-Offset-Gain (COG) model. Instead, we use here a more general polynomial model that can handle more difficult situations. We use an iterative non-linear least squares fitting algorithm to find an estimation of the best numerical model.

Raw fitting with all the data points will be polluted by the many outliers of the dataset. These come from matching errors in the disparity map (especially in the occluded area). To improve the fitting step, we add an outlier removal step. Two approaches were experimented. The first is an iterative method where we progressively remove points that have an error value that is above a threshold, which is computed dynamically using mean and standard deviation of all the error values.

The second method is a “single-step” outlier removal technique. It is based on the idea that there must be an optimal value for the coefficient of the standard deviation of errors. We provide in the experimental section the description of a short algorithm that can be used to determine this coefficient.

Experimental results

Experiments were processed on both artificially color altered image pairs and on real image pairs. For performance criterions, we computed both the Peak Signal to Noise Ratio (PSNR) and the mean value of the CIEDE2000 image metric that claims to reflect human perception (we seek to have the same average value in luminance and chrominance in both images). Experiments show high quality results. For the polynomial fitting, experiments show that limiting the degree to 4 is sufficient because 5 is sometimes too close to the dataset.

We test or algorithme on several colorspaces (RGB, YUV, YCrCb, XYZ, sRGB, CIELab, HSV). The experiments show that the best colorspaces to work on are the classical RGB and the YCrCb. Some of these are close (RGB and sRGB or YUV and YCrCb …).

On a set of real images using the RGB space the mean value of PSNR was 16.68058dB for the blue channel, 18.60726dB for the green channel and 21.17428 dB for the red while the mean of the CIELab Delta 2000 was 10.24934. After correction we obtain 19.39023dB for the blue, 20.17619 dB for the green and 21.28199dB for the red while the mean of the CIELab Delta 2000 is 6.40992.

The ratio of the Left/Right luminance and the ratio of the both Left/Right chrominance are also well scaled. From 0.89415 to 1.00110 and from 0.96216 and 1.06693 to 0.99789 and 1.00319.

The described method is used to compute a Look Up Table for the different channel of the image in the desired color space and then regularly update coefficients.

Further work will study the benefits of using multiple frames as input data instead of a single pair.

References

* Ilie and Welch, “Ensuring color consistency across multiple cameras”, ICCV 2005.

* Hasan, Stauder, Tremeau, “Robust Color Correction for Stereo”, CVMP 2011.

* Tehrani, Ishikawa, Sakazawa, Koike, “Iterative colour correction of multicamera systems using corresponding feature points”, JVCIP 2010.

* Wang, Yan, Yuan, Li, “Robust color correction in stereo vision”, ICIP 2011.

* Mei, Sun, Zhou, Jiao, Wang, Zhang, “On Building an Accurate Stereo Matching System on Graphics Hardware”, Third Workshop on GPUs for Computer Vision, Barcelona, 2011.

9013-1, Session 1

Temporal consistent depth map upscaling for 3DTV

Sebastian Schwarz, Mårten Sjöström, Roger Olsson, Mid Sweden Univ. (Sweden)

1.“What is the addressed scientific topic or problem?”

The ongoing success of three-dimensional (3D) cinema fuels increasing efforts to spread the commercial success of 3D to new markets. The possibilities of a convincing 3D experience at home, such as three-dimensional television (3DTV), has generated a great deal of interest within the research and standardization community.

A central issue for 3DTV is the creation and representation of 3D content. Scene depth information plays a crucial role in all parts of the distribution chain from content capture via transmission to the actual 3D display. This depth information is transmitted in the form of depth maps and is accompanied by corresponding video frames, i.e. for Depth Image Based Rendering (DIBR) view synthesis.

Acquiring scene depth information is a fundamental task in computer vision, yet complex and error-prone. Dedicated range sensors, such as the Time-of-Flight camera (ToF), can simplify the scene depth capture process and overcome shortcomings of traditional solutions, such as active or passive stereo analysis.

2.“What are the challenges and barriers faced?”

Stereo analysis is a common approach to scene depth extraction. Feature and area analysis between two camera views allows for the reconstruction of depth information based on the camera geometry. However, if parts of the scene are occluded in one view or areas have low or repetitive texture, stereo matching produces erroneous results.

Continuous-wave ToF sensors can overcome most of these shortcomings. They capture scene depth in real-time, independent from texture structure and occlusions. Admittedly, currently available ToF

sensors deliver only a limited spatial resolution. Thus we have seen a multitude of ToF depth upscaling proposals in the recent years. Many of these approaches utilize additional guidance information in the upscaling process, such as texture or edge information from video sources, depth reading reliability and ToF noise models.

Yet, one major guidance source is widely ignored: The previous results from time-consecutive sequences. Since ToF sensors can provide depth readings in 60 frames per second or more, previous depth readings were treated as expired and are discarded. However, for 3DTV scenarios it is beneficial to achieve temporal consistency in depth to avoid flickering artifacts in the DIBR view synthesis. The big question is how you determine which parts should be consistent and which parts are allowed to change.

3."Why this is important for the 3D community?"

Guided ToF depth upscaling simplifies scene depth acquisition for 3DTV and is therefore a highly active research area. By adding temporal consistency constraints to the upscaling process, we can reduce disturbing depth jumps and flickering artifacts in the final 3DTV content.

In summary, guided ToF depth upscaling allows for simplified 3D content generation, leading to more available content. Temporal consistency in depth maps enhances the 3D experience, leading to a wider acceptance of 3D media content. More content in better quality can boost the commercial success of 3DTV.

4."What is the original method proposed to address this problem or issue?"

The most common guided ToF upscaling algorithms base on joint-bilateral filtering, Markov random fields or error energy minimization problems. In 2012 we proposed and upscaling routine based on error energy minimization, weighted with edge information from an accompanying video source. In the original contribution we enforced spatial smoothness between neighboring pixels and assumed that edges in texture correspond to object transitions in depth. To avoid blurring at these important transitions, we weight the spatial smoothness constraints with texture edges, so that pixels at edges are less constrained to be similar. Since texture sources usually contain more edges than actual depth transitions, we mask the texture information with low resolution depth edge information. Combining the spatial smoothness constraint for a high resolution frame with the low resolution ToF depth, yields spatial error energy Qs. Minimizing this energy yields a high resolution depth map.

For this article we will introduce the additional temporal consistency constraint. The temporal error energy Qt is based on the value difference of a depth map pixel in two consecutive frames. But flat, constant temporal consistency is not what we are looking for. This way we would end up with a still image instead of a 3DTV sequence.

To decide which parts of a depth map should be consistent in relationship to the previous depth map, we utilize optical flow information. The optical flow between the accompanying texture frames is calculated and used as a weight for the temporal error energy Qt. Pixel with a high optical flow are considered moving and are therefore less constrained to be similar to their temporal neighbors. Minimizing the combination of spatial error energy Qs and temporal error energy Qt yields a temporal-consistent, high resolution depth map.

5."What is the novelty comparing to the state of art?"

The main novelty of this research is the introduction of temporal consistency to the ToF depth upscaling problem.

Extracting a good temporal weight from the optical flow information is crucial for this approach. And so is the right ratio between spatial and temporal error energy in the final optimization step. We determined these two factors based on a non-linear system identification process.

6."What is the efficiency of the method (presentation of results and comparison with the state of art)?"

We evaluate our results based on a per-pixel flicker measure presented Schmeing et al. in 2011. We upscale depth maps with different approaches and compare view syntheses with upscaled depth to original camera footage. In the first preliminary results we used the test sequences "Street" and "Hall2" from Poznan University. Our results show that our original implementation creates about 5% depth flicker, while with the addition of temporal consistency, we have only 3%. For comparison, joint-bilateral upscaling (JBU), a standard ToF upscaling solution, creates 9% flicker.

With our new temporal constraint we can reduce the video flicker due to depth inconsistency by about one, in comparison to our previous proposal, or even two thirds, in comparison to JBU.

9013-2, Session 1

Feature enhancing aerial lidar point cloud refinement

Zhenzhen Gao, Ulrich Neumann, The Univ. of Southern California (United States)

1. What is the addressed scientific topic or problem?

Aerial LiDAR (Light Detection And Ranging) point clouds captured using commercial laser range scanners are invariably noisy, mostly caused by scanner artifacts and alignment errors between scans. Additionally, occluded or sharp regions (such as boundaries, ridges, ravines, crest lines, etc.) are often under-sampled. Due to noise and under-sampling, a direct rendering or surface reconstruction of raw points produces grainy surfaces, gaps, and irregular boundaries. These artifacts are especially visually-disturbing for buildings where planar surfaces and straight lines are universal, as shown in Figure 1 (a) (http://www.screencast.com/t/0cIe3ibt) a rendering example of a raw building roof.

Refining raw points in a pre-process is one effective way to alleviate above visual symptoms. As buildings are the most important objects in urban scenes, this research focuses on refining raw aerial LiDAR building points.

2. What are the challenges and barriers faced?

In contrast to point data gathered by other scanners such as terrestrial LiDAR, aerial LiDAR building points bear two unique properties that make the refinement more challenging. Firstly, as captured by scanners equipped on a low-flying aircraft, the sampling resolution of aerial LiDAR points is relatively low. Compared to other point data that have hundreds or even thousands of points/m^2, a typical aerial LiDAR data only has 10-20 points/m^2. Most refinement approaches rely on dense data to infer and reconstruct the latent object. The low sampling rate makes those approaches inapplicable. Due to lack of information, the goal of accurately recovering the underlying surfaces for aerial LiDAR points is impractical.

Secondly, aerial LiDAR data can be characterized as being 2.5D in nature. For a building, the sensor is only able to capture detailed roofs but few samples on vertical walls, leaving roof boundary points discontinuous in the positions. This property pose a great challenge to any energy minimization approach as the neighbors of a boundary point only covers part of the environment, missing energy contribution from the rest of the environment. As far as we know, existing point-based refinement approaches can only regularize normal-discontinuous-features such as ridges, ravines and crest lines, few can handle such boundary points which we call position-discontinuous-features.

3. Why this is important for the 3D community?

Our refinement approach provides an effective way to alleviate visual artifacts while preserving and enhancing both normal- and position-discontinuous-features of raw aerial LiDAR building points. The refinement is a beneficial pre-process to improve quality of many other 3D applications that rely on raw points, such as direct rendering and surface reconstruction of urban data.

4. What is the original method proposed to address this problem or issue?

To make the problem tractable, we take as input a raw oriented aerial LiDAR point cloud of a single building. The building consists of either a single roof block or several roof blocks of different heights. The only assumption we make is piece-wise smoothness of roofs regardless of shape and complexity. The output is a new set of points with

smoothed noise, filled gaps, and enhanced both normal- and position-discontinuous-features.

The presented refinement approach extends recent developments in geometry smoothing and up-sampling to accommodate unique properties of aerial LiDAR building points. Importantly, position-discontinuous-features, i.e., boundary points, are explicitly regularized. The feature-aware two-step approach is guided by normal and boundary direction (direction of the underlying line of a boundary point). The first smoothing step applies a two-stage robust bilateral filtering, which first filters normals and then updates positions under the guidance of the filtered normals. A separate similar pass explicitly filters boundary directions and then updates boundary positions to match the new directions. Importantly, boundary directions are regularized in two parts involving different neighbors: on the x-y plane and along z-axis. This smoothing step effectively removes noise as well as preserves features. Figure 3 (http://www.screencast.com/t/3lyNNxWF) shows the process of the smoothing step. The second up-sampling step detects gaps on the latent surface with a local detector, and then fills with interpolated new points via a feature preserving bilateral projection operator. Gaps on boundaries are detected and filled explicitly under a similar process. A global uniform density can be achieved, which is adjustable by a single global parameter of a neighborhood radius. Finally, features can be enhanced by a simple extension of up-sampling only in the vicinity of features.

5. What is the novelty comparing to the state of art?

The presented point cloud refinement operates directly on points. It is fast, robust, simple to implement, and easily extensible. Comparison with previous work:

* Previous point-based refinements are limited to regularize normal-discontinuous-features, the presented approach can handle both normal- and position-discontinuous-features.

* A recent surface reconstruction method fits planes to points and locates creases/corners as plane intersections. By re-sampling planes back to points, the point representation can be preserved. However, geometry fitting is known to be time-consuming and sensitive to noise. Our approach does not require fitting to regularize features while preserving the point representation throughout the process.

* Geometry modeling can regularize position-discontinuous-features as intersections of fitted planes. They have shown to be effective for straight boundaries. But they suffer two limitations: i) Points are converted to lines/curves that cannot be taken as input for point-based applications. ii) Because of the planar assumption on intersecting surfaces, curved boundaries are not well handled. Without any geometric assumption, our approach provides smoother boundary which is also more accurate compared to the ground truth image.

6. What is the efficiency of the method (presentation of results and comparison with the state of art)?

Both synthetic and real data are tested. Through visual comparisons and quantitative measurements, experiments show that:

* The approach is robust for various noise levels

* The refinement quality is stable for vastly different point density

* The approach is applicable to buildings with various roof shapes and complexity, The refinement provides smoother planar areas, filled gaps, and sharpened both normal- and position-discontinuous-features (either straight or curved boundaries).

* The approach is easily extensible to other objects (e.g. ground) as well as to a larger data with several buildings.

* The refinement is fast in the sense that every step takes only a few seconds for all test data running on a single-core consumer-level computer. For instance, to refine a building with 23.5k points, the smoothing step takes 2.07 seconds, the up-sampling step takes 5.84 seconds to increase the point number to 52.3k, and the feature-enhancing up-sampling takes 2.57 seconds to finally output 57.5k points.

9013-3, Session 1

VOLUMNECT: Measuring Volumes with KinectTM

Beatriz Quintino Ferreira, Miguel Griné, Duarte Gameiro, Univ. Técnica de Lisboa (Portugal); João Paulo Costeira, Instituto de Sistemas e Robótica (Portugal) and DEEC, Instituto Superior Técnico, Lisboa (Portugal); Beatriz Sousa Santos, Univ. de Aveiro (Portugal)

1. INTRODUCTION

This paper presents an approach to volume measurement using Microsoft Kinect TM, which has several application scenarios such as warehouses or distribution and logistics companies, where it is important to promptly compute package volumes. For these environments, instead of using the standardized and pricey laser technology (several thousands of dollars), we propose a low-cost solution using the widely available plus quite affordable Microsoft Kinect TM,. In such situations, where the laser precision is not mandatory, we can automatically detect boxy objects from the depth camera data and compute their volume, paving the way for further space optimization. The proposed methodology, based on simple computer vision and image processing methods, provides a new and cost-effective solution for the industry. The attained results are promising, with accuracy close to the laser based volume measurement commercial systems.

In the following sections, a brief description of the used methodology is presented, as well as the main results and conclusions.

2. METHODS

To determine the volume of a cuboid shaped object it is just necessary to measure three edges (length, weight and height); our approach focuses on finding the inliers of the planes of the scene and, subsequently, the vertices that connect adjacent planes so that it is possible to determine each edge length. The application was developed in MATLAB® 7.9.0 (R2009b).

The main image processing methods used to measure the volumes with KinectTM were: RANSAC (“RANdom SAmple Consensus”) to find the planes contained in the scene, connected components to segment the object faces, morphological operations to filter the faces and also the Harris corner detector to find possible vertices.

Next, a system workflow is presented to illustrate the application operation, in which the methods are further discussed.

Considering a known environment with a cuboid object, for instance a box, with three visible edges, both RGB and depth images are acquired, as seen in figures 1 and 2. One should note that this implies a previous calibration of the KinectTM, using the camera model.

From the acquisition, the depth information is used to compute the 3D coordinates (X Y Z), using, again, the camera model, and a 3D point cloud representation is obtained. During this process, the invalid point readings are removed from the set. Having the point cloud, the inliers of the planes can be determined using the RANSAC method.

Once the planes of the object are delimited, the angle between each pair of neighboring planes is computed. From the obtained angles a set of three adjacent and orthogonal planes is selected so that we can determine the right edges to compute the volume.

The processing proceeds with the projection into 2D of each of the three orthogonal planes delimited in a binary image, using labels of connected components. Due to some faults that occur during the capture with the camera, morphological operations (mainly closure) are applied, as well, to filter these 2D images. This filtering is rather important for the next phase, which consists in applying the Harris corner detector to every segmented face, in order to find possible vertices.

The Harris corner detector parameter fine-tuning is crucial to find the pursued vertices. The system should determine the common vertex to the three orthogonal planes, among the corners found to be possible vertices after applying the detector, using Euclidean distances (minimization problem). Having identified the common vertex (emphasized in figure 4), the vertex on the other end of each edge is determined, using the maximum Euclidean distance among the corners found for each couple of planes. With the 3D coordinates obtained mapping the sets of two vertices (the common vertex of the three orthogonal planes plus the vertex on the opposite side of the edge) it is now possible to measure the three lengths required to compute the volume of the object, as represented in figure 5.

These lengths are computed through Euclidean distances between each couple of vertices, and the object volume is, thereafter, readily, calculated.

When the common vertex for the three orthogonal planes is not found, it is not possible to calculate the volume. Therefore, an error message is displayed asking the user to move the object so that this vertex can be computed. This improves our solution robustness, since no inconsistent value is obtained as the final result.

3. RESULTS

In order to evaluate the solution, an accuracy and precision study, using a data set comprising several cuboid objects, was conducted. Figures 6 and 7 show the histograms corresponding to one of the edge lengths and the volume of a specific box, for which we performed 20 tests. In the case presented, the box measurements were 0.375m, 0.145m and 0.118m. Looking at figure 6, which represents the results corresponding to the 20 measures obtained for first edge, it is possible to conclude that the maximum error is 0.008m (2%); figure 7 shows the volume computed, revealing that the maximum error is 0.0010m3 (15,6%), as the real volume was 0.00642m3.

In the final version of this paper, further detail concerning the test set (number and type of objects) and experimental results will be included.

4. CONCLUSIONS

This article presents a methodology to compute volumes of cuboid objects based on the data acquired by the depth camera of the Microsoft Kinect TM. The volumes of the objects from our data set were successfully computed with an accuracy suggesting an adequate compromise between performance and cost regarding the identified application scenarios.

Based on the experiments performed, we believe that better accuracy and precision could be obtained acting upon the following processing steps: exploring background subtraction and improving the Harris corner detector parameters.

9013-4, Session 1

3D-mesh indexing based on structural analysis

Meha Hachani, Lab. d'Informatique de Robotique et de Microelectronique de Montpellier (France); Azza Ouled Zaid, National Engineering School of Tunis (Tunisia); William Puech, Lab. d'Informatique de Robotique et de Microelectronique de Montpellier (France)

This paper presents a novel pattern recognition method based on Reeb graph representation. The main idea of this approach is to reinforce the topological consistency conditions of the graph-based description. This approach enfolds an off-line step and an on-line step. In the off-line one, 3D shape is represented by a Reeb graph associated with geometrical signatures based on parametrization approaches. The similarity estimation is performed in the on-line step. It consists to compute a global similarity measure which quantifies the similitude degree between any pair of 3D-models in the given dataset. The experimental results obtained on the SHREC 2012 database show the system effectiveness in 3D shape recognition.

In this paper, we presented a novel technique for content-based 3D model retrieval. Our contribution was to exploit the Reeb graph theory properties to define an efficient shape descriptor. A thorough experimental evaluation has shown the robustness and the performance improvement of our method, compared to other related works.

9013-5, Session 2

Real-time 3D human pose recognition from reconstructed volume via voxel classifiers

ByungIn Yoo, Changkyu Choi, Jae-Joon Han, Changkyo Lee, Wonjun Kim, Sungjoo Suh, Dusik Park, Samsung Advanced Institute of Technology (Korea, Republic of); Junmo Kim, Korea Advanced Institute of Science and Technology (Korea, Republic of)

This paper presents a human pose recognition method which simultaneously reconstructs a human volume from a depth image in real-time. The human pose recognition has recently gained a considerable attention to provide realistic experiences for various applications such as telepresence, game, virtual fitting, pose correction, rehabilitation, and real 3D contents control on real 3D displays.

However, it is a difficult topic incurred by a limited view of a camera which captures only visible surfaces of a human body. For example, when a user is dancing in a profile view, it is hard to recognize an accurate human pose due to serious self-occlusion. Moreover, if the recognized result is a skeletal pose without any volume, the user may feel a sensational discrepancy when touching virtual objects on real 3D displays. Accordingly, a human pose recognition method which simultaneously reconstructs a human volume is strongly required to manipulate virtual 3D objects more precisely.

Previously, researchers have focused on human pose estimation. Shotton et al. utilize randomized decision trees to recognize visible body parts, which are trained by synthetic depth images. They extend their work to infer self-occluded joint positions by aggregating offset votes of the joints from the visible body parts. Since then an incorporation of a global latent variable that encodes torso orientation or human height is employed to distinguish multiple distributions at leaf nodes of decision trees. Although they present breakthrough performance on human pose estimation, their foci aren't related to human volume reconstruction. From a viewpoint of human volume reconstruction, multiple cameras are generally utilized. Bergh et al. present human pose estimation from reconstructed 3D hull using a boosted classifier based on 3D Haarlets features. They use complex multiple camera configuration to acquire a human volume. Thus, there may be difficulties to be mounted on a living room. Izadi et al. enable real-time 3D reconstruction of targets by moving a depth camera. However, it still requires several image frames with different directions.

Main contributions of our approach are summarized as follows; firstly, we present a voxel classifier which reconstructs invisible body voxels while recognizing visible body voxels from a single depth image. The reconstructed voxels are used to infer a more accurate human pose. Secondly, we generate multi-layered training data which consist of synthetic visible and invisible voxels based on virtual ray-casting in a 3D graphics rendering engine. Lastly, Bayesian estimation method is employed to find optimal joint positions from the reconstructed voxels.

In detail, the massive synthetic voxels are generated by casting virtual rays from a virtual camera toward a 3D human model for training the voxel classifier. Whenever the casted ray penetrates each polygon of the human model, it generates a synthetic voxel which has the 3D position and the body part class. Note that the first voxel layer is identical to the visible body surfaces, while posterior voxel layers are the invisible surfaces. All the generated voxels form training data for the voxel classifier. An ensemble of decision trees is used as the voxel classifier. Since only visible body voxels can be observed in the real camera setup, the class information of the visible voxels is used for splitting. Once each tree is fully grown, each leaf node stores a normalized histogram of the visible voxels as well as histograms of the invisible voxels. In addition,

mean relative positions of invisible voxel layers from the corresponding visible voxel layer are computed to store in the same leaf node.

To reconstruct a human volume, each pixel x in the observed depth image is recognized as a maximum likely class for the visible voxel from the voxel classifier. In addition, invisible voxels associated to the visible voxel are similarly recognized by using the relative positions and the classes of the invisible voxels stored in the leaf node. Therefore the above procedure reconstructs the human volume in 3D space which consists of maximum likelihood estimates of the voxels.

And then a human pose is estimated from the reconstructed human volume. Since the reconstruction method sometimes yields multimodal voxel distributions for the same body part class, multiple candidates are tested by a Bayesian method to find optimal one. The Bayesian testing employs body part probabilities, silhouette matching accuracies and body parts continuities which consider connectivity relationship of kinematic chains. The maximum a posteriori spatial mean of voxel distribution for the body part is calculated as the joint position and it constructs the user's skeleton representation.

To demonstrate effectiveness, the proposed voxel classifier was trained from 42,030 human volumes which include various poses such as dancing, sports, and stretching activities. A supercomputer which connects 1,000 CPUs by Message Passing Interface was used for training the voxel classifier. We compared our proposed method (Whole voxel reconstruction: WVR) with the previous work (Visible voxel classification: VVC). In terms of body voxel IDs, the recognition rate of WVR is 83.1%, while the one of VVC is 56.8%. As for voxels position accuracy of WVR, an average error distance is 35.6 mm from ground truth. Considering a pose recognition to successful if error distances of all the estimated joint position are less than a 100mm, the recognition rate of WVR is 90.3%, while the one of VVC is 75.5%. Therefore the proposed WVR method has advantages respectively for more precise pose estimation.

In conclusion, the proposed voxel classifier reconstructs both a human volume and a pose from a single depth image at once in real-time. The overall run-time speed is above twenty frames per second on a single CPU. The novel multi-layered representation containing both visible and invisible synthetic voxels successfully provides the ground truth training data for the occluded body surfaces. In addition, due to augmenting the class IDs and its associated relative positions of the invisible voxels to the visible voxel in each leaf node, the ensemble decision trees as the voxel classifier reconstructs the human volume from any single depth image. The proposed voxel classifier is therefore successfully estimates complex human poses including severe self-occluded poses such as dancing and gymnastics activities.

9013-6, Session 2

Model-based 3D human shape estimation from silhouettes for virtual fitting

Shunta Saito, Keio Univ. (Japan); Makiko Kouchi, Masaaki Mochimaru, National Institute of Advanced Industrial Science and Technology (Japan); Yoshimitsu Aoki, Keio Univ. (Japan)

We propose a model-based 3D human shape reconstruction from 2 silhouettes. First, we synthesize a deformable body model from 3D human shape database provided by Digital Human Research Center, AIST, Japan. The database consists of homologous human model[1] which has the same topology and same number of vertices among the models. We perform principal component analysis on the database and synthesize an Active Shape Model(ASM)[2]. Then, we can deform the model's body type by back projecting parameters from PCA subspace to original dimension space. The pose changing can be achieved by reconstructing the skeleton from the model finished the body type deformation by ASM. Applying the pose change after body type deformation, the model can represent various body types and any pose.

We use the model for estimating the 3D human shape from front and side silhouette. The projected silhouette of the model is comparing with the input silhouettes. The model parameters are optimized to minimize the difference between corresponding silhouettes.

1. what is the addressed scientific topic or problem?

Then, we have to address the non-linear optimization problem in several tens of dimensions. The human shape reconstruction from only 2 silhouette is originally ill posed problem. There are many local solutions which can fit model's and input's silhouettes. It is challenging to find the global solution in this problem with low calculation cost. Therefore, we propose the method for addressing this optimization by using CMA-ES[3], a state-of-the-art optimization method, and a low calculation cost evaluation function representing silhouette difference.

2. what are the challenges and barriers faced?

Considering the demands from apparel venders in electronic commerce market, we cannot use the special equipments, for example, laser scanner, projector and camera system or multiple camera environment requires accurate calibrations, for estimation of consumer's body shape. Those equipments are expensive and cannot be used by consumers at their home, and it is not available to obtain the 3D shape as point clouds. Because, to use the 3D shape data for not only virtual fitting but also analyzing the consumer's body shape distribution to optimize the production line, the data have to be represented as models which have unified format for statistical analysis. Therefore, the easier and low cost method for estimating the 3D body shape and representing the result as a model is required.

3. why this is important for the 3D community?

We propose a model which has anatomical information. This information makes the extraction of many measurements of body parts and statistical analysis of body shape easily. In this research, we show how to build a human model which has those functions from relatively small dataset compared with CAESAR[4] or SCAPE[5], and use it for model-based shape-from-silhouette method. It is important for 3D community especially focusing the application of 3D human body model.

4. what is the original method proposed to address this problem or issue?

We construct a human body model which is implanted functional joint centers[6]. It allows the pose change based on skinning method with joints to be appropriate for reconstruction of reachable radius of body parts. When the model synthesized from a dataset which has not large number of data, it is important to consider body type changing and pose changing into separated ones for maximizing the representational variety of the model. If the separation is not enough, the resulting model will not be able to represent particular shapes not included in the dataset. In this research, we show the importance of functional joint centers to separate those features, body type and pose.

5. what is the novelty comparing to the state of art?

There are many researches of model-based shape-from-silhouette method for human body. The accurate shape reconstruction problem is considered as a different problem from pose estimation from images, so that many methods assumed the subject's pose is almost unified. However, if the model database has not large number of models, the pose variation when the subject is captured in a photo and it is used for estimation system is not ignorable. Therefore, we propose a deformable human model synthesized from comparatively small dataset and consider the pose variation when the subject is ready for capturing images with the model. Furthermore, there is a non-linear optimization problem in the model-based shape-from-silhouette method. Then, we apply a state-of-the-art optimization method, CMA-ES, for this human shape estimation problem.

6. what is the efficiency of the method (presentation of results and comparison with the state of art)?

Our method aims at a low cost and easy system for customers of online shopping service. In the present stage, the method is only tested in simulation experiment, but the system based on our method will allow the customers to know their 3D body shape with only an ordinary image capturing device.

[1] Mochimaru, M., Kouchi, M., Miyata, N., Tada, M. et al., "Dhaiba: Functional Human Models to Represent Variation of Shape, Motion and Subjective Assessment," SAE Technical Paper 2006-01-2345, 2006, doi:10.4271/2006-01-2345.

[2] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active Shape Models-Their Training and Application, Computer Vision and Image Understanding, Volume 61, Issue 1, January 1995, Pages 38-59, ISSN 1077-3142, 10.1006/cviu.1995.1004.

[3] Hansen N., Kern S., Evaluating the CMA evolution strategy on multimodal test functions, Parallel Problem Solving from Nature - PPSN VIII, pages 282–291. Springer, 2004.

[4] K. Robinette and H. Daanen, The CAESAR Project: A 3D Surface Anthropometry Survey, Second International Conference on 3-D Imaging and Modeling, 1999

[5] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: shape completion and animation of people. ACM Trans. Graph. 24, 3 (July 2005), 408-416

[6] Aoki, K., Kouchi, M., Mochimaru, M., and Kawachi, K., "Functional Shoulder Joint Modeling for Accurate Reach Envelopes Based on Kinematic Estimation of the Rotation Center," SAE Technical Paper 2005-01-2726, 2005, doi:10.4271/2005-01-2726.

9013-7, Session 2

3D face recognition via conformal representation

Junhui Han, Chi Fang, Xiaoqing Ding, Jian Sun, Tsinghua Univ. (China); Xianfeng Gu, Stony Brook Univ. (United States)

PROBLEM

Face recognition is a technology trend in recent years. 2D face recognition has made significant advances through the efforts of researchers, but it still bears limitations due to pose, illumination, occlusion, etc. [1]. 3D face which is composed of 3D surface and textures is quite different from 2D face. It is expected that 3D face recognition can make up for the shortage of 2D face recognition. This paper focuses on the algorithm of 3D face recognition.

CHALLENGES

Similar to 2D face, the inter-class variation of 3D face is also small. At the same time, some intra-class variations, mainly expression, are very large. These problems make 3D face recognition very difficult.

IMPORTANCE FOR THE 3D COMMUNITY

Apparently, there will be a great demand for 3D face recognition in authentication, access control, image retrieval, etc. in the near future. It is necessary for 3D community to develop 3D face recognition techniques. Furthermore, since all 3D objects have some similarity, ideals and techniques in 3D face recognition can be referred by other 3D research areas.

ORIGINAL METHODS

Existing methods to 3D face recognition can be put into two categories: feature analysis methods and deformable model methods. Feature analysis methods intend to make use of the rigid information of 3D face. Rigid information refers to the information which is invariant to expression variation. It is mainly contained in rigid regions, such as the nose region. The general process of this type of methods is that 3D face is firstly described as features like curvature, Gabor, LBP, etc. Discriminant analysis (DA) is then applied on these features to obtain recognition features. In [1], Mohammadzade et al. first use an iterative closest normal point (ICNP) algorithm to find the corresponding points between a generic reference face and every input face. They then choose the normal vectors of the aligned 3D faces as their features and apply LDA to get the recognition features for recognition. Maurer et al. [3] use Iterative Closest Point (ICP) algorithm to align faces. They then compute the distance between two faces on every pixel to form a distance map. This indicates that they use the coordinates of 3D face as their features. The distance map is then analyzed to obtain recognition features. Instead of using the rigid information of 3D face, the deformable model methods try to recover the neutral face from nonneutral face through deformable model. The deformable model is learned using training subjects with both neutral and nonneutral faces. This type of methods can reduce the impact of expression variation on nonrigid regions such as the mouth region, but they weaken the discriminant ability of rigid regions at the same time. In [4], Al Osaimi et al. build a PCA subspace using the shape residues between the pairs of neutral-nonneutral faces which are aligned using the ICP algorithm. They then use the learned patterns to morph out the expression from nonneutral faces for recognition.

NOVELTY OF OUR METHOD

Our method is a kind of feature analysis method. To make the most of the rigid information of 3D face, we use conformal representation as our features. Conformal representation is defined as the pair of mean curvature (MC) and conformal factor (CF) [6]. It is a complete description of 3D facial surface, which is the main novelty of our method. Considering that different regions of face deforms unequally due to expression variation, 3D face is divided into five parts through which we can deal with each part separately. LDA is then applied to each part to reduce the intra-class variation and enlarge the inter-class variation.

Specifically, facial surfaces are firstly preprocessed through topological denoising, hole filling and remeshing. Conformal maps of facial surfaces are then computed through Riemann Mapping. Conformal maps are 2D images with unit circles as their boundary. Mean curvature and conformal factor are then computed over conformal maps and are encoded to grey images, called Mean Curvature Image (MCI) and Conformal Factor Image (CFI). MCIs and CFIs are then divided into 5 parts: Cheek, Eye, Forehead, Mouth and Nose. A regularization process centered at the nose tip is then performed on each part, so that the same part of MCIs and CFIs has the same size. We obtain the conformal representation features of each part by concatenating its mean curvature features and conformal factor features. After that, LDA are adopted on each part to obtain the recognition features. Finally, we fuse the five parts on the distance level.

EFFICIENCY OF OUR METHOD

To prove the efficiency of our method, we carried out some experiments on the BU-3DFE database. BU-3DFE is a facial expression database, which makes our results more persuasive. BU-3DFE database contains 2500 textured 3D models of 100 subjects. For each subject, it includes one neutral and six prototypical non-neutral expressions (Happiness (HA), Angry (AN), Fear (FE), Surprise (SU) and Disgust (DI)) with 4 intensity level.

In our experiments, 50 subjects are randomly selected to form the training set, and the rest 50 subjects form the test set. The test set is divided into two parts: gallery and probe. The neutral expression models of the 50 subjects in the test set form the gallery, while the rest models in the test set form the probe. Therefore the gallery has 50 models and the probe has 1200 models. To make the recognition results stable, the experiments in this paper were repeated by 100 times.

Since no experiment results on this database have been reported, we compare our results with that of ICP, which is the baseline [2] algorithm in 3D face recognition. ICP is performed on the original 3D faces represented by 3D coordinates, which has been divided according to MCIs and CFIs. The results show that the accuracy of our method is 98.9%, while the accuracy of ICP is 83.5%. This verifies the efficiency of our method.

REFERENCE

[1] Mohammadzade, H., Hatzinakos, D., " Iterative closest normal point for 3D face recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence. Papers 35(10), 381-397 (2013).

[2] Chang, K.I., Bowyer, K.W., Flynn, P., "Multiple nose region matching for 3D face recognition under varying facial expression," IEEE Transactions on Pattern Analysis and Machine Intelligence. Papers 28(10), 1695-1700 (2006).

[3] Maurer, T., Guigonis, D., Maslov, I., Peseti, B., Tsaregorodtsev, A., West, D., Medioni, G., "Performance of geometrix activeid 3D face recognition engine on the FRGC data," Proc. Computer Vision and Pattern Recognition, 2005.

[4] Al Osaimi, F., Bennamoun, M., Mian, A., "An expression deformation approach to non-rigid 3D face recognition," International Journal of Computer Vision. Papers 81: 302-316 (2009).

[5] Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J., "A 3D facial expression database for facial behaviour research," Proc. Automatic Face and Gesture Recognition, 211-216 (2006).

[6] Lui, L.M., Wen, C.F., Gu, X.F., "A conformal approach for surface inpainting," arXiv:math.DG/1212.0981.

9013-9, Session 3

Real-time 3D shape measurement system with full temporal resolution and spatial resolution

Kai Zhong, Zhongwei Li, Guomin Zhan, Xingjian Liu, Yusheng Shi, Congjun Wang, Xiaohui Zhou, Cheng Wu, Huazhong Univ. of Science and Technology (China)

Real-time 3D measurement is becoming increasingly important in areas such as entertainment, manufacturing and medical science. Digital coded images projection techniques have shown their potential for dynamic 3D shape measurement. Many real-time 3D measurement systems have been developed with the advancement of digital-light-processing (DLP) technology and powerful hardware.

At present, the one-shot techniques credit with full temporal resolution, but suffer from poor spatial resolution. However, in many application areas, such as entertainment and medical science, full spatial resolution is very critical. Therefore, in very recent years, many researchers still focus on employing the well-known three-step phase shifting algorithm to archive full spatial resolution with high accuracy and robustness. As the three-step phase shifting algorithm requires a phase unwrapping processing to obtain absolute phase value before 3D reconstruction, they have extensively explored spatial or temporal phase unwrapping methods to solve this problem. However, the spatial methods cannot measure the object with discontinuous surfaces or multiple isolated objects, and the temporal methods which need additional images deduce the temporal resolution. Therefore, although a number of methods have been proposed in the recent decadal years, measuring arbitrary shape dynamic objects (especially the object with discontinuous surfaces or multiple isolated objects) with full resolution, including temporal resolution and spatial resolution, is still a big challenging problem.

This paper presents a real-time 3D shape measurement system based on phase measurement profilometry (PMP). This system mainly includes one DLP projector and two cameras. The projector can be regarded as the inverse of a camera, and it projects images instead of capturing them.

Typically, a series of vertical sinusoid wave images with three step phase-shifting amount are projected onto the measured objects, and the cameras have to capture the distorted fringe images synchronously. To synchronize with a DLP projector, the exposure time of the cameras must override the whole modulation time, which is for generating gray intensities by digital micro-mirror device (DMD). In order to correctly design the synchronous timing between the projector and the cameras, the delay of DMD flipping and delay of camera photosensitive response is evaluated, and then the delay of trigger signal, which is sent to cameras, can be calculated. According to experiment, the system can achieve precise synchronization acquisition rate of 220 per second.

Before measurement, this multi-view system has to be calibrated. And the calibration accuracy usually determines the measurement accuracy. In this system, the calibration procedure contains two steps: (1) calibrating the internal parameters of each camera individually; (2) determining the pose of each camera. For single camera calibration problem, traditional planar target calibration method is used to obtain the initial values, and these values are optimized by the bundle adjustment strategy to eliminate the influence of the inaccuracy of the calibration target. Then the world coordinate is located on the camera coordinate, the pose of each camera is optimized by epipolar constraint, and the calibrated translation vector has to be revised to remove the scaling factor. Experimental results showed that this method has a good calibration accuracy and stability in practice.

During image processing, any adjacent three images in the continuous capturing stream of each camera can be used to calculate the wrapped phase value by phase-shifting algorithm. Benefited from the multi-view constraint, each pixel can explore its corresponding point independently only in the wrapped phase-map and 3D point can be reconstructed by triangulation method, so arbitrary shape including the discontinuous or step-like surfaces can be measured with full spatial resolution. Moreover, as any adjacent three images per camera in the continuous capturing stream can be used to reconstruct one 3D frame, the 3D acquisition speed can be as fast as the capturing speed of camera. So our system can also perform full temporal resolution 3D measurements. In general cases, multiple fringe number and phase error can lead to multi-candidates and false correspondence. Aiming at searching the corresponding points in an accurate, robust and fast manner, a two-stage strategy, using disparity range constraint and mixed consistency check successively, is used to reject the wrong candidates. The disparity range is constrained by measurement volume, and it is very similar to the epipolar line segment in space. And an optimal fringe number is used to decrease the candidate number, and the wrong candidates can be effectively rejected. The mixed consistency bases on the observation that not only phase value but also gray intensity of the correct corresponding point should be consistent. So the remaining candidates will be checked by mixed consistency constraint for selecting the unique correspondence. This method is simply and easy to apply, and it is very suitable to perform fast 3D measurements.

The large amount of data intensive computing is always the bottleneck of real-time dynamic 3D measurement. In order to improve the 3D calculation efficiency, a hybrid parallel computing architecture based on CPU and GPU is proposed. For large amount of data intensive computing tasks, a massive fine-grained GPU parallel computing model is designed to accelerate the whole calculation process, and this model mainly contains three kernel functions, which can realize image rectification, phase wrapping and 3D reconstruction respectively. And CPU is mainly responsible for logical operations to connect the three kernel functions. After properly optimizing the GPU memory assessment, the experiment shows that our system can achieve real-time 3D measurement by employing only mainstream chip, which is very important in real-time 3D measurement applications. The computation speed in the GPU is nearly 400 times faster than that in the CPU.

On the basis of the above technologies, a dynamic 3D shape measurement system is developed. This system can achieve 220 3D frames per second fast 3D measurement and 60 3D frames per second real-time 3D measurement, and its spatial resolution is 1024 ? 768. It should be noted that if cameras with higher capturing speed and resolution was employed, the temporal resolution and spatial resolution of 3D measurement would also be higher. Compared with state-of-art real-time measurement systems, the developed system can measure arbitrary shape dynamic objects (especially the object with discontinuous surfaces or multiple isolated objects) with full temporal resolution and spatial resolution, which is very important in scientific research and industrial applications.

9013-10, Session 3

New concept of technology chain for 3D/4D content generation and display

Malgorzata Kujawi?ska, Robert Sitnik, Tomasz Kozacki, Warsaw Univ. of Technology (Poland)

Over the last decade 3D/4D technologies gain continuously social interests in applications as 3D television and cinema, education, gaming, virtual reality (VR), augmented reality (AR). The ideal features of future 3D cinema, TV or display include:

- viewer can freely move camera (or himself) around a scene,

- realism of the scene is as natural as real life and visualization is as natural as possible,

- as many as possible scene elements are interactive ones enabling to viewer impressions of living world,

- possibility of mixing as many different content sources as possible,

- displaying a content which can be viewed by multiple users.

To introduce a new visual content production technology which would be fulfilling all or most of the requirements, a complex solution is required starting from capture and ending on visualization/ display.

The most futuristic vision of 4D (3D plus time) systems (often referred to as 3D video) requires capturing of real world 3D objects or scenes, and delivering at the observer side a ghostlike, high-quality optical replica of a moving object or a scene, which is floating in a space and can be viewed freely from different angles. This vision is often referred to as True 3D which is realized by holographic capture and display. It shall include features of interactive VR and AR with the capability of mixed content obtained from real and virtual worlds. The development of technology which implement True 3D vision will break the grounds in multimedia and other applications using 3D/4D imaging.

Therefore the main goal of this paper is to introduce the concept of next-generation of 3D and 4D imaging technology, which is capable of multi-source capture, multi-representation processing, multi-data-content mixing and provides a holographic display with ultrahigh resolution. The realization of the new concept of the 3D technology chain will allow for production and display of color holographic data content that is composed of real life and synthetic scenes. This includes dynamic objects of human sizes, far distant backgrounds and synthetic scene elements. The diversity, complexity and extreme high resolution of this 3D and 4D imaging system requires the development of tools that compose the holographic content captured from a multi-source system. The combination of these tools is capable producing a holographic video with a resolution above 1 billion pixels.

Nowadays consumer expects very high quality content, which cannot be successfully delivered digitally by holographic means only. Therefore it is proposed that the modern holographic content (Fully 4D) is acquired by several types of capture systems and/or can be generated by computer. The diversity of the content is related to the nature of data capture and scene requirements. The innovative 3D/4D imaging approach makes the full use of two different 3D data representations, namely surface and holographic (amplitude-phase) ones. This data after appropriate conversion are subjected to processing, mixing and interaction, and finally are composed into holographic representation ready to be displayed.

The concept of the novel fully 3D and 4D imaging technology basing for on two 3D/4D data representations is presented in more details in Fig.1.

The accomplishment of the main objective of the presented concept is based on several original proposals which have actually are step by step implemented in the area of 4D imaging. The auxiliary objectives and ground breaking aspects of the proposed 3D/4D technology refer to:

- development of sophisticated 3D/4D acquisition techniques and prototypes,

- development of innovative data and content processing technology,

- development of advanced, color holographic displays: dynamic with spatio-temporal multiplexing and updatable one.

The realization of this concept will pave the path for the next generation of 4D technologies based on multi-source data and future holographic display that provides generation, management and display of contents in multiple applications ranging from 3D cinema and interactive television supported by Augmented Reality.

9013-11, Session 3

Low-cost structured-light based 3D capture system design

Jing Dong, Purdue Univ. (United States); Kurt R. Bengtson, Hewlett-Packard Co. (United States); Jan P. Allebach, Purdue Univ. (United States)

1. Addressed Scientific Topic:

Currently, three-dimensional measurement is a very important and popular topic in computer vision [1], [2]. Due to its accuracy and time-efficiency, 3D capture is broadly applied in a number of areas, such as object recognition, dental and facial imaging, 3D map building, and manufacturing. 3D capture techniques may be divided according to contact and non-contact methods. The problem with contact methods is the slowness and high cost of the probing process. In addition, the probe may cause damage to the surface of the object. Non-contact methods can be classified into passive and active methods. For passive methods, two or more views are captured by the cameras and the correspondences between the images need to be found to reconstruct the 3D surface. Thus, passive methods are limited to reconstructing relatively smooth 3D surfaces due to the complexity of finding correspondences. To solve this, structured-light methods (included in active methods) are currently more widely used. One of the cameras is replaced by a projector, and a coded pattern is projected. By decoding the captured encoded image, one can find the correspondence between the projected image and the captured image. Thus, the 3D object can be reconstructed using triangulation. Lanman and Taubin [3] provide a tutorial course for beginners to build their own 3D capture system. They explain the mathematics of the triangulation, camera and projector calibration, and develop a classic structured light scanning system and a laser line scanning system. Geng [4] presents a review of the recent advances in 3D surface imaging technologies. He focuses on noncontact surface measurement techniques based on structured light and categorizes and compares different coding strategies. He also discusses the calibration techniques and numerous applications of 3D surface imaging techniques.

2. Challenges and Barriers Faced:

Most of the 3D capture products currently in the market are high-end and pricey. They are not targeted for consumers, but rather for research, medical, or industrial usage. Unlike research, medical or industrial object capture, object capture for the home hobbyist or small business does not require very high accuracy. It is a challenge to balance the trade-off between accuracy and system cost. We can sacrifice part of the accuracy to lower the cost of the 3D capture system, on the one hand. On the other hand, we do not want to lose so much accuracy that the object cannot even be recognized. Therefore, our goal is to maintain sufficient accuracy while keeping the system cost within a range that is suitable for home or small business use. A number of factors need to be considered when choosing the system components. The first group of these includes the resolution, light-level output, geometric distortion, and working distance for the projector. The working distance must be chosen such that the whole object station is completely included in the field of view of the camera. Another issue with the projector is the throw ratio. The throw ratio is the ratio between the throw distance and the width of the screen, where the throw distance is the distance from the screen to the projector. Most projectors on the market have a relatively large throw ratio. Therefore, given a desired projected image size, the large throw ratio requires a large throw distance, which causes the whole 3D capture system not to be compact. The second group of factors is related to the camera. These include resolution, sensitivity, geometric distortion, and color fidelity [5]. The latter is important if a second frame is to be captured under normal illumination without the structured light pattern, to provide the surface reflectance information that is to go with the object shape. To solve these issues, we introduce the novel dual projector system described in Sec.4 below.

3. Importance for the 3D Community:

Most of the current 3D capture devices in the market are targeted for research, medical, or industrial usage. Very few aim to provide a solution for home and small business applications. Our goal is to fill in this gap by only using low-cost components to build a 3D capture system that can satisfy the needs of this market segment. Therefore our work is important for the 3D community; so that a broader group of users can have an affordable device and sufficient resolution to capture their objects in 3D.

4. Original Method Proposed to Address this Issue:

In this paper, we present a low-cost 3D capture system based on the structured-light method. The system is built around the HP TopShot

LaserJet Pro M275. For our capture device, we use the 8.0 Mpixel camera that is part of the M275. We augment this hardware with two 3M MPro 150 VGA (640?480) pocket projectors. Both the M275 and the MPro 150 are low-cost products. To further reduce the cost and complexity, we use a single shoot of the binary M-array structured light pattern designed by Lei et al. [6]. We use the Matlab Toolbox developed by Bouguet [7], which is an implementation of the well-known Zhang's [8] camera calibration method for the camera and projector modeling and calibration, since the projector can be modeled as an inverse camera.

5. Novelty Compared to the State of Art:

We present a novel design using low-cost, readily available components to build a 3D capture system that has sufficient resolution for home and small business use. Compared to the state of art, the image capture and reconstruction processes are also time efficient since only a single shot of the projected M-array pattern is required. We also describe an analytical approach to predicting the achievable resolution of the reconstructed 3D object based on differentials and small signal theory, and an experimental procedure for validating that the system under test meets the specifications for reconstructed object resolution that are predicted by our analytical model. The experimental procedure is based on comparing results from the analytical sensitivity analysis, the results from the simulation codes [6], and analysis of the image data captured by our prototype system.

6. Efficiency of the Method:

As a preliminary validation of the effectiveness of our system design, we conducted an experiment with a first-generation system (not the one described in Sec. 4 above) that consisted of a single AAXA M2 XGA (1024?768) projector and a Logitech C905 2.0 Mpixel webcam. We compared simulation results [7] with image data captured by this system. Using the simulation codes, we calculate the camera image plane pixel shift caused by unit pixel shift on the projector image plane. From the simulation, we obtain that a shift of 1 pixel on the projector image plane results in an average of 1.054 pixels shift in the x direction and 0.0495 pixels shift in the y direction on the camera's image plane. Our analysis of the relation between the captured image data and the projected image gives us an average of 1.0498 pixels shift in the x direction and 0.0473 pixels shift in the y direction on the camera's image plane. By comparing our experimental measurements from the camera-projector system with the simulation results based on the model for this system, we conclude that our prototype system has been correctly configured and calibrated and that with the analytical models to be presented in the full paper, we will have an effective means for specifying system parameters to achieve a given target resolution for the reconstructed object. The full paper will show results for the second-generation system described in Sec. 4 above.

REFERENCES:

[1] Blais, F., "Review of 20 years of range sensor development," Journal of Electronic Imaging 13(1), 231–243 (2004)

[2] Gueziec, A., Hummel, R., "Exploiting Triangulated Surface Extraction Using Tetrahedral Decomposition," IEEE Transactions on Visualization and Computer Graphics 1(4), 328-342 (1995).

[3] Lanman, D. and Taubin, G., "Build your own 3d scanner: 3d photography for beginners," [SIGGRAPH'09: ACM SIGGRAPH 2009 courses], ACM, New York, 1–87 (2009).

[4] Geng, J., "Structured-light 3d surface imaging: a tutorial," Adv. Opt. Photo 3(2), 128–160 (2011).

[5] Lei, Y., Majewicz, P., Bengtson, K. R., Li, L. and Allebach, J. P., "Composite Target for Camera-Based Document/Object Capture System, " Journal of Imaging Science and Technology, to appear.

[6] Lei, Y., Bengtson, K. R., Li, L. and Allebach, J. P., "Design and decoding of an M-array pattern for low-cost structured light 3D reconstruction systems, " IEEE International Conference on Image Processing (2013).

[7] Bouguet, J., "Camera Calibration Toolbox for Matlab," 9 July 2010, http://www.vision.caltech.edu/bouguetj/calib_doc/.

[8] Zhang, Z., "A flexible new technique for camera calibration," IEEE Transactions on Pattern Analysis and Machine Intelligence 22(11), 1330–1334 (2000).

Conference 9014: Human Vision and Electronic Imaging XIX

Monday - Thursday 3 –6 February 2014

Part of Proceedings of SPIE Vol. 9014 Human Vision and Electronic Imaging XIX

9014-1, Session Key2

Seven challenges for image quality research *(Keynote Presentation)*

Damon M. Chandler, Oklahoma State Univ. (United States)

Image quality assessment has been a topic of recent intense research due to its usefulness in a wide variety of applications. Owing in large part to efforts within the HVEI community, image-quality research has particularly benefited from improved models of visual perception. However, over the last decade, research in image quality has largely shifted from the previous broader objective of gaining a better understanding of human vision, to the current limited objective of better fitting the available ground-truth data. In this paper, I argue that, despite the explosive growth in image-quality research, we are today only marginally closer to understanding how humans perceive artifacts in images than we were 30 years ago. I specifically discuss seven open challenges which stem from lack of complete perceptual models for: natural images; suprathreshold distortions; multiple distortions; interactions between distortions and images; images containing nontraditional distortions; and enhanced images. I also discuss challenges related to computational efficiency. The objective of this paper is not only to highlight the limitations in our current knowledge of image quality, but to also emphasize the need for additional studies beyond those commonly cited, and the need for alternative theories and techniques beyond those commonly employed.

9014-2, Session Key2

Images shared in social media: a window into human sentiment and emotion *(Keynote Presentation)*

Shih-Fu Chang, Columbia Univ. (United States)

No Abstract Available

9014-3, Session Key2

Digital coloring books *(Keynote Presentation)*

Patrick Hanrahan, Stanford Univ. (United States)

No Abstract Available

9014-4, Session 1

Auditory modulation of visual motion perception *(Invited Paper)*

Mark E. McCourt, North Dakota State Univ. (United States)

No Abstract Available

9014-5, Session 1

Modulation of luminance contrast perception by background acoustical noise *(Invited Paper)*

Laura Rossi, Istituto Nazionale di Ricerca Metrologica (Italy)

No Abstract Available

9014-6, Session 1

Audiovisual focus of attention and its application to UHD video compression *(Invited Paper)*

Martin Rerábek, Hiromi Nemoto, Jong-Seok Lee, Touradj Ebrahimi, Ecole Polytechnique Fédérale de Lausanne (Switzerland)

No Abstract Available

9014-7, Session 1

Influence of audio triggered emotional attention on video perception *(Invited Paper)*

Freddy Torres, Hari Kalva, Florida Atlantic Univ. (United States)

Perceptual video coding methods attempt to improve compression efficiency by discarding visual information not perceived by end users. Most of the current approaches for perceptual video coding only use visual features ignoring the auditory component. Many psychophysical studies have demonstrated that auditory stimuli affects our visual perception. In this paper we present our study of audio triggered emotional attention and it's applicability to perceptual video coding. Experiments with movie clips show that the reaction time to detect video compression artifacts was longer when video was presented with the audio information. The results reported are statistically significant with p=0.024.

9014-8, Session 1

3D sound and 3D image interactions *(Invited Paper)*

Jonathan Berry, Durham Univ. (United Kingdom); Nicolas S. Holliman, The Univ. of York (United Kingdom)

No Abstract Available

9014-9, Session 1

WWII cargo cults and their implications for the future of digital cinema imaging and sound *(Invited Paper)*

Harry Mathias, San José State Univ. (United States)

No Abstract Available

9014-10, Session 1

Does movie genre or the scene statistics dominate in the balance of audio and video bitrates on perceived overall quality? *(Invited Paper)*

Poppy Crum, J. Tobin, Dan Darcy, Scott J. Daly, Dolby Labs., Inc. (United States)

No Abstract Available

9014-11, Session 2

Roughness versus contrast in natural textures

Pubudu M. Silva, Thrasyvoulos N. Pappas, Northwestern Univ. (United States); Huib de Ridder, René van Egmond, Technische Univ. Delft (Netherlands)

No Abstract Available

9014-12, Session 2

An investigation of visual selection priority of objects with texture and crossed and uncrossed disparities

Darya Khaustova, Jérôme Fournier, Emmanuel Wyckens, Orange SA (France); Olivier Le Meur, Univ. de Rennes 1 (France)

Context

Emergence of 3DTV created the necessity to study human visual perception of stereoscopic content and its difference to 2D content. Some of these studies [1-5] aim at paving the way for the design of computational models of visual attention for 3D content. The introduction of binocular disparity is the most promising approach. However, the influence of depth cue on the way we look at a picture is still an open debate.

Most studies have reported that observers take longer time to look at closer areas in stereo conditions. However, there are some contradictions between different results concerning eye movements in 3D. Some groups [1, 2] claim that fixation are more widespread in 3D mode, others [4] assert the opposite. On this basis, it seems that the influence of the binocular depth cue on visual attention is dependent on the content itself [3]. In our previous study [6] we found out that 2D and 3D visual strategies are very similar in a context of uncrossed disparity. It is however noticeable that there is a significant difference between 2D and 3D case for uncrossed disparity for average saccade length. We didn't find any evidences that discomfort can influence the way we observe images.

Objective

The objective of the present study is to compare recorded eye-tracking data for different 2D and 3D conditions considering effects of major visual factors as scene complexity and extent of disparity. The present study consists of the two following experiments.

The goal of the first experiment is to continue experiment described in [6], where we studied visual attention in 3D using stimuli with uncrossed disparity. In order to complete the previous study we designed a new experiment using stimuli with crossed disparity. The change of scene complexity within each particular scene is done by modifying textures of some objects and/or of the background. Three conditions are introduced to examine the influence of depth on results: 2D images i.e. no binocular depth, 3D images ensuring comfortable visualization and 3D images inducing visual discomfort. These conditions are reached by controlling the amount of depth of images (Depth of Focus (DoF) criteria).

The aim of the second experiment is to verify hypothesis that texture and contrast (2D criteria) are more influential factors in guiding our gaze than the amount of depth (3D criteria). As stimuli we used images, which contained four spheres on a gray background. In stereoscopic condition any of the spheres could have a different position in depth: in front of the display plane, behind the display plane and on the display plane. In 2D condition all four spheres seemed to be located on the display plane. In order to study the influence of texture some of the spheres had the same gray level as the background, while the others had the check board texture. By performing the eye-tracking experiment and studying gaze plots of the observers we aimed to found out the priority in selection of the spheres in depth and to compare results to the 2D condition.

Method

Both experiments consist of three steps: (1) estimation of camera baseline and convergence distance (In order to create comfortable and uncomfortable viewing conditions for the first experiment and to position sphere in different depth levels in the second experiment), (2) generation of CGI 3D scenes with different amounts of depth and textures, and (3) psychophysical tests, using an eye-tracker.

The amount of depth presented to observers is controlled by changing the virtual 3D camera parameters. The determination of camera parameters is done with the Stereo Calculator developed in Orange Labs, allowing a precise calculation of a camera baseline and a convergence distance for the creation of stimuli with different DoF. The calculations can be accomplished when viewing environment (display size and resolution, viewing distance), camera parameters (focal length, sensor size), scene parameters (foreground and background distance of a scene) and human visual attributes (depth of focus, inter-pupil baseline) are defined.

The scenes are generated with Blender software, which allows to measure foreground and background distances of the scene and to control stereoscopic camera parameters. For the first experiment it was decided to use 4 different scenes. Each scene has two texture variation and three depth range: 2D, 3D comfortable and 3D uncomfortable. For the second experiment 14 different scenes were generated. Each scene has different combinations of textured and non-textured spheres and four depth ranges: 2D, crossed disparity, uncrossed disparity and combination of crossed and uncrossed disparities.

The psychophysical test set-up consists in a Tobii eye-tracker and a 3D display. The observers have to perform a free viewing task. Each scene instance is presented only once to each observer in order to prevent memorizing and hence top-down visual mechanisms. The analysis is accomplished exploiting fixation data of observers, saliency maps and regions of interest.

Conclusions

This study was launched to generalize the studies in understanding the difference in spatial observation of 2D and 3D content depending on texture and amount of depth. In the first experiment 51 observers participated in the test. Considering scenes with crossed disparity we find out that objects located in front of the display plane are the most salient, even if observers experience discomfort. In the second experiment we aimed to generalize previous studies of visual attention. 28 observers were watching the scenes with crossed and uncrossed disparities. We evaluated features influencing the saliency of the objects in stereoscopic conditions by using content with low-level visual stimuli. We detected that texture is the most important feature for selection of objects. Objects with crossed disparity are significantly important for selection process as well. However, objects with uncrossed disparity are less important for visual attention.

We believe that these experiments can help to refine existing models of visual attention in 3D.

References

1. Hakkinen, J., et al. What do people look at when they watch stereoscopic movies? 2010. San Jose, California, USA: SPIE.

2. Jansen, L., S. Onat, and P. König, Influence of disparity on fixation and saccades in free viewing of natural scenes. Journal of Vision, 2009. 9(1).

3. Huynh-Thu, Q. and L. Schiatti. Examination of 3D visual attention in stereoscopic video content. 2011. San Francisco Airport, California, USA: SPIE.

4. Ramasamy, C., et al., Using eye tracking to analyze stereoscopic filmmaking, in SIGGRAPH '09: Posters2009, ACM: New Orleans, Louisiana. p. 1-1.

5. Wismeijer, D.A., et al., Depth cue combination in spontaneous eye movements. Journal of Vision, 2010. 10(6).

6. Khaustova, D.y., et al., How visual attention is modified by disparities and textures changes? 2013: p. 865115-865115.

9014-13, Session 2

Memory texture as a mechanism of improvement in preference by adding noise

Yinzhu Zhao, Naokazu Aoki, Hiroyuki Kobayashi, Chiba Univ. (Japan)

Memory colors refer to colors associated with certain familiar objects in our daily lives. For example, an apple is associated with red as a banana is associated with yellow.1 Memory colors are formed in everyday life when we experience familiar objects and situations. As an individual's standard of recollection for familiar objects,2 memory colors have an important characteristic that people living in the same environment tend to have similar memory colors because they are surrounded by the same objects. In terms of color reproduction in photography, it is found that preserving actual colors of sky, vegetation and other natural objects has high preference, while reproducing memory colors of facial skin color is more preferable. Not only do memory colors vary greatly from one locale or nation to another, they also vary within sex, generation, and economic conditions.3

In this study, we propose "memory texture" as another standard of recollection for familiar objects and that it is formed by the same mechanism as memory color. The concept of memory texture was obtained after noticing that there are images whose preference was actually increased by adding image noise, which has heretofore been considered unnecessary and better to be eliminated.4 Specifically, people feel the texture of familiar objects as more preferable when they present memory textures after adding noise to the images. First, pictures of seven familiar objects with different textures were selected. Then, memory texture was quantified by selection of the most accurate one from pictures of different noise levels. After that, the features of memory texture and the relationship between memory texture and image preference were investigated.

EXPERIMENT

Stimulus

Seven objects were included in the memory texture experiment (Fig.1). The study added 1/f noise to each of the objects with RMS granularity values at regular intervals so each object has six photographs. All images were printed after they were adjusted to meet the actual size of the objects.

Observers

Twenty-four subjects in their 20s participated in this experiment. All experiments were carried out under the observation of the Toshiba natural color evaluation lamp (5000K, 700lx), and in the distance of distinct vision.

Procedure

Photographs were laid out on the table by category. Then, the observer carried out the following four-step experiment:

Step 1. Recall the texture of the object without seeing real object.

Step 2. Rearrange the photographs in order of accuracy to memory texture.

Step 3. After seeing the real object, rearrange the photographs in order of accuracy to real-life texture.

Step 4. Rearrange the photographs in preference order.

RESULTS AND DISCUSSION

1. Relationship between memory texture and real-life texture

This study investigated the relationship between memory texture and real-life texture.

The vertical axis shows average RMS values of the photographs chosen as the most accurate by subjects. The results can be divided into two groups.

Subjects recalled the textures of cardboard, tissue paper, newspaper, onion, and wooden chopsticks as grainier than in real life while objects like potato and wooden board were recalled as less grainy than in real life.

There is no consistency between memory texture and real-life texture, meaning that granulated-related memory texture is not related to the type of object. This phenomenon matches the characteristic of memory color.2 In the memory color experiment, which was conducted with color patches, memory color had a higher saturation recall. However, the experiment conducted with real objects had no consistency between memory color and actual object color.

2. Relationship between memory texture and preference of images

The horizontal axis is the normalized rank of preference, which means that the greater the value, the higher the preference. Each mark—'*' and '**'—indicates significant differences with p-levels of 5% and 1%, respectively. The numbers written above the square makers indicate the ranking of memory texture for the six photographs.

Both of these objects showed higher preference and higher ranking of memory texture in RMS5, RMS10 than the original images. Similar results were obtained for the other objects.

After the four-step experiment was complete, the same experiment was rerun, adding white noise to the photographs to examine the difference with regard to noise type.

As shown in the graph above, when adding white noise to photographs, memory texture also differs from real-life texture. The result of preference experiment was also similar with 1/f noise.

SUMMARY

Memory texture experiments have been carried out to quantify memory texture and examine the relationship between memory texture and real-life texture. This study also found that memory texture and preference of image are correlated. From this research, it can be concluded that memory texture differs from real-life textures. Furthermore, there was a correlation between the graininess of memory texture and the subject's preference for graininess in the object's image.

9014-14, Session 3

Computer vision enhances mobile eye-tracking to expose expert cognition in natural-scene visual search tasks

Tommy P. Keane, Nathan D. Cahill, Rochester Institute of Technology (United States); John A. Tarduno, Robert A. Jacobs, Univ. of Rochester (United States); Jeff B. Pelz, Rochester Institute of Technology (United States)

Many advanced fields of study and creation are limited in their success by the uniqueness of their experts. Expertise in a field is a complex and vague notion, especially when being considered to be passed along in edification. For our research we are approaching the question of visual search task expertise for geology and the geo-sciences, as a way to further improve educational practices and technologies. When presented with a prima facie view of a scene whose structure, topology, and perceivable qualities have been formed by geological processes, what kinds of visual cues are discovered by observers, and how does that differ based on their expertise and experience in the geo-sciences? Also, what kind of search path for visual perception, both spatially and

temporally, do observers take when presented with a novel scene view? And, ultimately, how does their visual search, and its relationship to visual cues, lend to the accuracy of their geo-scientific understanding of a scene? Investigating these questions, we believe that we can further the understanding of that indescribable "stroke of genius" that affects the experienced expert, and which the novice has not yet gained exposure.

Our research builds on modern advances in visual perception that further investigate the role of cognition and how the process of perception is greatly influenced by many co-occurring sources of information in the brain and mind. Geology is well-understand as being a discipline where expertise comes from the breadth and variation of a scientist's field exposure, not necessarily the length of their career. This has solidified to us the notion that the expertise in visual search tasks isn't just about what is seen, but about what is being seen in relation to every other field experience the expert has had. Novices typically have no experience, certainly not anything trained or geologically directed, and so their visual search of a scene is scattered and unsure. Experts are directed and purposeful, and they somehow know seemingly innately what to avoid and what to focus on, without any aid or direction. This has lead us towards a need for further considering Gestalt processes when positing statistical analysis of cognitive data. Specifically, this means moving beyond fixation counts and durations or heat-maps, and into spatio-temporal analysis, inter-expert gaze pattern comparisons, and the relationships between scenes' content structure and the observers' gaze patterns.

To obtain this data we are using mobile eye-tracking technology in conjunction with a ten-day geo-science field trip, where a dozen scene views are recorded for a dozen subjects each year in our 5-year study. These are sites formed by geological processes requiring an understanding of plate tectonics, vulcanology, glaciology, etc., thus testing the breadth and variations amongst experts and novices in different scenarios, but with the shared goal understanding the formation. Subjects' gaze patterns are recorded for up to two minutes of unguided searching, and then audio and continued eye-tracking occurs during a lecture by a geology professor that follows. With subjects' eye-tracking data all being captured simultaneously, but separately, we needed to develop a common mathematical space for analysis. Without that, the gaze patterns as mapped by the mobile eye-tracking technology only occur within each subjects' scene video's image space. So we also capture high-resolution panoramic imagery of the scene to be used for recording of the view, educational display of these exemplar scenes, and as the ideal unified image space for the gaze analysis.

The complexity of this analysis and the source of novel advancements lie in the fact that all our data is captured in natural settings with organic and complex structures in the imagery and almost no control over lighting. Basic computer vision algorithms have been found to be limited in registration and mapping of our disparate mobile eye-tracker scene videos and observers' "point of regard" into the common space. And while an equirectangular panorama mapping provides a useful and aesthetically interesting full view of the scene, it complicates further the mathematical complexity and obfuscates the natural geologic structure of the scenes. So we have moved ahead into 3-dimensional (3-D) processing, mapping, and visualization of our data and non-linear mapping techniques like point-set registration.

Once the data is in the same mathematical space, we have been able to statistically investigate the spatio-temporal relationships not only within each observers' gaze paths, but between observers, and between the observers and the scene they were viewing, in a 3-D computational image space with OpenGL. In the necessary research into, and application of, these algorithms to arrive at data that can be successfully and intelligently analyzed, we have developed of a unified visualization, research, and interactive analysis framework. Our software tools are open-source, custom libraries built on OpenGL and developed entirely in python for universality (and with cython for computational efficiency).

This work is supported by NSF Grant No. 0909588, and we aim to not only investigate education, but develop and share novel educational materials. All of our data and software will be made publicly available in 2014, but for ongoing information please visit http://geovis.cis.rit.edu.

9014-15, Session 3

An adaptive hierarchical sensing scheme for sparse signals

Henry Schütze, Erhardt Barth, Thomas M. Martinetz, Univ. zu Lübeck (Germany)

Compressive sensing (CS) is a signal acquisition technique by which an unknown signal of interest is simultaneously compressed and sampled. Each sensing value is obtained as the inner product between the signal and a predefined test function. The objective is to perform as few sensing actions as possible while capturing as much information as possible in order to enable an accurate reconstruction of the signal from the sensing values. If the signal of interest is k-compressible, in a specific orthogonal basis, and the test functions satisfy the restricted isometry property, then the number of sensing values required for a faithful reconstruction of the signal is of order O(k log n), where n is the signal dimension. With CS, the reconstructed signal is typically obtained by solving a convex optimization problem, i.e. seeking for the sparsest solution of an underdetermined system of linear equations.

Here we present a novel sensing algorithm that directly samples the coefficients of the signal in the orthogonal basis in which the signal is sparse. The algorithm is based on a binary decision tree that we call 'sensing tree'. Each node of the sensing tree is assigned to a subset of basis functions which form a particular test function that is constructed as a linear combination of the basis functions in the subset. If the absolute sensing value for a particular node is higher than a threshold, then the sensing is continued with the two descendants of that node. Each node represents a subset of basis functions that is half the size of the subset of its parent node and is disjoint to the subset of its sibling node. Else, if the threshold is not exceeded by the absolute sensing value, the sensing for the complete subtree rooted at that node is terminated.

The new algorithm is highly adaptive because the selection of the next test function depends on the previous sensing values. Moreover, solving an optimization problem for reconstruction is unnecessary, because the reconstructed signal is simply a sum of projections onto the orthogonal basis functions.

For k-sparse signals, our algorithm performs O(k log n) sensing actions like in CS. Fortunately, the value of k does not need to be known. We present results of our algorithm for test images that are ksparse in the DCT and Haar wavelet domain. Moreover, we present a rate-distortion analysis for three different sensing algorithms evaluated on a number of standard test images (Cameraman, Lena, Pirate).

The results show that our adaptive sensing algorithm performs better than (i) sensing by projections on randomly selected Haar basis functions and reconstruction from those coefficients, and (ii) CS with random (real valued) noiselet measurement ensemble and subsequent reconstruction by L1-norm minimization in a Haar basis.

The sensing scheme is relevant to known phenomena in human vision since the initial sensing values provide a 'gist' of the scene based on which further samples are acquired and a more detailed representation of the scene is obtained. Furthermore, we will relate the sensing strategy to eye movements and the corresponding sensing strategy of the human visual system.

9014-16, Session 3

Referenceless perceptual fog density prediction model

Lark Kwon Choi, The Univ. of Texas at Austin (United States); Jaehee You, Hongik Univ. (Korea, Republic of); Alan C. Bovik, The Univ. of Texas at Austin (United States)

The perception of outdoor natural scenes is important for understanding the environment and for executing visual activities such as object

detection, recognition, and navigation. In bad weather, visibility can be seriously degraded due to the absorption or scattering of light by atmospheric particles such as fog, haze, and mist [1]. Since the reduction of visibility can dramatically degrade an operator's judgment in a vehicle and induce erroneous sensing in remote surveillance systems, visibility estimation and enhancement methods on fog images have been widely studied [2~9]. Current fog visibility estimation methods require a corresponding fogless image to compare visibility [2], or salient objects in the image such as lane markings or traffic signs to supply distance cues [3]. Hautiere et al.'s method [4] depends on side geographical information obtained from an onboard camera, so such methods work only under limited conditions and are not necessarily applicable to general fog scenes. Regarding visibility enhancement of fog images, diverse defog algorithms [5~9] have been introduced. Early on, the performance of algorithms was only evaluated subjectively due to the absence of any appropriate assessment tool. Recently, gain parameters [10] were compared before and after a defog algorithm, or modified image quality assessment (IQA) tools were applied [5]. However, the subjective evaluation approach is not useful for large remote or mobile data, gain comparison method requires both the original and foggy images, and IQA methods are generally inappropriate since they are designed to measure distortion levels rather than visibility of a fog image.

We propose, for the first time (to our knowledge), a perceptual fog density and visibility prediction model based on natural scene statistics (NSS) [11] [12] and "fog aware" statistical features applied to a referenceless, single fog image. This model can predict the visibility in a foggy scene without reference to a corresponding fogless image, without side camera information, without training on human-rated judgments, and without dependency on salient objects in the fog image. The proposed fog density predictor only makes use of measurable deviations from statistical regularities observed in natural fog and fog-free images. This proposed model can be applied to assess the performance of defog algorithms by quantifying the visibility of defogged output images.

The proposed perceptual fog density prediction model is based on the construction of a collection of fog aware statistical features and fitting them to a multivariate Gaussian (MVG) [13] model. The fog aware statistical features are derived from a corpus of 160 natural fog and fog-free images [14-17], respectively, based on a space domain regular natural scene statistic (NSS) [11] model and observed fog characteristics including low contrast, fainted color, and shifted intensity. The spatial NSS model involves computing local mean subtracted contrast normalized coefficients (MSCN) [18]. Ruderman observed that such normalized luminance values strongly tend towards a unit normal Gaussian characteristic on good quality photographic images [11]. In addition, the divisive normalization is related to the contrast-gain masking process in early human vision [19], [20]. Although fog images are natural images, the variance of the MSCN coefficients is found to increase as the fog density increases. Furthermore, the pairwise product of neighboring MSCN coefficients along vertical orientation exhibits a regular structure. Hence, we use both as fog aware features. Additional fog aware statistical features are extracted as the local mean and (coefficient of) variance for sharpness [12], contrast using edge energy, image entropy [21], dark channel prior [6], color saturation, and colorfulness [22]. A total of 9 local fog aware statistical features are computed for each P ? P partitioned image patch. For a collection of fog aware statistical features, only a subset of the patches are used because every image is subject to some kind of limiting distortion [23], while, for a test fog image, all patches are used.

The 'fog level' of a test fog image is then expressed as Mahalanobis-like distance a multivariate Gaussian (MVG) fit of fog aware statistical features extracted from the test fog image, and an MVG model of the fog aware features extracted from the corpus of 160 natural fog-free images. Similarly, the 'fog-free level' of a test fog image is also expressed as distance between MVG model of fog aware statistical features extracted from a test fog image and from a corpus of 160 natural fog images. Finally, the perceptual fog density of a test fog image is calculated as the ratio of fog level to fog-free level.

To test the performance of the proposed perceptual fog density prediction model, a subjective study is performed by using another 100 fog images consisting of newly recorded fog images, widely used fog images in defog algorithms [6-9], and corresponding defogged output images. Test images were presented for 8 seconds on the center of a LCD monitor at a resolution of 1920 ? 1080. A total of 20 naïve subjects rated the visibility of fog images using the single-stimulus continuous quality evaluation (SSCQE) [24] method.

We use Pearson's linear correlation coefficient (LCC) and Spearman's rank ordered correlation coefficient (SROCC) to evaluate the performance of the proposed model. The predicted perceptual fog density scores are passed through a logistic non-linearity [14] before computing LCC and SROCC relative to visibility scores on fog images from the human subjects. The performance of the proposed model was compared using diverse patch sizes ranging from 4 by 4 to 160 by 160 pixels. The LCC and SROCC of the proposed model for a patch size 8 by 8 is 0.8748 and 0.8578, respectively, while the LCC and SROCC across a wide range of patch sizes show stable performance over all test fog images.

In addition, we validate the possibility of the proposed model as a tool to assess the performance of defog algorithms by comparing the predicted visibility of defogged output images [6-9] with the corresponding measured visibility from the human subjects. The high correlation between the predicted visibility of the proposed model and the measured visibility in a human subjective study indicates that the proposed model can reasonably evaluate the performance of defog algorithms.

9014-18, Session 3

Dynamics of Backlight Luminance for Using Smartphone in Dark Environment

Nooree Na, Jiho Jang, Hyeon-Jeong Suk, KAIST (Korea, Republic of)

People spend many hours looking at digital displays in their daily lives. Unlike ordinary objects, which are seen through recognizing reflected lights, digital displays emit light directly, resulting in more stress to the human eyes. In case of smartphone displays, this is more problematic due to their high mobility and social functionalities. Users sleep with their smartphones, and often wake up in the middle of the night to check the time or messages, during which they experience discomforting glare and strain to their eyes. This particular behavior is only seen with smartphones, and for this situation, providing an ideal backlight luminance is necessary. Currently, the solution most smartphones provide is called auto brightness. However, this function primarily operates in daily environments, and its features are inappropriate for pitch dark conditions such as when lying in bed at night, despite the frequent usage of smartphones prior to sleeping. Considering the two contexts as shown in Fig 1, this study investigates the ideal luminance levels for the initial viewing and the constant viewing of smartphone displays in dark environments.

In order to identify the optimal level of luminance for comfortable viewing of smartphones in the night time, two levels of luminance were investigated; the optimal level for initial viewing to avoid a harsh glare flashing into users' eyes, and the optimal level for constant viewing, which provides comfort, but is bright enough for constantly reading the displayed material. Accordingly, two types of stimuli were created for initial viewing, and constant viewing. For initial viewing, a white screen was shown on the smartphone. For the constant viewing, reading materials with black texts and white background were shown instead. In total, 5 levels of backlight luminance were used for viewing: 10cd/m2; 40cd/m2; 70cd/m2; 100cd/m2; 140cd/m2.

Fifty participants took part in the user test. There were a total of 5 sets, and each set mandated from the participants an evaluation for the initial viewing context and the constant viewing context. To find the ideal luminance for initial viewing, participants were first dark adapted for 5 minutes. After the adaptation, they were asked to look at a smartphone containing the white background with one of the five backlight luminance. Their facial squint was recorded, and evaluated by the experimenters based on 3 levels: neutral; slight squinting; squinting until eyes closed. Participants also rated how glaring the stimuli were based on the De

Boer rating scale (1: unbearable; 5: just acceptable; 9: just noticeable). Following the user study for initial viewing, ideal luminance for constant viewing was explored.

Participants were given reading materials to read for five minutes with the same backlight luminance from the initial viewing. In addition to having their eye blink frequencies recorded, participants also assessed their subjective preference on the stimulus using a 5 point scale. The process was repeated 5 times with randomly chosen luminance and reading materials.

For the facial squint evaluation, a repeated measure ANOVA test found significant differences between the facial squint levels under each luminance ($p < .01$). There were close to no changes in the participants' facial expression for stimuli with backlight luminance of 10 cd/m2 and 40 cd/m2, whereas the squinting of their eyes were observed for when the luminance was over 70 cd/m2 as shown in Table 1. A similar result was acquired in the subjective glare evaluation. The glare evaluation score fell rapidly as the luminance increased. For 70 cd/m2 and over, the score was less than 5 points on average, indicating that 70 cd/m2 is the acceptable limit for glare brightness ($p<.01$). These results confirmed that a luminance greater than 40 cd/m2 is inappropriate for initial viewing on a smartphone display. On the other hand, 10 cd/m2 helped users avoid eye fatigue and/or discomforting glare. Also, it was observed that there were significant differences between the eye blink frequencies under each luminance ($p=.05$). According to Hecht, the blink frequency under a non-stress condition and glare condition are about 10 blinks/min and 13 blinks/min respectively, and it increases when people feel visual discomfort. In this regard, it can be assessed that as the luminance increases, especially when it is over 100 cd/m2, it is not easy to use a smartphone in a dark environment. Lastly, the preference scores for each luminance were significantly different ($p<.01$). Luminance of 40 cd/m2 was most preferred, whereas 10 cd/m2 and 140 cd/m2 were the least preferred. Considering both blink frequency and preference, it is identified that backlight luminance with 40 cd/m2 is the optimal condition for constant viewing on smartphone displays.

By using the empirical results of the user test, the dynamics of backlight luminance was developed as shown in Fig 2.The luminance starts from 10 cd/m2 for avoiding glare, and as the user uses the smartphone, the luminance gradually increases to 40 cd/m2 for visual comfort. The luminance changes for 40 seconds because it takes approximately 40 seconds for light adaptation to be completed in indoor environments.

In this study, the dynamics of backlight luminance, which gradually changes luminance as time passes, was developed. By implementing this result, users would not have to feel the sudden glare when initially looking at a smartphone display in dark environments. Moreover, with the gradual increase in luminance, users can comfortably adapt to a backlight luminance that is bright enough for long uses in the dark.

9014-19, Session 3

Effects of image size and interactivity in lighting visualization

Michael J. Murdoch, Mariska G. M. Stokkermans, Philips Research (Netherlands)

Computer graphics-based visualizations are crucial in the development of lighting systems because they can clearly illustrate concepts as well as accurately convey the visual characteristics of a proposed lighting system. In our previous work, we introduced a novel approach of comparing the perceptual attributes of a virtual scene to those of a real scene, and we have shown a good match between renderings, photographs, and the real scene[1]. This and related previous studies were conducted using static images on large (42-46 inch) televisions [2] [3]. Building on this, we are interested in how the perceptual accuracy is affected by screen size: either with a smaller, portable screen such as a laptop or tablet, or with a larger (nearly life-size) projected image. Additionally, in an unpublished study involving a more complicated scene, we found that multiple camera views provided a more accurate perceptual impression of the room than a single wide-angle overview, presumably because this offered observers the opportunity to interactively explore scene elements. We hypothesize that both larger images and interactivity in viewing will lead to greater understanding of the scene and more accurate perceptual results.

1. EXPERIMENT

Our research track of proving the perceptual accuracy of visualizations of lighting systems with respect to a real scene was extended in new experiments investigating the effect of image size and interactivity. Using the same scene as in previous experiments[1], which exists both as a physical Light Lab as well as a detailed 3D model, we showed a series of fifteen lighting conditions varying in spatial distribution, color temperature, and brightness. Renderings were made of each light condition in two formats: a static, wide-angle view of the whole room and a "panoramic" full-sphere view from a single viewpoint which was navigable by the observer with a computer mouse. Both formats were viewed on three displays: a 15" laptop screen, a 46" TV, and a projected 138" image, as seen in Figure 1. Viewing distances were fixed to keep the visual angular widths constant between displays. Each of 24 observers viewed four of these six presentations (2 formats x 3 displays) in a balanced incomplete block design. Analogously, 28 observers viewed the same lighting conditions in a real room experiment in the physical Light Lab.

A questionnaire was used by all observers in both virtual and real room experiments. For each lighting condition, observers rated the scene on a 1-7 scale according to the ten perceptual attributes Overall Pleasantness, Overall Brightness, Overall diffuseness, Contrast, Uniformity, Shadow Visibility, Coziness, Liveliness, Tenseness, and Detachment. These attributes comprise overall impressions, physical uniformity, and perceived atmosphere[4]. Additionally, observers were asked to rate each format/display combination on five items related to presence and quality; these may be found in the legend of Figure 2.

2. RESULTS

With the observers' ratings of the perceptual attributes, we performed several analyses using linear mixed models (LMM), which are appropriate for incomplete and mixed between-within subject designs[4]. Firstly, looking for the effects of format and display within the present experiment, separate LMMs were computed for each of the 10 attributes. We found significant effects ($p<0.05$) of format for attributes Brightness, Contrast, Shadow Visibility, Liveliness, Tenseness, and Detachment; and significant effects of display for attributes Pleasantness, Brightness, Diffuseness, Contrast, Coziness, and Liveliness. These effects show that observers indeed perceived the room differently depending on the presentation.

A between-subjects comparison with the baseline real room experiment results allows an assessment of how accurately these different presentation styles convey the perceptual attributes of the real room. Separate LMMs were computed for each perceptual attribute with factors of light condition, display, and format, including the real room as a unique display/format combination. Full results including effects of lighting condition will be explained in the paper; here main effects for the various attributes are shown in Table 1, which shows differences in estimated marginal means between the real room and each presentation style using the model. The table shows notable robustness to the display and format (average absolute delta mean of 0.125 on 7 point scale), with most perceptual attributes consistently not significantly different – this overall high level of accuracy between the visualizations and the real room is the result of the success of our choices in 3D model, rendering, and tonemapping. Differences arise primarily for Brightness, which gets worse with panoramic format and with the projection, and for Uniformity, which gets worse with the laptop and in the static TV case.

An analysis of a post-experiment questionnaire reveals how impressions of presence and quality relate to the above, indirect measures. Using LMMs for each of 5 questions with factors display and format, display was found to be significant for all questions, while format provided a nuanced result to be discussed in the full paper. The effect of display is shown in Figure 2 where all the questions show a peak for the TV and lower values for the laptop and projection.

Returning to our hypotheses, neither is proven clearly. However, setting

aside Brightness, which throughout our research is consistently difficult to convey, both large-size projection and interactive viewing show an advantage. For other aspects, the effect of display size is a peak for TV, which correlates with image quality, but because overall the renderings give a good impression of the real world, using a laptop (e.g., for convenience) or a projector (e.g., for a large audience) does not risk losing perceptual accuracy in most attributes. Regardless of presentation, additional research is warranted for how visualizations affect brightness perception in general and for unusual, i.e., dimmed, lighting conditions.

REFERENCES

[1] Engelke, U., Stokkermans, M. G. M., and Murdoch, M. J., "Visualizing Lighting with Images: Converging Between the Predictive Value of Renderings and Photographs," Proc. SPIE 8651, 1-10 (2013).

[2] Salters, B., Murdoch, M. J., Sekulovski, D., Chen, S., and Seuntiens, P., "An evaluation of different setups for simulating lighting characteristics," Proc. SPIE 8291, 1–13 (2012).

[3] Stokkermans, M. G. M., Murdoch, M. J., and Engelke, U., "Preference for key parameter of tone mapping operator in different viewing conditions," in Experiencing Light, 1–4 (2012).

[4] Seuntiens, P.J.H. and Vogels, I.M.L.C., "Atmosphere creation: the relation between atmosphere and light characteristics" in 6th Conference on Design & Emotion, Hong Kong, (2008)

[5] McCulloch, C. E. and Searle, S. R., [Generalized, Linear, and Mixed Models], John Wiley & Sons, Inc., New York, USA (2001).

9014-20, Session 3

On the delights of being an ex-cataract patient

Floris L. van Nes, Technische Univ. Eindhoven (Netherlands)

Visual experiences before and after cataract operations, and what they indicate

1. INTRODUCTION

When I started school it was discovered that I was myopic (about –3D) and therefore I got glasses. After that age, my eyesight stayed more or less constant until, while trying to read from the projected slides during HVEI in 2011, I realized that my acuity had diminished considerably.

However, I also noted with surprise that when I was anywhere else, say walking in the woods with our dog, everything there looked entirely normal. Then I discovered – not as a philosophical1,2, or evolutionary concept3, but rather a personal reality, that human perception is only indirectly related to what is imaged on our retinas, but rather is inferred with our 'knowledge of the world'1,2 from the sensory input at that moment.

It turned out that in both eyes I had beginning cataract. My acuity clearly further deteriorated with its progress, but my "brain" for a long time managed to keep up the illusion of a sharp world where I lived and moved in – until that illusion collapsed; apparently the difference between sensory input and 'memory' had become too large. So I prepared for having my eye lenses excised. I chose to become 'emmetropic' with my future new implanted lenses (and use reading glasses for nearby viewing), even though I was told I could not reach 20/20 vision without any corrective distance glasses because of astigmatism.

I was operated in May (left eye) and June 2012 (right eye). Those cataract operations triggered most of the experiences described in this paper, on sudden changes in color- and depth perception, and of 'the sensation of super-sharpness'. Some similar experiences have been collected through interviews and literature.

2. COLOR

When the eye lenses get clouded they usually also get yellowish4. But since this happens very slowly, it is not to be expected that people will notice the resulting 'loss of blue' – so one would expect, at least from the retinal input, that after a cataract operation the ex-patient will notice with that eye a sudden change in color of all, or many objects.

The paper will present data from the author and a number of other ex-patients including (deceased) painters, for example Turner5, on the effects of the yellowing and its post-operation cancelation.

3. STEREOPSIS

The essence of experiencing 'true binocular stereopsis' is hard to describe, since it is easily confounded with inferences of depth from motion parallax and other monocular cues. It is perhaps the only 'new feature' that is contributed to visual perception through two eyes instead of one.

It is not uncommon after a cataract operation on one eye to lose binocular stereopsis temporarily, since basically with that eye now much more detail is seen than with the other, unoperated eye. That effect in my case was exacerbated by the fact that the operated eye had become emmetropic, and the unoperated one was still myopic. Yet I was entirely unprepared for the experience I had one week after the second operation when, after the morning ritual of removing the little protective cap over the just operated eye, preventing any rubbing in it during sleep, and dripping disinfecting and antibiotic fluids in it, I looked out of the window in our garden and.....saw it in "super-depth", with a depth impression stronger than anything I could remember ever having seen. Another novel feature of this experience of very strong stereopsis was its range, at least several tens of metres, and much farther than I was familiar with.

This experience was evoked several times, spontaneously, at other spots; but always in our garden, that is essentially a stretch of woodland, having many shrubs and trees of various sizes. And therefore a suitable surround for creating depth impressions.

After a few days it turned out to be impossible to evoke the same 'wowness', with its strong accompanying emotion. However, occasionally, the evocation of 'super stereopsis ' to a certain extent still remains possible.

4. VISUAL ACUITY

A few months after my second cataract operation I acquired (1) reading glasses, that give me very good near vision, and (2) 'glasses for other distances', in particular far away, providing me with an acuity of well over V=1. However, since I greatly enjoyed being able to do all non-nearby daily tasks, such as driving a car, without glasses, which I had absolutely needed for those tasks all of my life, I regarded the second glasses as a kind of telescope, reserving their use for special occasions, when I really wanted to see as 'far and sharp as possible'.

It was not before many months after obtaining the distance glasses that I discovered that, after putting them on, I may obtain a "super-sharpness impression", with an accompanying sensation of joy and feeling of presence. Under certain conditions there also appears to be an interaction between sharpness and stereopsis: immediately after putting the glasses on the depth of field seems to be enormous; i.e. much larger then I normally experience.

What causes all these perceptions and sensations ? I will try to give an answer in my paper, partly based on the concept of 'Bayesian reweighting of perceptual criteria'6,7.

REFERENCES

1. Helmholtz, H. von (1866) Concerning the perceptions in general. In: Treatise on physiological optics, Vol. III, 3rd edition. 1925 Opt. Soc. Am. Section 26, reprinted in New York: Dover, 1962.

2. Gregory, R.L. (1997) Knowledge in perception and illusion. Phil. Trans. R. Soc. Lond. B 352, pp. 1121-1128

3. Koenderink, J. (2011) Vision as a User Interface. Proc. SPIE-IS&T 7865, pp. 786504-1 – 786504-13

4. Pokorny, J., Smith, V.C. and Lutze, M. (1987) Applied Optics 26, pp. 1437-1440

5. Carinna E. Parraman, personal communication, 2013

6. Tyler, C.W. (2004) Theory of texture discrimination based on higher-order perturbations in individual texture samples. Vision Research Vol. 44, pp. 2179-2186

7. Tyler, C.W. and Gill, Navdeep (2013?) Stereo Hysteresis as Bayesian Cue Reweighting. To be published

9014-45, Session PTues

Color visualization of cyclic magnitudes

Alfredo Restrepo, Viviana Estupiñán, Univ. de los Andes (Colombia)

We exploit the perceptual circular ordering of the hues for the visualization of cyclic variables. The correspondence between hues and values of the cyclic or circular variable is called a color code. By assigning the 4 unique hues to 4 cardinal values of the cyclic variable, 8 color codes are possible. Cyclic orderings are mathematically characterized with the work of Megiddo: Partial and complete cyclic orders. Nimrod Megiddo; Bull. Am. Math. Soc.; Vol. 82,

No. 2, pp. 274-276; March, 1976. Examples of displays are given. A version for dichromatic people is also given.

9014-46, Session PTues

Quality evaluation of stereo 3DTV systems with open profiling of quality

Sara Kepplinger, Technische Univ. Ilmenau (Germany); Nikolaus Hottong, Hochschule Furtwangen Univ. (Germany)

Results from several studies show that the influences by different stereo 3DTV devices as well as different laboratory settings are not considerable for every research question investigated (e.g. [1], [2], [3]). However, some issues require a standardized setting excluding as much degrees of freedom as possible. Often, only available stereo 3DTV devices are used for quality assessment activities and their technical characteristics are reported in short. If different tests are made with different devices the question is how comparable the results may be. Herein, product related influence factors are not mentioned at all or even are regarded explicitly, like for example glasses and their comfort, or the device design and appeal. These make it hard to allow comparability of evaluation results. Hence, the authors investigate the question towards which extent stereo 3DTV systems as presentation device should be considered as influencing variable. Are there significant differences between technology wise different stereo 3DTV devices after sublimation of all other possibly differentiating factors? Therefore, two quality evaluations of stereo 3DTV systems are conducted.

Herein, the challenges are the

• Presentation devices themselves and their way of visual representation, as well as their calibration.

• Glasses.

• Test environment, viewing position (preferred viewing distance (PVD), viewing angle…).

• Test stimuli and content to show (high quality, device demanding content) as well as flawless and artifact free play out.

• Missing reference and insufficient defined quality describing attributes.

Therefore, the method used is Open Profiling of Quality (OPQ) which is a mixed method approach which allows to combine quantitative quality rating of quality acceptance (binary, yes/no) and overall quality rating (continuous scale, between 0-bad and 100-good) with the development of quality describing vocabulary [4]. These by the user developed attributes give a picture why something is rated as good or bad. Once the attributes are defined by the user, a rating of the quality with these attributes happens on a continuous scale between 0-minimum of appearance and 100-maximum of appearance.

For the test three commercial stereo 3DTV devices of (almost) the same size representative for the current developments were used:

• A: LCD, 46-inch-diagonal, LED-backlight, active shutter glasses (infrared), 1920x1080, time-sequential

• B: PDP, 50-inch-diagonal, active shutter glasses (bluetooth), 1920-1080, time-sequential

• C: LCD, 47-inch-diagonal, LED-backlight, passive polarizing glasses, 2x 960x1080, field-sequential

Within the first evaluation an elaborate calibration of all three displays was done in order to allow a general comparability. This involves the control and measurement of almost all other additional marginal conditions (i.e., light and color temperature of the viewing environment, light absorption by glasses) to exclude possible influencing factors on the experiment. This includes the development of glasses which are not differentiable, have the same weight, and the same design, independent from their technology based on the display used. Within the second evaluation all the conditions were kept the same, but the displays were not calibrated against each other. Instead, the factory settings were used investigating each system's strength.

The viewing environment for the experiments is based on the standard ITU-R BT.500 for studios and laboratory environments [5].

As the displays and not the quality of test stimuli should be assessed, optimal stereo 3D content is used. This means, that no quality problems are available from a technical, directorial, or perceptual psychology point of view. Therefore, 10 second long uncompressed master files of two different 3D cinema movies were used as test stimuli. They have the format 1080p/24/RGB, and are uncompressed and separated for left and right. The depth budget of the stimuli is adjusted to 50-inch displays available and converted to the HD709 color space. With these conditions the test stimuli are used without any further recoding. In order to meet the specific representation characteristics and quality degrading criteria of the displays, the test stimuli were extracted from the available movies based on following conditions:

• Object and scene tempo: high, low (Criteria: display based motion blur)

• Shown 3D depth: high, low (Criteria: depth simulation)

• Absolute scene contrast: high, low (Criteria: left/right separation, ghosting)

The play out to the displays happens through a HDMI-1.4 interface without any picture conversion in the 2x1080p/24/RGB/frame packet format, and with the software "Stereoscopic Player" [6]. The average data rate was herein around 300 MByte/s wherefore SSDs with high performance were necessary.

Each of the two tests itself takes about 90 minutes maximum per test participant. The test itself contains of several parts. In a first step an introduction into the whole procedure and overall questions including a stereo eye test [7] and a simulator sickness questionnaire (SSQ) [8] which is repeated at the end of the experiment. In a second step, a settling in period of two minutes happens with a representative set of used test stimuli. After that, the individual vocabulary definition by finding quality describing attributes happens as described by Strohmeier [4]. With these defined attributes the assessment period is conducted, wherein the combination of quantitative quality acceptance rating, overall quality rating, and judgment via quality describing attributes takes place based on a randomized presentation of test stimuli in randomized order on the three different displays.

Results of both experiments show that the overall acceptance of the display representation quality was rather high for all displays. There is a trend noticeable, that display "C" was rated as acceptable more often. Within the first experiment with calibrated displays as well as within the second experiment with the factory settings a significant difference in overall quality ration of the representation quality between display "A" and "C" ($p<0.05$), as well as between display "B" and "C" ($p<0.05$) could be detected with a Wilcoxon signed rank test. This is confirmed by analyzing the ratings of the different test stimuli per display and comparing them. The perceived difference is also reflected by the quality describing attributes after a Procrustes analysis.

These two experiments show that it is necessary to consider the different presentation devices used when examining comparison studies. However, in a further step up to now not regarded influences should be considered as well as a forced-choice-experiment with the presented defined quality describing vocabulary.

REFERENCES:

[1] Pinson, M., Janowski, L., et al., "The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study",

IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 6, pp. 640-651, 2012.

[2] Li, J., et al., "Cross-lab study on preference of experience in 3DTV: Influence from display technology and test environment", Proc. 5th International Workshop on Quality of Multimedia Experience (QoMEX 2013), 2013, pp. 16.

[3] J. Li, O. Kaller, F. De Simone, J. Hakala, D. Juszka, P. Le Callet, "Cross-Lab Study on Preference of Experience in 3DTV: Influence from Display Technology and Test Environment" Proc. 5th International Workshop on Quality of Multimedia Experience (QoMEX 2013), 2013.

[4] Strohmeier, D., "Open Profiling of Quality: A Mixed Methods Research Approach for Audiovisual Quality Evaluations", doctoral thesis, Technische Universität Ilmenau, 2011.

[5] ITU Recommendation: ITU-R BT.500-13. International Telecommunication Union, Geneva, Switzerland, 2012.

[6] Stereoscopic Player. Online available: http://www.3dtv.at/Downloads/Index_de.aspx. Last seen on 18th July 2013.

[7] The Randot SO-002 Manual. Online available: http://www.stereooptical.com/instruction-downloads. Last seen on 18th July 2013.

[8] Kennedy, R., Lane, N., Berbaum, K., Lilienthal, M. „Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness", Int. Journal on Aviation Psychology, 3(3), pp. 203-220, 1993.

9014-47, Session PTues

Saliency map computation in the MPEG-4 AVC compressed stream

Marwa Ammar, Mihai P. Mitrea, Marwen Hasnaoui, Télécom SudParis (France)

Introduction:

Visual perception is not limited to a simple photography of the world. Indeed, with its characteristics, our visual system provides us with knowledge about objects and events around us. Vision depends not only on the ability to perceive objects by the ratio between their size and the distance between the eye and the screen, but also on other visual, cognitive or semantic factors.

Such features of the human visual system (HVS) can be considered for the specification of a vision model for video processing (compression, watermarking ...). In this respect, perceptual masking and saliency map are two related yet different approaches.

On the one hand, perceptual masking is a model representing the perceptual characteristics of human eyes with three filters T, L, C (susceptibility artefacts, luminance perception, and contrast perception). On the other hand, a saliency map is a 2D topographic map representing the regions of an image/video at which the human visual system will mostly focus on. In order to highlight all the regions of interest, this map exploit the spatial and the temporal properties of the image/video.

State-of-the-art:

Several incremental studies, from still images to uncompressed video, already considered saliency maps in order to improve performances for a large variety of applications, such as the processing of rapid scenes, the selective video encoding, the prediction of video surveillance and the rate control.

For still image, the static saliency map is composed of three conspicuity feature maps: the intensity conspicuity map, the color conspicuity map and the orientation conspicuity map. These three maps correspond to different physical realms. First, the intensity map corresponds to the retina sensibility to the intensity of the light. Secondly, the color map relates to the sensibility to colors presented in each image by (r, g, b) because in the human vision the cone cells of the retina are denoted as red-sensitive, blue-sensitive and green-sensitive. Finally, the orientation map is given by the four orientations (0, 45°, 90°, 135°) for which neuronal sensitive features exist.

For videos, the static saliency map can be completed with the motion saliency map obtained by exploiting the motion vectors fields and highlighting regions having the highest motion energy since human vision tends to focus on moving objects and to ignore static objects when watching videos.

Paper objective:

The challenge of the present study is to extract the information required by the saliency map (intensity, color, orientation and the motion) directly from the MPEG-4 AVC stream syntax elements, with minimal decoding operations. This way, any compressed domain video processing algorithm can be matched to the HVS peculiarities without the need for complex (hence time consuming) decoding/encoding operations.

Paper main contribution:

First, an a-priori in-depth study on the MPEG-4 AVC semantic is carried out so as to identify inside each GOP (Group Of Pictures) the elements attracting the visual attention. Secondly, the MPEG-4 AVC reference software is completed with software tools allowing the parsing of these elements and their subsequent usage in objective benchmarking experiments. This way, it was demonstrated that an MPEG-4 saliency map is given by a combination of static and motion saliency maps.

Some details about the MPEG-4 AVC saliency map computation:

Intensity conspicuity map

It is computed based on an energy selection criterion: the intensity conspicuity map is given by the residual blocks elements whose energies are larger than the average value in that block. This criterion was heuristically set and then a posteriori validated by information-theory based methods.

Color conspicuity map

It is paired designed with the Intensity conspicuity maps.

Orientation conspicuity map

It is extracted based on the variation of the type of the Intra prediction mode considered when encoding that block with respect to its neighbours.

Motion conspicuity map

It is defined by taking into consideration the ratio of the current motion vector Euclidian norm to the maximal motion vector norm in that frame.

Merging formula

Several formulas for combining the four above conspicuity maps into a global MPEG-4 AVC saliency map are investigated: MEAN, MAX, AND, OR, …

Experimental results:

This MPEG-4 AVC saliency map can be used in order to optimally (under HSV constraints) selected the 4x4 blocks which can carry the mark in a watermarking application. When included in an m-QIM watermarking method, for fixed robustness against attacks and for the same amount of the inserted information, the saliency map resulted in improved transparency properties: an increase by 24% in PSNR and decreased by 71% in DVQ (the Watson's Digital Video Quality metric). These quantitative results were obtained out of processing 2 hours of heterogeneous video content.

Conclusion:

To the best of our knowledge, the present paper advances the first saliency map directly computed from the MPEG-4 AVC syntax elements. Experiments carried out under the framework of robust watermarking validate this approach. Further work is scheduled on exploiting this saliency map within intelligent video surveillance applications (selective encoding, region of interest detection, etc).

9014-48, Session PTues

Visual manifold sensing

Irina Burciu, Adrian Ion-Margineanu, Thomas M. Martinetz, Erhardt Barth, Univ. zu Lübeck (Germany)

We present Manifold Sensing (MS) as a novel method for sequential, adaptive sensing. MS localizes data points on manifolds of increasing but low dimensions, thereby limiting the total number of sensing measurements required to solve a particular recognition task. The method involves a two-fold adaptation process: (i) the algorithm adapts to particular data sets, and (ii) every new measurement (sample) depends on the previously acquired measurements. MS is related to Compressive Sensing [1], in the sense that every measurement is a weighted sum of the original unknown signal (the world to be sensed). MS, however, involves the above two-fold adaptation process.

MS is based on the assumption that natural images lie on non-linear low-dimensional manifolds and aims at optimizing both the sensing and the efficient representation of the visual world.

For a given data set, we first learn a manifold of dimension N, typically 2 or 3, by using Locally Linear Embedding (LLE) [2]. Any new data point is then projected to this manifold by using the pseudo-inverse matrix that minimizes the mean projection error over the adaptive data set. The adaptive data set is a subset of the original data set, which is embedded either in a higher dimension N+1, or in the same dimension N. The process is iterated up to a dimension N_max. Thereby, any new measurement depends on the previous ones because the adaptive data set used to make a new measurement is a neighbourhood defined by the previous measurements. There are three parameters: the number of neighbors for the local topology in the LLE algorithm, the (decreasing) size of the adaptive data set, and the dimensions N of the manifolds.

We evaluated the performance of MS on three benchmarks, two for face-recognition and one for the recognition of everyday objects. Thus, the information gathered during sensing is quantified by the recognition performance that it enables. In other words, the acquired samples are mainly assessed by how much they contribute to a particular task and not by how accurate they represent the world. However, we also evaluate the distance to the nearest neighbor in the data set, a measure that corresponds to a reconstruction error.

The first benchmark is the UMIST database with faces (twenty different persons with different poses and a total of 1000 images of size 256x256 pixel). We use MS to compute (i) the Signal to Noise Ratio (SNR) between the test image and the image that is the nearest neighbor on the learned manifold, and (ii) the person recognition rate (test images assigned to the class of the nearest neighbor). With only 30 measurements, i.e. a compression ratio greater than 2000, we obtain an average SNR over 20 test images (one per class) of 22.70 dB (the best possible average SNR for the database, when the correct nearest neighbor is identified in all cases, is 22.73 dB). On the same data, PCA with 30 components yields an average SNR of 22.60 dB. The recognition rate for MS with 30 measurements is 100%. With only 3 measurements (and N=3), the recognition rate is 75%. We obtained similar results on a different database, which contains face images of twenty different persons in different poses (a total of 1300 images of size 640 x 480 pixel) [3].

We also evaluate Manifold Sensing for object recognition on the ALOI database (Amsterdam Library of Object Images). This database contains twenty different everyday objects imaged with different rotation angles and different illuminations resulting in a total of 1400 images of size 192 x 144 pixels. With only 38 measurements, i.e. a compression ratio greater than 700, we obtain an average SNR of 15.3994 dB (in this case the best possible SNR is 15.4973, PCA with 38 components yields 15.1545 dB) and a 100% recognition rate.

Besides providing an effective sensing strategy for known environments, e.g. in robotics, we expect these results to provide new insights in explaining visual processes such as retinal and cortical projections [5], peripheral vision, gist, and eye movements. One could argue that human vision employs similar strategies since we are often capable of providing a “gist” of the scene before processing the details. In this context, it seems particularly striking that acceptable face recognition is possible with only 3 measurements.

References

[1] E. J. Candès, and M. Wakin, An introduction to compressive sampling. In IEEE Signal Processing Magazine, volume 25(2), pages 21-30, 2008.

[2] S. Roweis, and L. Saul, Nonlinear dimensionality reduction by locally linear embedding. In Science, volume 290(5500), pages 2323-2326, 2000.

[3] http://robotics.csie.ncku.edu.tw/Database.htm

[4] J. Romberg, Imaging via Compressive Sampling. In IEEE Signal Processing Magazine, volume 25(2), pages 14-20, 2008.

[5] W. Coulter, C. Hillar, G. Isley, and F. Sommer. Adaptive compressed sensing: A new class of self-organizing coding models for neuroscience. In IEEE International Conference on Acoustics Speech and Signal Processing, volume 5370, pages 5494-5497, 2010.

9014-49, Session PTues

Visually lossless coding based on temporal masking in human vision

Velibor Adzic, Howard S. Hock, Hari Kalva, Florida Atlantic Univ. (United States)

Abstract

In this paper we present methods to apply temporal masking phenomena of the human visual system (HVS) to improve compression performance of video encoders. The proposed approach is robust and can be implemented with minimal changes in any modern hybrid coder. Algorithm is evaluated using the most recent standard (HEVC/H.265) in order to show potential benefits for future applications. Experiments show savings of up to 10% in bitrate without any loss in quality as confirmed by two sets of subjective experiments using DSCQS and ACJ methodology.

Introduction

Psychophysical experiments are used to develop models of various perceptual phenomena observed in the HVS. Some of the models are implemented in modern image and video coding algorithms but many more are left unexplored. Visual masking is one such phenomenon that was originally reported at the end of the ninetieth century. This phenomenon is exhibited in either spatial or temporal domain. While researchers considered spatial masking for image coding and perceptual extensions, temporal visual masking received much less attention from engineering community. It wasn’t until couple of decades ago that temporal masking was explored for potential benefits in video coding.

Model

Temporal masking phenomenon can be used for video coding if we develop a model that correlates results from psychophysical experiments with parameters of video coding. Temporal masking occurs when one visual stimulus masks the other which is presented in a close temporal sequence. High tolerance for visual impairments just before or after a significant change between the frames of the video sequence is natural consequence of HVS limitations. Such sequence of stimuli can be observed during a scene change or locally in the areas of the pictures where luminosity is significantly changed between consecutive video frames.

Taking into account the characteristics of the frames (resolution, framerate, texture and luminosity) we developed a model that fits experimental data to a logistic function representing strong pattern masking that follow monotonically increasing curve of target visibility as the duration of stimulus onset asynchrony increases.

For the purposes of bitrate savings we are interested in the exact amount of additional quantization that can be introduced without being noticed by observers. This is clearly inversely proportional to target visibility and hence can be modeled by logistic function of a following form:

$$\Delta QP = s + (\Delta QP_{max} - s)/(1 + \exp(L - 2.5*(F/k)))$$

In this formula ΔQP is increase for quantization parameter (QP), ΔQP_{max} is a difference between maximum QP value allowed by the encoder and the QP value used by a standard encoder for that frame, s is initial value for QP increase (starting offset), F is the sequence number of the frame in a ramp, L is logistic parameter and k is used as normalizing coefficient for different number of frames in the ramp (number of frames divided by 5). Ramp can be extended to any number of frames by adjusting parameter k. Increasing the ramp increases bitrate savings but begins to

introduce perceptible distortion. In all of our experiments we use ramp with 5 frames only.

The backward masking model is implemented in the reference encoder for HEVC/H.265 standard (HM). All ramps are set to 5 frames before each scene cut. Values of QP increase for each frame in the ramp are determined using sampled values from continuous function.

This model is further extended to account for forward masking and local masking within the frames.

Experiments and results

The only way to confirm visually lossless performance is by using subjective testing. We conducted experiments with 20 subjects using adjectival categorical judgment (ACJ) method and then using double-stimulus continuous quality-scale (DSCQS) method, both as specified in ITU BT-500. All recommendations from the BT-500 specification were implemented in order to obtain valid results. We decided to use high definition clips of popular video sequences that were chosen based on YouTube popularity.

Only the results of the completed tests for a backward temporal masking (BM) as compared to HEVC (HM) are presented in this abstract. Additionally, we prepared control case sequences in which the same amount of quantization that is used for BM is inserted at random positions during the scene; we denote these sequences as RM.

In ACJ methodology sequences are presented in randomized pairs after which the subject is asked to assign a score to the sequence that is second in order. Results show that subjects didn't notice any statistically significant difference between HM and BM sequences. However, there is significant difference in quality ratings between both HM and BM sequences as compared to RM sequences.

Results of the DSCQS test underwent ANOVA analysis and were calculated for a 95% confidence interval indicate that no impairments were noticed in the BM sequences. Moreover, we additionally extracted and analyzed pairs of DSCQS comparisons between HM and BM sequences and couldn't find any statistically significant difference in scores. It is clear from the presented results that RM sequences again underperformed as compared to HM and BM sequences at the same QP level, which validates the claim of strong effects of backward masking that occludes impairments in BM sequences.

Results similar to ones reported previously for the AVC/H.264 sequences by the same authors. Savings are dependent on the number of scene cuts in the sequence and the available room for QP increase.

Savings in bitrate for the same subjective quality range from 2 to 10 percent, depending on the sequence. Average savings for the whole dataset are above 5 percent in all cases.

Extension of the model that includes forward masking parameters and local masking in the regions of the pictures that change significantly between two frames is analyzed. Correlation between content characteristics and masking effects gives strong indication for additional savings in bitrate.

9014-50, Session PTues

Some further developments on a neurobiologically-based two-stage model for human color vision

Charles Q. Wu, Stanford Univ. (United States)

This paper describes some further developments on a neurobiologically-based two-stage model (presented at Human Vision and Electronic Imaging XVII in 2012) for human color vision. In this model both stages are trichromatic -- particularly, the second stage is trichromatic, consists of complementary colors (instead of Hering's opponent colors), and directly corresponds to color appearance. In the present paper, I first present a conceptual analysis of Hurvich and D. Jameson's hue cancellation experimental paradigm and a quantitative analysis of their experimental results. Such analyses yield a convincing conclusion that the relevant data are consistent more with an explanation of complementary colors than with the opponent-colors theory as propounded by Hering-Hurvich-Jameson. In this regard, this author's analyses and conclusions are along the same line as those presented by K.A. Jameson and D'Andrade (1997), Pridmore (2011), and several other researchers. Second, I present the new model in terms of a computer simulation of two layers of units: The first layer represents the human retina with three sets of receptors corresponding to S, M, and L cones; and the second layer represents the spiny stellate neurons in layer 4C of the primary visual cortex. A crucial component of the model is that color composition is achieved through synchronized firings of such neurons. The model has been run to successfully simulate complementary afterimages and some variants of the afterimage phenomenon.

9014-51, Session PTues

Face detection on distorted images using perceptual quality-aware features

Suriya Gunasekar, Joydeep Ghosh, Alan C. Bovik, The Univ. of Texas at Austin (United States)

We use tools from perceptual image quality assessment (IQA) to better understand and to improve the performance of automated face detection algorithms when operating on distorted images. We first study the performance degradation of a widely-used generic face detector when the perceived image quality is degraded by three common distortions: additive white gaussian noise (AWGN), gaussian blur (GBlur) and JPEG compression (JPEG). We then propose a new face detector based on "QualHOG" features which augment face-indicative HOG features with perceptual quality-aware features to provide increased tolerance against distortions. We quantify the resulting improvements on a large, "Distorted Face Database" which we specifically created to enable our novel series of studies, and which is also being made publicly available.

There has recently been high interest in the development of automated No Reference (NR) image quality algorithms that aim to accurately predict end-user quality-of-experience. NR algorithms, as opposed to Full Reference (FR) and Reduced Reference (RR) algorithms, do not need to access or have information about undistorted "reference' images, and hence have a wider scope for application in practical settings where references are seldom available. The existing literature on the effects of image quality on biometric tasks has largely focused on the impact of scene-dependent challenges such as occlusion, illumination, etc., on face detection and recognition. In contrast, we focus on "quality impairments" arising from imperfect image capture, compression, processing, transmission, etc., while studying their effect on face detection performance. This is an important line of inquiry as in many communication systems, the effect of such quality impairments on detection/recognition can often be mitigated, e.g., by reallocating resources such as bandwidth.

We begin by exploring the question of whether the perceptual quality of facial images is a good predictor of the success of a face detection algorithm. Early work in this direction by Rouse et. al. show that perceptual FR image quality algorithms correlate strongly with the "recognizing threshold" of human observers. However, the effects of quality on machine vision algorithms have not been evaluated and moreover FR algorithms have limited applicability. We use an easily computable NR model called NIQE as a proxy for visual quality and study the performance degradation of a widely used HOG-based face detector as a function of NIQE scores. It is observed that the detection algorithm degrades rapidly within a fairly narrow range of NIQE scores. This means that if the visual quality of an image is in this range, a modest increase in quality can drastically improve face detection performance. These results can be fruitfully used to guide resource allocation in a communication systems associated with face detection tasks.

Secondly, we propose a new face detector based on QualHOG features that augment face-indicative HOG features with perceptual quality-aware Spatial Natural Scene Statistics (NSS) features. The motivation behind appending quality-aware features to HOG features is that the optimal

decision boundary in the HOG vector space varies non-trivially with input image quality and the Spatial NSS features in QualHOG potentially model a quality dependent boundary shift in the HOG feature space. We implement a fast computational algorithm using integral images to speed up Spatial NSS feature computation in a scanning window approach.

The new Distorted Face Database was curated from high quality images available freely on the internet. A total of 150 training images and 65 test images with one or more frontal faces were crawled, containing a total of 1231 and 393 faces respectively. These images were modified to introduce varying amounts of three types of distortions considered: AWGN, GBlur and JPEG, to create a variety of training/test image sets.

The positive face samples from each set of training and test images were manually annotated and extracted. For each set of training images, a random subset of negative patches from the non-face parts of the images was selected initially. Soft-margin linear SVMs were first trained using QualHOG features extracted on the positive and negative samples from different combinations of the training datasets described above. As a baseline, analogous classifiers were trained using the HOG features. For each detector, the non-face regions of training images were searched exhaustively for false positives ("hard negatives"). The classifiers were then retrained using the initial samples and the hard negatives to produce final detectors. For each test dataset, an exhaustive set of negative samples were extracted from the non-face regions. The area under precision recall curve (AUPR) was used as the evaluation metric.

In the results we first established that an effective NR algorithm, NIQE, can be used to evaluate the trade-off between face detection performances against image impairments arising from three distortion types considered. The performances of generic HOG-based face detectors degrade rapidly for NIQE scores greater than 4. It was also observed that for NIQE scores in the 5-8 range, a modest improvement in perceived image quality measures drastically improves face detection performance. This evaluation could be further used to guide resources in constrained settings.

We then show that QualHOG features are more effective at learning a face detector that is robust to the distortions considered. The QualHOG based detectors typically produced reliable results (AUPR>0.8) for test datasets with NIQE scores of up to 6.5, while HOG based detectors provide equivalent performance on images with NIQE scores only up to 5. Another way of examining the results is to measure the performance within a narrow range of NIQE scores, and here too, substantial gains were observed. For example, within a NIQE range centered at 6.0, QualHOG trained on distorted images achieves an AUPR of 0.86 while the AUPR achievable without the augmenting NSS features is only 0.72. Finally, for completeness the performance of QualHOG and HOG based face detectors were also compared against ground truth distortion levels and trends observed were similar to those observed again NIQE scores. Thus, the QualHOG based face detectors are able to achieve acceptable face detection performance at much higher levels of visual impairment than what is currently possible.

9014-52, Session PTues

Consciousness and stereoscopic environmental imaging

Steve Mason, Yavapai College (United States)

CONTEXT: This paper is concerned with the question of human consciousness that has intrigued philosophers and scientists for centuries: its nature, how we perceive our environment, how we think, our very awareness of thought and self – it has been suggested that stereoscopic vision is "a paradigm of how the mind works" (1) – in depth perception, laws of perspective are known, reasoned, committed to memory from an early age – stereopsis, on the other hand, is a 3D experience governed by strict laws but actively joined within the brain – one sees it without explanation.

OBJECTIVE: Much has been written about the process with few definitive conclusions – our awareness of the experience has only resulted in varying ways to calculate and create space on a flat surface – how do we, in fact, process two different images into one 3D module within the mind and does an awareness of this process give us insight into the workings of our own consciousness?

METHOD: How does this play out in imaging? As I reported at the 2007 SPIE conference in San Jose, one method of imaging employs only one image to create a truly 3D space, relying on the differing wave-lengths of light to determine the two images by using glasses (ChromaDepth™ 3D glasses) that refract the light in a different direction for each eye – colors of differing wavelengths will appear at varying distances from the viewer resulting in a 3D space – the latter involves no calculation, no manufacture of two images or views -- thus, spontaneous creation and instant recognition are possible which opens 3D imaging up to the creative artist.

RESULTS: Using this method, I initiated the creation of environmental spatial imaging – I generated 3D images that literally surround the viewer, a Star Trek space that one can truly enter, not just view through a window frame – the image is printed and then adhered to a circular mount – the viewer then enters the circle to experience colors suspended in a 3D space – I found that circular mounts needed to be at least 12 to 24 feet in diameter and 2 or3 feet tall to be effective and 8 feet tall or more for overwhelming effect – within the space the viewer experiences a loss of surface, or picture plane, upon which the image is created – unlike traditional 3D imaging methods, ChromaDepth 3D™ glasses can produce the space directly from the colors, even peripherally – the viewer is made acutely aware of the experience and thus the experience becomes the explanation.

NOVELTY: The aesthetic potential for 3D imaging has become limitless, yet few artists have pursued a truly fine art application – after the 2007 conference and the rising popularity of 3D, I expected a far greater reaction from artists but as I visit the museums of New York, LA, San Francisco, I see nothing to reflect this truly powerful and mind altering aesthetic – since then I have discovered far more potential for perception-based imaging than I could ever have anticipated – by focusing on our awareness we gain a deeper understanding of how the brain works, how we see – art can use the visual experience, not simply an explanation, to further the viewer's awareness of the process of seeing, — I present to you here one such example (see illustration).

9014-53, Session PTues

Bivariate statistical modeling of color and range in natural scenes

Che-Chun Su, Lawrence K. Cormack, Alan C. Bovik, The Univ. of Texas at Austin (United States)

Natural scene statistics (NSS) have been proven to be important ingredients towards understanding both the evolution of the human vision system and the design of image processing algorithms. Extensive work has been conducted towards understanding the luminance statistics of natural scenes, and the link between natural scene statistics and neural processing of visual stimuli. The natural scene statistics and models of 2D images have been applied to various image and video applications with success, e.g., image denoising and image/video quality assessment.

There have also been several recent efforts conducted on exploring 3D natural scene statistics. Potetz et al. examined the relationships between luminance and range over multiple scales and applied their results to shape-from-shading problems. Yang et al. explored the statistical relationships between luminance and disparity in the wavelet domain, and applied the derived models to a Bayesian stereo algorithm. Recently, Su et al. proposed robust and reliable statistical models for both marginal and conditional distributions of luminance/chrominance and disparity in natural images, and further incorporated those models in a chromatic Bayesian stereo algorithm yielding significantly improved performance relative to using luminance only.

However, no work has been reported on bivariate statistical modeling of luminance/chrominance and range data in natural scenes. In this paper, we aim to fill this gap by analyzing the high-definition, high-quality color images and corresponding ground-truth range maps from LIVE Color+3D Database Phase-1. We first transformed 2D images into the perceptually relevant CIELAB color space, and utilized an over-complete wavelet transform, steerable pyramid decomposition, to perform multi-scale, multi-orientation decomposition which mirrors the band-pass filtering that occurs in area V1 of the primary visual cortex. Next, we performed the perceptually significant process of divisive normalization transform (DNT) on the image wavelet coefficients at all sub-bands. For range data, we first took reciprocal of all ground-truth range maps, and then performed the same multi-scale, multi-orientation wavelet decomposition and divisive normalization transform to obtain the range coefficients after DNT.

To perform bivariate statistical modeling of these image and range coefficients after DNT, we examined two cases of spatially adjacent pixels, the horizontally and vertically adjacent pixel locations. Specifically, for horizontally adjacent statistics and modeling, we collected pairs of luminance/chrominance and range wavelet coefficients after DNT at sampled pixel locations (x,y) and (x+1,y) from all color images and ground-truth range maps in the database, while for vertically adjacent case, we sampled pairs at (x,y) and (x,y+1).

First, we examined the joint empirical statistics of spatially adjacent wavelet coefficients at each sub-band for both luminance/chrominance and range. We found that the bivariate joint histograms of spatially adjacent wavelet coefficients after DNT, which no longer have Gaussian-like properties at particular sub-band orientations, can be well fitted by bivariate generalized Gaussian distributions (GGD). Moreover, we discovered that there exist both scale- and orientation-dependencies in the joint distributions of spatially adjacent luminance/chrominance and range wavelet coefficients.

To be more specific, spatially adjacent pairs of wavelet coefficients in both natural images and ground-truth maps are extremely correlated when decomposed by a band-pass filter whose orientation is parallel to their spatial relationship; however, these pairs of wavelet coefficients become almost uncorrelated when their spatial relationship and the sub-band orientation are orthogonal. We also examined the bivariate conditional distributions of luminance and chrominance wavelet coefficients given different values of co-located range wavelet coefficients. In particular, we found that when conditioned on different values of range wavelet coefficients, both luminance and chrominance maintain constant correlation coefficients between spatially adjacent wavelet coefficients. These joint scale- and orientation-dependencies, as well as the conditional invariant correlation are completely reflected by the parameters of the corresponding bivariate GGD models.

We believe that these bivariate statistics and models will prove useful in understanding the processing of three-dimensional visual stimuli in vision systems, and will also benefit various image/video and computer vision applications, e.g., 3D quality assessment, 2D-to-3D conversion, etc.

9014-21, Session 4

X-eye: a reference format for eye tracking data to facilitate analysis across databases

Stefan Winkler, Subramanian Ramanathan, Advanced Digital Sciences Ctr. (Singapore)

Datasets of images annotated with eye tracking data constitute important ground truth for the development of saliency models, which have applications in many areas of electronic imaging. While comparisons and reviews of saliency models are quite common, similar comparisons among the eye tracking databases themselves are much less common.

In an earlier paper, we reviewed the content and purpose of over two dozen databases available in the public domain and discussed their commonalities and differences. A common issue with using the data from the various datasets is that their formats vary a lot owing to the nature of tools used for eye movement recordings, and often specialized code is required to use the data for further analysis. In this contribution, we therefore propose:

1. A common reference format for eyetracking data.

2. Conversion routines for 13 existing image eyetracking databases to that format.

3. An analysis of center bias across databases as an example application facilitated by X-Eye.

REFERENCE DATA FORMAT

We first define a uniform data structure for image eyetracking data. The header part contains the following information:

1. Image name

2. Image width

3. Image height

4. User ID

5. Number of fixations

This is followed by the list of individual fixations:

1. Fixation number

2. Fixation location (x,y) in image coordinates

3. Fixation begin time

4. Fixation end time

5. Fixation duration

6. Inter-fixation time

DATA CONVERSION

Since the different databases include eye movement information recorded using different eye tracking hardware and software, their file formats vary considerably. For example, some eye trackers record eye movements based on the image coordinates, while others record the same with respect to screen coordinates. The coordinate system is one key aspect that needs to be accounted for, while employing eye tracking data for further analysis. We wrote code to convert each dataset to the reference data format described above. The conversion routines as well as the outputs for the different datasets in the above reference format will be shared with the scientific community upon acceptance of this paper.

ANALYSIS ACROSS DATASETS

Once data is available in a common reference format, different types of analysis are possible across databases. As an example, we analyzed 13 datasets for center bias, which relates to the tendency of users to fixate around the image center. Consistent with previous works, we defined the center region using rectangles around the image center, that is of the same aspect ratio and encompassing (a) 11%, and (b) 25% of the image area. We analyzed the average number of fixations falling in these central regions, and found 3 out of 13 databases to have the majority of fixations within the central 11%, and only 2 out of 13 databases with the majority of fixations outside the central 25% region.

Since early fixations can be more likely to fall around the image center, we also considered the fixations (a) over the entire stimulus presentation time, and (b) within the first 500ms of stimulus onset for our analysis. These results will be shown in the full paper.

The full paper will also contain more detailed graphical representations of the center bias analysis. For example, we are computing heatmaps of all fixations over all images for each dataset. Furthermore, the histograms of fixations as a function of the distance from the center will also be shown. Finally, we are able to identify the most and least center-biased images from each dataset along with their eye fixation distributions. This allows us not only to draw conclusions about which databases contain the most center bias, but also to determine which type of content is most prone to center bias effects.

9014-22, Session 4

Modeling the leakage of LCD displays with local backlight for quality assessment

Claire Mantel, Jari Korhonen, DTU Fotonik (Denmark); Jesper M. Pedersen, Soren Bech, Bang & Olufsen A/S (Denmark); Ehsan Nadernejad, Nino Burini, Søren O. Forchhammer, DTU Fotonik (Denmark)

Up until now research in image and video quality metrics has mainly focused on assessing the effect of transmission and compression artifacts on quality and considered displays as artifact free. This is for example reflected in the ITU or VQEG recommendations (ITU-BT512, VQEG2010), where minimum characteristics are given for the displays so that they are basically considered 'good enough' to have negligible effect on the displayed quality.

The importance of the display was investigated with the arrival of Liquid Crystal Displays (LCDs) to evaluate the change they would imply compared to Cathode Ray Tubes (CRTs) and the switching from Standard resolution (1280x720) to High Definition (HD) resolution (1920x1080) (Tourancheau2007, Tourancheau2009). The resulting effect from those studies is that the recommended viewing distance switched from 8H (H being the height of the display) for SD content to 3H for HD (according to VQEG2010). It is known that LCDs changed viewing experience compared to CRTs: the sharpness was increased but the color gamut differed and their temporal characteristics led to motion blur (Elze2007). The continuous progress to avoid motion blur and the stop of CRTs production recently led vision scientists to reconsider LCD as 'good enough' for their needs (Lagroix2012).

The arrival of LCD screens with local backlight dimming sheds a new light on that matter as the backlight dimming is bound to have an impact on image quality and there is no standard for computing a backlight. Brunnström et al have indeed shown that local backlight dimming has an impact on the perception of motion blur (Brunnström2010). The two major spatial artifacts that can result from backlight are clipping when Liquid Crystal cells (LCs) do not receive enough light to achieve the targeted brightness and leakage when LCs receive more light than they are able to block and reach a higher brightness than desired (Seetzen2004). An example of how the displayed image is produced, including leakage, is shown in Fig. 1.

Several algorithms have been developed to try and exploit best the possibilities offered by local backlight. Using simple image statistics (Seetzen2004), multiple histograms of the image (Zhang2012, Nadernejad2013) or a model of the display (Albrecht 2009, Burini2013), the algorithm decides on the intensity of each Light Emitting Diode (LED) according to its own strategy. For exactly the same input image, the rendered image displayed on the screen will be different and vary in quality, so there is an additional step in the quality assessment chain.

Indeed, backlight dimming implies that the screen needs to be modeled as there is no signal representing the displayed image (as the decompressed image can be to evaluate the effect of compression). A previous study has shown that it is possible to model the displayed image (Korhonen2011), which can then serves then as an input for objective metrics.

The performance of local backlight dimming algorithms is evaluated most of the time through contrast measure and there is no dedicated metric. A previous study by the authors investigated the efficiency of 'classic' image quality metrics on images displayed with several local backlight dimming algorithms (Mantel2013). Results showed that several metrics performed satisfyingly for bright images but fewer with contrasted dark images. This kind of image is indeed the most challenging for local backlight dimming as there is no perfect display solution but a trade-off must be chosen between providing enough light to pixels (i.e. to avoid clipping) and not providing too much light to dark areas so they are truly black (i.e. to avoid leakage).

Many characteristics of a display could play a role in the resulting quality of an image. The aim of this paper is to investigate the importance of leakage for such a model. More precisely: the vision angle and the precision of the leakage factor. To assess the role of those features we compare the correlation of subjective evaluations (our ground-truth) to quality metrics applied on frames computed with different models.

We set-up a behavioral experiment to obtain subjective ratings of videos displayed on a LCD screen with an LED backlight composed of 16 segments: 8 rows and 2 columns. To evaluate the role of the vision angle, the experiment is split in two parts: in the first one the participants are located right in front of the display while in the second they are viewing the screen with a 15° angle on the left from the center. 16 observers participated in the experiment; each of them performed both parts with a 3-weeks interval in between.

A previous experiment (Mantel2013) showed that the impact of local backlight dimming was mainly visible on dark videos with high contrast; moreover, by definition leakage is mainly perceivable on videos with dark areas, so we focused on that kind of videos for this experiment. Four sequences were selected 'Stars', 'Titles' (both from Sita), 'Volcano' (CDVL2012) and 'Uboat'; frames of the first three are visible in Fig. 2, the fourth one cannot be shown due to its copyright.

Seven different algorithm ms were chosen to compute the backlight dimming for each frame of each video: 'Gradient-Descent' with two leakage factor (Burini2013), 'Nadernedjad' (Nadernedjad_SPIE2013), 'Albrecht' (Albrecht2009), 'Cho13' (Cho2013), 'Zhang' (Zhang2012) and 'Full' which means setting all LEDs at full intensity. A too fast variation in the LED intensity produces a flash called flicker artifact. As those algorithms are designed for images and not videos, a filter was applied to remove flicker when necessary (Nadernedjad_MMSP2013).

The participants were asked to rank the algorithms in the order they preferred. For each sequence, they could see all versions as many times as they wanted before ranking them (but had to see them all at least one time).

The displayed frames were computed using the model described in Burini2013 changing two parameters: the leakage factor and the spatial variation of leakage. We used three different values of leakage factor: 0, which corresponds to ignoring leakage, 0.00047 and 0.00068 which are the leakage factors on the display used for the experiment measured at 0° and 15° vision angles, respectively. The leakage was either considered constant over the whole screen or varying horizontally (using linear interpolation and leakage measures made from 0° to 30°).

The corresponding computed images were used as inputs for the metrics are MSE (computed on all pixels and on pixels presenting leakage only), PSNR, SSIM, IWSSIM (Wang 2011), HDRVDP (Mantiuk2011) and PSNR-HVS-M (Ponomarenko2009). The number of sequences used is too low and their content too specific to draw any conclusion regarding the metrics performance. The issue in this paper is solely to investigate which leakage model seems better adapted to visual perception.

For all images metrics, 10 different temporal pooling methods were applied: the average, the average over the worst 10%, the best 10%, the first 2s, the last 3s, Minkowsky summation with powers from 3 to 5, the FIR described by Hamberg and DeRidder (Hamberg1995) and the asymmetrical pooling introduced by Ninassi et al. in Ninassi2009. The temporal pooling method had an effect on the correlation values but not the type of model or the leakage factor, therefore to clarify results only the best pooling (i.e. the one with highest correlation) is presented in this analysis. For MSE and MSE on leaking pixels the best pooling method is the average over the worst 10%, for the other metrics it is the one introduced by Ninassi et al. in Ninassi2009.

The correlations of subjective data with objective metrics using the different leakage models are visible in Fig. 3. As previously stated, the sequences number and content are too few to draw any conclusion regarding the metrics performance. The correlations using constant and horizontally-varying leakage modeling seem really close for all metrics at 15° and some (MSE on leakage, SSIM, IWSSIM and HDRVDP) at 0°. Indeed the Williams test (Howell2010) on Pearson's correlation coefficient show that for those values (in gray in Fig. 3) the correlations with constant and horizontally varying leakage are not statistically different ($p>0.05$). This indicates that the horizontally varying model might prove useful for viewings at 0° but is not at 15°.

Figure 4 shows the correlations with the different leakage values. Results show that not considering the leakage in the display model

decreases the metrics performance: the correlations are negative, these differences are statistically significant as verified using Williams' test on CC values ($p<0.05$). The other comparison was between the 0° and 15° leakage value: it seems that the 15° leakage factor allows metrics to perform better, even for predicting subjective grades at 0°. However, this difference is significant only for some metrics: MSE, PSNR and PSNR-HVS-M.

To conclude, results show that for dark sequences accounting for the leakage artifact in the display model is definitely an improvement. Approximating that leakage is constant over the screen seems valid when viewing from a 15° angle while using a horizontally varying model might prove useful for 0° viewing.

9014-23, Session 4

On improving the pooling in HDR-VDP-2 towards better HDR perceptual quality assessment

Manish Narwaria, Matthieu Perreira Da Silva, Patrick Le Callet, Romuald Pepion, Institut de Recherche en Communications et en Cybernétique de Nantes (France)

Traditional capture and display devices can only support a limited dynamic range (contrast) and color gamut given the hardware limitations. As a result, the real physical luminance present in a natural scene cannot be captured by these. However, with the recent advancements in the related software and hardware technologies, it is now possible to capture or reproduce higher contrast and luminance ranges. Such scene-referred visual signals are known as High Dynamic Range (HDR) signals. Obviously, they are visually more appealing as they can represent the dynamic range of the visual stimuli present in the real world. Not surprisingly, the emergence of HDR is seen as an important step towards improving the visual quality of experience (QoE) of the end users. Even though subjective assessment of visual quality remains the 'gold' standard, its deployment is difficult in some situations (eg. real-time HDR compression). Thus, there is obviously a strong need to develop objective computational models that can predict the perceptual quality of HDR signals in an objective manner. Such models will be extremely useful in an HDR processing pipeline for predicting the visual quality of processed HDR images/videos. Unfortunately, the conventional objective visual quality prediction methods do not take into account the luminance range and typically assume that the input pixel values are perceptually uniform. As a result, these cannot be used in case of higher luminance conditions as is usually the case with HDR visual signals. Recently, the HDR-VDP-2 algorithm has been proposed. It is an extension of the Visible Differences Predictor (VDP) algorithm. The HDR-VDP-2 uses an approximate model of the human visual system (HVS) derived from new contrast sensitivity measurements. Specifically, a customized contrast sensitivity function (CSF) was employed to cover large luminance range as compared to the conventional CSFs.

HDR-VDP-2 is essentially a visibility prediction metric. That is, it provides a 2D map with probabilities of detection at each pixel point and this is obviously related to the perceived quality because a higher detection probability implies a higher distortion level at the specific point. Nevertheless, in many cases, it is crucial to know an overall quality score (rather than just the local distortion visibility probability). Pooling is a crucial aspect in converting local error distribution into a single score that denotes the perceptual quality and the human visual system (HVS) can do it easily and accurately. But it is much more difficult to realize that in an objective quality prediction model given the underlying complexities and lack of knowledge of the HVS's pooling mechanisms. It is believed that multiple features jointly affect the HVS's perception of visual quality, and their relationship with the overall quality is possibly nonlinear and difficult to be determined apriori. The proposed work seeks to address the issue of feature pooling for HDR quality assessment.

Our work is novel in that there does not exist a comprehensive treatment of the aforementioned issue of pooling in the current literature especially for HDR quality assessment. It is also challenging to realize and model the actual perceptual pooling mechanisms for quality assessment. Therefore, the main contribution of the proposed work lies in analyzing and exploring a more effective pooling solution towards more accurate and objective HDR perceptual quality measurement. Our effort is also amongst the first to introduce an HDR image database with subjective scores. This will be of immense value to the research community given the lack of publicly available databases for HDR content quality evaluation.

In its original implementation, the authors of HDR-VDP-2 tried over 20 different combinations of aggregating (or pooling) functions. These included maximum value, percentiles (50, 75, 95) and a range of power means (normalized Minkowski summation) with the exponent ranging from 0.5 to 16. The aim was to maximize the value of Spearman's correlation coefficient in order to find the best pooling function and its parameters. While HDR-VDP-2 is fairly comprehensive method for HDR quality assessment, we identify the following 2 issues with regards to pooling in HDR-VDP-2 which in our opinion deserve further analysis:

1. Parameter optimization: the parameters of the pooling function in HDR-VDP-2 were found by maximizing (optimizing) correlation using existing LDR image databases. Therefore, its effectiveness in predicting the visual quality of HDR images is questionable given the different characteristics LDR and HDR images especially in terms of distortion visibility and overall visual appeal.

2. Selected form of pooling function: the original HDR-VDP-2 uses a parametric logarithm based function. However, the choice of such function has not been justified and more sophisticated pooling function is expected to lead to better objective prediction performance.

In this work, we aim to address the two mentioned issues. The first and foremost requirement to that end is the development of a comprehensive image database with distorted HDR images and their subjective quality ratings. We conducted careful subjective studies using a total of 140 distorted HDR images and 27 observers. For displaying the HDR images, SIM2 Solar47 HDR display was used. More details will be included in the final manuscript. We then used the said in-house developed database to improve pooling in HDR-VDP-2. More specifically:

1. We optimized the pooling function in HDR-VDP-2 on subjectively rated HDR images to improve the prediction accuracy as compared to the original pooling function. More specifically, HDR contains much more fine details and conveys more precise scene information. This can be accounted for via proper selection of the parameters in the pooling function by the use of HDR stimulus thereby leading to more accurate modeling of the pooling in the HVS.

2. As mentioned, the perceptual pooling of the HVS is highly non-linear We therefore use a machine learning based non-linear pooling methodology. This leads us to find a theoretically better grounded pooling function because the features from HDR-VDP-2 comprises the input while the subjective quality the target values. Such data-driven feature pooling helps in avoiding ad-hoc functional form as well as increasing the quality prediction accuracy.

9014-24, Session 5

Theory and practice of perceptual video processing in broadcast encoders for cable, IPTV, satellite, and internet distribution *(Invited Paper)*

Sean McCarthy, ARRIS (United States)

Digital video plays an important part of people's lives every day by providing vital information and enjoyable entertainment. Most of the time, viewers are not aware of the massive and complex infrastructure that exists beyond the TV screen that is needed to deliver the bits that make up the video. But sometimes – perhaps more often than any of us would like – viewers are aware of visual distortions and compression artifacts that distract from the program they are watching. Those artifacts are a

result of a mismatch between the bit rate the video program needs for acceptable video quality and the bit rate that can be delivered over the viewer's connection.

The potential for frequent bit rate mismatches – and resulting deficit in video quality – is becoming more of a concern all the time. Consumer demand for digital video is sky rocketing, yet video service providers (cable, IPTV, satellite, and now quite often the Internet) can build out new channel capacity only so fast. Moreover, many service providers rely on legacy MPEG-2 and H.264 decoders that are too costly to replace or upgrade quickly with, for example, the latest video compression standard, High Efficiency Video Coding (HEVC). As a result, service providers are becoming ever more sophisticated about the video processing that surrounds the core encoding process. One of those new sophisticated video processing technologies leverages human visual perception.

This paper describes a perceptually-inspired video processing (PVP) technology that improves compression efficiency over what can be achieved with MPEG-2, H.264, and HEVC by themselves. Data that we will present show that compression efficiency can be improved by up to 50%, though 10-25% is more usual in actual practice. This PVP technology has been recently and successfully productized to become a component of professional video encoders being deployed and used by major cable, IPTV, satellite, and Internet video service providers worldwide.

A key element of the perceptual video processing technology described in this paper is the formation of spatial maps that predicted how likely it might be that a free viewing person would look at particular area of an image or video. We will illustrate how the predicted eye-tracking attractor maps evolve over time depending on the particulars of the content being viewed, and address special-handling situations such as hard-cut scene changes, inverse telecine, and other visual phenomenon that occur in video but not natural scenes.

We will present in this paper a novel model and supporting theory used to calculate the eye-tracking attractor maps. We will show how the underlying perceptual model was inspired by electrophysiological studies of the amphibian retina, and we will explain how the model incorporates statistical expectations about natural scenes as well as a novel method for predicting error in signal estimation tasks. We will present electrophysiological data to illustrate the correlation between the retina-inspired model and measured responses of amphibian retinal neurons. And we will show that the model provides useful eye-tracking attractor information despite the differences between amphibian and human vision.

This paper will also describe how the eye-tracking attractor maps are created and utilized in real-time to modify video prior to encoding so that it is more compressible but not noticeably different than the original unmodified video. We will illustrate how the modifications preferentially impact bit allocation in regions of the video that are both hard the compress (high entropy) and also likely to be difficult to track visually.

9014-25, Session 5

Temporal perceptual coding using a visual acuity model *(Invited Paper)*

Velibor Adzic, Florida Atlantic Univ. (United States); Robert A. Cohen, Anthony Vetro, Mitsubishi Electric Research Labs. (United States)

Overview:

Psychophysical studies of the human visual system (HVS) revealed that through a very sophisticated and complex cascade of sub-systems it filters input information in such a way that only "important" components are kept. What this effectively means is that significant portion of the visual signal presented to our eyes never reaches higher processing levels in our brain. It is not before the received signal passes through HVS that the description of its quality can be given.

In general, for image and video coding systems, in order to obtain better quality we have to spend more bandwidth or use more bits to describe the original signal. However, having limited resources requires us to spend bits in a careful and optimized manner. Spreading bits uniformly across spatial and temporal dimensions of a compressed signal is often redundant because of the way HVS processes visual stimuli. If we can filter out the same information that is not kept by HVS we can achieve visually lossless compression while reducing the bitrate or size of the compressed data. For visually lossy compression, perceptual techniques also can be applied to obtain coding gains over non-perceptual coders.

While some of the aspects of the HVS have been considered in the design of the modern hybrid coders, many other characteristics are not exploited. Although newly released video standards such as HEVC provide significant improvements over state-of-the-art codecs such as H.264/AVC, the reference encoders still make decisions based on non-perceptual metrics such as sum of squared error. Recent research has incorporated perceptual distortion models into these codecs, in which the quantization process is modified so that coarser quantization can be used in areas where the HVS is less sensitive to spatial distortion. While reducing the overall bitrate, such methods do not eliminate the need for certain data to be signaled to the decoder. Furthermore, the existing perceptual coding models do not incorporate the concept of visual acuity as affected by motion. As described in the next section, research has shown that the sensitivity the HVS to objects containing higher frequency components is affected by the velocity of the objects.

This abstract summarizes a method for using a temporal visual acuity model in a video coder to eliminate the need to signal higher-frequency transform coefficients. The end effect is that blocks corresponding to areas with faster motion are coded using fewer coefficients, thus reducing the overall bitrate of the compressed data. Experimental results are shown for when this model is integrated into HEVC.

Temporal Visual Acuity Model:

For the visual acuity model we are using results from the experiments by Kelly and Eckert and Buchsbaum. Kelly established the thresholds of contrast sensitivity at different velocities of moving sine-wave gratings. His results for the maximum spatial frequency resolved by the HVS for a fixed low contrast level are used to establish velocity threshold. After separation of the model for horizontal and vertical components we a maximum perceptible frequency:

$$K_i = (K_{max} * v_c) / (v_{Ri} + v_c)$$

where i denotes the x or y component, K_max is the highest perceptible frequency (32 cycles/deg in our paper) for a static stimulus, v_Ri is the x or y component of the retinal velocity of a stimulus and v_c is the corner velocity, i.e. where the spatiotemporal sensitivity of the HVS is greatest, fit to Kelly's data. In this model v_c = 2 deg/s.

A reasonable estimate for the number of pixels per degree of viewing angle, under typical high-definition viewing conditions, is 64 pixels/deg. To avoid the need to signal additional side information indicating the velocity of objects, we use motion vectors as an approximation, as the decoder already has motion vectors available before the inverse quantization process is applied. The velocity on the image plane of a block can then be computed given its motion vector length and the frame-rate of the video. We use a model from Daly, which computes the eye velocity as a function of target velocity on the image plane and then subtracts the eye velocity from image plane velocity to obtain retinal velocity v_R.

For each block of transform coefficients that have an associated motion vector, the value K_i is used to compute horizontal and vertical position thresholds, for which all transform coefficients located in rows or columns above those thresholds are not signaled. This thresholding is equivalent to setting the removed coefficients to zero. The distortion used in cost functions is also modified so as not to be penalized by the removal of these coefficients, according to the perceptual model.

Experimental Results:

The HM 11.0 software was modified to incorporate the temporal visual acuity model. The unmodified codec was run for a given QP value, and then the modified codec was run using the same QP. For this abstract, 100 frames of BasketballPass and 50 frames of PartyScene were coded.

Additional simulation results will be provided in the full paper. For QPs of 20 and 30, the bit-rate reductions for PartyScene were 23.4% and 12.2% respectively. For BasketballPass, the bit-rate reductions were 9.0% and 1.7%.

Summary:

This abstract summarizes work showing how a temporal visual acuity model can be used to improve the coding performance of HEVC by eliminating the need to signal coefficients based upon the frequency content and velocity of blocks. Reductions in bit-rate of up to 23% were obtained. Future research will include combining it with spatial perceptual models, which have been shown to produce additional gains in compression efficiency.

9014-26, Session 5

Characterizing perceptual artifacts in compressed video streams *(Invited Paper)*

Kai Zeng, Tiesong Zhao, Abdul Rehman, Zhou Wang, Univ. of Waterloo (Canada)

Please see attached file. Thanks

9014-27, Session 5

Zero shot prediction of video quality using intrinsic video statistics *(Invited Paper)*

Anish Mittal, Nokia Research Ctr. (United States); Michele A. Saad, Intel Corp. (United States); Alan C. Bovik, The Univ. of Texas at Austin (United States)

We propose a no reference (NR) video quality assessment (VQA) model. Recently, `completely blind' still picture quality analyzers have been proposed that do not require any prior training on, or exposure to, distorted images or human opinions of them cite{niqe}. We have been trying to bridge an important but difficult gap by creating a `completely blind' VQA model. The principle of this new approach is founded on intrinsic statistical regularities that are observed in natural vidoes. This results in a video `quality analyzer' that can predict the quality of distorted videos without any external knowledge about the pristine source, anticipated distortions or human judgments. Hence, the model is zero shot. Experimental results show that, even with such paucity of information, the new VQA algorithm performs better than the full reference (FR) quality measure PSNR on the LIVE VQA database cite{livevqadatabase}. It is also fast and efficient. We envision that the proposed method is an important step towards making makes real time monitoring of `completely blind' video quality feasible.

9014-28, Session 6

Personalized visual aesthetics *(Invited Paper)*

Edward A. Vessel, New York Univ. (United States)

People derive pleasure from looking at visual media, and even static images can be intensely moving. How is visual information linked to aesthetic experience, and what factors determine whether an individual finds a particular visual experience pleasing? Here, we present findings from two behavioral experiments and one brain imaging (fMRI) experiment that address the degree to which individuals' aesthetic responses are determined by objective image features (such as contrast or the presence of specific visual structure) versus internal, subjective factors that are shaped by a viewer's personal experience. Importantly, our experimental approach emphasizes the measurement of preferences for individual observers.

In the first experiment (Vessel & Rubin, 2010), observers made preference judgments about a set of color photographs of real-world scenes and, in a separate session, about a set of novel, abstract images that contained a variety of low-level image features but lacked semantic content (e.g. fractals, kaleidoscopic images, etc.). A measure of agreement across observers was calculated by i) computing correlations between the preference scores for pairs of observers and ii) averaging these correlations across all pairs of observers. In agreement with previous findings (e.g. Kaplan & Kaplan 1995), observers showed a high degree of agreement in which real-world scenes were preferred ($r = 0.46$). However, agreement on the abstract images was significantly lower ($r = 0.20$). This suggests that individuals' preferences for real-world scenes were not driven by low-level visual features (present in both stimulus sets) but rather by higher-level semantic associations. In the absence of such commonly-shared semantic associations, preferences for abstract images were highly individual. Interestingly, intermixing the two types of stimuli in a single session led to more individualistic preferences for the real-world scenes, likely a consequence of the de-emphasized importance of semantic information.

In a second experiment, a similar method was used to measure agreement across groups of observers making preference judgments for either natural scenes, faces, architecture or artworks. Given the different perceptual features upon which recognition of faces, scenes and buildings rely, it is reasonable to hypothesize that aesthetic appreciation of these stimulus domains may also show salient differences. Agreement for which stimuli were preferred was very high for faces ($r = 0.67$) and relatively high for landscapes ($r = 0.45$), but much lower for architecture ($r = 0.20$) and artwork ($r = 0.13$). This suggests that preferences for faces and landscapes, evolutionarily important stimuli, rely on similar information across people, whereas aesthetic appreciation for architecture and artwork (artifacts of human culture) relies on more individual aesthetic sensibilities.

The relatively strong agreement across observers for the set of natural landscapes permitted a test of the nature of the information most relevant for determining the average preferences. A model that captured landscape topography was better able to predict average preferences than a model that captured features relevant for human exploration (adj. R-squared of 0.18 vs 0.05, respectively; e.g. Kaplan & Kaplan, 1992). A model including several measures of low-level image quality (luminance contrast, mean luminance, range) also captured a significant portion of the variance in mean preference ratings (adj. R-squared 0.14), though less than the topographical measures.

Similarly, preferences for faces also showed high agreement. In addition, we found that agreement across observers was influenced by stimulus gender. Although heterosexual men judged both male and female models as being equally beautiful on average (and more beautiful than non-models), there was significantly higher agreement on which female faces were preferred than on which male faces were preferred ($r = 0.43$ vs 0.21, $p < 10e\text{-}6$).

In contrast, aesthetic preferences for architecture appeared to reflect individual differences in aesthetic style. A comparison between observers' preferences on interior and exterior images of architecture further hinted at the existence of individualized aesthetic styles: if two observers agreed in their preferences for indoor architecture, there was a tendency for them to also show similar preferences for outdoor architecture ($r = 0.39$; $p < 0.001$).

Amongst all of the stimulus domains tested, aesthetic preferences for artwork showed the greatest individual differences. The highly individual nature of aesthetic preferences for artwork presented a unique methodological advantage, in that it allowed for a dissociation to be made, when considered across a group of observers, between the objective features of any one artwork and subjective aesthetic appeal.

In a third experiment (Vessel, Starr, Rubin, 2012), we collected fMRI while observers made aesthetic judgments about artworks on a 4-point scale (4 being "most moving"). A comparison of highly-moving versus low-preference artworks revealed a network of brain regions sensitive to aesthetic appeal. In posterior (occipitotemporal) cortex, increasing preference led to linearly increasing fMRI signal. In contrast, regions of the frontal cortex showed a different, "step"-like pattern: no difference was observed for images that were rated as 1, 2, or 3 – only images rated

as “most moving” (4) showed an increase in signal. This supports the existence of a sensory/semantic preference analysis and an anatomically separate, higher level aesthetic/emotional process that is insensitive to low-level stimulus features.

One of the regions that showed this step-like pattern was the medial prefrontal cortex (MPFC), a region of the default mode network (DMN; Shulman et al., 1997; Gusnard & Raichle, 2001). The DMN, which is normally suppressed during engagement with external stimuli, has been hypothesized to support aspects of self-relevant mentation such as imagining oneself in a future scenario (versus another person; Buckner et al. 2008; Andrews-Hanna et al., 2010). The simultaneous activation of portions of the DMN and sensory cortices, as occurred when observers reported being highly moved by artwork, is rare, and likely represents a unique brain state of resonance between one’s perception of the external world and internally directed thoughts.

Taken together, these experiments suggest that while lower-level visual features play a crucial role in conveying visual information, a perceiver’s overall aesthetic experience is primarily determined by higher-level semantic associations and emotional responses, and in a manner that is linked to self-relevance.

9014-29, Session 6

Identifying image preferences based on demographic attributes

Elena A. Fedorovskaya, Daniel R. Lawrence, Rochester Institute of Technology (United States)

In today’s modern world, people collect and utilize digital images prolifically, whether for personal use or for business use. With the pervasiveness of digital media, special importance is now being placed on personalization, tailoring media material to individual viewers on the basis of descriptive and behavioral information about those viewers and the media content used to attract and hold the viewers’ interest and attention during interaction with the media.

The intent of this study is to determine what sorts of images are preferred by which demographic groups. In other words, the investigation attempts to identify images whose preference ratings are influenced by the demographic attributes of the viewer. To that end, we use the data from an experiment where 19 participants (10 women and 9 men) rated approximately 800 images based on “visual interest” or preferences in viewing images. In addition to gender, participants are classified with respect to their photographic expertise (expert vs. novice), age (25-39; 40-49; 50-older), and education (high school diploma; undergraduate college degree; graduate degree). The images were selected to represent the consumer “photospace” ? typical categories of subject matter found in consumer photo collections. They were annotated using perceptual and semantic descriptors.

In analyzing the image preference data, we emphasize the influence of specific demographic items on the analysis, so that the solutions are essentially confined to subspaces spanned by the emphasized item(s). This particular analysis of ratings (i.e., ordered-choice or Likert) data involves a multivariate procedure known as forced classification, a feature of dual scaling, a discrete analogue of principal components analysis (similar to correspondence analysis). In a conventional multivariate analysis, the investigator has no control over the latent dimensions ? which ones appear or how they appear. In fact, a dimension of specific interest might not appear at all. Ideally, the analysis of these ratings data would enable the investigator to emphasize (or focus on) the effect or influence of a specific item or collection of items ? in this case, demographic items that are “external” to the preference data. Using this technique, we can know definitively which images’ ratings have been influenced by the demographic item(s) of choice. Subsequently, image descriptors are evaluated and linked, on one hand, to the computable image features, and on the other hand, to the preferences associated with viewers’ demographic attributes.

9014-30, Session 6

Chamber QoE: a multi-instrumental approach to explore affective aspects in relation to quality of experience

Katrien De Moor, Norwegian Univ. of Science and Technology (Norway) and Univ. Gent (Belgium); Filippo Mazza, Ecole Centrale de Nantes (France); Isabelle Hupont, Instituto Tecnológico de Aragón (Spain); Miguel Rios Quintero, Technische Univ. Berlin (Germany); Toni Mäki, Martin Varela, VTT Technical Research Ctr. of Finland (Finland)

1. INTRODUCTION

Evaluating (audio)visual quality and Quality of Experience (QoE) from the user’s perspective, has become a key element to optimize users’ experiences and their quality. Traditionally, the focus lies on how multi-level quality features are perceived by a human user. In this context, quantitative psycho-perceptual evaluation methods are commonly used. The interest has however gradually expanded towards human cognitive, affective and behavioral processes that may impact on, be an element of, or be influenced by QoE and which have been under-investigated so far.

In 2012, a new, broadly supported definition of QoE was introduced [1], defining QoE as an emotional state, i.e., ‘the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user’s personality and current state’. It may be influenced by factors at the human, system and context level [1]. There is however a major discrepancy between this holistic conceptualization and traditional, standardized QoE assessment: influencing factors are insufficiently taken into account and QoE is not evaluated in terms of experienced affect.

In this paper, we present results from a follow-up lab study [2] on video quality (N=27), aimed at going beyond the dominant QoE assessment paradigm and at exploring the relation between perceived overall quality and experienced affect. We therefore complement the use of ‘traditional’ QoE self-report measures with ‘alternative’, emotional state- and user engagement-related self-report measures and explore how they relate to each other. In addition, we collected EEG (physiological) data, gaze-tracking data and facial expressions (behavioral) data. In this paper, the focus is primarily on (but not limited to) the self-report data. In the full paper, we also reflect on methodological challenges related to research on QoE and affect.

2. METHODOLOGY

2.1 Test setup and procedure

The test content consisted of three 10-minute movie excerpts with different error profiles and the test procedure was as follows: In the introduction (part 1), the participant was welcomed and given a number of instructions by the test leader. After filling in a short pre-test questionnaire, the EEG headset was fitted on the participant’s head and electrode signals were checked. The questionnaire included the Pick-a-Mood (PAM) scale [3] and questions to collect contextual information about the participant’s TV viewing behavior and basic socio-demographical profile.

Part 2 consisted of four viewing sessions, all preceded by a standard calibration of the participant’s gaze. During the first viewing session, a neutral video of approximately two minutes was shown to the participant. Next, the three 10-minute video clips (the actual test content) were shown one by one and after every test clip, the participant filled a during-test questionnaire: Participants were asked to rate the overall quality (5-point ACR scale) and its acceptability (binary scale). In addition, it contained the following ‘alternative’ self-report measures: the pictorial 9-point Self-Assessment Manikin (Pleasure, Arousal, Dominance) [4], three constructs from the Differential Emotions Scale [5] (Joy, Surprise, Interest; 5-point scales) and 2 adapted constructs from the User Engagement scale [6] (Focused attention, Felt involvement, 5-point scales). The last part contained questions related to expectations, content likeability and

familiarity, distortions (in case perceived: description, link to content, degree of annoyance).

During part 3, the participant filled in a short post-test questionnaire, which contained the PAM-scale [3] again, as well as two questions to explore the possible influence of previous experiences/memories on the experienced affect. The possible influence of the test setup was also briefly evaluated. Next, the EEG headset was removed and every participant received two cinema tickets as compensation. The test duration was around 55-60 minutes and in total, 27 people participated.

As mentioned above, three other types of data were collected. We used the Emotiv EPOC, a consumer-grade wireless neuro-headset with 14 channels to collect EEG data [7] and the Tobii T60 eyetracking system [8] for tracking the participant's gaze during the viewing sessions. Finally, the participant's face was recorded by means of a webcam to enable post-processing for facial expression analysis purposes, using the Noldus FaceReader software (extraction of emotional states) [9].

2.2 Test material and test environment

Based on a multi-staged process, three 10-minute excerpts from an action movie, with similar temporal and spatial complexity were selected. They were encoded in H.264 and AAC in full HD 1080p with an average bitrate of 25mbits/s for video and 384kbits/s audio. Packet loss-based errors were inserted, using a 4-state Markov model [10, 11], with slicing as the resulting error type. Audio and video were multiplexed. The three error profiles were as follows:

- no error

- constant error (average packet loss rate= 0,125%, constant visual slicing error between minute 2 and 9)

- strong intermittent error (average packet loss rate= 0,25%, slicing error every 2 minutes with 1 minute duration)

The errors were randomized across the sequences and the narrative structure was respected when presenting the sequences. Contents where displayed on the 17" Tobii T60 TFT monitor, with a 1280x1024 resolution. The aspect ratio of the sequences was not modified to fit the SXGA resolution of the monitor. The tests took place at the VTT Multimedia Lab, where the user was isolated. Room luminosity was between 21-25 Lux from the back of the screen and sound was reproduced in desktop speakers set to approximately 60 dB.

3. PRELIMINARY RESULTS

Preliminary findings with respect to the self-report data indicate that there are significant differences ($p < .01$) between the three error profiles in terms of QoE when considering the traditional measures. When considering the 'alternative' measures, the data point to a significant difference in terms of pleasure and the degree to which the overall quality meets expectations (both at the $p < .01$ level). In addition, felt involvement is significantly higher for the unimpaired clips versus those with errors ($p = .05$). Comparing the traditional and alternative measures, we found a significant, yet low correlation between the engagement-related attributes and the perceived overall quality. The latter is also positively related to pleasure: higher overall quality scores and higher reported pleasure go hand in hand. However, we found no correlation between overall quality and the self-reported joy and surprise, which are essential in the literature on 'delight' (which is fragmented across different disciplines/fields, e.g., psychology, customer behavior/loyalty, service (quality) and marketing research). Although our findings are still incomplete and preliminary, they support the claim that traditional measures of QoE will need to be reconsidered and extended with measures of delight or frustration.

9014-31, Session Key3

The science of social interactions on the web *(Keynote Presentation)*

Ed H. Chi, Google (United States)

Social interactions have always been an important part of human learning and experience. We now know that social interactions are critical in many knowledge and information processes. Research has shown results ranging from influences on our behavior from social networks [Aral2012] to our understanding of social belonging on health [Walton2011], as well as how conflicts and coordination play out in Wikipedia [Kittur2007]. Interestingly, social scientists have studied social interactions for many years, but it wasn't until very recently that researchers can study these mechanisms through the explosion of services and data available on web-based social systems.

In this talk, I plan to illustrate a model-driven approach to researching social interactions on the Web. Our research methods and systems are informed by models such as information scent, sensemaking, information theory, probabilistic models, and evolutionary dynamic models. These models have been used to understand a wide variety of user behaviors, from individuals interacting with social bookmarks in Delicious to groups of people working on articles in Wikipedia. These models range in complexity from a simple set of assumptions to complex equations describing human and group behaviors. By using this model-driven approach, we further our understanding of how knowledge is fundamentally constructed in a social context, and a path forward for further social interaction research.

9014-32, Session 7

Alone or together: measuring users' viewing experience in different social contexts

Yi Zhu, Technische Univ. Delft (Netherlands); Ingrid Heynderickx, Technische Univ. Eindhoven (Netherlands); Judith A. Redi, Technische Univ. Delft (Netherlands)

Online video services are booming these days, and this market is only destined to grow in the coming years [1]. In this context, an objective model which can estimate the Quality of the user's Visual Experience (QoVE) is badly needed, especially for multimedia service providers wishing to optimize video delivery towards fulfilling user expectations. QoVE is supposed to reflect the level of satisfaction of a user with a video, but has been mainly linked to artifact visibility for a long time [2]. In fact, the objective approaches to measure QoVE have been mostly based on an estimation of the visibility of artifacts generated by video signal impairments at the moment of delivery (e.g., blockiness, blurriness, and transmission errors) and on a prediction of how annoying these artifacts are for the end user [3].

Recently, it has been shown that these approaches have limitations, because user satisfaction doesn't purely depend on artifact visibility. Other factors, e.g., user interest [4] and personality [5] are proved to have a crucial influence on QoVE too. In other words, QoVE is the result of the interaction of a set of factors, not necessarily independent and not limited to the visibility of artifacts [2,6]; defining these factors and quantifying them are therefore necessary prerequisites to QoVE estimation. In [6], three classes of factors are suggested to influence QoVE. Human factors are related to the user's personal characteristics, e.g. age, gender, personality and interest. System factors usually refer to properties that determine the technically produced quality of online video service, such as media configuration, device and genre. Finally, context factors describe all aspects of the user's environment, e.g., social interaction, location, or economic status. Little research has been devoted to factors other than the system ones, and within the context factors, especially social context has been poorly considered so far. It is interesting to look into social context as it is well known that user experience is affected by the interaction the user has with other people [7]. Social context can consist of a user's family, friends or even strangers. It has been shown that users love to watch television as a social activity (e.g., watching World Cup soccer games with friends) [8], co-viewers enjoy each other's company and group viewing can increase a user's overall enjoyment [9]. However, the direct link between social context and eventual QoVE still remains unexplored.

This research aims at understanding how a social context influences a user's QoVE when watching online videos. By social context we specifically indicate here the group of people sharing a viewing experience while being physically co-located (i.e., presence/absence of

co-viewers). In particular, we are interested in understanding whether a user's tolerance to visual artifacts in video (Perceived Visual Quality, PVQ) changes depending on the social context. To answer these questions, we will conduct an experiment comparing two different real-life viewing situations, as shown in Figure 1: one with a single person watching a video alone, and a second one with multiple friends watching a video together. Participants involved in one viewing situation (e.g., alone) will not see videos in the other situation (e.g., in group); as such we follow a between-subjects design. The number of people co-viewing videos will be limited to three at a time. Participants will be seated on a couch in front of a screen (41" LCD), similar to an environment they normally watch videos in. The viewing distance will be 6 times the height of the screen (approximately 3 meters) in order to satisfy the preferred viewing distance [10] and to allow the three viewers located in different positions on the couch to have roughly the same viewing conditions. The rest of the environmental settings will follow the ITU-R BT.500-13 Recommendations and will be the same for all participants.

We will select 2 different video clips from 3 different genres, each of them to be encoded at two different bitrates: high (3500kbps) and low (1000kbps). The reason to choose only two bitrate levels is that we are not interested at this stage in the changes of user QoVE with different bitrate level; rather, we want to focus on the changes in QoVE in different viewing situations, given a certain bitrate level. The genres typically watched alone or in group, will be selected by means of a pilot survey. All the video clips used in the experiment will be at least 5 minutes; to allow participants to engage with the video content. Participants within each viewing situation will be further divided into two sub-groups, each of which will watch the video content at either a high or low bitrate level. After watching each clip, participants will be asked to indicate their PVQ on a 5-point scale (bad to excellent), according to [10]. Furthermore, since there is no standardized approach to measure QoVE, we will measure it based on 4 different aspects, being satisfaction, involvement, enjoyment, and endurability. Finally, we are interested in checking whether the different social contexts can influence the user's level of information assimilation, being also a crucial component of the viewing experience [11].

The experiment will output individual scores and Mean Opinion Scores (MOS) for each video clip, each bitrate level and each viewing situation used in the experiment. First we will check whether there is a significant difference in MOS between the two viewing situations for the same video clip. We predict that the presence of co-viewers will increase the user's tolerance to perceived artifacts and his or her overall satisfaction. Second, for each viewing situation, we are interested in determining whether there is a joint effect of viewing situation and video genre on PVQ. We predict that the effect of social context on PVQ is more pronounced for specific genres (e.g., for those that people prefer to watch with friends). The resulting findings will be reported at the conference and in the final conference paper. They can provide insights in the relationship between social situation and QoVE in multimedia services. As such, they might help identifying genres for which group viewing is more commonly preferred and for which more distortions can be tolerated, so that resources can be saved when streaming them; conversely, videos that are watched typically alone might need more bandwidth and a less severe compression.

9014-33, Session 7

Is there a "like" effect?: considering social biases in image quality evaluation

Filippo Mazza, Ecole Centrale de Nantes (France); Matthieu Perreira Da Silva, Patrick Le Callet, Univ. de Nantes (France)

Multimedia quality evaluation evolved from the mere technical parameters analysis to the consideration of human factors in content perception. Later on, image aesthetic appeal has been introduced to deepen the analysis. Research in image aesthetics showed good results, allowing to build automatic aesthetic assessment systems[1] evaluating low-level features. These features were mainly taken from photography literature, as contrast, composition and colors. Many researches on color as factor on human behaviours have been carried out; a critical review on color psychology is present in Whitfield's review.[2] However, a gap is still present, as low-level computable features alone are not enough to express completely the impact of multimedia on users. Moreover, there is a large variability between people in aesthetic canons.

High level human-oriented semantics are required.[3] These should take into account more complex mechanism in human content perception. Content semantics are not new, and are for example used with content recognition in automatic image tagging.[4] Anyway there is more than the simple perception of what is the content. Content is not only perceived but processed by deeper mechanisms in human brain involving many factors, as memories, personal experiences, belief, and cultural background; we refer to this fact as "content cognition". First steps in this direction have been done: image psychology has been suggested as a potential focus for further research by Fedorovskaya[5] and recently familiarity within image context[6] has been addressed.

In this paper we consider image psychology elements proposed by Fedorovskaya focusing on social factors as particular cognitive biases may be triggered. Cognitive biases are deeply investigated in psychology due to their impact on the decisional process. Their importance is underlined by the fact that they occur at an unconscious level, so that people are even not aware of being influenced. Between cognitive biases, an important category is regarding the social elements impacting our perception; these are called social biases.[7] We think that image evaluation is impacted by this bias by different social clues. In this work we explore social clues conveyed aside the image in evaluations gathered online, as in some cases previous quotes are shown with the image itself; this is the case for example of Flickr and Facebook, while for others as Photo.net - used also by Datta[3] - it is not.

We expect that image likeability will be influenced by others opinions, especially in those cases when evaluation is not on scales extremes (i.e. very low or high quality). Influence of social interactions have been demonstrated to create a bias in some user's ratings about products.[7] We expect this behaviour also for image ratings.

Researching social biases impact for multimedia quality evaluation is interesting from multiple perspectives.

First it can give us clues regarding the best practices designing subjective evaluations. Secondly, if analyticallyevaluated, it can be considered for building improved quality evaluation models. This is especially true for evaluations done remotely via internet: nowadays social networks are a huge resource and their social engagement features - e.g. "Likes", "Share" or friends suggested elements - are always present. How this factor can be exploited and/or corrected has not been dealt at the moment in our knowledge. Furthermore from a practical point of view it can open the path to further research in product advertising as it can influence social impact.

Further works will focus on social biases conveyed inside images, that is to say conveyed by the content itself.

This is especially true when people are present in the image, as more elements can be understood as the situation depicted. A "follow the crowd" mechanism can arise, pushed by social biases, accepting as correct informations provided by others. Social relationship features in an image have been investigated through image processing techniques to improve aesthetic evaluation of images.[8] Also other studies on psychology of attraction showed a 'social proof' behaviour;[9] this shows again how content cognition is influenced by social clues.

REFERENCES

[1] Datta, R. and Wang, J., "ACQUINE: aesthetic quality inference engine-real-time automatic rating of photo aesthetics," Proceedings of the international conference on . . . , 1–4 (2010).

[2] Whitfield, T. W. and Wiltshire, T. J., "Color psychology: a critical review.," Genetic, social, and general psychology monographs 116, 385–411 (Nov. 1990).

[3] Datta, R., Li, J., and Wang, J., "Algorithmic inferencing of aesthetics and emotion in natural images: An exposition," Image Processing, 2008. ICIP 2008. . . . , 8–11 (2008).

[4] Fu, H., Semantic Image Understanding : From Pixel to Word, PhD thesis (2012).

[5] Fedorovskaya, E. a. and De Ridder, H., "Subjective matters: from image quality to image psychology," 8651, 86510O–86510O–11 (Mar. 2013).

[6] Chu, S. L., Fedorovskaya, E., Quek, F., and Snyder, J., "The effect of familiarity on perceived interestingness of images," 8651, 86511C–86511C–12 (Mar. 2013).

[7] Wang, C. A., Zhang, X. M., and Hann, I.-h., "Social Bias in Online Product Ratings : A Quasi-Experimental Analysis," WISE 2010, St. Louis, MO (February 2010), 1–35 (2010).

[8] Li, C. and Gallagher, A., "Aesthetic quality assessment of consumer photos with faces," Proceedings of IEEE . . . , 3–6 (2010).

[9] Jones, B. C., DeBruine, L. M., Little, A. C., Burriss, R. P., and Feinberg, D. R., "Social transmission of face preferences among humans.," Proceedings. Biological sciences / The Royal Society 274, 899–903 (Mar. 2007).

9014-34, Session 7

Assessing the impact of image manipulation on users' perceptions of deception

Valentina Conotter, Duc Tien Dang Nguyen, Giulia Boato, Maria Menendez, Univ. degli Studi di Trento (Italy); Martha A. Larson, Technische Univ. Delft (Netherlands)

1. CHALLENGE

Generally, we expect images to be an honest reflection of reality. However, this assumption is undermined by the new sophisticated and wide-spread image editing technology, which allows for easy manipulation and distortion of digital contents. Modified data may influence people's opinions, for example, altering their attitudes to past events depicted in images. As a consequence, the image editing debate is growing in importance, and it has a significant impact on how we communicate information both in the public and the private sphere.

Our work is an initial investigation analyzing how different types of image manipulation impact users' percep- tions. We are motivated by the observation that expectations of images truthfulness are in some sense related to context. The extreme cases are clear: expectations of truthfulness of images are radically different if the image is hanging in an art gallery or being used as evidence in a court case. What is not well understood, however, are the specific characteristics of images and contexts that influence users perceptions of deception and how strong and how consistence these perceptions are. This paper reports the results of a study that provide evidence that different types of manipulation and different contexts have different influences on users perceptions of the decep- tiveness of modified images. This evidence is collected with a study that applies crowdsourcing methodologies to elicit the opinion of a large number of people about different types of manipulations, different image contents, and different reasons for motivations, and different contexts of use.

To the best of our knowledge, this work represents a first approach to gain more insights on the relation between the level of modification in images and its perception on users. Currently very little related work is present in the literature, mainly involving small scale user studies focused on the impact of a specific editing techniques.

2. PROPOSED APPROACH

As a first step, we create a dataset of 275 pairs of original-manipulated images. The applied manipulations reflect possible forgeries to which we may be exposed in our daily-life, ranging from common and popular image processing operators to more technically complex and time-consuming image editing. The images in the dataset fall into four categories based on the level of the applied manipulation: (a) Similar: images are visually similar, with differences only in resolution or format, (b) Enhanced: images are slightly modified on global low- level attributes (e.g., contrast enhancement, light balancing, etc.), (c) Retouched: images are not only different on global characteristics but also altered locally, (d) Doctored: images are manipulated via advance editing techniques (e.g., splicing) changing the semantics of the original picture and/or creating something different from the original data. The groups for modifications track the decreasing similarity of the manipulated image with respect to the original image, measured in terms of the Structural Similarity (SSIM) index. This full- reference metric ranges from 1 for very similar contents to 0 for very different images. As reported in Figure 1(a), we can see, for example, that images falling in the category "similar" present, on average, a higher value of SSIM when compared to the other categories.

Next, we aim at studying how people perceive the manipulations in the created images. In order to do that and collect users' opinions about each pair of images in the dataset, we carefully design a Human Intelligent Task (HIT) and publish it on Amazon Mechanical Turk (mturk.com). 553 workers participated in this crowdsourcing user study, performing 2503 accepted annotations (quality controls have been run to filter out unreliable answers).

3. MAIN RESULTS

3.1 Perceived difference between original and manipulated images

First, people are asked to rate the perceived level of difference between the pair of original-manipulated images in a 7-point Likert scale (1 indicates that images are perceived as very similar while 7 stands for very different images). As a first result, we deduce that the perceived difference tracks the increasing level of manipulation (Figure 1(a)). Moreover, in Figure 1(b) users' evaluations are plotted (blue crosses) together with the average SSIM of the images falling within the same ranking (red line). This confirms that the perceived difference tracks also the objective quality of the considered images, measured with a traditional quality metric.

3.2 Perceived deception of the manipulation

Next, users are asked to evaluate how honest they think would be to use the manipulated version of image, in order to get insights about the perceived deception. Answers goes from 1 indicating that "The image does not distort reality", increasing up to 5 where "The image can be considered to directly manipulate peoples feelings or opinions". Moreover, users were asked to rank the difference in terms of perceived deception between the original and manipulated image, using a 7-point Likert scale. We verify that both the perceived level of deception and the perceived deception difference between un-modified and modified images track the level of manipulation (Figure 1(a)).

Since we are interested in the influence of the context on the use of the manipulated images, we ask the workers to select a role from a list (e.g., photographer, blogger, journalist, etc.) and play that role while answering the questions about image deception. The main reason to ask the workers to play a role is to encourage them to answer freely concerning possible non-honest use of an image, disregarding hesitation that they might feel if we were to ask their opinion directly. The results of the study shows that the choice of the photographer role has significant correlation with the perception of an image manipulation as being honest. Our observations provide significant insight onto the extent to which people understand the motivation that drives an editor to change images (the motivations were clearly different from one editor "role" to another.)

4. OUTLOOK

By using different analysis metrics and analyzing more answers obtained from the conducted crowdsourcing user study, in the final paper we will deeply analyze this study, also assessing the influence on user perception of deception played by the main motivations that lead to manipulate the content.

9014-35, Session 8

Spectral compression: weighted principal component analysis versus weighted least squares

Farnaz Agahian, Brian V. Funt, Simon Fraser Univ. (Canada); Seyed Hossein Amirshahi, Amirkabir Univ. of Technology (Iran, Islamic Republic of)

Principal Component Analysis (PCA) is one of the most widely used techniques in compression of large spectral images and guarantees the best possible representation of the high-dimensional spectra in the low-dimensional eigenvector sub-space. This compression method gives an equal treatment to all wavelengths throughout the spectrum and tries to minimize the squared reconstruction errors between the actual and reconstructed spectra. On the other hand, in many applications such as digital image archives, electronic commerce, art conservation science and telemedicine, it is required to somehow compress the spectral data to preserve as much color information as possible. Indeed, the compression technique should be modified so as to consider individual wavelengths differently depending on their relative importance to the human visual system.

In this paper, the idea of weighted spectral compression is examined by making a comparison between the two weighting strategies, i.e., Weighted Least Squares (wLS) versus Weighted Principal Component Analysis (wPCA). However, the common goal of both methods is minimizing weighted errors, in former a series of weighted linear equations is solved to calculate the principal component coefficients but in the later this goal is achieved by applying an appropriate weighting function on spectral data before forming the correlation matrix and extracting the principal eigenvectors. In the other words, in the case of wLS, the spectral sub-space is formed by non-weighted ordinary eigenvectors calculated from the non-weighted spectral dataset and weighting is performed in the process of fitting a linear model to the spectral data.

In addition, a comparison is made among seven different weighting functions incorporated into ordinary PCA/LS to give selectively more importance to the wavelengths that correspond to higher sensitivity in the human visual system. Weighted compression is performed on reflectance factors of 3219 colored samples (including Munsell and NCS data) and spectral and colorimetric errors are calculated in terms of CIEDE2000 and root mean square errors.

The results show that except when the compression involves a large number (>10) of basis vectors, non-weighted (NW) compression always leads to higher colorimetric errors, although its spectral accuracy is better. For spectral compression involving more than three basis vectors, wLS is more successful than wPCA in reproduction of the color information under both A and D65 illuminants. Comparison of the different weighting functions indicates that, incorporating weights based on the square root of the principal diagonal of matrix R reduces the colorimetric errors more than the other weighting functions and lead to the best reproduction of color information under both illuminants particularly when using a low number of basis vectors.

It should be mentioned that matrix R is an orthogonal projector developed by Cohen and Kappauf for decomposing the color stimulus into its fundamental and metameric black. This matrix is calculated from matrix A defined as the product of a set of color matching functions (in this study, CIE 2°-1931) and a given illuminant:

R=A?(A^’ A)?^(-1) A’ (4)

If matrix A is defined such that its columns are formed using just three color matching functions, then the diagonal of R will represent the squared magnitude of spectral lights in color space.

As matrix R is a function of the illuminant, in this study we make three different matrices corresponding to three different illuminants and subsequently find three different sets of weighting functions for each combination of illuminant and observer. To do so, in addition to two common illuminants A and D65, a synthetic illuminant that is a linear combination of D65 and A is also employed. In this way, the properties of both illuminants were incorporated in a single synthetic one. The effect of this contribution is evident in both colorimetric as well as spectral errors. This method can be extended to more than two lights. In addition, in some circumstances where different lights have different importance, the contribution of various illuminants can be simply controlled by replacing the ordinary mean by a weighted one

9014-37, Session 8

Creating experimental color harmony map

Christel Chamaret, Fabrice Urban, Josselin Lepinel, Technicolor S.A. (France)

Harmony is naturally surrounding human and directly related to naturalness. Man-made creations are potentially a source of disharmony since the practical aspect is usually the purpose. Mainly, this paper focuses on color aspect of harmony even if other image characteristics seem to be involved in harmony feeling such as edge contrast or lightness [1]. Several definitions and precisions about the color harmony topic appeared the last decades [2]. Most of them agreed on the relationship of pleasantness and harmony, concluding that good aesthetic pictures are harmonious. However, Schloss and Palmer [1][4] recently demonstrated the need for differentiation between preference and harmony through aesthetic assessment of color pair combinations, leading to a clarification of previous experiments and conclusions [6][7].

This paper focuses on the perception of color harmony. Rather than studying aesthetics, we propose a framework for assessing color harmony in pictures. Can harmony assessment be enlarged to more than two or three color combinations? Is color palette representative enough or does spatial composition of color matter [17]? Can we design a ground truth for assessing computational models of color harmony in pictures [9]?

Experimental studies have been extensively conducted to assess color harmony by presenting color combinations of doublets and triplets by means of simple patterns [6][8]. In addition, users have defined their favorite color palette through an online tool [10] and O’Donovan et al [11] have learnt preferred and potentially most aesthetic color combinations (5 colors) from large datasets. Nevertheless, there is a lack for assessment of more complex color stimuli and the impact of spatial composition.

This paper addresses the question of perception of color harmony in pictures. Main contributions involve the selection of pictures with different color distributions and two subjective experiments:

- a search-task experimental protocol for eye-tracking
- a pair-wise comparison of still pictures

These two complementary experiments are proposed to measure the degree of color harmony present both locally in a picture (with the experimental fixations map) and globally, relatively within the complete dataset (with the pair-wise annotation). The same 30 color pictures have been used for the two experiments and extracted from flickr© by means of a multicolor search engine [14]. Two or three main hue components have been defined for the request and successively turned along the hue wheel to vary the color distribution of requested images. Finally, the dataset has been refined by keeping only images with color distributions that best fit different harmony templates defined by Matsuda [1]. Consequently, the pictures have specific color distributions that handle complementary, analogous and orthogonal properties of color harmony [2]. Semantic content that could potentially bias the observers assessment have been minimized by preferring the selection of abstract pictures.

The eye-tracking experiment has been conducted with an Eye-link 1000 Hz with around twenty observers. The protocol was composed of two passes. The first pass in free viewing was intended to get the viewer familiar with the content. In the second pass, the user had “to search for the most disharmonious areas in the picture”. In order to keep observers concentrated on the task, we asked them after each harmony search task to point out with the gaze which color was not harmonious on a hue wheel picture. Both passes presented grey images with a cross randomly located on the screen between each stimulus for two seconds and the viewing time was five seconds per image.

The pairwise comparison also involved around twenty observers that have to designate “the most harmonious picture” when watching a pair for five seconds on a monitor. The methodology exposed in [12] has been employed in the context of color harmony task in order to increase the number of different stimuli while guaranteeing a reliable ranking with the Bradley-Terry model [13].

The resulting scale is so far consistent with the perception of color harmony. The eye-tracking experiment provides fixations map that locally highlight color disharmony when watching globally the picture. This dataset is of high interest for any color image processing application that would locally retouch colors [16] and especially algorithms that aim at harmonizing pictures [15].

[1] E. Fedorovskaya, C. Neustaedter, and W. Hao. Image harmony for consumer images. In Image Processing, 15th IEEE International Conference on, pages 121–124, 2008.

[2] Westland S., Laycock K., Cheung V., Henry P.M. & Mahyar F., 2007. Colour Harmony, Colour: Design & Creativity, 1 (1), 1, 1-15.

[3] Schloss, K. B. & Palmer, S. E. (2011). Aesthetic response to color combinations: Preference, harmony, and similarity. Attention, Perception and Psychophysics, 73, 551-571.

[4] Schloss, K. B., & Palmer, S. E. (2010). Aesthetics of color combinations. Proceedings of the SPIE, 2010.

[5] Y. Matsuda. Color design. Asakura Shoten, 1995.

[6] Ou, L.-C. and Luo, M. R. (2006), A colour harmony model for two-colour combinations. Color Res. Appl., 31: 191–204.

[7] Ou, L.-C., Luo, M. R., Woodcock, A. and Wright, A. (2004), A study of colour emotion and colour preference. Part I: Colour emotions for single colours. Color Res. Appl., 29: 232–240.

[8] X. Guan, PhD thesis, Hong Kong Polytechnic University (2007).

[9] W.G. Kuo, Y.C. Wei and S.M. Lin, Investigation on various color-harmony models in predicting color harmony for color-apparel images, AIC 2013.

[10] Adobe Kuler. http://kuler.adobe.com/

[11] Peter O'Donovan, Aseem Agarwala, Aaron Hertzmann. Color Compatibility From Large Datasets. ACM Transactions on Graphics, 2011.

[12] Jing Li ; Marcus Barkowsky ; Patrick Le Callet; Boosting paired comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs. Proc. SPIE 8648, 2013.

[13] Bradley, R. and Terry, M., Rank analysis of incomplete block designs: I. the method of paired comparisons,"Biometrika 39(3/4), (1952).

[14] http://labs.tineye.com/multicolr

[15] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. 2006. Color harmonization. In ACM SIGGRAPH 2006 Papers (SIGGRAPH '06).

[16] Erik Reinhard, Michael Ashikhmin, Bruce Gooch and Peter Shirley, Color Transfer between Images, IEEE CG&A special issue on Applied Perception, Vol 21, No 5, pp 34-41, September - October 2001.

[17] C. Chamaret, F. Urban, No-reference harmony-guided quality assessment, CVPR 2013 Workshop Visual Analysis and Beyond.

9014-38, Session 8

Exploring the use of memory colors for image enhancement

Su Xue, Minghui Tan, Yale Univ. (United States); Ann McNamara, Texas A&M Univ. (United States); Julie Dorsey, Holly E. Rushmeier, Yale Univ. (United States)

Introduction

Memory colors refer to those colors recalled in association with familiar objects. The deficiency with existing research in this area is that a) screen memory colors have not been rigorously established and b) existing studies do not include extensive human judgments when evaluating image edits based on memory colors. We first perform a context-free perceptual experiment to establish the overall distributions of screen memory colors for three pervasive objects (skin, sky and grass). Then, we use a context-based experiment to locate the most representative memory colors. Finally, we show a simple, yet effective, application using representative memory colors to enhance the color reproduction of digital images.

Establishing Memory Colors

First, as a context-free perceptual experiment, we repeated the memory color experiment conducted by Bartleson et al. in 1960 using digitized color chips based on the Lu'v' color gamut over a uniform neutral gray background, participants were crowd sourced. Lu'v' values were converted to sRGB for display. No attempt to detect or adjust for the type, size or settings of the monitors of the participants was made. This was deliberate to enable discovery of the most general representative memory colors. Figure 2 (left) in the supplemental document shows the results. We then use three 2D Gaussian functions to approximate these elliptically shaped distributions, see Figure 2 (right) in the supplemental document. Due to the nature of crowd sourcing, a wide range of participants, cultures, monitors, viewing conditions, and etc., are sampled. Despite all the potential variations in the stimuli, we find that the results are very close to those from existing psychophysical experiments, which were executed under controlled settings. This is a nice result as it validates the usage of crowd sourcing techniques for color evaluation.

Second, we perform a context-based experiment to refine the memory colors identified in the context-free study. We take natural images that include regions of skin, sky and grass, and generate a set of stimuli by shifting the colors of individual regions toward a few candidate colors (based on the memory colors established in the context-free experiment). Then, we ask viewers to rate these manipulated images, which contain different combinations of candidate colors. The ratings reveal humans' preferences for the joint distribution of three memory colors in the context of natural images. In this manner, we thereby locate the representative memory colors based on these preferences. As before, we use crowd sourcing to elicit the judgments on stimuli, while keeping results as general as possible. Participants rate images based only on the quality of the color reproduction, ignoring all other image elements such as content, composition etc. The generally high positive correlations demonstrate that, despite different image contexts, the preferences to memory colors are rather consistent.

Color Enhancement

Given a raw photo as input, people often seek a simple solution to making the component colors more pleasing. Color can be manipulated in many dimensions; the domain of possibilities is overwhelming, making this a challenging task for non-expert users. Notably, the representative memory colors, which we located using human preference via context-free and context-based experiments, serve promisingly as candidate "standards" for color reproduction. For images containing regions of memory objects, arbitrary color correction is converted to 1D color shifting: moving colors of memory objects toward (or away from) a target, that is, the representative memory color. An adaptive scaling of colors is followed in order to enhance the color reproduction. We found, through a third experiment, that several scenarios exist, in which this simple, yet effective, color manipulation technique can significantly enhance the perceived color reproduction of images. See the supplemental document for details and more examples.

9014-39, Session 8

Perceptual evaluation of colorized nighttime imagery

Alexander Toet, Michael J. de Jong, Maarten A. Hogervorst, TNO Defence, Security and Safety (Netherlands); Ignace T Hooge, Utrecht Univ. (Netherlands)

We recently presented a look-up-table based color transform that produces fused multiband nighttime imagery with a realistic (intuitive) and constant (stable) color appearance. To assess the practical value of this color fusion transform we performed two experiments in which we compared human visual scene recognition for grayscale nighttime imagery, color transformed nighttime imagery and standard daytime color photographs. The stimulus set represented 30 different scenes. Each

scene was registered both in grayscale at night (in darkness) using both a digital image intensifier and a thermal camera, and in daytime using a standard digital color camera. Colorized nighttime imagery was produced by applying our newly developed color transform to the set of grayscale intensified images. In the first experiment we investigated the amount of object- and scene-level information observers can perceive in a short time span (the gist of the scene). Participants watched brief presentations of the stimuli on a computer monitor and were then asked to provide a full report of what they had seen, using a free-recall method to collect responses. Subjects were given an unlimited amount of time to write down their responses, and were asked to reply as accurately as possible. The results were quantified in a precision-recall framework. Precision is scored as the fraction of detections that are true positives, while recall is the fraction of true positives that are detected. Evidently, the best image representation provides the largest number of correctly perceived details and the smallest amount of false alarms. The ground truth was a list of scene elements (key elements) for each scene that was constructed by three experts. Our results show that standard daytime color photographs and color transformed nighttime images yielded the highest precision and recall measures, while both grayscale intensified and thermal images yielded significantly lower values. There was no significant difference between the performance with standard daytime color photographs and color transformed nighttime images. In the second experiment, we measured the eye fixations of participants who were instructed to freely explore the images. Fixation duration was longer and fixation rate (fixations/s) was higher for both grayscale intensified and thermal images than for daytime color photographs and color transformed nighttime images. This indicates that observers have more difficulty extracting information from grayscale nighttime imagery than from either colorized nighttime imagery or daytime photographs. Summarizing, we conclude that colorized nighttime imagery produced with our newly developed color image fusion transform yields enhanced gist perception with reduced cognitive effort. This suggests that application of our color transform to conventional monochrome nighttime imagery may help to improve scene understanding and recognition, reaction time, and object identification.

9014-40, Session 9

Reaching into pictorial spaces

Robert Volcic, Istituto Italiano di Tecnologia (Italy); Dhanraj Vishwanath, Univ. of St. Andrews (United Kingdom); Fulvio Domini, Istituto Italiano di Tecnologia (Italy) and Brown Univ. (United States)

While binocular viewing of 2D pictures generates an impression of 3D objects and space, viewing a picture monocularly through an aperture produces a more compelling impression of depth and the feeling that the objects are “out there”, almost touchable. Here, we asked observers to actually reach into pictorial space under both binocular- and monocular-aperture viewing. Images of natural scenes were presented at different physical distances via a mirror-system and their retinal size was kept constant. Targets that observers had to reach for in physical space were marked on the image plane, but at different pictorial depths. We measured the 3D position of the index finger at the end of each reach-to-point movement.

Observers found the task intuitive. Reaching responses varied as a function of both pictorial depth and physical distance. Under binocular viewing, responses were mainly modulated by the different physical distances. Instead, under monocular viewing, responses were modulated by the different pictorial depths. Importantly, individual variations over time were minor, that is, observers conformed to a consistent pictorial space. Monocular viewing of 2D pictures thus produces a compelling experience of an immersive space and tangible solid objects that can be easily explored through motor actions.

9014-41, Session 9

A framework for the study of vision in active observers

Carlo Nicolini, Istituto Italiano di Tecnologia (Italy); Carlo Fantoni, Istituto Italiano di Tecnologia (Italy) and Univ. degli Studi di Trieste (Italy); Giovanni Mancuso, Robert Volcic, Istituto Italiano di Tecnologia (Italy); Fulvio Domini, Istituto Italiano di Tecnologia (Italy) and Brown Univ. (United States)

We present a framework for the study of active vision, i.e., the functioning of the visual system during actively self-generated body movements. In laboratory settings, human vision is usually studied with a static observer looking at static or, at best, dynamic stimuli. In the real world, however, humans constantly move within dynamic environments. The resulting visual inputs are thus an intertwined mixture of self- and externally-generated movements. To fill this gap, we developed a virtual environment integrated with a head-tracking system in which the influence of self- and externally-generated movements can be manipulated independently. As a proof of principle, we studied perceptual stationarity of the visual world during lateral translation or rotation of the head. The movement of the visual stimulus was thus parametrically tethered to self-generated movements. We found that estimates of object stationarity were less biased and more precise during head rotation than translation. In both cases the visual stimulus had to partially follow the head movement to be perceived as immobile. We discuss a range of possibilities for our setup among which the study of shape perception in active and passive conditions, in which the same optic flow is replayed to stationary observers.

9014-42, Session 9

Shadows in pictorial space

Maarten W. A. Wijntjes, Huib de Ridder, Technische Univ. Delft (Netherlands)

No Abstract Available

9014-43, Session 9

3D space perception as embodied cognition in the history of art images

Christopher W. Tyler, The Smith-Kettlewell Eye Research Institute (United States)

Depth perception and the evocation of a sense of space has a long provenance in the history of images, especially in the context of the perspective representation of distance through size recession and the solidity of protruding objects through shading, shadows and highlights. A goal of these depth representation techniques is not simply accurate 3D representation but also to give the viewer a sense of direct involvement with the objects depicted in the scene. This kind of involvement has recently been recognized as a form of mental function termed “embodied cognition”, or experiencing a depicted scene as though one is a physical participant rather than just intellectually understanding the nature of the scene being depicted.

Embodied cognition is usually framed in terms of action observation, in the sense that observing the actions of others is experienced similarly to performing the action oneself, and numerous studies have shown corresponding similarities in the patterns of brain activation in the two situations. Here I will focus on the spatial experience of embodied cognition and its relation to the ecological optics of J.J. Gibson. This view introduces the concept of a three-dimensional ‘ground’ as the stage for the ‘figure’ of the action(s) depicted in the image. Figure/ground

categorization is typically considered as a two-dimensional categorization problem, in which one region of an image is identified as the figure of current interest and the remainder of the image territory is categorized as ground. The categorization is well-known to be a transient, attention-driven process that can rapidly fluctuate among different figural elements and their components, as in the classic Rubin face/vase alternating image.

It is also well-known that there is a surface continuity property to the ground assignment, which is perceived to continue behind the figure in a separate 2D layer (known as 'amodal completion'). Thus, the classical figure/ground conceptualization has a proto-3D aspect to it in terms of the relative depths of the assignation of figure as foreground and the ground as background (although this can be reversed in the case of depictions of holes). In this assignation, little attention is usually paid to the structure of the perceived background beyond this aspect of the continuity of the amodal completion behind the figure. Consideration of the depiction of pictorial space in the history of art images, however, reveals that the figure/ground relationships in elaborated scenes incorporate a full 3D component in the figure/ground assignment. The image region designated as figure is perceived as a solid object with depth structure that is completed by the amodal completion process. That is, the invisible back side of a protruding object is understood to have a 3D structure dictated by the configuration of the visible front side, a form of 3D amodal completion. Similarly, the continuity of the background behind this 3D figure is understood to be a continuation of the depth structure of the visible regions of the ground. Rather than being simply the uniform plane of the typical figure/ground demonstrations, the backgrounds in art images form rooms, courtyards, landscapes, and so on, with elaborated 3D structure. In this way, the figure/ground assignment is applied to provide a full sense of the 3D configuration of the pictorial space.

The important aspect of this conceptualization is that the 3D spatial structure of the ground is maintained in the form of the pictorial space even though the ground is de-emphasized by the figure/ground assignment process. There is thus a tension between the background region considered as 'negative space', or non-processed territory, and the background region considered as the extended 'lebensraum' that forms the pictorial space occupied by the 3D figure(s). Although it is well-understood that the boundary between figure and ground is perceptually 'owned' by the figure and dissociated from the ground, it can now be seen that this does not imply that the ground is free of properties. Under the new interpretation, the term 'negative space' is a serious misnomer: the ground is negative in terms of figure but positive in terms of space, and should in fact be termed 'negative figure'. Indeed, space is itself a negative concept in 3D terms, receding rather than advancing, concave rather than convex, but, qua space, it opens up a wide range of possibilities for figure placement, even when totally empty. Whereas the figure crystallizes a particular action, space represents the limitless possibilities for action. In the elaborated usage through the history of art, pictorial space may contain an array of structures, from valleys to colonnades, that provide a more structured concept of the available space, without reaching the level of figural objects. This elaborated, even theatrical, concept of space may appropriately be termed 'action space'.

The appreciation of an action space is thus quite different from an empty canvas, or the uniform background of the figure/ground demonstrations. In order to appreciate it as such, the viewer has to mentally project their body into to the depicted space and imagine the possibilities of action within it. For the successful evocation of such action space, a painting recruits the perceptual processes of spatial appreciation through the full array of Gibsonian depth cues, in a form that is generally understood as providing an allocentric sense of the 3D space within the image. However, the level beyond the perceptual appreciation of 3D space as such is the sense of being drawn into the pictorial space to participate as a player in the depicted scene. This is a form of embodied cognition different from the identification with an observed action; rather than identifying one's body with the body depicted in the image, it is more a question of identifying the action space surrounding one's body with the space depicted in the image. The experience of the depicted space then becomes one of the egocentric space inhabited by one's own body, with its limitless array of action possibilities. This may be considered a spatial form of 'situated cognition' in which the cognition, instead of being situated in the context of the real objects being cogitated, is projected into the artificial pictorial space of the image. The historical development of this form of involvement with pictorial space will be analyzed.

9014-44, Session 9

Learning to draw and how it changes the brain

Lora T. Likova, The Smith-Kettlewell Eye Research Institute (United States)

A week of training in unique memory-training paradigm based on non-visual drawing that I have recently developed led to dramatic improvements in memory-guided spatiomotor performance in both congenitally blind and sighted subjects (Likova 2012, 2013). In a total absence of simultaneous sensory input from the images (already) memorized through tactile exploration, massive activation was found in the primary sensory cortex for vision (area V1) that was topographically organized even in the congenitally blind subjects who lacked any visual experience. This led us to propose a re-conceptualization of the classical key module for 'visuo-spatial' working memory as being not visual but independent of sensory-modality, i.e., 'amodal-spatial' in nature, though implemented in V1. Moreover, the behavioral improvements successfully transferred to a range of untrained tasks.

These remarkable results lead to the question: Does learning to draw affect other memory and object processing areas, providing a way to both change the learning brain, and provide further insights into brain organization? In particular, recent studies are questioning the classic model of memory that attributes only a restricted initial role to the hippocampus. Our novel memory paradigm has the potential to dissociate perception/memory processes that are difficult to disentangle in visual stimulation paradigms. Our paradigm engages the full 'perception-cognition-action' cycle through a typical 'visual' task (drawing) performed non-visually.

Conference 9015: Color Imaging XIX: Displaying, Processing, Hardcopy, and Applications

Monday - Wednesday 3 –5 February 2014

Part of Proceedings of SPIE Vol. 9015 Color Imaging XIX: Displaying, Processing, Hardcopy, and Applications

9015-1, Session 1

Improving information perception from digital images for users with dichromatic color vision

Omid Shayeghpour, Daniel Nyström, Sasan Gooran, Linköping Univ. (Sweden)

Color vision deficiency (CVD) is the inability or limited ability to recognize colors and discriminate between them. A person with this condition perceives a narrower range of colors compared to a person with normal color vision. A growing number of researchers are striving to improve the quality of life for CVD patients. Finding cure, making rectification equipment, providing simulation tools and applying color transformation methods are among the efforts being made by researchers in this field.

In this study we concentrate on recoloring digital images in such a way that users with CVD, especially dichromats, perceive more details from the recolored images compared to the original image. The main focus is to give the CVD user a chance to find information within the picture which they could not perceive before. However, this transformed image might look strange or unnatural to users with normal color vision.

During this color transformation process, the goal is to keep the overall contrast of the image constant, while adjusting the colors that might cause confusion for the CVD user. First, each pixel in the RGB-image is converted to HSV color space in order to be able to control hue, saturation and intensity for each pixel. Next step is to find colors that are safe and perceived perfectly normal by dichromats and those that could be problematic.

The method for recognizing the safe and problematic hue ranges was inspired by a condition called “unilateral dichromacy” in which the patient has normal color vision in one eye and dichromacy in another. A special grid-like color card is designed, having constant saturation and intensity over the entire image, while the hue smoothly changes from one block to another to cover the entire hue range. The next step is to simulate the way this color card is perceived by a dichromatic user and finally to find the colors that are perceived identically from the two images and the ones that differ too much. This part makes our method highly customizable and we can apply it to other types of CVD, and even personalize it for the color vision of a specific observer.

The resulting problematic colors need to be dealt with by shifting the hue or saturation. Our next step is to write a set of rules to perform a color transformation that keeps the basic appearance of the original image while adjusting the troublesome colors. For instance, colors with very low saturation (neutral grays) or those with very low intensity (black) should remain unchanged.

The results for the method have been evaluated both objectively and subjectively. First, we simulated a set of images as they would be perceived by a dichromat and compared them with the simulated view of our transformed images. The results clearly show that our recolored images can eliminate a lot of confusion for the user and convey more details. Moreover, an online questionnaire was created and over 40 users with CVD confirmed that the transformed images allow them to perceive more information compared to the original images.

9015-2, Session 1

Spectral analysis of omnidirectional color signals in natural scenes

Shoji Tominaga, Daiki Watanabe, Keita Hirai, Takahiko Horiuchi, Chiba Univ. (Japan)

The light source forming color images in a natural scene is not a single light source, but a mixture of a point light source such as sunlight and of an area light source with a spatially changing spectral distribution. Moreover, note that an object surface in a natural scene is illuminated not only by such a light source but also by all the reflected lights from the surrounding object surfaces. Spectral analysis based on omnidirectional observations in a natural scene is a relatively new topic.

In previous paper, the authors proposed a method for analyzing omnidirectional color signals in a natural scene, which contained different illuminations of daylights and indirect illuminants of the reflected lights from different object surfaces. We used a multiband omnidirectional imaging system.

In this paper, we expand the previous method to analysis of the omnidirectional color signals in a variety of natural scenes with seasonal and temporal changes. A new version of the multiband imaging system with six spectral channels is used for capturing high-resolution high-dynamic range images in the omnidirectional observations at a particular point in a natural scene. The spectral-power distributions of color signals are recovered using the Wiener estimator from the captured six-band images. The distributions are represented by 61-dimensional vectors, which have spectral features at 61 points with an equal interval of 5nm in the visible wavelength range [400, 700nm].

We investigate the spectral compositions of the omnidirectional illumination based on the principal component analysis of the whole set of color signals acquired at the same location in different seasons and different times of day.

The collections of omnidirectional images were conducted at three locations A, B, and C on campus in Chiba University during 2010-2012. The locations A, B, and C are Katarai no Mori forest, Medicinal plant garden, and the roof of a building, respectively. It is found that all the omnidirectional color signals collected in different seasons and different times of day at each location can be expressed in a linear combination of only three principal components. This property has the potential for high data compression. Moreover, we compare the spectral features of the principal components among three data sets of color signals observed at three different locations. Then it is found that the first three principal components have similar spectral features in each other, regardless of different scenes in different locations.

Therefore, the averages of the first three principal components are considered as the invariant bases of the color signals in natural scenes. The quantities are expected to be independents of such factors as time, seasons, and location.

The reliability of the proposed analysis method is examined from various points of view. As an application, it is demonstrated that the invariant bases of color signals are useful for image rendering in a natural environment.

9015-3, Session 1

Realistic fetus skin color processing for ultrasound volume rendering

Yun-Tae Kim, Samsung Advanced Institute of Technology (Korea, Republic of); Kyuhong Kim, SAMSUNG Electronics Co., Ltd. (Korea, Republic of); Sung-Chan Park, Samsung Advanced Institute of Technology (Korea, Republic of); Jooyoung Kang, Jung-Ho Kim, SAMSUNG Electronics Co., Ltd. (Korea, Republic of)

1. Background, Motivation, and Objective

The evolution of ultrasound imaging from 2D to 3D imaging has enhanced diagnostic accuracy. In particular, 3D ultrasound screening

for the detection of congenital anomalies in a fetus has been widely adopted at the Department of Obstetrics and Gynecology. The objective of this study is to enable realistic fetus volume rendering with coloring similar to a real baby's skin in order to provide psychological stability and enjoyment to pregnant women.

2. Methods

We propose a realistic fetus skin color processing method that uses both a 2D color map and an S-curve tone mapping function (TMF) for ultrasound volume rendering.

First, the gamut model of skin color is optimized as the color distribution of the database consisting of baby images.

Sample patch values are drawn in Cb-Cr plane using RGB-YCbCr conversion equation. For polygon modeling of the drawn Cb-Cr data, Cb and Cr values are converted to the chroma and hue and the maximum chroma values in each hue are selected. For modeling the gamut in the lightness-chroma plane, the chroma value is converged to 0 from the center of the maximum chroma according to the increase or decrease of the lightness.

This is used to create the 2D color map for the tone mapping of the ray casting. The intensity and chroma of the 2D color map correspond to the shading, and the hue is given by the depth in the volume rendering. In addition, the lightness value of the 2D color map is inverted to give a translucent effect. The black blob generated by the speckle is represented as a bright gray value; therefore, the image rendered by the lightness-inverted color map appears more clear and translucent.

Second, to enhance the contrast of the rendered image, the luminance, color, and tone-curve parameters of the TMF with the S-curve shape are controlled by the 2D Gaussian function, depending on the lighting position. The purpose of this algorithm is as follows: translucent effect is represented if the lights on the back of the object are existed, the high contrast image is shown and if the lights in front of the object are located, luminance values are reduced. The translucent effect is presented by controlling the tone curve if the object is lit from behind.

3. Results and Discussion

In the experiments, we used an NVIDIA GTX 560 Ti graphics card and CUDA parallel coding to implement the volume rendering based on ray casting. The resolution of the volume data is 200x200x200 voxels. The speed of the volume rendering is the 19.5fps (frame per second). We generated 10 color maps and rendered images for different values of hue and chroma. Ten observers evaluated the skin-color likeness of the images. According to the user satisfaction results, the hue range is 130.4~135.3 and the chroma range is 76.2~92.6. The best sMOS (scaled Mean Opinion Score) results are obtained as 84.6%. The experimental results show that the proposed method achieves better and more realistic skin color reproduction than the conventional methods.

9015-4, Session 1

What impacts skin color in digital photos?

Albrecht J. Lindner, Stefan Winkler, Advanced Digital Sciences Ctr. (Singapore)

Skin colors are an important topic in image processing and are used for various tasks such as face detection/recognition, tracking of body parts or color correction. However, skin colors in images are influenced by many different factors ranging from the pictured person (e.g. skin type, degree of tanning) over surrounding factors (e.g. illumination, geographic location) to technical factors (e.g. camera type, flash setting). In this paper we pursue a holistic approach to assess the impact of different factors on skin colors in images.

We download a large database of images from the web and also record their technical and semantic metadata. This includes camera type, flash setting, geolocation, keywords and depicted person. Images that share one metadata attribute in common are then summarized in one separate dataset.

We then use a statistical framework to analyze the downloaded images, where each image is described by a three-dimensional histogram in CIELAB color space. We use the Mann-Whitney-Wilcoxon test to assess whether certain histogram bins have a significantly higher or lower bin count in face images in comparison to other non-face images. This results in a significance distribution in color space that culminates in the bin with the most prominent color in face images. A complete assessment of the entire distribution indicates which colors are likely to be skin color and which colors are not. We run a different test for each metadata attribute to acquire a specific significance distribution per attribute.

We demonstrate the precision of the estimated skin color distributions at the hand of various examples such as skin colors in images from different geolocations, skin colors in images annotated with 'sunburn' or skin colors in images depicting 'Barack Obama' or 'Nicole Kidman'.

The distributions from the statistical analysis do not only provide the CIELAB values of skin color given an attribute, but also how significant the result is. This allows to compare the impact of different metadata attributes on skin colors in images. Our comparison demonstrates that technical metadata has a comparatively weak impact. In contrast, semantic metadata -- keywords and depicted person -- has a more significant impact. This suggests that semantic metadata provides strong clues about the expected skin color distribution for a given image and should be considered for related applications.

One such application is skin detection. We show at the example of keywords that knowing an image's keywords can improve the precision of detected skin maps. The experiment we conduct uses images annotated with different keywords and we create manual ground-truth for 50 images per keyword. The remaining images are used to estimate a keyword-specific skin color distribution. We then measure the precision-recall curves for all images using the respective keyword-specific color distributions. We then repeat the same experiment but only use a general skin color model instead. The skin detection with the keyword-specific color models clearly outperforms the standard detection with a general color model.

9015-5, Session 2

Microscale halftone color image analysis: perspective of spectral color prediction modeling

G. M. Atiqur Rahaman, Ole L. Norberg, Per Edström, Mid Sweden Univ. (Sweden)

The basic technique for most regression based color prediction models is to use the measured average reflectance of an area of halftone ink dots and of unprinted paper to estimate the model parameters. In this case, the assumption is that the colors of ink and paper are uniform and constant. But, in practice, ink density becomes zero and unevenly absorbed towards the boundary. Most significantly lateral light scattering takes place around the solid or fractional ink boundary causing optical dot gain. The consequence is the change in effective ink coverage and color saturation between solid ink and optically gained area. So each ink dot can be divided into two sub areas: full ink and optically gained. However, the amount of ink spreading, absorption or light scattering depends on a particular combination and physical properties of paper, inks, halftone and printing techniques. Hence, the model parameters vary as a function of the print media combination. But little efforts have been made to correlate the parameters to the relevant physical attributes. Analysis of microscopic image can give comprehensive understanding of multiple media interactions. In this study, a simple approach has been proposed to segment a microscale halftone color image in full tone ink, optically gained and paper area. Using the segmentation results a wide range of print samples have been analyzed for dot gain. Finally, we propose a general idea to expand the classical Murray Davies model to improve accuracy by incorporating the influence of optically gained area.

The recent approach to study dot gain is to capture the microscopic image in reflectance and transmittance mode, assuming transmitted

image has no lateral light scattering. But the assumption may have arguments, and capturing or processing transmitted images require special expertise and illumination-camera setups. In our approach, we segment only the reflection images into full ink, optically gained and paper area to calculate corresponding fractional coverage.

The k-means clustering algorithm has been applied to segment the single ink color halftone images. The major advantage is to avoid single threshold by clustering the pixels based on relative color values. The samples include both AM and FM halftoned prints by offset and prepress ink jet printer on commercially available multiple paper grades using the four primary color inks. The microscale halftone patterns were captured by a modified microscope with RGB camera.

Samples printed in a precisely controlled prepress ink jet printer give, in most cases, negative mechanical and high optical dot gain, showing a close relation between effective and reference ink coverage. But the offset samples give more uniform curves demonstrating the dot gain performance. The full article includes illustrations of the segmentation performance, and corresponding relation among effective, full ink and optical coverage against reference ink coverage for each set of sample. Based on the results and experimental observations, we propose an idea to extend the classical Murray-Davies model incorporating the influence of optically gained area. The continuation of the study will be to validate the model.

9015-6, Session 2

Reproducing the old masters: applying colour mixing and painting methodologies to inkjet printing

Melissa K. Olen, Univ. of the West of England (United Kingdom); Joseph Padfield, The National Gallery (United Kingdom); Carinna E. Parraman, Univ. of the West of England (United Kingdom)

Considerable advancements have been made in aqueous inkjet technology for the fine art market, yet there are still significant opportunities to investigate ulterior digital printing methods that specifically address the working methods of artists. This research investigates multi-channel inkjet printing methods, which deviate from standard colour management workflows by reflecting on art historical processes, including the construction of colour in old master works, to reproduce specific colour pigment mixes in print. This is approached by incorporating artist colour mixing principles relevant to traditional art making processes through direct n-channel printing and the implementation of multiple pass printing. By demanding specific ink colourants to be employed in print, as well as the application of mixing colour though layering, we can mimic the effects of the traditional processes. These printing methods also generate colour through a variety of colour mixtures that may not have been employed or achieved by the printer driver. When printing with multiple passes, utilising a single ink colour with each pass, the print order can be controlled, which has shown to have considerable effects on the final colour produced. For example, the result of a multi-pass, layered yellow and cyan print produces a different final colour when printed cyan over yellow compared to yellow over cyan, or even a single pass combination of the two.

The objective of this research is to explore colour mixing and layering techniques in the printing of inkjet reproductions of original artworks that will maintain subtle colour transitions in dark shadow regions. Details in the darker regions of original works are maintained by slight colour distinctions, which are barely decipherable except upon close inspection. While these colours are lost in traditional inkjet reproduction, by using direct n-channel editing capabilities to reproduce a painted original with high dynamic range we can improve colour variation in the shadow regions. By limiting the use of black ink we aim to better reproduce more saturated dark colours, controlling the mixture to create these dark colours with increased density.

This research will be jointly undertaken with the National Gallery, London, where the application of developed methodologies will be applied in colour matching historic artists' pigments for print reproduction. The aim of this research is to lead to better methods of reproducing traditional artists' colour pigments utilising current and expanding technologies. The objective is to create novel approaches for the inkjet printing process to address current limitations of digital reproduction, and open up new avenues of inkjet production.

9015-7, Session 2

Color prediction modeling for five-channel CMYLcLm printing

Yuan Yuan Qu, Paula Zitinski Elias, Sasan Gooran, Linköping Univ. (Sweden)

In printing, halftoning techniques saturate the area on the substrate with ink dots in higher or lower degree with the goal of obtaining darker or lighter shades. Nevertheless, in the case of lighter shades, the individual cyan (C) or magenta (M) dots are then in a sparse pattern, undesirably perceptible against the white background. One of the solutions is to use less pure versions of cyan and magenta inks - light cyan (Lc) and light magenta (Lm). By removing the harsh cyan or magenta dots and using Lc and Lm in light and near neutral areas, it is possible to obtain a more viable reproduction.

When Lc and Lm are added to the usual CMY printing, the characterization and separation of CMYLcLm printers becomes more challenging. As reported in Son et al., 2011 and Agar, 2001, the correlated processes of characterization and separation of CMYLcLm printers are achieved by firstly characterizing the CMY printing using a three-channel color prediction model and then converting C to (C and Lc) and M to (M and Lm) using strategies such as color smooth transition by lightness values or dot visibility analysis.

In this paper, we aim to involve Lc and Lm in printer characterization. In theory, it is feasible to characterize the printer by Yule-Nielsen modified spectral Neugebauer model using 25 Neugebauer primaries. However, we consider that Lc and Lm are associated with C and M. We aim to characterize the five-channel CMYLcLm printer using a three-channel color prediction model by taking the properties of Lc and Lm into account.

By plotting the CIELAB color values of prints using Lc, C, Lm and M with different coverage from 0 to 100%, we noticed that Lc and Lm cannot represent halftoned C and M directly. In our research we treated the combinations of Lm and M (and Lc and C) in the characterization as compound inks.

The proposed characterization of CMYLcLm is based on our previous CMY color prediction model (Qu and Gooran, 2013), which can be executed using either CIEXYZ tri-stimulus values or spectral data. We treat C + Lc and M + Lm as new compound inks to replace C and M in our previous color prediction model.

In the experiments, the printed samples include training and testing patches. For a better evaluation, we divide the test samples into different groups: CLcMLm, CLcY, MLmY. The prediction errors within each group are acceptable, based on both calculations using CIEXYZ measurements and using spectral data. The calculated mean and maximum color difference ?E94 is 0.58 and 3.48 respectively over 596 CLcMLm samples. The ongoing research is about results of CMYLcLm values and will be included in the full paper.

References:

U. Agar (2001), "Model Based Color Separation for CMYKcm Printing", 9th Color Imaging Conference, Scottsdale, AZ, USA.

Y. Qu and S. Gooran (2013), "Simple Spectral Color Prediction Model using Multiple Characterization Curves", Proc. TAGA (Technical Association of the Graphic Arts), February 2013, Portland, Oregon, USA.

Ch. Son, H. Park and Y. Ha (2011), "Improved Color Separation Based on Dot-Visibility Modeling and Color Mixing Rule for Six-Color Printers", Journal of Imaging Science and Technology, Vol. 55, No. 1, pp. 10505-1-10505-16(16).

9015-8, Session 2

Physical and optical dot gain: characterization and relation to dot shape and paper properties

Mahziar Namedanian, Daniel Nyström, Paula Zitinski Elias, Sasan Gooran, Linköping Univ. (Sweden)

The tone value increase in halftone printing commonly referred to as “dot gain” actually encompasses two fundamentally different phenomena. Physical dot gain (also known as mechanical dot gain) refers to the fact that the size of the printed halftone dots differs from their nominal size. Physical dot gain is closely related to the printing process, including the nip pressure in offset, the ink-transfer and ink-setting processes, and paper properties. Optical dot gain (also known to as the Yule-Nielsen effect) originates from light scattering inside the substrate, causing light exchanges between different chromatic areas, and making the dot appear bigger than its physical size when it is perceived or measured. Due to their different intrinsic nature, physical and optical dot gains need to be treated separately.

In this study, we characterize and compare the dot gain properties for offset prints on coated and uncoated paper, using AM and first and second generation FM halftoning. Spectral measurements by a spectrophotometer are used to compute the total dot gain. Microscopic images, with a resolution corresponding to 2 ?m/pixel, are used to separate the physical and optical dot gain, using a previously proposed method. The microscopic images are further used to study ink spreading and ink penetration, and to compute the Modulation Transfer Function (MTF), describing the light scattering properties for the substrate, by using a knife-edge approach.

The experimental results show that the total dot gain is larger for the uncoated paper. However, when the physical dot gain is separated, it is found to be larger for the coated paper. Microscopic images of the prints reveal that the ink penetrates into the pores and cavities of the uncoated paper, resulting in inhomogeneous dot shapes. For the coated paper, the ink spread on top of the surface, giving a more homogenous dot shape, but also covering a larger area, and hence larger physical dot gain. The results also show that the physical dot gain reaches its peak at the ink level where the halftone dots start to overlap, and that by selecting different dot shapes, this peak can be shifted.

The reason that the total dot gain is larger for the uncoated paper is because of the optical dot gain. The effect of optical dot gain depends on the ratio of the length of the lateral light scattering within the substrate to the size of the printed halftone dots. It has also been shown that the optical dot gain is related to the dot perimeter, which is larger for the uncoated paper because of the inhomogeneous dot shapes. The full paper will thoroughly examine how the optical dot gain is related to both the dot perimeter and the light scattering, using the physical dot shapes and the MTFs for the substrates.

When comparing the different halftones, it is clear that the optical dot gain is inversely proportional to the halftone dot size, because the ratio between the lateral light scattering and the dot size increases for smaller dots. This also affects the volume of the color gamut, where the gamut increases with decreasing dot size, illustrating the fact that optical dot gain is not all bad, but also increasing the gamut of color prints.

9015-9, Session 3

Gamut mapping in a high-dynamic-range color space

Jens Preiss, Technische Univ. Darmstadt (Germany); Mark D. Fairchild, James A. Ferwerda, Rochester Institute of Technology (United States); Philipp Urban, Fraunhofer-Institut für Graphische Datenverarbeitung (Germany)

Two sequential transformations are usually applied for displaying high-dynamic-range (HDR) images on low-dynamic-range (LDR) output devices (such as displays or printers): 1. HDR tone mapping and 2. Color gamut mapping. Such mappings may leave some room for improvement, particularly because most tone mapping operators (TMOs) disregard color information and gamut mapping algorithms (GMAs) (operating on LDR color spaces) may misinterpret the magnitude of perceived color contrasts within HDR scenes.

In this paper, we present a novel approach of tone mapping as gamut-mapping in an HDR color space. HDR and LDR images as well as device gamut boundaries can simultaneously be represented within such a color space. This enables a unified transformation of the HDR image into the gamut of an output device (in this paper called HDR gamut mapping). The recently introduced hdr-CIELAB and hdr-IPT color spaces [1] are HDR extensions of CIELAB and IPT. The modifications consist mainly in a simple replacement of the spaces’ non-linearities by an appropriately parametrized Michaelis-Menten function. We adopt similar modifications to the hue linear LAB2000HL color space [3] designed to improve CIELAB and IPT with respect to perceptual uniformity. The resulting new HDR color space is denoted as hdr-LAB2000HL.

Since the concept of HDR color spaces is new, no experience within imaging applications has been gained so far. Therefore, an additional aim of this paper is to investigate the suitability of hdr-LAB2000HL to serve as a working color space for the proposed HDR gamut mapping. For the HDR gamut mapping, we used a recent approach that iteratively minimizes an image-difference measure subject to in-gamut images [7]. The resulting image is within the LDR gamut and has a minimal difference to the original HDR image with respect to the measure. We used the improved Color-Image-Difference (iCID) measure [4] as an objective function, which compares two images with respect to local lightness/chroma/hue differences as well as lightness/chroma-contrast and -structure deviations. Visual experiments revealed that iCID-based gamut-mapping optimizations were judged to be perceptually more similar to the original image than results of state-of?the-art spatial GMAs [4]. In this paper, we employ the same optimization algorithm but use an hdr-LAB2000HL representation of the images’ pixels. For transforming the HDR image to hdr-LAB2000HL, we computed the adapting luminance by a geometric mean of the image pixels. The image’s luminance values were then linearly scaled to map the computed adapting luminance to the middle gray value within hdr-?LAB2000HL.

As starting image of the iterative optimization we use an in-gamut image that is derived from the original HDR image by applying an existing TMO and GMA. We use three different TMOs: 1. Reinhard’s bilateral TMO, 2. Drago’s TMO (both from the HDR toolbox by Banterle [5]), and 3. tone mapping by iCAM06 [6]. For gamut mapping we used an existing GMA incorporated in the USNewsprintSNAP2007.icc profile. We chose this small newspaper gamut to better illustrate the differences between results. Another gamut mapping we tested is the color space transformation from XYZ values to sRGB.

In a psychophysical experiment, HDR-gamut-mapped images obtained by minimizing the iCID measure were compared to the reference tone-and gamut-mapped images. On an HDR display, both LDR representations were shown left and right to the original HDR image. The observers were asked to choose the LDR image which is perceptually more similar to the HDR image. Twelve natural images with indoor and outdoor scenes were used with the mentioned three TMOs and two device gamuts (newspaper gamut and sRGB). For the small newspaper gamut, about 95% prefer the optimized to the reference LDR im?age – the concept of replacing tone mapping by gamut mapping in an HDR space performs well. If the output gamut is sRGB, only 52% of the observers prefer the optimization. This result is worse because the optimization shows drawbacks for those images which have very bright highlights. A further analysis reveals that a better (local) luminance adaptation would improve the results. Note that only a global luminance adaptation was considered in this work.

We have shown that tone mapping is only a special case of gamut mapping if high-?dynamic-range images are represented in a high-dynamic-range color space. Thus, tone and then gamut mapping can be replaced by one transformation. Further research should consider local luminance adaptation for the hdr-LAB2000HL representation

and an (image independent) encoding within device profiles for faster processing.

References

[1] M. D. Fairchild and P.-H. Chen. Brightness, lightness, and specifying color in high?dynamic-range scenes and images. Proc. SPIE 7867, Image Quality and System Per?formance VIII, 2011.

[2] E. Reinhard. Tone Reproduction and Color Appearance Modeling: Two Sides of the Same Coin? In 19th Color and Imaging Conference, 2011.

[3] I. Lissner and P. Urban. Toward a Uni.ed Color Space for Perception-Based Image Processing. IEEE Transactions on Image Processing, 21:1153–1168, 2012.

[4] Preiss, Jens and Fernandes, Felipe and Urban, Philipp. Improving the Color-Image-Di.erence (CID) Measure. IEEE Trans. Image Processing (submitted), 2013.

[5] Francesco Banterle. HDR Toolbox for processing HDR images into MATLAB and Octave. [Online]. Available: https://github.com/banterle/HDR_Toolbox. [Accessed: Apr-19-2013].

[6] J. Kuang, G. M. Johnson, and M. D. Fairchild. iCAM06: A re.ned image appear?ance model for HDR image rendering. Journal of Visual Communication and Image Representation, 18(5):406–414, 2007.

[7] J. Preiss and P. Urban. Image-di.erence measure optimized gamut mapping. In 20th Color and Imaging Conference, 2012.

9015-10, Session 3

Color preservation for tone reproduction and image enhancement

Chengho Hsin, Zong Wei Lee, Zheng Zhan Lee, Shaw Jyh Shin, Feng Chia Univ. (Taiwan)

Many applications such as high dynamic range (HDR) tone reproduction and low dynamic range (LDR) image enhancement is based on manipulating luminance values, rather than work directly on the red(R), green(G), and blue(B) components of a given color image. To reconstruct a color image R'G'B' from the luminance output, a constraint of preserving the original hue and saturation is often required, that is R:G:B=R':G':B'. The common approach to realizing this color preservation constraint is by letting R'/R = G'/G = B'/B = L'/L, where L and L' are the luminance values before and after processing, respectively. This approach generally performs well under the conditions that the input is a gamma-corrected and LDR image. However, if either condition is violated, the resultant image may become highly colorful (This is often mistakenly termed as over-saturated). The exceedingly large chroma produced by the great ratio L'/L causes a highly colorful phenomenon. To overcome this problem, Tumblin and Turk [1] proposed that the color channels of the output image are proportional to that of the gamma-like-adjusted input image. This can be described by R'/(R^s) = G'/(G^s) = B'/(B^s) = L'/(L^s), where the exponent "s" is usually chosen smaller than 1 and can be viewed as the inverse of a gamma parameter. Mantiuk et al. [2] conducted a series of subjective appearance matching experiments to measure the change in image colorfulness after contrast compression and enhancement. By fitting these experimental data, a formula is derived to determine the exponent "s" for a tone-mapping operator. Although the nonlinear relation between the output R'G'B' and the input RGB specified by the exponent "s" may reduce the exceedingly large chroma to a normal range, it modifies the original hue and saturation so that the color preservation is no longer maintained.

The goal of this work is to model the adjustment in chroma that needs to be made after tone reproduction or image enhancement and still preserve the original hue and saturation of the input image. Our goal is different from that of appearance models which try to predict perceived colors accounting for both the visual system and the viewing environment. Recovering the photographic colors that could be expected under optimal camera exposure settings is our objective. We assume that the input and output trichromatic values are linear with respect to radiance (not gamma-corrected). All results are transformed to the sRGB color space for display. The output luminance L', in general, is obtained by a nonlinear mapping of the input luminance L. Let (L')^r define the predicted linear version of the output luminance under an optimal camera exposure setting, where the exponent "r" is called an image gamma. With this definition, we propose a novel color preservation constraint described by R'/R = G'/G = B'/B = ((L')^r)/L. The image gamma is estimated by the nonlinear mapping between the input and output luminance values. Approximating the nonlinear mapping by a power law, L' = L^u, where "u" is a mapping exponent, the relation between the image gamma "r" and the mapping exponent "u" can be modeled. For strong luminance compression (u < 0.4), the output luminance L' is viewed as the lightness, so its predicted linear luminance is given by setting the image gamma r = 2.5. If luminance compression is small (0.4 < u < 1), the chroma needs to be reduced mildly. Hence, the predicted linear output luminance is obtained by specifying the image gamma slightly smaller than the inverse of the mapping exponent, r < (1/u). For luminance expansion (u > 1), the output luminance itself is considered to be linear, and thus r = 1. In summary, the image gamma "r" used for computing the predicted linear output luminance is modeled by a hyperbolic tangent function of the mapping exponent "u". Furthermore, we apply the maximum differential entropy principle to arrive a simple solution for the mapping exponent "u" which is expressed by the ratio of the mean of the logarithm of the output luminance to that of the input luminance.

Both qualitative and quantitative experiments were conducted to validate the effectiveness of the proposed color preservation approach. We also compared the proposed method with other schemes. Four HDR and five LDR image databases were used in experiments. A sigmoid tone mapping operator was chosen to produce an ideal output luminance image. For quantitative evaluation, we propose a lightness difference index and a colorfulness difference index to measure the quality of the recovered output color image. The mean of L* in CIELAB color space is used to represent the lightness. The lightness difference index essentially computes the variation of the lightness between the original and the recovered color images. Similarly, the weighted sum of the chroma mean and its standard deviation developed by Hasler and Susstrunk [3] is adopted to represent the colorfulness. The variation of the colorfulness before and after tone mapping is measured by the colorfulness difference index. The values of both indices ranged from -25 to +25 are considered to be appropriate. This defines a square region in a plane spanned by the coordinates specified by the two indices. The indices of 99% of images tested by the proposed method located inside this appropriate square region, which is overwhelmingly better than other methods.

The major contribution of this work is to devise the predicted linear output luminance so that a novel color preservation constraint could be implemented without causing a highly colorful phenomenon.

[1] J. Tumblin and G. Turk, "LCIS: A boundary hierarchy for detail-preserving contrast reduction," In Siggraph 1999, Computer Graphics Proceedings, pp. 83–90, 1999.

[2] R. Mantiuk, R. Mantiuk, A. Tomaszewska, and W. Heidrich, "Color correction for tone mapping," EUROGRAPHICS, vol. 28, no. 2, pp. 193-202, 2009.

[3] D. Hasler and S. Susstrunk, "Measuring colourfulness in natural images," in Proc. SPIE 5007, Human Vision and Electronic Imaging VIII, 87, 2003.

9015-11, Session 3

Color signal encoding for high dynamic range and wide color gamut based on human perception

Mahdi Nezamabadi, Scott Miller, Scott J. Daly, Dolby Labs (United States); Robin Atkins, Dolby Canada Corp. (United States)

In the image display chain the signal transfer function describes how to convert signal representations such as film density, voltages, and

digital codes to optical energy. For the reference standard display this is known as electro-optical transfer function (EOTF). The EOTF curves based on power functions have traditionally been used for digital cinema and television displays with standard dynamic range and peak brightness values of about 50 to 100 Cd/m2. As stated in ITU BT.2246, two threshold functions, called Schreiber and Barten, have been used to illustrate the performance of the power-based (gamma) EOTFs regarding to visual detection thresholds for 10-bits and 12-bits implementations. Todays' displays have much higher dynamic range and peak luminance levels and traditional power –function based EOTFs would result in quantization artifacts especially at low luminance levels. A new EOTF, based on human perception, called PQ, was proposed in a previous work (SMPTE Mot. Imag. J 2013, 122:52-59) and its performance was evaluated for a wide range of luminance levels and encoding bit-depth values (Please see Figure 1 in the appendix for comparison between Rec 1886 and the proposed PQ). In the previous work simulations and visual experiments were performed to compare efficiency of the proposed PQ to the traditional power based transfer functions in encoding luminance channel of the signal. This paper is an extension of that previous wok to include the color aspects of the PQ signal encoding. The efficiency of the PQ encoding and bit-depth requirements were evaluated and compared for standard color gamuts of Rec 709 (SRGB), and the wide color gamuts of Rec 2020, P3, and ACES for a variety of primary spaces such as RGB, YCbCr, and XYZ.

In a selected color space for any potential local gray level, 26 color samples were simulated by deviating one quantization step from the original color in every directions of the corresponding color space dimensions (please see Figure 2 in the appendix for an example of the RGB space). The quantization step sizes were simulated based on the PQ and gamma curves for different bit-depth values and luminance ranges for each of the color gamut spaces and representations. Color differences between the gray field and the simulated color samples were computed using CIE DE2000 color difference equation. The maximum color difference values were used as a metric to evaluate the performance of the corresponding EOTF curve (Figure 3 and 4 in appendix shows a few examples). While not obvious, the increased color gamuts require more bits to keep quantization below threshold, even for luminance modulations.

Visual experiments are being performed to verify the simulation results for the proposed PQ and the traditional power based transfer functions for a display having options for both the standard 709 gamut and the wider P3 gamut. Regarding the visual detection thresholds, the minimum bit-depth required by the PQ and gamma encodings are evaluated and compared through these visual experiments.

9015-12, Session 4

Design for implementation of color image processing algorithms

Jamison Whitesell, Dorin Patru, Eli Saber, Rochester Institute of Technology (United States); Gene Roylance, Brad Larson, Hewlett-Packard Co. (United States)

High-level modeling languages, such as Matlab, and their associated simulation environments, process data represented as floating-point numbers because they have access to the necessary computational resources to do so. In contrast, in embedded, handheld, or mobile systems computational resources are limited, and therefore for example one would not represent an 8-bit integer as a floating-point number, nor perform operations using a floating-point unit.

The design of color image processing algorithms at the high-level usually focuses exclusively on functional correctness. As a result, optimal representations and efficient operations selections are left to the implementation phase. Based on multiple color space conversions encountered in the custom hardware implementation of a segmentation algorithm, in this paper we argue that, and show how, optimal representations and efficient operations can and should be selected during the high-level modeling.

In a first example we have changed an exponent of from 2.4 to 2.5, i.e. a square and a square root. In a second example we have replaced a division by 12.92 with a division by 16, i.e. four shift rights instead of a costly division. In a third example we have converted a division by a constant denominator into a multiplication with its inverse. And finally, in a fourth example, we have approximated a cube root with a 4th and 16th roots. In all these cases, the correlation coefficients between the output image results of the original and modified algorithms are better than 0.9.

In the paper we first describe the representation ranges and operations that were found necessary in the original, high-level design of the algorithm. Then we show in detail how we have optimized representation ranges and modified operations for efficient hardware implementation. At each step we report the correlation with the output of the original algorithm. Finally, we describe a set of rules and guidelines, which if followed will result in algorithms truly designed for implementation (DFI).

Most often the same individual or group does not perform the high-level algorithm development and its implementation. The latter will follow the former verbatim to maintain functional correctness, missing optimization opportunities.

We believe that our findings are very valuable to be disseminated in this venue for two reasons: first, it will alert designers to optimizations possible during the high-level modeling of color image processing algorithms, and second, these optimizations will result in shorter implementation times. Ultimately, algorithms will run faster while consuming less hardware resources and less power.

9015-14, Session 4

Dynamic histogram equalization based on gray level labeling

Bongjoe Kim, Samsung Electronics Co., Ltd. (Korea, Republic of); Gi Yeong Gim, Hyung Jun Park, Samsung Digital City (Korea, Republic of)

Histogram equalization is the one of the well-known methods for contrast enhancement in a variety of applications. Numerous histogram equalization methods have been proposed and they generally shown good performance on almost all type of images. However, they still show some problems such as washed out appearance, gradation artifact and detail loss for certain classes of images which have some gray levels with very high frequencies.

This paper proposes a novel dynamic histogram equalization method to overcome such drawbacks of the histogram equalization methods. The proposed method first divides the histogram of input image into a proper number of sub-histograms based on gray level labeling method. Then each sub-histogram is assigned to a new dynamic range according to the number of pixels in each sub-histogram, thus a sub-histogram with a larger number of pixels will occupy a bigger portion of the dynamic range. Finally, Bi-histogram equalization is applied to each sub-histogram for a better overall contrast enhancement. Bi-histogram equalization decomposes sub-histogram into two parts with respect to its mean value and applies conventional histogram equalization to each of parts independently.

Different from existing dynamic histogram equalization methods, the proposed method employs a sub-histogram division operation over not the histogram domain but image spatial domain. Dynamic histogram equalization method based on histogram domain first smoothes the input histogram by using a one dimensional smoothing filter. The smoothed histogram is partitioned into sub-histograms based on the local minima. However, the local minima are highly sensitive to smoothing filter scale. Moreover, as it does not consider image spatial information, the gradation artifact may occur. To handle these problems and use image spatial information, we partition input histogram based on gray level labeling method. Gray level labeling method first constructs multi-resolution image pyramid that preserves image structure and contains representative gray level. Multi-resolution image pyramid defines input image as the finest resolution, and successively coarse resolutions are

generated by vector median filtering followed by down-sampling. After construction of a multi-resolution image pyramid, region growing and labeling are performed through image pyramid top-down to the finest resolution in the image pyramid. For region growing and labeling, we start at the top resolution and generate an empty label image. Successively, the lower label image is generated according to gray level similarity and spatial connectivity. Finally, histogram partition is performed based on gray level labeling.

Experimental results show that the proposed method produces better contrast enhanced images than several state-of-the-art methods without introducing several side effects, such as washed out appearance gradation artifact and detail loss etc. Unlike the other methods, the proposed method is free of parameter setting and can be applied to a wide range of image types.

9015-15, Session 4

Slide localization in video sequence by using a rapid and suitable segmentation in marginal space

Sheetal Rajgure, Vincent Oria, New Jersey Institute of Technology (United States); Pierre Gouton, Univ. de Bourgogne (France)

Segmentation techniques have been always an area of interest for researchers. Many techniques have been proposed since its origin. This paper focuses on different images in color space and proposes a well suited algorithm for the scenario.

The color distributions for image having more color predominance has more scattered distribution. The images which are not predominant in color their distribution are limited to one axis and hence not that predominant in color. Since grayscale image loses the color information necessary for segmentation. In this paper we use both DCT energy axis (axis 1) and color axis (axis 2) for the analysis. This space preserves the non-correlation of data and total energy. We have used the translation of RGB space to DCT technique by R Kouassi.

Grayscale images capture the details but disregard any color information. Image where color information is not that predominant or the color distribution is limited to one axis, marginal space is more useful. HSI space provides a better de-correlation of information in the psychovisual sense. In this space, color information can be reduced to a composite monochrome image by Carron's criterion[1] which digitally merges Intensity , Saturation and Hue information into a single magnitude . Thus segmentation techniques developed for gray-scale images can be used.

K-means is a classical technique widely used to that classifies the input data into multiple classes based on their inherent distance from each other. In this paper we use K-means technique to segment the image into different clusters. We use the color histogram and using Tsai's moment preserving method [3] using multiple thresholds to fix the initial cluster centroids. For DCT, K-means clustering is performed on axis 1 and axis 2 separately and then both the regions are merged.

We begin with the discussion of the analysis. Classical K-means approach uses random centroids initially. Thus the results vary according to chosen centroid. We fix the centroid by moment preservation [3] which is better than the random selection. It is quite evident from the results that grayscale loses lot of information when the image has very predominant color information. Best results are obtained by combining DCT axis 1 and 2. Marginal Images show intermediate results but are better than grayscale. With marginal we preserve color information on one axis.

[1] T. Carron, Segmentation d'images couleur dans la base Teinte-Luminance-Saturation: approche numeHrique et symbolique, Ph.D.Thesis, Univ. Savoie, France, 1995.

[2] RK Kouassi, P. Gouton M. Paindavoine "Approximation of the Karhunen Loeve transformation and its application to color images", Image Communication, vol.16 n°6, Elsevier Publishing, pp 541 - 551 February 2001.

[3] W.H.TSAI "Moment-Preserving Thresholding: A New Approach", Computer Vision. Graphics, and Image Processing 29, 377-393 (1985)

9015-16, Session 4

SVM-based automatic scanned image classification with quick decision capability

Cheng Lu, Purdue Univ. (United States); Jerry Wagner, Brandi Pitta, David Larson, Hewlett-Packard Co. (United States); Jan Philip Allebach, Purdue Univ. (United States)

Multifunction printer (MFP) products with scan, print, and copy capability are now widely used both in the office and at home. One major issue for these devices is copy quality. In order to achieve as high quality as possible for every input document, multiple processing pipelines are provided in the product hardware and firmware. Every processing pipeline is designed specifically for a certain class of document, which may be text, picture, or a mixture of both. Significantly lower quality output will result if a given document is processed by a pipeline that was not designed for it. For example, if a text document is processed by picture pipeline, then the output will contain text with blurry edges, which is undesirable. MFP products commonly provide an option on the front control panel for the user to choose which pipeline (mode) is to be used for each input document. However, most users simply choose the default mode without any change. Users always prefer that product can work in a "push the button and be done" fashion. So an automatic classification process is necessary before sending the input document to its corresponding processing pipeline. In our work, two sets of parameters are applied for mono and color documents, respectively, because it is assumed that the users will be required to choose from these two modes.

The proposed classification algorithm follows a binary-tree, two-stage decision structure. Each SVM classifier extracts two features from the input document and makes a decision based on them. Given its targeted real time application in low-end MFP products, two pre-processing procedures need to be done for every original scanned document to reduce the computational load. We first transform the original scanned document into a gray scale image by averaging the three RGB channels. This makes it easier for hardware implementation. Then, this gray level image is down-sampled to 75 ppi by block-averaging. So the final input to the subsequent classifiers is a low-resolution gray-level image, which makes the classification task more challenging.

9015-17, Session 5

Optimal color temperature adjustment for mobile devices under varying illuminants

Kyungah Choi, Hyeon-Jeong Suk, KAIST (Korea, Republic of)

With the development of display devices, great attention has been paid to display color reproduction. However, dissatisfaction caused by the chroma adaptation has not been yet studied enough in spite of complaints about viewing yellowish or bluish display, as shown in Fig. 1. In this regard, the purpose of this study is to find the optimal color temperature for the smartphone display by observing the relationship between the illuminant conditions and the ideal whites users perceive. The study is composed of two parts. Part I focuses on the effect of illuminant color temperature. Part II investigates the effect of illuminance.

A total of 95 participants with an average age of 22.01 years and a standard deviation of ±1.94 years were recruited. For visual examination, a total of 16 stimuli with different color temperatures varying from 2800-25000 K were presented on the Samsung Galaxy SIII smartphone. The subjects were asked to evaluate the optimal level of the display color temperatures under 19 illuminants simulating daily lighting experiences using a five point Likert scale. The illuminants ranged from 2500-20000 K in color temperature and from 30-3000 lx in illuminance.

In Part I, the total scores of the optimal level for the 16 stimuli were calculated for each of the 11 illuminants with different color temperature but equal illuminance (1000 lx), as listed in Table 1.

A positive correlation is observed between the optimal display color temperature and the illuminant color temperature ($r=0.90$, $p<0.05$). The optimal display color temperatures range from 6000-11000 K while the color temperatures of illuminant vary from 2500-20000 K. The display color temperature higher than the surrounding illuminant is perceived to be ideal, up until the illuminant reaches 10000 K. When the illuminant is above 10000 K, 11000 K is the marginal value of the ideal whites perceived.

Nonlinear regression analysis was performed in order to predict the optimal display color temperature (TD) by taking the illuminant color temperature (TI) as independent variables ($R2=0.91$, $p<0.05$). The derived formula is as follows: $TD=2814.47*\ln(TI)-16422.55$

In Part II, the total scores of the 16 stimuli were calculated under the 8 illuminants with different illuminances. At lower illuminance (30 lx), the color temperatures perceived as the most optimal are within a more restricted range: from 7200-9000 K. Moreover, the optimal display color temperature draws closer to the illuminant color temperature as the illuminance increases up to 3000 lx.

In conclusion, the study reveals that the display color temperature perceived to be ideal increases as the illuminant color temperature rises; however, the optimal color temperatures are restricted within the range of 6000-11000 K. Moreover, under lower illuminance, the correlation is relatively weaker confirming previous color reproduction theories. These findings can be used as the theoretical basis for designers and manufacturers when designing a color strategy for smartphone display.

9015-18, Session 5

Primary selection for uniform display response

Carlos E. Rodriguez Pardo, Gaurav Sharma, Univ. of Rochester (United States); Xiao-Fan Feng, Sharp Labs. of America, Inc. (United States)

The process of selecting primaries for display systems usually depends on a balance between different requirements and constraints that includes power consumption and color gamut. Different methodologies have been proposed to optimize the design and usually, they are based on the computation of parameters that characterize globally the display performance, like maximum luminance, gamut volume, or gamut coverage, sacrificing in some cases the performance in different regions of the gamut. In this paper, we present a methodology for display design that optimizes its global response while reducing the difference with the local performance. The methodology considers the gamut volume, the optical power and white luminance as global parameters, while the parameters for local performance are based on the display luminance and the luminance of the set of optimal colors. The methodology is applied for three and multiprimary displays.

9015-19, Session 5

Moire reduction using inflection point in frequency domain

Dae-Chul Kim, Wang-Jun Kyung, Bong-Seok Choi, Yeong-Ho Ha, Kyungpook National Univ. (Korea, Republic of)

Recently, an attempt of aliasing reduction method without OLPF was introduced to reduce the manufacturing cost of digital camera. This method converts interpolated RGB image signals to luminance and chrominance signals and applies low-pass filters to them, respectively. However, detail loss in the image is existed with the use of low-pass filter. In this paper, a moire reduction method with preserving the detail information in case of using no OLPF is suggested. The proposed method firstly analyzes the SFR(spatial frequency response) of camera and the patterns corresponding to the SFR by using resolution chart. Then, these patterns are determined by using the difference of intensity between current pixel and neighborhood pixels in horizontal and vertical directions, and are used to detect the moire region of luminance channel. Next, to reduce the moire, the moire regions are analyzed in the frequency domain. The maximum values per each frequency are calculated to detect the moire component in the frequency domain. Then, the moire component is determined as inflection point among maximum values per each frequency that are located in the frequency domain between high frequency and DC components. Therefore, the moire is reduced by removing its frequency component. Finally, color moire is corrected by multiplying the chromaticity ratio. Through the experimental results, it shows that the proposed method reduces the moire in luminance and color channels with preserving the detail information.

9015-20, Session 5

Recalling white point of smartphone under varying illuminants

Kyungah Choi, Jiho Jang, Hyeon-Jeong Suk, KAIST (Korea, Republic of)

Color constancy ensures that the perceived color of objects remain constant despite changing illuminants. The phenomenon of color constancy is highly dependent on the illuminants mainly due to the viewers' chroma adaptation. Facilitated by the built-in RGB sensor, some smartphones are now able to read the chromatic properties of illuminants and therefore, can auto-adjust the display color temperature. In this regard, this study aims to investigate the color constancy in white displays rendered under varying illuminants.

A total of 58 participants with an average age of 21.38 years and a standard deviation of ±3.12 years were recruited. For visual examination, 6 standard colors were presented on the Samsung Galaxy S3 smartphone display, varying from 6000 K to 11000 K. 15 comparison colors were chosen from 2700 K to 20000 K, including the standard colors. The subjects began an asymmetric color-matching session with a training phase, during which they were instructed to remember a standard color. In the test phase, the subjects selected one color among the comparison colors by recalling the standard from the training phase. During the test period, 12 test illuminants were randomly presented that were chromatically different (2500 K ~ 20000 K) but of equal illuminance (600 lx). A multiple-training, short-delay procedure was adopted; the subjects rechecked the standard color every 6 matches.

The total counts of the selected colors were calculated for each of the 6 standard colors under the 12 illuminants. The recalled colors were obtained for each standard by extracting the highest counts, as underlined in Table 1.

The effect of the illuminant change on the subjects' asymmetric matches is shown schematically in Fig. 1. In the plot, each solid line represents the recalled colors for the 6 standards. The recalled colors are displaced from the standard colors indicating that a shift in color memory has occurred. The directions of the shift in memory are generally headed toward higher color temperature. The 8000 K standard color, closest to the default white color temperature of the Samsung Galaxy S3, tends to be remembered as a higher color temperature when the illuminant color temperature is above 6000 K. However, the 6000 K and 7000 K standard colors show lower shifts in memory.

Multiple regression analysis was performed in order to predict the color temperature of the memory color (TM) by taking the standard color (TS) and illuminant color temperature (TI) as independent variables ($R2=0.89$, $p<0.05$). The derived formula is as follows: $TM = 10038.47*\ln(TS)+0.17*TI-82978.80$

In conclusion, the study observes the changes in color memory under varying illuminants. The directions of the shift in memory were generally toward higher color temperature. Moreover, colors with low color temperature show smaller shifts in memory. The empirical results of this

study can be used to propose a guideline for designing a color strategy for smartphone displays.

9015-21, Session 5

Evaluation of static color breakup for natual images on field sequential displays

Jae Uk Kim, Chang Mo Yang, Jang Hyeon Bae, Choon-Woo Kim, Inha Univ. (Korea, Republic of); Ho Seop Lee, Daesik Kim, Samsung Electronics (Korea, Republic of)

In order to reduce power consumptions, field sequential color (FSC) displays represent color images by sequentally controling backlights and modulating light transmission. FSC LCD displays without color filter can reduce power consumptions. However, sequential way of displaying would generate undesirable color artifacts often called as color breakup (CBU). There are two types of the CBUs on the FSC displays : dynamic and static. This paper presents a new method to determine the optimum control scheme for backlights and tone levels to reduce static CBU appeared on FSC LCDs.

9015-22, Session 5

Preserving color fidelity for display devices using scalable memory compression architecture for text, graphics, and video

Fritz Lebowsky, Marina M. Nicolas, STMicroelectronics (France)

High-end monitors and TVs based on LCD technology continue to increase their native display resolution to 4k by 2k and beyond. Subsequently, uncompressed pixel amplitude processing becomes costly not only when transmitting over cable or wireless communication channels, but also when processing with array processor architectures. For motion video content, spatial preprocessing from YCbCr 444 to YCbCr 420 is widely accepted. However, due to spatial low pass filtering in horizontal and vertical direction, quality and readability of small text and graphics content is heavily compromised when color contrast is high in chrominance channels. On the other hand, straight forward YCbCr 444 compression based on mathematical error coding schemes quite often lacks optimal adaptation to visually significant image content. We present a block-based memory compression architecture for text, graphics, and video enabling multi-dimensional error minimization with context sensitive control of visually noticeable artifacts. As a result of analyzing image context locally, the number of operations per pixel can be significantly reduced, especially when implemented on array processor architectures. A comparative analysis based on some competitive solutions highlights the effectiveness of our approach, identifies its current limitations with regard to high quality color rendering, and illustrates remaining visual artifacts.

9015-23, Session 5

Simplifying irradiance independent color calibration

Pouya Bastani, Brian V. Funt, Simon Fraser Univ. (Canada)

An important component of camera calibration is to derive a mapping of a camera's output RGB to a device-independent color space such as the CIE XYZ or sRGB. Commonly, the calibration process is performed by photographing a color checker in a scene under controlled lighting and finding a linear transformation M that maps the chart's colors from linear camera RGB to XYZ. When the XYZ values corresponding to the color checker's patches are measured under a reference illumination, it is often assumed that the illumination across the chart is uniform when it is photographed. This simplifying assumption, however, often is violated even in such relatively controlled environments as a light booth, and it can lead to inaccuracies in the calibration.

Conventional least-squares regression takes into account both the direction and magnitude of RGB vector, and thus, any intensity gradient on the color chart arising from the illuminant irradiance variation will result in an inaccurate calibration. Funt and Bastani [2,3] suggested a calibration method based on minimizing the angular difference between the camera RGB and CIE XYZ coordinates of patches on the color checker. This method, however, requires a non-linear optimization solver, such as Nelder-Mead, that in addition to being complex to implement on some devices, has a high computational cost.

In this paper, we suggest an alternative calibration scheme that is efficient to compute as well as being independent of irradiance. Specifically, we first normalize the set of camera RGB and CIE XYZ color vectors, thereby removing all dependence on the scene irradiance. We then minimize the squared difference between each pair, leading to a least-squared minimization problem on the unit sphere, which can be solved using the Moore-Penrose pseudo-inverse. The calibration is thus a least-squares minimization based on the normalized color vectors.

In this paper, we perform several experiments using both real and synthesized image to compare the performance of the proposed calibration technique to that of least-squares and angle minimization. Our experiments show that the proposed method is completely unaffected by any irradiance gradient and finds an accurate mapping of colors from camera linear RGB to CIE XYZ. While its performance, as measured in terms of CIEDE2000 color difference between mapped and measured XYZ coordinates, is similar to that of angle minimization, the new technique can be computed much more efficiently. The computational efficiency and ease of implementation of this method lends itself well to such applications as measurement of soil color using mobile devices [4], where the color correction matrix may need to be computed multiple times to account for different illumination conditions even thought the computational resources are limited. Another advantage of the new technique over angle minimization is that it can easily be combined with higher order regression techniques, such as root-polynomial color correction [1] to yield more accurate irradiance-independent calibration methods.

REFERENCES

[1] G. Finlayson, M. Mackiewicz, Anya Hurlbert, "Root polynomial color correction," IS&T/SID Nineteenth Color Imaging Conference, pp. 115–119, 2011

[2] B. Funt and P. Bastani, "Intensity Independent RGB-to-XYZ Colour Camera Calibration," Proc. AIC 2012, pp. 128–131, Taipei, Sept. 2012.

[3] B. Funt and P. Bastani, "Irradiance-Independent Camera Color Calibration," 2013 (in review)

[4] Luis Gómez-Robledo, Nuria López-Ruiz, Manuel Melgosa, Alberto J. Palma and Manuel Sánchez-Marañon, "Mobile phone camera characterization for soil colour measurements under controlled illumination conditions," Proc. AIC 2013, Newcastle Upon Tyne, July 2013.

9015-24, Session 5

Using statistical analysis and artificial intelligence tools for automatic assessment of video sequences

Brice Ekobo Akoa, Emmanuel Simeu, TIMA Lab. (France); Fritz Lebowsky, STMicroelectronics (France)

This paper presents a new approach to developing video quality measurement of video encoders and decoders. Our Video Quality Measuring Tool (VQMT) includes a HVS (Human Visual System) model trough statistical means to qualify and quantify the quality of a given video, or decide among two video which one has a better visual quality. The idea is founded on firstly selecting fundamental no-reference video

quality metrics which have been calibrated across subjective test scores (DMOS), and secondly applying advanced statistical analysis and artificial intelligence methods providing suitable output decision (evaluation, rating or classification). Such methods can be more easily applied to different objective quality metrics in a generic way. The efficiency of our statistical approach is demonstrated by a comparison with existing research work based on synthetic video artifacts. One method uses the classification criteria of the nearest neighbor in a Euclidean distance. The other method exploits machine learning by a neural network. The results obtained by each method are compared with a database of scores derived from human judgment (perceived overall video quality).

9015-25, Session 6

Hybrid halftoning using direct multi-bit search (DMS) screen algorithm

Kartheek Chandu, Mikel J. Stanich, Ricoh Production Print Solutions, LCC (United States); Chai Wah Wu, Barry M. Trager, IBM Thomas J. Watson Research Ctr. (United States)

Nearly every printer technology cannot print continuous-tone (contone) data directly. Therefore it must rely on Digital halftoning technology to convert contone image to a lower bitdepth image suitable for printing given the bitdepth limitations of the target printer. Ink jet technology produces discrete dots of different sizes based on the amount of ink ejected for each drop size. The number of dot sizes produced defines the bit depth of the printer, which is some cases is not an integer power of two. The objective of Digital halftoning is to select the correct dot size and spatial arrangement so that when printed and blurred by the Human Vision System (HVS), resemble the desired contone image. For example, a printing device having only black and white dots when combined can display various gray levels. Single drop size or single exposure capability is referred to as “binary” or “bi-level” printing. Digital printers were initially black and white devices having low resolution. To satisfy increased image quality requirements digital printers have evolved to high resolution CMYK color, having “multi-bit” or “multi-tone” capabilities to print with more than one bit of data per pixel. The new capability requires multi-bit halftoning to select at each pixel among multiple drop sizes or exposure levels. Multi-bit halftoning provides the potential to increase overall image quality beyond binary halftoning. Image quality of a printer increases more rapidly with bit depth than resolution, further promoting increased bit depth rather than higher resolution. Conventional digital halftoning techniques employ either dot size “Amplitude Modulation” (AM) or dot density, “Frequency Modulation” (FM) to define the spatial arrangement of the PELs. Generally, AM halftoning is used by Electro-Photographic (EP) printers to provide the best printed dot stability, while FM halftoning is used in inkjet printers where dot stability is not an issue. FM halftoning has advantages of higher spatial resolution and resistance to moiré artifacts, when used for color printing. AM/FM ``hybrid” halftoning algorithms provide a combination of the advantages by combining the aperiodic grid from FM screening with dot size modulation from AM screening. Hybrid halftoning produces stable dots like AM halftones, while avoiding moiré through irregular dot placement like FM halftones. They are also less sensitive to color mis-registration and can easily combined to create rosette free color halftone masks. Many binary aperiodic clustered dot halftone frameworks have been proposed in the past, but it is our view that there are no high quality multi-bit hybrid screen algorithms. In our paper, we propose the framework for an algorithm to use Direct Multi-bit Search (DMS), which is an extension to Direct Binary Search, to create multi-bit aperiodic clustered dot screens. A pixel validation map is provided to the DMS algorithm to guide the formation of homogeneous clusters. The DMS algorithm operates without any user defined guidance, iteratively choosing the best drop size (absorptance level), using an array of valid pixels constraint. This process is repeated with added constraints including wraparound and stacking condition, to halftone a set of constant tint patches. The patches created span the entire range of gray levels creating visually pleasing multi-bit halftone screen. The resulting patterns for the range of patches are converted to a set of threshold arrays that can be used for point operation halftoning. The result is a point operation screening having smoother appearance and improved detail rendering, compared to conventional clustered dot halftoning. Much of the improvement originate from the improved sampling of the aperiodic screen.

9015-26, Session 6

A Riesz energy based approach to generating dispersed dot patterns for halftoning applications

Chai Wah Wu, Barry M. Trager, IBM Thomas J. Watson Research Ctr. (United States); Kartheek Chandu, Mikel J. Stanich, Ricoh Production Print Solutions, LLC (United States)

FM and Hybrid AM/FM halftoning solutions depend on the ability to generate pleasant patterns of dispersed dots. Many approaches to generate such patterns are based on heuristics to minimize an energy function. The energy can be based on models of the human visual system or potential fields. In most such algorithms the generated pattern is tied to the initial grid, and to generate a pattern for a different grid requires rerunning the algorithm. In this paper we propose an algorithm based on Riesz energy minimization to generate disperse patterns for halftoning applications. One novel aspect of this algorithm compared with other algorithms is that it does not depend on the underlying grid and the same optimization result can be used on different grids depending on application. In particular, we propose to use a nonlinear optimization algorithm to minimize an objective function consisting of the Riesz energy of the configuration. The desired pattern is obtained when the optimization algorithm converges. The points are snapped to grid points or used without snapping as dot centers for a hybrid screen design. In halftoning applications, the points are generally on a plane. In applications where the resulting pattern is used in a dither mask, the points live in the 2-D torus to account for the wrap-around effect in tiling the masks. In both cases, the gradient and Hessian can be easily computed explicitly. We solved the optimization algorithm using two approaches. In the first approach, we computed the gradient and Hessian explicitly and use MATLAB’s unconstrained large-scale optimization algorithm. One problem with this approach is that as the number of points becomes large, the size of the Hessian because too large to fit in system memory. In the second approach, rather than computing the full Hessian matrix explicitly, we use a limited memory Quasi-Newton update (L-BFGS) to generate a low rank update of the Hessian matrix. On a PC workstation with a 2.53GHz Xeon quad-core processor and 16 GB of RAM, we were able to run a problem with 6400 points in about 75 minutes using the first approach, employing gradient and Hessian explicitly and using MATLAB’s unconstrained large-scale optimization algorithm. The same 6400 point problem using the same platform, took 20 minutes with the second approach using limited memory Quasi-Newton update (L-BFGS) to generate a low rank update of the Hessian matrix. We compared this algorithm with other algorithms such as K-means and Direct Binary Search and found that the proposed method is superior in terms of the uniformity of the dots. This is measured both with the variance of the Voronoi region areas and the variance of the minimal distance among dots. One novel feature of the proposed algorithm is that the points are generated on the plane and are snapped to the grid as a final step. This means that the result of the Riesz energy minimization can be used in different grids simply by applying the snapping operation to a different grid. In addition, the running time for algorithms such as DBS will depend on the size of the grid and the number of points, whereas in the proposed algorithm, the running time of the optimization is independent of the grid size.

9015-27, Session 6

Yule-Nielsen effect in halftone prints: graphical analysis method and improvement of the Yule-Nielsen transform

Hebert Mathieu, Univ. de Lyon (France) and Institut d'Optique Graduate School (France)

The color calibration of printing processes and the design of color hardcopies can be considerably eased by using a spectral reflectance prediction model calibrated from a small set of color patches printed with the considered printing system, support and inks. One of the most accurate models available today is the Yule-Nielsen modified spectral Neugebauer (YNSN) model combined with the method by Crété & al. for estimating the effective surface coverage of the inks. The YNSN model predicts the spectral reflectance of halftone raised to the power 1/n as a linear combination of the spectral reflectances of the fulltone colors (i.e. colors made by covering the whole surface with one or several inks, also called Neugebauer primaries) also raised to the power 1/n, where n is a tunable parameter which generally increases as the halftone screen frequency increases or as the ink dot size decreases. The weights attributed to the different fulltone colors are the effective surface coverages of these colors in the halftone. The power 1/n transform, characteristic of the Yule-Nielsen transform, empirically models the nonlinear relationship between the spectral reflectances of halftones and fulltones due to the internal propagation of light by scattering into the printing support, a well-known phenomenon also called "optical dot gain" or "Yule-Nielsen effect". In this paper, we propose a graphical method showing this non-linear relationship in a very intelligible way for single-ink halftones (e.g. cyan ink printed at 0.5 nominal surface coverage). The graph represents the reflectance of the single-ink halftone as a function of the reflectance of the same ink fulltone. If the spectral measurement contains k wavebands, then k points can be displayed in the graph and it is easy to check how well the Yule-Nielsen transform matches them. The measured points generally follow a curved line which fairly coincides with the plot of the Yule-Nielsen transform computed with the appropriate n value and effective ink surface coverage. In some cases however, the coincidence is not perfect and alternative transforms (based on other functions than the power 1/n function) can be found. We propose a few ones which provide appreciable gain in prediction accuracy, this latter being measured in terms of average CIELAB deltaE 1994 values over several tens of printed colors. The graphical method is also very interesting to display fluorescence effects by the inks in the visible spectral domain, whereas these effects are rather difficult to detect by looking directly at the spectral reflectances. When an ink is fluorescing, the points in the graph are distributed along two distinct curved lines instead of one. Since no transform can render two lines, no perfect matching between the model and the measured points can be achieved and specific fluorescence model is necessary if the prediction accuracy of the model is too bad. Although it provides no solution to this problem, the method that we propose is a good diagnostic tool to understand why the YNSN model is not accurate in some cases.

9015-28, Session 6

Irregular clustered-dot periodic halftone screen design

Chuohao Tang, Purdue Univ. (United States); Alex Veis, Hewlett-Packard Scitex, Ltd. (Israel); Robert A. Ulichney, Hewlett-Packard Labs. (United States); Jan P. Allebach, Purdue Univ. (United States)

Screening is one of the most widely used halftoning methods [1],[2]. Screens that generate periodic, clustered-dot halftone textures are especially popular because of the inherent stability of these halftone textures in the presence of non-ideal behavior of the marking engine, and the fact that such screens allow better control of moiré due to the superposition of the different colorant planes. With commercial offset printing in which the plates are written using a high-resolution, laser-marking system, the screens for the four colorants C, M, Y, and K can all be based on a single screen that is rotated to the ideal angles of 0, 30, 45, and 60 degrees. In addition, the screen frequency can be specified arbitrarily within a certain range.

However, with digital printing systems, the achievable screen angles and frequencies are limited by the finite addressability of the marking engine. In order for such screens to generate dot clusters in which each cluster is identical, the elements of the periodicity matrix must be integer-valued, when expressed in units of printer-addressable pixels. Such screens are called regular screens. They tile the plane into identical microcells that align with the printer-addressable lattice. They can be used to generalize the traditional rotated screens used in commercial offset printing by not having each screen in the screen set be a rotated version of a single screen, and by not having the rows and columns of dot clusters be oriented at 90 degrees with respect to each other. Such screens are called non-orthogonal screens. With a non-orthogonal screen, the elements of the periodicity matrix are integer-valued, but the two column vectors are not orthogonal. Non-orthogonal regular screens [2],[3] are widely used as the basis for clustered-dot, periodic screens in laser, electrophotographic printers and other digital printing systems. However, such screens still have only a limited ability to approximate the rotated screen sets used for commercial offset printing.

To achieve a better approximation to the screen sets used for commercial offset printing, irregular screens can be used. With an irregular screen, the elements of the periodicity matrix are rational numbers, when expressed in units of printer-addressable pixels. Irregular screens tile the plane into microcells that are not aligned with the printer-addressable lattice. When each printer-addressable pixel is assigned to a unique microcell, the resulting discrete-parameter microcells are not identical. Since the elements of the periodicity matrix for an irregular screen are rational, the overall microcell structure, and thus the shape of the dot clusters will repeat over a certain number of microcells. This defines the minimum supercell size for the irregular screen. However, this supercell is not necessarily rectangular. With a larger number of repetitions of the microcell, it is possible to achieve a repeating block that is rectangular, It is called the Basic Screen Block (BSB). Although irregular screens are used in some high-end digital printing systems, procedures for their design have not to our knowledge been discussed in the scholarly literature.

In this paper, we describe a procedure for design of high-quality irregular screens. We treat the BSB as our supercell, and use the Direct Binary Search algorithm [4] to determine the order in which dots or holes will be added to the microcells within the BSB. Then, we start with the design of the midtone 50% level halftone pattern. Within each microcell, we define one vertex as a hole-cluster center and the center of the microcell as a dot-cluster center. These coordinates are quantized to the printer-addressable lattice. We partition the lattice of printer-addressable pixels into labeled regions using a Voronoi diagram. Each such region will contain either a dot-cluster or a hole-cluster. Then, we propose an algorithm to refine the label assignments of pixels in the BSB to improve the quality of the 50 percent level halftone texture. This is done via a heuristic search to reduce a cost function. Our goal is to make the lengths of the boundaries of dot and hole regions and their areas as similar as possible. We also want the dot and hole regions to be as compact as possible. So our choice of the cost function mostly depends on the lengths of the boundaries and the areas of the dot and hole regions. After obtaining the 50 percent tone level halftone pattern, we then propose an algorithm to determine how to add dots from midtone to shadow and to remove dots from mid-tone to highlight. These processes are based on the relation between the added dots or holes and the centroids of the affected hole and dot areas. Finally, we present experimental results illustrating the progress of the design procedure in terms of the metrics for halftone quality on which the cost function is based. We also compare images halftoned with regular screens that are designed to yield the closest approximation to the target screen angles and frequencies; images halftoned with irregular screens in which dot and hole growth are controlled by fixed-shape templates, in a manner similar to that described in [5]; and images halftoned with irregular screens that have the same angles and frequencies, but which are designed using our new approach.

References:

[1] S. G. W. C. Haines and K. Knox, "Digital Color Halftones," in Digital Color Imaging Handbook. Boca Raton, FL: CRC Press, 2002, pp.385-490.

[2] C. Lee and J. P. Allebach, "The Hybrid Screen—Improving the Breed," IEEE Trans. on Image Processing, Vol. 19, pp. 435-450, February 2010.

[3] Y. Y. Chen, M. Fischer, T. Kashti, D. Shaked, and J. P. Allebach, "The Lattice-Based Screen Set: A Square N-Color All-Orders Moiré-Free Screen Set," Color Imaging XVII: Displaying, Processing, Hardcopy, and Applications, SPIE Vol. 8292, R. Eschbach, G. Marcu, and A. Rizzi, Eds., San Francisco, CA, 23-26 January 2012.

[4] D. J. Lieberman and J. P. Allebach, "A dual interpretation for direct binary search and its implications for tone reproduction and texture quality," IEEE Trans. on Image Processing, pp. 1950–1963, November 2000.

[5] J. P. Allebach, "Random Nucleated Halftone Screen," Photogr. Sci. Engrg., Vol. 22, pp. 89-91, March-April 1978

9015-29, Session 6

Effect of image capture device on the accuracy of black-box printer models

Jason Youn, Jian Sun, Yanling Ju, Purdue Univ. (United States); Tamar Kashti, Tal Frank, Dror Kella, Hewlett-Packard Indigo Ltd. (Israel); Mani Fischer, Hewlett-Packard Labs. Israel Ltd. (Israel); Robert Ulichney, Hewlett-Packard Labs. (United States); Guy Adams, Hewlett-Packard Labs. (United Kingdom); Jan Allebach, Purdue Univ. (United States)

Digital halftoning is the process of representing a continuous-tone image with a device that can render only two or a few different levels of absorptance. In the process of electrophotograpic (EP) printing, the deposition of toner to the printer-addressable pixel is greatly influenced by the neighboring pixels of the digital halftone. This is due to a number of phenomena that arise during the writing of the latent image on the organic photoconductor drum, transfer of toner to the drum, further transfer of this toner to the media, and fusing of the toner to the media surface. As a consequence, the printed halftone image can differ significantly from that which would be predicted by a simple point-to-point transfer from the digital halftone to each corresponding printer-addressable pixel. To account for these effects, printer models can either be embedded in the halftoning algorithm, or used to predict the printed halftone image at the input to an algorithm that is used to assess print quality.

We have recently developed a series of tabular equivalent grayscale models that account for the influence of a 5x5 neighborhood on the printed absorptance of a given printer-addressable pixel with stochastic, dispersed-dot [1], and stochastic, clustered-dot [2] halftone textures, and a larger 45x45 neighborhood that is intended to capture the effect of long-path scattering of light from the point where it is incident on the surface of the media to the point where it finally exits [3]. Most recently [4], we developed a series of six new models to more accurately account for local neighborhood effects and the influence of a 45x45 neighborhood of pixels on the central printer-addressable pixel. These new models are divided into three classes, and have a variety of computational structures that allow system designers to choose the model that is best-suited to their particular application. They also offer varying degrees of accuracy. With these models, we demonstrated results for irregular, clustered-dot, periodic halftones. We refer to all these models as black-box models, since they are based solely on measuring what is on the printed page, and do not incorporate any information about the marking process itself.

All these models depend on a training process that is based on analysis of images of specially designed test pages, printed using the target device for which the model is being developed. Thus, it is important to understand the impact of the image capture device on the effectiveness of the models that are developed from the images acquired with that particular capture device. The characteristics of these devices will vary according to frequency resolution (MTF), tone reproduction, noise, geometric distortion, and the spatial uniformity of these factors across the field of view [5], [6]. Comparing the performance of different capture devices in this context poses the interesting conundrum that we don't have absolute ground truth information. That is, just because models developed from test pages acquired with a given capture device are good at predicting images of halftone prints captured with the same device, does not mean that the models are accurately describing what is on the printed page. Our solution to this problem is to separately evaluate the characteristics of the image capture devices that are being compared, using our composite test page [6] or a similar set of target patterns, and then to refer all model results to the capture device that has the highest quality. One potential drawback of this approach is that a given capture device may score well according to some attributes and not so well for other attributes, whereas the situation could be reversed with a different capture device. Then, it may not be obvious which capture device should be treated as the reference device. An additional factor to be considered is the size of the field of view of the capture device. With a flat-bed scanner, it is possible to capture an entire training page for the model in one step. With camera-based capture devices, it is necessary to acquire multiple frames in a step-and-repeat fashion, either manually, or by using a computer-controlled positioning system.

In this paper, we will compare black-box models developed with three different capture devices: (1) an Epson Expression 10000XL flatbed scanner operated at 2400 dpi with an active field of view of 309.88 mm x 436.88 mm, (2) a QEA PIAS-II camera with resolution 7663.4 dpi and a field of view of 2.4 mm x 3.2 mm, and (3) Dr. CID, a 1:1 magnification 3.35 micron true resolution Dyson Relay lens-based 3Mpixel USB CMOS imaging device developed at Hewlett-Packard Laboratories – Bristol. Our target printer is an HP Indigo 5000 Digital Press. In addition to comparing the accuracy of the black-box model predictions of print microstructure using models trained from images captured with these three devices, we will also compare their ability to predict metrics of print quality.

References

[1] P. Goyal, M. Gupta, D. Shaked, C. Staelin, M. Fischer, O. Shacham, R. Jodra, and J. Allebach, "Electrophotographic Model Based Halftoning," in Color Imaging XV: Displaying, Hardcopy, Processing, and Applications, SPIE Vol. 7528, R. Eschbach, G. Marcu, S. Tominaga, and A. Rizzi, Eds., San Jose, CA, 17-21 January 2010.

[2] P. Goyal, M. Gupta, C. Staelin, M. Fischer, O. Shacham, T. Kashti, and J. P. Allebach, "Electro-Photographic Model based Stochastic Clustered-Dot Halftoning with Direct Binary Search," Proceedings of ICIP 2011 IEEE International Conference on Image Processing, Brussels, Belgium, 11-14 September 2011

[3] Y. Ju, D. Saxena, T. Kashti, D. Kella, D. Shaked, M. Fischer, R. Ulichney, and J. P. Allebach, "Modeling Large-Area Influence in Digital Halftoning for Electrophotographic Printers," Color Imaging XVII: Displaying, Processing, Hardcopy, and Applications, SPIE Vol. 8292, R. Eschbach, G. Marcu, and A. Rizzi, Eds., San Francisco, CA, 23-26 January 2012.

[4] Y. Ju, T. Kashti, T. Frank, D. Kella, D. Shaked, M. Fischer, R. Ulichney, and J. P. Allebach, "Black-Box Models for Laser Electrophotographic Printers – Recent Progress," Proceedings NIP29: IS&T's 29th International Conference on Digital Printing Technologies, Seattle, WA, 29 September – 3 October 2013.

[5] X. Zhang, T. Kashti, D. Kella, T. Frank, D. Shaked, R. Ulichney, M. Fischer, and J. P. Allebach, "Measuring the Modulation Transfer Function of Image Capture Devices: What Do the Numbers Really Mean?" Image Quality and System Performance IX, SPIE Vol. 8293, F. Gaykema and P. D. Burns, Eds, San Francisco, CA, 23-26 January 2012.

[6] Y. Lei, P. Majewicz, K. R. Bengtson, L. Li, and J. P. Allebach, "Composite Target for Camera-Based Document/Object Capture System," Journal of Imaging Science and Technology, to appear.

[7] G. B. Adams, "Handheld Dyson Relay Lens for Anti-Counterfeiting," 2010 IEEE International Conference on Imaging Systems and Techniques (IST 2010), 6 pp. 2010.

Research supported by Hewlett-Packard Indigo, Ltd. Rehovot. ISRAEL.

9015-30, Session 6

Ink-constrained halftoning with application to QR codes

Marzieh Bayeh, Univ. of Regina (Canada); Erin Compaan, Univ. of Illinois at Urbana-Champaign (United States); Theodore Lindsey, Univ. of Kansas (United States); Nathan Orlow, Univ. of Illinois at Urbana-Champaign (United States); Stephen Melczer, Simon Fraser Univ. (Canada); Zachary Voller, Iowa State Univ. (United States)

Since its introduction in 1994, the popularity of the QR (Quick Response) code system has exploded. Despite their widespread and ubiquitous use in advertising, most examples of QR codes do not feature any significant human recognizable information. In this project we examine a novel method for halftoning that enables us to incorporate visually significant data into the two dimensional Black-and-White barcodes, without affecting their machine readability. The inherent property of this halftoning method, which involves halftoning under a constraint on the amount of ink used, can be deployed and utilized in other areas.

The QR code is composed of Black-and-White elements called modules, which are interpreted by a reading device as 0-or-1 bits. In this work we embed a given grayscale image into a QR code by partitioning the image into blocks corresponding to the QR modules. Processing each image block, we calculate the equivalent grayscale image block which minimizes the error to the desired image *under the constraint* that the average grayscale value must be at least 80% black in regions which correspond to black QR modules (and at least 80% white in regions which correspond to white QR modules). The error term is based on the square of the L2 norm, and uses a linear filter approximating the human visual system in order to create more visually pleasing images. As each image block entry is treated as a real number in [0,1], and both the constraint on the average grayscale and the human visual system are linear, this results in a convex optimization problem which can be solved directly and efficiently.

Once the grayscale minimum is determined for a block, we present three strategies for a final conversion to a Black-and-White image. The first uses Bayer matrices to perform an ordered dithering of the block, while the second uses the DBS algorithm for halftoning. The final approach takes a pixel and splits it into a k-by-k region of sub-pixels, on which the error diffusion of Jarvis, Judice, and Ninke is performed. In each case, the result is a binary image which *stays within the constraints* imposed by the QR code.

We also investigate the amount of information lost from the original image by computing the entropy of the image distribution, taking into account the added information from the QR code. This leads to a tradeoff curve between robustness of the code and the information density.

Our methods readily adapt to varying resolutions, leading to their ability to be implemented in a wide variety of settings. Numerous examples of QR codes with embedded images will be shown, and the methods discussed above are compared.

9015-31, Session 7

ColorChecker at the beach: dangers of sunburn and glare *(Invited Paper)*

John J. McCann, McCann Imaging (United States)

In High-Dynamic-Range (HDR) imaging, optical veiling glare sets the limits of accurate scene information recorded by a camera. But, what happens at the beach? Here we have a Low-Dynamic-Range (LDR)scene with maximal glare. Can we calibrate a camera at the beach and not be burnt? We know that we need sunscreen and sunglasses, but what about our camera? The effect of veiling glare is scene-dependent, so when we compare RAW camera digits with spotmeter measurements we find significant differences. As well, these differences vary, depending on where we aim the camera. When we calibrate our camera at the beach we get data that is valid for only that part of that scene. Camera veiling glare is an issue in LDR scenes in uniform illumination with a shaded lens.

9015-32, Session 7

The bright future of metameric blacks *(Invited Paper)*

Philipp Urban, Fraunhofer Institute for Computer Graphics Research IGD (Germany)

No Abstract Available

9015-33, Session 7

Can color management and anaglyph 3D images be friends once more? *(Invited Paper)*

Andrew J. Woods, Curtin Univ. (Australia)

No Abstract Available

9015-34, Session 7

Feeling edgy about color blindness *(Invited Paper)*

Reiner Eschbach, Stephen C. Morgana, Xerox Corp. (United States); Anna Quaranta, Cristian Bonanomi, Alessandro Rizzi, Univ. degli Studi di Milano (Italy)

No Abstract Available

Conference 9016: Image Quality and System Performance XI

Monday - Wednesday 3 –5 February 2014

Part of Proceedings of SPIE Vol. 9016 Image Quality and System Performance XI

9016-1, Session 1

Just noticeable differences in perceived image contrast with changes in displayed image size

Jae Young Park, Sophie Triantaphillidou, Ralph E. Jacobson, Univ. of Westminster (United Kingdom)

Previous studies concerning the identification of image attributes that are most affected by changes in the displayed image size have indicated that sharpness and contrast are the most significant. In previous work we evaluated the just noticeable difference in perceived sharpness when changing displayed image size. In this paper an evaluation of the degree of change in the perceived image contrast with respect to changes in displayed image size was carried out. This was achieved by collecting data from a series of psychophysical investigations that used techniques to match the perceived contrast of displayed images of five different sizes.

The paper first describes a method employed to create a series of S-shaped filters for contrast manipulation. The filter's shape is based on the phenomenon of simultaneous lightness contrast. A total of twenty-four 6th order polynomial S-shape functions were applied to every original test image to produce test images with different contrast levels. The objective contrast related to each function was evaluated from the gradient of the mid-section of the curve (gamma). The manipulation technique took into account published gamma differences that produced a just noticeable difference (JND). The filters were designed to achieve approximately half a just noticeable difference in perceived contrast between images viewed from a certain distance and having a certain displayed image size, whilst keeping the mean image luminance unaltered. The processed images were then used as test series in a contrast matching experiment, exploring the perceived changes in contrast with changes in five different displayed image sizes.

For the creation of a test image database, a Canon EOS 30D digital SLR camera, with a quality zoom lens was used for recording natural scenes with varying scene content under various illumination conditions. In the psychophysical investigation, a total of sixty-four original test-images were selected from the database; they were resized using bi-cubic interpolation to five different sizes (372 x 280 pixels, 449 x 338 pixels, 526 x 396 pixels, 635 x 478 pixels and 744 x 560 pixels). The smallest size was based on prevalent image dimensions of LCDs of DSLR. The medium-size and large-size images were 2 times and 4 times larger respectively than the small-size images, covering 9.0% and 18.1% of the display. Observers were asked to match each time the perceived contrast of two images of the same scene, displayed at different sizes. They adjusted the contrast of the small version to match the contrast of the larger reference image, using a slider-bar. Each step in the slider-bar range corresponded to a different filtered image, i.e. to approximately half a just noticeable difference in contrast. A total of twenty expert and non-expert observers took part in the contrast matching experiment.

Validation of the just noticeable difference step was carried out by a second psychophysical investigation, involving a forced choice paired comparison and using all sixty-four scenes and three expert observers.

Results showed that,overall, the perceived contrast increased proportionally with decreased image size. The degree of change in perceived image contrast between images of different sizes varied with original scene content, but not considerably. The average just noticeable difference between large and medium-large was 0.22, between large and medium was 0.27, between large and medium-small was 0.31 and between large and small was 0.54.

9016-2, Session 1

The subjective importance of noise spectral content

Donald J. Baxter, STMicroelectronics Ltd. (United Kingdom); Jonathan B. Phillips, NVIDIA Corp. (United States); Hugh Denman, Google (United States)

The influence of noise on subjective quality depends on the magnitude, spectral content and type of the noise. There are many studies on the human visual system's sensitivity to noise. Missing is a quantitative subjective study about the spectral frequency content of luminance and chrominance noise in the context of consumer photography content. This is required to validate Camera Phone Image Quality (CPIQ) Initiative Visual Noise metric [1] quality mapping function in terms of Overall Quality Loss JNDs. The closest subjective studies are by Keelan [1] and Johnson [3]. The Keelan study does not include noise spectral content variation. The band pass filtering in the Johnson study is not representative of low frequency chrominance noise, Color Mottle. One unanswered question is the relative importance of luminance and chrominance noise. This paper describes a subjective pilot study probing 3 axes of noise—namely amplitude, spectral content and noise type—and the lessons learned for the full visual noise subjective study.

The subjective pilot study used the ISO 20462 softcopy copy ruler protocol to yield the required subjective ranking in secondary Standard Quality Scale (SQS) JNDs and to enable correlation with Keelan's results. For the described study, a noise simulator was designed. The simulator models the image sensor's noise characteristics, and applies variable frequency-based noise shaping in the AC1C2 opponent color space. The frequency filtering is selectable between a low pass Butterworth filter and a Log Gabor band pass filter. Four softcopy ruler images scenes were selected as a baseline.

The pilot study contained 4 different categories of noise masks:

1. The electron level series
2. The low pass filtered series
3. The band pass filtered series
4. Aptina noise study correlation images.

The aim of the pilot study was to explore the wider experimental space in order to determine the key items for the final full visual noise subjective experiment.

The electron level series has no frequency filtering applied. The noise level is controlled by setting the maximum sensor signal in electrons. This is equivalent to changing the image sensor's analogue gain (ISO) setting. The advantage of this definition is a direct relationship with the camera's specification.

The low pass filter series contains 3 different axis of experimentation: 1) two different noise levels to the access the sensitivity to the noise amplitude at different strengths of filtering, 2) low pass filter cut off frequencies of 2, 4, 8 and 16 cycles/degree (CPD), and 3) three different noise types: luminance only noise, chrominance only noise and both luminance and chrominance noise.

The band pass filter series is motivated the Johnson study [3]. The original objective was to provide insight about the shape of contrast sensitivity function (CSF) under the softcopy ruler experimental conditions. This series involved 5 different band pass filter center frequencies (2, 4, 8, 16, and 32 CPD) and 2 different noises types(luminance only noise and chrominance only noise).

The final correlation series contained the luma only and RGB noise images kindly provided by Aptina from the Keelan noise study. The aim was to re-confirm the Keelan results and provide a correlation point.

A total of 4 observers from NVIDIA and STMicroelectronics participated in rating the 160 images. For the correlation images, the average rating

had the same trend as the Keelan study but with a different absolute level. As Figure 1 shows, the results are noisy due to the small number of observers. Participant feedback indicated that the characteristics of the band pass filtered images were not typical of cell phone cameras. Figure 4 shows a CSF-like relationship versus spatial frequency of noise with a higher sensitivity to luminance rather than chrominance noise.

The full visual noise subjective study is currently in progress at Google, NVIDIA and STMicroelectronics and results will be shared. For the full study, the band-pass filter series has been removed and the number of ruler scenes increased to 6. The scene selection includes both images rich in detail and images with significant uniform regions to probe variability in scene susceptibility to noise.

REFERENCES

[1] Baxter, D.J., and Murray, A., "Calibration and adaptation of ISO visual noise for I3A's Camera Phone Image Quality initiative," Proc SPIE 8293, (2012)

[2] Keelan, B.W., Jin, E.W and Prokushkin, S., "Development of a perceptually calibrated object metric of noise", Proc SPIE 7867, p786708 (2011)

[3] Johnson, G.M., and Fairchild, M.D., "The Effect of Opponent Noise on Image Quality," Proc SPIE 5668, p 82-89 (2004)

9016-3, Session 1

Spatial contrast sensitivity and discrimination in pictorial images

Sophie Triantaphillidou, John Jarvis, Gaurav Gupta, Univ. of Westminster (United Kingdom)

This paper describes continuing research concerned with the measurement and modelling of human spatial contrast sensitivity and discrimination functions using complex pictorial stimuli, with and without the presence of noise. In last year's conference we discussed the contrast detection and discrimination model frameworks developed by Barten and showed that they have a potential to provide a sound starting position for our modelling purposes. We presented progress in the choice of contrast metrics for defining contrast sensitivity; apparatus, laboratory set-up and imaging system characterization; stimuli acquisition and stimuli variations; spatial decomposition; methodology for subjective tests and initial observations. Based on this work, we have now refined the experimental paradigm and we measure the following visual response functions:

i) Isolated Contrast Sensitivity Function (iCSF), describes the ability of the visual system to detect any spatial signal in a given spatial frequency band (visual octave) in isolation and is the closest equivalent to a conventional sinewave CSF. For a given frequency the isolated contrast sensitivity is referred to as i-detect.

ii) Contextual Contrast Sensitivity Function (cCSF), describes the ability of the visual system to detect a spatial signal in a given octave contained within an image. A comparison between isolated and contextual band conditions indicates the extent that spatial information outside of the band of interest acts as a source of signal masking (noise). For a given frequency the contextual contrast sensitivity is referred to as c-detect.

iii) Isolated Visual Perception Function (iVPF), describes visual sensitivity to changes in suprathreshold contrast of any spatial signal in a given spatial frequency octave in isolation. For a given frequency the isolated contrast discrimination is referred to as i-discrimination.

iv) Contextual Visual Perception Function (cVPF), describes visual sensitivity to changes in suprathreshold contrast in an image. For a given frequency the contextual contrast discrimination is referred to as c-discrimination.

We have experimented with images of a fixed angular field size, under free-view conditions. The experimental set-up is based on a 2AFC paradigm, where standard and test stimuli are presented with a brief temporal separation. Stimuli include three pictorial images, with different scene content and image spectra, two different contrast levels per image and three levels of dynamic (shot) noise. Results, so far, from all four experimental conditions indicate the following:

Measurements of i-detect produce a spatial iCSF which closely follows the expected profile and magnitude of a sine-wave CSF appropriate for the stimulus size and mean luminance employed. This is not unexpected, since the stimuli used for i-detect experiments contain pictorial information only from a narrow band (not dissimilar to images of gratings), on an otherwise uniform background. The measured cCSFs are found to be lower than the iCSFs and flatter in profile. iVPFs, cVPFs and cCSFs are also shown to be similar in profile.

Barten's contrast sensitivity function model is shown to successfully predict iCSF. For a given frequency band, the reduction or masking of c-detect sensitivity compared with i-detect sensitivity is predicted from the linear amplification model (LAM). This model indicates that the masking is generated primarily from pictorial information contained within a plus and minus one octave frequency range from the band itself. If, for a given band, c-detect is known, then our previously described discrimination model (an extension of Barten's discrimination model) offers a good prediction of the observer's ability to discriminate spatial information in a real complex scene (c- discrimination).

The ability of the LAM to predict visual sensitivities in the presence of noise is currently under examination.

We conclude with a discussion of the implications of the measured and modelled profiles of cCSF and cVPF to image quality modelling.

9016-5, Session 1

Evolution of slanted edge gradient SFR measurement

Don Williams, Image Science Associates (United States); Peter D. Burns, Burns Digital Imaging (United States)

The well-established Modulation Transfer Function (MTF) is an imaging performance parameter that is well suited to describing certain sources of detail loss, such as optical focus and motion blur. As performance standards have developed for digital imaging systems, the MTF concept has been adapted and applied as the spatial frequency response (SFR). However, since being introduced as an ISO standard for measuring image resolution devices more than a decade ago, the slanted edge-gradient analysis protocols of ISO 12233 and 16067-1 have evolved. Out of necessity, practitioners modified minor elements of the standard method to suit specific camera characteristics, unique measurement needs, or computational shortcomings in the original method. Some of these adaptations have been documented and benchmarked, but a number have not. In this paper we describe several of these modifications, and how they have improved the reliability of the resulting system evaluations.

Our approach borrows from statistics as we describe how the slanted-edge SFR method can be thought of as a way of estimating the underlying imaging performance parameter. As with any estimate, careful sampling of available observations is important. In our case this involves intelligent region-of-interest (ROI) selection around the edge feature to be used. A second notion is bias error that is introduced by characteristics, such as optical distortion, which conflict with underlying assumptions of the method. We describe how the influence of such spatial effects can be minimized by improved edge detection. Following a discussion of how elements of the SFR method have been adapted to other imaging characteristics, such as optical flare and edge-raggedness, we describe ways in which the method might be made more useful in the future, and outline several current challenges.

9016-6, Session 2

Evaluation of perceptual resolution of printed matter (Fogra L-Score evaluation)

Thomas Liensberger, BARBIERI electronic snc (Italy); Andreas Kraushaar, Fogra-Forschungsgesellschaft Druck e.V. (Germany)

The evaluation of perceived image quality in prints is an active field of research. Definitions of measurements of print quality attributes that correlate with visual perception by technology-independent means, even across many printing technologies, is under current scrutiny. It is influenced by a number of different quality attributes that are, for convenience only, categorized into colour and surface finish, homogeneity, perceptual resolution and artefacts.

ISO 29112 states that printer resolution, a quantification of the ability of a digital printing system to depict fine spatial detail, is a perceptually complex entity with no single, simple, objective measure. It further defines 5 print quality characteristics namely: native addressability, effective addressability, edge blurriness, edge raggedness, and the printing system modulation transfer function (MTF) that somehow contribute to the perceived resolution. However previous research at Fogra indicated, that a test target comprising a reasonable mount of patches, each of which contains spatial patterns (e. g. concentric circles) of varying line width and contrast, provides a very high correlation with the perceived resolution.

In this case the RIT Contrast-Resolution Target was tested. The performed psychophysical experiments were done with print samples produced on different printing systems covering offset lithography, electrophotography and inkjet. The question asked to the observer was: “Mark and count the circles where you see the circles – without distracting artefacts such as aliasing patterns”. The inter- and intra-observer variability was found to be very consistent. The correlation of the so derived index (from 0 “worst” to 100 “perfect”) correlated very well with the results of a rank order experiment based on more than 20 print samples asking to sort all prints by means of their highest perceptional resolution, i.e. details sharpness. This paper addresses the objective assessment of this index by means of a scanner-based method followed by a computer-aided evaluation.

The underlying idea of the objective assessment is to scan first the printouts with a qualified, colorimetrically characterized scanner (e.g. ICC scanner profile) preferably as CIELAB-TIFF. A following image check for further processing (sizing, rotation, artefacts) has to be performed before applying a spatial (Gaussian) filtering to the lightness channel that corresponds with the human visual system response (at 50cm – viewing distance). A normalized 2D cross correlation between the computed test target (TIFF-file) and the processed image (TIFF-file) is the core of the algorithm. After applying statistical evaluation such as threshholding or heuristical plausibility check a correlation between the findings and the results of psychophysical experiments (counted circles) will be done. The final step is a transformation (bracing) of the data into a reasonable scale (L-scale: 0 – 100).

In order to develop and test the algorithm (computer program) a training set of representative samples was needed that span the entire range of the perceptual continuum (of interest). Here 20 print samples from Fogra database have been selected that were printed with inkjet and electrophotographic printing systems on different kinds of media like PVC, paper or textile. The ground truth (termed here “counting index”) has been derived by performing psychophysical experiments where 20 observers rated the 20 print samples to derive percentiles for each of the 100 circles. It was found that the majority of the counting indices ranged from 40 to 80. In order to establish a meaningful psychophysical scale (“attribute JND”) that facilitates the range from 0 to 100 a transformation has been used. Here a simple cumulated Gaussian distribution was used since it is limited between 0 and 1 and provides a good correlation with the experimental findings. It must be noted that the L-Score does not reflect the simple marking and cumulating of the circles anymore. The correlation coefficient between the L-Score (“Matlab Evaluation”) and the visual data can be considered high (correlation = 0,978).

The performance of the L-Score was validated by means of a new set of 5 print samples that were not part of the training data. Here the prints were measured as defined by the L-Score and the findings were compared with the visual findings derived with 5 observers only. The correlation can also be considered reasonably high (correlation = 0.987). In addition a rank order experiment was conducted and it was found that the L-Score always corresponds to the mean ranks.

Based on an existing test target (RIT Contrast-Resolution Test Target) an abridged method for determining the perceptual resolution termed “L-Score” of printed matter was developed and will be proposed for further analysis and inclusion of ISO 15311 and ISO 18621 to be part of the system check evaluation.

9016-7, Session 2

Automated measurement of printer effective addressability

Brian E. Cooper, Ahmed H. Eid, Edward E Rippetoe, Lexmark International, Inc. (United States)

1. Introduction

1.1 ISO TS 29112

Effective addressability is one of five characteristics defined in ISO/IEC TS 29112, a technical specification that describes the objective measurement of printer resolution for monochrome electrophotographic printers. Other characteristics include edge definition (defined by edge blurriness and edge raggedness), spatial frequency response (defined by the printer’s modulation transfer function), and native or physical addressability. Because printer resolution is a broad, complex quality that defies a single, objective measure, this collection of attributes seeks to quantify its various facets.

To facilitate the measurement of these attributes, ISO/IEC TS 29112 describes a procedure for using an ordinary reflection scanner to quantify most of the attributes. (Native addressability is the exception, requiring visual analysis.) In addition, the specification offers measurement algorithms implemented in Matlab.

Thus, the effective addressability measurement procedure offers an automated way to experimentally determine the addressability of the final, printed output.

1.2 Definition

Since the vast majority of printers are digital devices, addressability is a key consideration. Addressability defines the maximum number of spots or samples within a given distance, independent of the size of the spots when printed. The native or physical addressability is the addressability that the user can access directly, commonly reported by the rendering firmware of the printer. However, some printers may support a higher addressability below the control of the user. Thus, effective addressability is the addressability demonstrated by the final, printed output. It is the minimum displacement possible between the centers of printed objects.

2. Test target design

Separate test targets exist for each the horizontal and vertical effective addressability, although they follow the same design. The test target contains two types of elements, repeated to fill the page: fiducial lines and line segments. The fiducial lines are longer, extending vertically across the page. Between the fiducial lines are several shorter line segments. The position of each line segment shifts incrementally in the horizontal direction across the page. In all cases, the lines should be thick enough to render reliably on the given printer.

The fiducial lines serve as a relative reference for the incremental displacements of the individual line segments, providing a way to tolerate larger-scale physical distortions in the printer, such as skew or misalignment.

The test target automatically adapts to the page size reported by the printing device. It also selects an initial resolution for the target that is twice the native addressability reported by the printer. The user can

change the target resolution, if needed, as explained in the measurement section below.

3. Measurement procedure

The user should place the printed test target onto the flatbed of a scanner, after physically rotating the printed page at approximately ±45°. The exact angle is not critical, but the rotation allows the scanner to measure an effective addressability beyond the scanning resolution.

The algorithm binarizes the scanned image and ignores any objects that are too small or too large. It computes the relative distance dx_i between each line segment i and its corresponding fiducial line. Then it computes the difference between each successive pair of relative distances:

Dx_i = |dx_i - dx_{i-1}|

The distribution of Dx_i values provides the estimate of effective addressability.

If the printer can properly render the target resolution, then all line segments will shift by this incremental amount. In this case, the histogram will be unimodal, with a mean equal to one divided by the target resolution.

On the other hand, suppose that the target resolution exceeds the printer's capabilities. Some of the line segments will shift by the printer's actual addressability, while others will not shift. The distribution of Dx values is bimodal. The mean of the second (non-zero) peak indicates the effective addressability. If the histogram is unimodal, then the test is inconclusive. The printer successfully achieved the target resolution. Thus, the user must increase the target resolution (e.g., double it), reprint a new sample, and repeat the measurement procedure.

4. Automatic histogram analysis

Although visual interpretation of the histograms is always recommended, the algorithm automatically estimates whether the histogram is unimodal or bimodal.

First, the algorithm applies the Lilliefors statistical test for unimodal, normal distributions. If the distribution fails this test, then the algorithm fits the histogram to a Gaussian mixtures model with two classes. The mean of the second (non-zero) peak is the estimated effective addressability.

Ideally, the peak or peaks are narrow and distinct. The algorithm reports the 90% confidence interval of the non-zero peak to indicate the quality of the measured histogram. For bimodal distributions, it also computes the percentage of overlap between the two peaks. Experimentally (based on comparative visual analysis), an overlap of roughly 5% or less clearly indicates a bimodal distribution. Several legitimate bimodal distributions occurred above this limit, but the choice was not always clear.

5. Results

While refining the test target and the measurement procedure, we conducted several experiments. After finalizing the test target, we conducted two round-robin experiments with the ISO committee plus an evaluation of approximately 80 samples across 20 printers.

In all but one case, the measured effective addressability matched the resolution reported by the printer's RIP. One printer included in the 2010 ISO round-robin evaluation reported 600x2400 dpi resolution. However, the effective addressability measurement application showed a unimodal histogram, even with a 4800 dpi target. Without access to this printer, we could not repeat the measurement with a 9600 dpi target. However, measurements with a high-resolution camera confirmed valid 4800 dpi addressability in the available sample, suggesting valid behavior for the measurement procedure.

Four printers (including the one above) showed a different addressability (both reported by the RIP and measured experimentally) in the cross-track and in-track directions. Otherwise, all other printers had the same addressability in both directions.

For the entire set of approximately 100 samples, we are currently conducting a more detailed evaluation of the measured effective addressability. In addition to the data described previously relating to the quality of the estimation, we are measuring the consistency of the addressability.

9016-8, Session 2

Perceptual metrics and visualization tools for evaluation of page uniformity

Minh Q. Nguyen, Purdue Univ. (United States); Renee Jessome, Stephen Astling, Eric Maggard, Terry Nelson, Mark Shaw, Hewlett-Packard Co. (United States); Jan P. Allebach, Purdue Univ. (United States)

Uniformity is one of the issues of most critical concern for laser electrophotographic (EP) printers. Typical non-uniformity defects include mottle, grain, pinholes, and finger prints. Among these defects, mottle (low spatial frequency) and grain (high spatial frequency) are the most commonly observed. Those defects may occur locally or globally within a letter-size printed page with distributions that are random. It is a real challenge to make a Print Quality (PQ) assessment due to the large coverage of a letter-size, constant-tint printed test page and the variety of possible non-uniformity defects. For instance, a printed page with 95% of good coverage and 5% of damage could receive an overall score from an expert observer as bad.

In order to tackle the problem of quality assessment, we propose, in this paper, a novel method that uses a block-based technique to analyze the page both visually and metrically. With a print-to-scan method, we use a grid of 150 pixels x 150 pixels (? inch x ? inch scanned at 600 dpi) square blocks throughout the scanned page. For each block, we examine two aspects: behavior of its pixels within the block (metric of graininess) and behavior of the blocks within the printed page (metric of uniformity). Both delta E (CIE 1976) and the L* lightness channel are employed. For an input scanned page, we create eight visual outputs, each displaying a different aspect of uniformity by using various color codings. Our method has received very positive feedback from engineers charged with printer development in terms of its ability to facilitate a better understanding of the presence of defects of uniformity and graininess on the printed page.

In order to apply machine learning to predicting PQ, we train scanned pages of different 100% solid colors: cyan, magenta, red, and green separately by using the support vector machine (SVM) algorithm. We observed that using the lightness solely results in a 10% higher accuracy than using delta E. This is due to the fact that the spatial frequency contrast sensitivity of the human visual system is higher for the lightness channel than for either of the two chrominance channels (red-green and blue-yellow). We use four metrics as features for the SVM: variation of page lightness, variation of page graininess, average of page graininess and maximum of page graininess. Our results show that we can predict, with 90% accuracy, the assignment by a human expert of one of the four grades (A, B, C, and D where A is the best grade and D is the worst) of uniformity to the print. Hence, our work shows promise to be applicable to automatic PQ assessment.

9016-9, Session 2

On the analysis of wavelet-based approaches for print mottle artifacts

Ahmed H. Eid, Brian E. Cooper, Lexmark International, Inc. (United States)

Print mottle is one of several attributes described in ISO/IEC DTS 24790, a draft technical specification for the measurement of image quality for monochrome printed output. It defines mottle as aperiodic fluctuations of lightness less than about 0.4 cycles per millimeter, a definition inherited from the latest official standard on printed image quality, ISO/IEC 13660. In a previous publication, we introduced a modification to the ISO/IEC 13660 mottle measurement algorithm that includes a band-pass, wavelet-based, filtering step to limit the contribution of high-frequency fluctuations including those introduced by print grain artifacts. This modification improves the algorithm's correlation with the subjective evaluation of experts who rated the severity of printed mottle artifacts.

Seeking to improve upon the mottle algorithm in ISO/IEC 13660, the ISO 24790 committee evaluated several mottle metrics. This led to the selection of the above wavelet-based approach as the top candidate algorithm for inclusion in a future ISO/IEC standard. Recent experimental results from the ISO committee showed high correlation between the wavelet-based approach and the subjective evaluation conducted by the ISO committee members based upon 25 to 30 samples covering a variety of printed mottle artifacts. In addition, we introduce an alternative approach for measuring mottle defects based on spatial frequency analysis of wavelet-filtered images. Our goal is to establish a link between the spatial-based mottle (ISO/IEC DTS 24790) approach and its equivalent frequency-based one in light of Parseval's theorem. Our preliminary experimental results showed a high rank correlation between the spatial and frequency based approaches.

9016-10, Session 3

MFP scanner diagnostics using a self-printed target to measure the modulation transfer function

Weibao Wang, Purdue Univ. (United States); Peter Bauer, Jerry K. Wagner, Hewlett-Packard Co. (United States); Jan P. Allebach, Purdue Univ. (United States)

In the current market, reduction of warranty costs is an important avenue for improving profitability. Our goal is to develop an autonomous capability for diagnosis of printer and scanner caused defects with mid-range laser multifunction printers (MFPs), so as to reduce warranty costs. In addition to reduction of warranty costs, the results of our work have the potential to increase customer satisfaction by allowing them to more quickly resolve their print quality issues with less effort on their part. If the scanner unit of the MFP is not performing according to specification, this issue needs to be diagnosed. If there is a print quality issue, this can be diagnosed by printing a special test page that is resident in the firmware of the MFP unit, and then scanning it. However, the reliability of this process will be compromised if the scanner unit is defective. Thus, for both scanner and printer image quality issues, it is important to be able to properly evaluate the scanner performance. In this paper, we consider evaluation of the scanner performance by measuring its modulation transfer function (MTF). The MTF is a fundamental tool for assessing the performance of imaging systems. It has been applied to a range of capture and output devices, including printers and scanners. Several ways have been proposed to measure the MTF, all of which require a special target, for example a slanted-edge target or a sine wave target. It is unacceptably expensive to ship every MFP with such a standard target, and to expect that the customer can keep track of it. To reduce this cost, in this paper, we examine two different approaches to this task. They are based, respectively, on a self-printed grille pattern target and a self-printed slanted-edge target. All the targets are printed using the laser printer in the MFP. We choose these two kinds of targets because they are easy to generate and print by a laser printer compared to a sine wave target. Also both approaches have their comparative advantages and disadvantages. For example, the slanted-edge method is a single measurement; while instead, using a grille pattern target need several measurements on different frequencies in order to get the MTF. For using a self-printed grille pattern target to calculate the MTF, we use two different methods, one is based on Fourier analysis; and the other is based on contrast, which means that the MTF is the contrast at a given spatial frequency relative to low frequencies. For using a self-printed slanted-edge target to calculate the MTF, we design and print a black and white slanted-edge target, which is similar to the standard QA-62 slanted-edge target. There are several aspects that will impact the results for the measurement of MTF using a self-printed slanted-edge target compared to using a standard target. So we make the target black and white to exclude the printing halftone patterns. Then we propose algorithms to improve the results using a self-printed slanted-edge target by preprocessing the edge and the black area in the target. Finally, we present experimental results for MTF measurement using self-printed targets and compare them to the results obtained with standard targets using two different scanners.

9016-11, Session 3

High-performance automatic cropping and deskew of multiple objects on scanned images

Ilya Kurilin, Samsung Advanced Institute of Technology (Russian Federation); Ilia V. Safonov, Nokia R&D Ctr. (Russian Federation); Michael N. Rychagov, Samsung R&D Institute (Russian Federation); Ho Keun Lee, Sang Ho Kim, Samsung Digital City (Korea, Republic of)

Sometimes user needs to scan several small documents/objects such as receipts, tickets, business cards, postal cards at one pass and to save each document in an individual file with skew correction. There are a number of issues in fully automatic segmentation and cropping of multiple objects. The first one is dealing with scanning bright or white objects. Such objects on white background of scanner are fragmented to pieces due to segmentation. Second issue results from closely placed or overlapped objects on scanner platen. Closely placed objects are often recognized as single one. To prevent such situation the limitation of minimal distance between objects is declared. Another issue relates to shading effect on scanning area boundaries which can be the cause of wrong segmentation.

Workflow of proposed algorithm includes automatic segmentation and subsequent skew correction.

Segmentation of multiple objects on a scanned image comprises of several basic stages. First stage is preprocessing. During preprocessing an input image is resized to 75 dpi to be independent from scanning resolution and enhanced for emphasizing scanned objects on white background. The resizing step is important for decreasing of computational complexity of segmentation algorithm as well. Second stage is segmentation itself. Downsized RGB image is converted to binary one by thresholding and then labeling of connected regions is applied. In our approach, first and second stages are combined in one-pass procedure to improve computational performance of the algorithm. After that, some objects are fragmented and represented by plurality of speckles, fragments of text, line segments, etc. Third stage applies classification of connected regions in order to merge fragmented objects by means of new heuristic procedure. Four classes of regions are supposed: foreground regions with rectangular shape; foreground regions with shape of line segments; separate area between rectangular regions and other non-rectangular foreground connected regions. The classified regions are arranged by area in decreasing order. Large rectangular objects are considered as most reliable and needless in merging with other closely placed regions. For that, separating areas are calculated between rectangular regions and line segments with different skews. We suppose that separating areas cannot be included into any regions group during regions merging. Such assumption allows preventing erroneous merging of closely placed rectangular regions. Region merging process starts in decreasing order from rectangular regions, then continues from line segments and proceeds until all adjacent regions are merged. Finally, coordinates of bounding boxes and skew angle are calculated for each group.

We suggest quantitative quality criteria for comparison of proposed approach with competitors. Criteria are based on calculation of errors numbers. We take into account two possible error types related to object skew estimation and boundary box detection. Test set consisted of 20 images of A4 size, 300 dpi scanning resolution. About half of test images contained several photos. Other images included documents more complex for segmentation: business cards, receipts, boarding pass, etc.

Comparison results demonstrate that quality of segmentation and deskew outperforms similar feature in competitors' scanning software. The average processing time is about 0.2 s on modern quad-core PC due to algorithmic optimization and parallelization by means of OpenMP. The proposed algorithm provides users with additional functionality and comfort. In addition the technique can be adopted for processing of document images captured by camera of mobile devices.

9016-12, Session 3

Visual quality of printed surfaces: study of homogeneity

David Nébouy, Mathieu Hébert, Thierry T. Fournel, Lab. Hubert Curien (France); Jean-Luc Lesur, Gemalto (Gémenos) (France)

The present work proposes a first attempt of automated procedure to estimate one relevant attribute for the quality of printed surfaces: the homogeneity of tones, i.e. the homogeneity of surfaces on which is printed an originally uniform digital color image. In contrast with other methods proposed in the literature [1-3], we make an important distinction between printing quality, which is related to the degradation of the original digital image when transferred to the paper or plastic support, and image quality, which includes the perceived quality of the original image in addition to the printing quality. Moreover, we intend to link objective data assessment with subjective judgments. Our approach for assessing the homogeneity of prints first consists in the digitalization of the printed surface using a high resolution scanner, then in processing the scanned image in order to compute a homogeneity index. The method for computing this homogeneity index relies on three main steps:

1) We perform some preprocessing comprising image resizing and filtering operations in order to simulate near vision (represented by a high resolution image), and far vision (low resolution image). This multi-scale vision approach, as well as the conversion of the RGB colors into CIELAB colors and low-pass filtering in order to simulate the human vision spatial bandwidth intend to transform the raw scanned image into several new ones, closer to human perception, which will be the starting point for the computation of the homogeneity index. After the preprocessing steps, four different gray-level images are obtained: two luminance images and two hue images, both corresponding to high and low resolutions.

2) Since the scanned image of the printed copy of a uniform colored patch can be considered as a texture, we chose to use a classical tool in texture characterization: the computation of Haralick's parameters from a gray-level co-occurrence matrix (GLCM), also called gray-tone spatial-dependence matrix. In contrast with first-order statistic tools (histogram, mean, standard deviation, skewness, kurtosis...) which are representative of gray-level distribution of the pixels in the image regardless of their spatial arrangement, the GLCM is a second-order statistic tool involving two pixels simultaneously. It represents the variations in gray level values between pixels distant from each other of a characteristic distance in horizontal and vertical directions (these two directions have been specially considered because they coincide with the orientations of classical defects of most printing systems). For each of the gray level images generated by the preprocessing step, we compute a GLCM. Four matrices are thus obtained, then normalized and we compute from each one a value inspired of Harlick's homogeneity parameter where the sensitivity gray-level variations in the image can be tuned.

3) The combination of the four values obtained from the four images, where a double weight is attributed to the luminance images (as it is commonly done in video compression), yields our homogeneity index.

In order to verify that this index is consistent with visual perception, we compare it with the empirical evaluation by observers based on experiments. We printed several versions of a same digital image on different supports (office paper, glossy photo paper, calendered paper, white polymer), with different printing techniques (laser, inkjet, retransfer) and we asked observers to place them on a graduated line in respect only to their perceived homogeneity. The index computed according to the method that we propose for the assessment of the homogeneity of printed colors is in accordance with the empirical assessment by the observers. In this sense, the algorithm that we have developed seems to be consistent with the human visual perception. This homogeneity index is a first attribute which should be combined with other attributes in order to get a global quality score for printed natural images. It may also help into comparing the performance of different printing systems in terms of visual quality.

[1] M. Pedersen, N. Bonnier, J. Y. Hardeberg et al., "Attributes of a New Image Quality Model for Color Prints," Color and Imaging Conference, 2009(1), 204-209 (2009).

[2] M. Pedersen, and S. A. Amirshahi, "Framework for the Evaluation of Color Prints Using Image Quality Metrics," Conference on Colour in Graphics, Imaging, and Vision, 2010(1), 75-82 (2010).

[3] M. Pedersen, Y. Zheng, and J. Hardeberg, [Evaluation of Image Quality Metrics for Color Prints] Springer Berlin Heidelberg, 30 (2011).

9016-13, Session 3

A computational texture masking model for natural images based on adjacent visual channel inhibition

Yucheng Liu, Jan P. Allebach, Purdue Univ. (United States)

Banding, as a common print defect has received a lot of attention from researchers. It differs from the common compression type image defects and appears as band-like color inconsistency varying in either high or low frequency, depending on the width of the bands, along the processing direction of the printer. Banding comes in a variety of intensities and patterns, and gives rise to a perceptually annoying defect that significantly threatens the quality of prints. For a high-end commercial printer like the HP Indigo Press, quality control is an important and yet costly issue. On one hand, customer has a much higher expectation and lower defect tolerance than for regular home or office printers; On the other hand, a fully automatic print quality evaluation tool is still not available, meaning that extra human labor is needed for print quality control, which induces high additional cost. An automatic perceptual evaluation tool is highly in desired.

Many researchers have reported results on diagnostic banding evaluation. But those works are mostly on the perceptibility of singular bands on constant grey-scale or color patches. However, in the real application setting, we are almost always working with natural images, text, and graphics in marketing collateral, photo-albums, and other documents. For natural image, there are some mature off-the-shelf perceptual image qualities metric like VIF, SSIM, and some improved strategies to tackle the sub-pixel misalignment problem for comparison between print and digital image, like SP-SSIM. Unfortunately none of them are optimized for such special defect structures like banding; and none of them produce an accurate perceptual quality prediction for a banding contaminated image. In this work, we develop a perceptual evaluation tool designed specifically for banding defects in the presence of customer content.

In our work, we consider two factors that affect the banding visibility: color and contrast masking. And accordingly, our method consists of two parts, each addressing one factor described above.

For the color dependent perception, we already have the well-known perceptually uniform color space CIELAB, and it would seem straightforward for us to just assume the same isotropic color change in this color space gives rise to an equivalent amount of perceptual difference. However, this color space is essentially based on low frequency color matching experiments conducted on uniform patches. Thus the same conclusion doesn't necessarily apply to a structural color variation like banding. To obtain data in support of this argument, we generate artificial banding of controlled pattern and severity on constant color patches with our defect simulator. The colors on which we generate the banding defects sample the whole RGB color space with reasonable intervals. The patches are evaluated by both a banding metric and human subjects in a psychophysics experiment. We first acquire a 1-D projected signal by averaging the luminance of the patch along the scan direction. Then the banding metric looks for band-like response in the 1-D projected signal in multiple scales followed by a cross scale analysis to find out the perceptual significant band. Finally a quality score is obtained by a 1-D pooling operation. In the psychophysics experiment, the subjects are asked to rate the overall quality of the patches. And by comparison of the quality score from the metric and the overall subject rating from the psychophysics experiment, we obtained a banding defect severity gain for the sample colors.

Contrast masking is another well-known mechanism that accounts for the fact the local smoothness, texture, and orientation of the content

affect the visibility of the defects. We model the response of human visual system with a gain control model. The inputs to the model are sub-band coefficients obtained by performing steerable pyramid filter decomposition on both the reference image and defect image. Since the banding defect is a highly oriented print defect with color variation just in one direction, we just look at the neural response for this specific orientation. This ensures that our model response more specifically for banding, and is also good for reducing model complexity. However, coefficients of all orientations will still be included for the local inhibitory pool, which models the masking effect of the neighboring pixels. Parameters in this model include weights for the sub-band coefficients representing the contrast sensitivity function of human visual system, and the non-linearity terms for the neural response and inhibitory pool. These parameters are further refined through a training process. The ground truth is obtained from a second round of psychophysics experiment, in which subjects are asked to mark locally the severity of the banding. We evaluate the performance of our model with sample images composed of a variety of typical commercial print content.

9016-14, Session 4

JPEG ringing artifact visibility evaluation

Sirui Hu, Zygmunt Pizlo, Jan P. Allebach, Purdue Univ. (United States)

No Abstract Available

9016-15, Session 4

Mobile phone camera benchmarking: combination of camera speed and image quality

Veli-Tapani Peltoketo, Sofica Ltd. (Finland)

When a mobile phone camera is tested and benchmarked, the significance of quality metrics is widely acknowledged. Color reproduction, sharpness and noise metrics are usually defined as the most significant quality parameters of the camera. Also several other quality metrics are available, for example IEEE P1858 working group has defined the lens distortion and lateral chromatic aberration as benchmarking components. However, the speed or rapidity metrics of the mobile phone's camera system have not been used with the quality metrics even if the camera speed has become more and more important camera performance feature.

Capturing the moment is an essential requirement in modern mobile phone cameras and it requires a great performance and speed from camera systems and pushes camera developers to find new innovative breakthroughs to fulfill users' needs. However, features like great pixel count, image stabilization, auto focus, auto exposure and auto white balance together with different mixtures of hardware and software based image pipelines and very complex image defect correction generates different and unpredictable delays to the image capturing. An excellent image quality may be ignored, if the camera is always too slow to capture the needed action. It is also notable that according to statistical analysis of large image databases, over 70% of digitally captured images contain human faces. Therefore, different face, smile and blink detection algorithms may cause extra delays to the image capturing pipeline.

As well as quality benchmarking metrics, there are existing methods to evaluate the camera speed. For example, very recently accepted standard ISO 15781:2013 defines several methods to evaluate the shooting time lag, shutter release time lag, shooting rate and start-up time. However, the standard is more suitable to compact and DLSR cameras and the methods can be difficult to use in mobile phone environment. It is also notable that the standard is focused only to the still image capturing measurements. The role of video recording is increasing even faster than still image capturing and therefore video performance measurements are also needed.

It is quite obvious that individual camera standards focus to specific quality or performance metrics. Therefore, combinations of different quality and speed metrics are missing. If mobile phone cameras are supposed to compare in comprehensive way, this kind of combined benchmarking metrics is needed.

There are several tasks in this work. Firstly, the most important image quality metrics are collected from the standards and papers. Secondly, the speed related metrics of a mobile phone's camera system are collected from the standards and papers and also novel speed metrics are identified. Thirdly, different combinations of the quality and speed metrics are validated using mobile phones in the market. Mobile phones are selected from the different prize categories and operating systems. The measurements are done using software based and automatic test system which is executed towards application programming interface of different operating system. This approach gives comparable speed measurement values between different operating systems and removes the influence of mobile phone specific camera applications. Finally, the most suitable combinations are selected according to the measurement values.

The result of this work gives detailed measurement results of mobile phone camera systems in the market. The paper defines also a proposal of combined benchmarking metrics for mobile phone cameras, which includes both quality and speed parameters.

9016-16, Session 4

Device and algorithms for camera timing evaluation

Lucie Masson, Frédéric Cao, Clément Viard, Frédéric Guichard, DxO Labs (France)

This article presents a novel device and algorithms to measure the different timings of a digital camera, for still-images and videos. This device (further referred as timer) is compliant with ISO 15781 Standard, and allows fully automatic protocols. Its design, with a large number of LEDs on 5 synchronized lines, allows more accurate measurements than other LED panels on the market, and is particularly well adapted for the measurement of electronic rolling shutter.

It enables the measurements defined in the ISO 15781 standard: shutter time lag and the shutter release time tag as well as several other measurements which have a high interest for the digital camera industry: exposure time (shutter speed), electronic rolling shutter (ERS), video frame rate, vertical blanking, counting of missing and duplicated frames.

The measurement of these different timings gives important information impacting the image quality of digital cameras. A poorly-selected shutter time will result in motion blur, or a noisy image, or bad image exposure. Some cameras tend to select a long shutter time in low light conditions in order to reduce photon shot noise, but this leads to blurry (and possibly unusable) images if the camera is hand held.

Another time-related measurement is the electronic rolling shutter (ERS), equipping most CMOS sensors. The pixels on the sensor are not exposed at the same time, and this leads to unpleasant effects on images if the scene or the camera is moving. A fast moving camera, or fast moving subjects with a speed similar to the rolling shutter, causes still-image distortions or "jello effect" vibrations in videos.

Frame rate is another important parameter. It must be high enough to produce a smooth motion, and shutter time should not be shorter than half the frame time. But in low light conditions, it may become necessary for the camera to increase its shutter time to reduce photon shot noise, which can cause a drop in frame rate, and missing or duplicated frames.

The timer, shown on Figure 1, is composed of five lines of one hundred LEDs. On each LED line, the LEDs light up sequentially. At any given moment, there is always one and only one lit LED on each line. When an LED is turned off, the next LED is turned on, cycling when the end of the line is reached. On any given line, each LED is lit for exactly the same amount of time, and the total time to light all of the LEDs is given in milliseconds by the 7-segment display on the right. Each LED remains on

exactly one 1/100th of the total displayed line time. Lines with the same period are synchronized: the first LED of these lines is lit at exactly the same time.

ISO 15781 standard requests a timing device with an accuracy of at least 1ms and capable of measuring times up to 10s. This device exceeds these requirements, as the lit time for LEDs is adjustable from 0.01ms to 100ms, leading to measuring times up to 10s with a maximum accuracy of 0.01ms on a single frame.

Each measurement is based on the number and the position of detected lit LEDs on image, as displayed on Figure 2.

The exposure time of the picture is deduced from the product of the number of visible LEDs with the lit time of each LED.

Rolling shutter, time lag, and frame rate are computed by comparing the position of LEDs on different lines in a single picture or the positions of LEDs on a single line in different frames. For rolling shutter measurements, the sensor lines (orthogonal to the ERS direction) have to be aligned with the timer lines.

The accuracy of the measurements depends on the selected period of the LED lines. Exposure time measurement has best accuracy when the line period is only slightly larger than exposure time. It is then very useful to keep lines with different periods, since a very good accuracy can be obtained on a single still image, or only one video, even if the value of the measured timing is not even approximately known before shooting. Note that ERS measurements need a pair of synchronized lines.

The automated measurement algorithms detect the timer either in a natural scene or in a lab scene working on both still-images and videos. The positions of the lit LEDs atre also automatically detected by the algorithm.

Furthermore, It can also be remotely controlled by a computer via USB, which makes shooting protocols easier to follow. The timer parameters can be remotely retrieved and adjusted.

The article describes the specifications of our new device, how it is used to compute the different timings, and the measurement accuracy depending of the selected periods for the LED lines.

In a companion article presented in this conference, we also present how the timer can accurately be synced with camera trigger in order to totally automate shooting plans of shutter time lag and shutter release time lag measurements.

The protocol and measurements have been validated on many different cameras of different types and in different lighting conditions and light source, with different camera settings (different image resolution, still or video etc…). We will present a selection of results.

9016-17, Session 4

Embedded signal approach to image texture reproduction analysis

Peter D. Burns, Burns Digital Imaging (United States); Donald J. Baxter, STMicroelectronics Ltd. (United Kingdom)

Image texture is the term given to the signal (information-bearing) fluctuations such as those for skin, grass and fabrics. Since image processing aimed at reducing image noise can also remove important texture, standard methods for evaluating the capture and retention of image texture are currently being developed. Concurrently, the evolution of the intelligence and performance of camera noise-reduction (NR) algorithms poses a significant challenge for these protocols. Cameras already exist for which the Dead Leaves [1] and low contrast radial sine-wave texture reproduction protocols [2] ratings are inconsistent with subjective texture quality. This paper reviews the requirements for a improved texture measurement protocol, and describes an alternative approach to texture reproduction analysis. This method is based on embedding periodic test signals within synthetic texture regions.

Many of the latest NR algorithms are ‘content-aware’, which can lead to different levels of NR being applied to various regions within the same digital image. The non-local (NL) class of NR algorithms [3] look for pixels with a similar local neighborhood within a larger search window. This approach dynamically finds the best region, line or curve over which to filter image pixel variations. This form of local decision-making reduces the effectiveness of the Dead Leaves texture evaluation method, in particular the separation of image (signal) texture fluctuations from noise, estimated from uniform areas. [4] Other methods for texture capture evaluation are also influenced by these noise-reduction methods. A radial sinewave target, also called a Siemens Star, is used to evaluate an effective signal MTF. However, some noise-reduction algorithms filter noise along the radial fingers of the target, which can lead to misleading results.

In developing an improved method for measuring the capture of image texture the following requirements were considered;

1. The test target should have texture fluctuations similar to natural materials such as grass, to ensure realistic NR algorithm behavior.

2. The analysis should result in a reliable texture MTF in the presence of image noise such as that observed with current equipment and high ISO (low-light) settings.

3. The influence of non-linear signal mapping should be minimal.

We describe a target that includes a synthetically-generated texture image that is based on the spatial auto-covariance and color-covariance of real grass and stone textures. Low-amplitude sine-wave signals are embedded in this texture image. Conceptually, this can viewed of as a low amplitude version of the IEC 61146-1 sine-wave multi-burst chart buried within the synthetic texture. The magnitude of the embedded sine waves is sufficiently low to avoid changing the visual nature of the synthetic texture, but of sufficient amplitude to be reliably extracted via power spectrum techniques.

The protocol comprises the following steps

1. OECF based linearization step

2. Conversion to luminance

3. Spectral 2-D identification of up to 9 embedded sine-wave components including compensation for the spectral window due to the finite data array

4. Texture MTF generation from the compensated spectral peaks.

The variation of the signal-extraction results is described for different images sizes, crop position and image tilt. Results are also compared to those for the low contrast Siemens Star method. The comparative performance of the embedded texture protocol is demonstrated under 3 different test conditions and levels; Gaussian blur, camera photon shot noise, and non-linear signal distortion. We also discussed the trade-off of reduced number of texture MTF sample points and signal detection.

9016-18, Session 5

Perceptual tools for quality-aware video networks

Alan C Bovik, The Univ. of Texas at Austin (United States)

No Abstract Available

9016-19, Session 6

On the definition of adapted audio/video profiles for high-quality video calling services over LTE/4G

Maty Ndiaye, Catherine Quinquis, lab’Orange (France); Mohamed-Chaker Larabi, Univ. de Poitiers (France); Gwenael Le Lay, lab’Orange (France); Abdelhakim Saadane, Univ. de Nantes (France); Clency Perrine, Univ. de Poitiers (France)

Context :

During the last decade, the important advances and the widespread availability of mobile technology (operating systems, GPUs, terminal resolution…) have pushed toward a fast development of voice and video services. While multimedia services have largely grown on mobile devices, the generated increase of data consumption leads to a saturation of mobile networks. In order to provide data with high bit-rates and maintain the performance as close as possible from the traditional network, the 3GPP (The 3rd Generation Partnership Project) worked on the design of the LTE/4G standard. LTE/4G integrated several levels of Quality of Service (QoS) to cope with the demand of interactive and real-time services in particular in terms of bandwidth and latency [1]. With this new standard, user satisfaction is becoming a major concern for mobile network operators, responsible of both efficient management of network resources and resource allocation across users. For the video calling service, the main focus of this study, four factors may affect the user satisfaction as described in [2]: 1) encoding scheme and the content of the audio/video data, 2) QoS of the network (bandwidth, throughput, delay, jitter and packet loss), 3) end-user terminal (screen size and resolution, camera performance, CPU power, audio capabilities) and 4) human perception. Consequently, offering a high user satisfaction level implies for the mobile network operator to reach a trade-off between these aforementioned factors in addition to handling operational and economical issues.

Objective :

The purpose of the present study is to draw recommendations related to audio and video media profiles (selection of audio and video codecs, bit-rate, frame-rate, audio and video formats) for a typical video-calling services held over LTE/4G mobile networks. These profiles defined according to targeted devices (smartphones, tablets), and QoS issues (available throughput, delay and jitter) so as to ensure the best possible QoE.

It is well known that several factors play a role on the user's QoE [3] e.g. audio spectrum, video frame-rate, video fluidity, compression scheme, jitter buffer management, audio and video synchronization. Hence, a careful study of these factors is tackled, as they may have a negative impact on the perceived quality of end user.

Method :

Our experimental study consists in the simulation of end to end calls through the LTE /4G network thanks to video-calling probes which are able to set-up and tear-down video calls using the SIP protocol together with recording QoS indicators and transmitted media files.

Therefore, once the connection to the LTE / 4G network is established, the framework simulates audio/video calls by injecting in the network reference files, representative sequences of a videoconferencing service with several levels of complexity.

The media (audio/video) coding scheme is also specified (codec, bit-rate, frame-rate, resolution).Therefore, for audio stream, several audio corpus are encoded using ITU-T standardized codecs (narrow-band and wideband), and Opus IETF codec. For video stream, x.264 implementation of the H.264/AVC coding standard [4] and WebM (VP8) are considered as they are widely used within currently deployed web 2.0 applications. The video image resolutions considered for this study are CIF and 4CIF. The video bit-rates range from 128 kbps up to 1 Mbps while the audio bit-rates range from 8 kbps up to 32 kbps,

For each call set-up and at the end of each communication, the framework records data related to the service performance (registration and establishment call times, number of send and received packets, negotiated codec) and the network performance (packets losses, jitter, one-way delay). The transmitted media files are also recorded so as to allow the analysis of the received files by setting-up subjective quality experiments.

One the one hand, audio quality subjective tests is run using the ITU-T Absolute Category Rating (ACR) method for narrow-band and wide-band signals. On the other hand, for video quality assessment, the ACR method describe in ITU-T P.911 [6] with CIF and 4CIF video sequences recorded in the framework.

The output of these subjective tests gives a wide picture of the perceived quality for audio and video signals when considering both low and high-bit rates in the scope of a video-calling service. This allows us to draw very important conclusions about the ranges of bit-rates, for both audio and video guarantying QoE.

Conclusion :

This study allows drawing recommendations in terms of image format, bit rate, compression algorithms based on subjective quality assessments. The context of use is targeting a video calling service in good radio conditions (RSRP ≈ -60 dB) over the LTE / 4G networks [800Mhz, 2600 MHz]. For a premium video calling service on a smartphone (CIF, 25fps), a VP8 video codec running at 640 kbps coupled with a G.722.1 audio codec at 32 kbps for audio will offer the highest quality, as it reaches a MOS of 4.1 representing a quality judged as good even very good. For premium video calling service on a tablet (4CIF, 25fps), the highest quality will be obtain by combining a VP8 video codec or an X.264 encoder running at 768 kbps coupled with a AMR-WB audio codec at 23.85kbps as it reaches a MOS of 4.

Several combinations can be proposed for medium or low-cost type level video call services relying on bit-rate tradeoffs. Additional works are currently undergone so as to find out the influence of the radio coverage variations due to displacement in the cell on the user experience on a video calling service.

9016-20, Session 6

Subjective quality assessment of an adaptive video streaming model

Samira Tavakoli, Univ. Politécnica de Madrid (Spain); Kjell E. Brunnström, Kun Wang, Acreo Swedish ICT AB (Sweden) and Mid Sweden Univ. (Sweden); Börje Andrén, Acreo Swedish ICT AB (Sweden); Narciso García Santos, Univ. Politécnica de Madrid (Spain); Narciso Garcia, Univ Politécnica de Madrid (Spain)

The streaming of the video contents is currently expanding rapidly as an alternative to traditional TV broadcast. However, because of the variable bandwidth of the networks used to deliver multimedia content, a smooth and high-quality playback experience could not be always guaranteed. Using segments (chunks) in multiple video qualities, HTTP adaptive streaming (HAS) of video content is a relevant advancement which offers significant advantages in terms of both user´s Quality of Experience (QoE) and resource utilization for content and network service providers.

By recent high usage of adaptive streaming technique, various studies has been carried out in this area which has been generally focused on the technical enhancement of HAS technology. On the other hand, all ongoing researches are currently providing the basis of HAS applications and how they can be improved for potentially supersede the previous proprietary adaptive streaming approaches. However, to optimize the QoE of HAS, still there is a big demand of studying the perceptual quality of the end-user. Taking this into account, the objective of this experiment was to study the QoE of a HAS-based video broadcast model. After discussion with the most dominant video streaming companies in Sweden, following factors were considered to study:

1. Adaptation scenarios

The different states in which HAS client should make the adaptation decision has been categorized in two classes: when increasing the quality and when decreasing the quality. For each of these states, different scenarios as the possible behaviour of the client were considered to be applied to the test video contents. Also, the QoE of each of the adaptive streams has been considered to be compared with the adaptation scenarios.

2. Chunk size

Currently, majority of the video service providers are using 10 seconds' duration chunk size when adaptive streaming. However, such a long duration chunk is not suitable for the live broadcast. On the other hand, using short chunks has low switching granularity. Considering these issue, in this experiment the effect of long and short chunk (10 and 2

seconds) on perception of adaption scenarios has been studied.

3. Content type

In total, 7 video sources in different content types, such as movie, sport, documentary, music video and news, have been considered. The characteristics of the contents are different containing from smooth to sudden motions, no scene change to fast scene change, and recorded using still, zoom or moving camera.

All video sources were originally in 1080p and in 24 or 25fps that all were played in the test in 25 fps.

4. Adaptive streams

To produce the adaptive streams, the characteristics of the streams which are in practice provided by the streaming companies for the living room platforms have been considered. Consequently, in total 4 streams encoded in 5 Mbps, 3 Mbps, 1 Mbps and 600 kbps have been considered as the adaptive streams. Following the same planning strategy, 720p was chosen for the resolution of adaptive streams.

Considering all above factors, 12 test conditions applied on 11 source videos led to generate 132 Processed Video Sequences (PVS).

The test was run in the lab of Acreo research institute in Sweden, while the environment was in accordance with Rec. ITU-R BT. 500-13. The 48" Hyundai display was used to display the video sequences in the viewing distance of 4 times of the display height.

A modified single stimulus Absolute Category Rating (ACR) scale procedure (ITU-T P.910) was used as the test methodology in this experiment (considering individual PVS, with no audio). To assess the quality of the PVSs, the subjects have been asked to vote for two questions: the overall quality of the test sequence and if they have perceived any change in the quality.

Prior to the subjective experiment, the observers were screened for visual acuity. Only one subject at a time was attending each test session. The total number of the observers was 23, including 6 females and 17 males, with the average age of 30, and 5 of them had subscription from the media service providers.

Different observations have been obtained by calculating the mean opinion scores of the collected experimental data. Considering the overall behavior of the test conditions in all the video contents, increasing the video quality has been perceived better than decreasing the quality, as one could expect. On the other hand, playing the constant quality of 5 Mbps and 3 Mbps streams have been preferred to the quality variation. However, the subjects preferred to have change in the video quality instead of constantly watching the stream with 1 Mbps or 600 Kbps quality, when in the later stream this became more severe. Effect of video content type and the scene on the perceptual quality of adaptation scenarios has been also perceived. Furthermore, it has been observed that the chunk size used to change the quality has had a significant influence on the test subjects' quality perception. Considering the adaptation scenarios, gradual changing the quality by 10 seconds chunk has been mostly perceived better than the other scenarios, while a reverse result has been observed about rapidly changing the quality when the chunk is 2 seconds. In many cases, the difference between the performances of the scenarios has been statistically significant, which will be comprehensively reported in the paper.

Since quality adaptations are temporary events when streaming the video, for the future work, considering a different experimental method which could make the event closer to what is happening in the real condition is addressed.

9016-21, Session 6

Video interpretability rating scale under network impairments

Thomas Kreitmair, Cristian Coman, NATO Communication and Information Agency (Netherlands)

The paper presents the results of a study on the effects of network transmission channel parameters on the quality of streaming video data. In NATO, the interpretability levels of imagery products are defined in the Standardization Agreement (STANAG) 7194, which reflects the National Image Interpretability Rating Scale (NIIRS) defined in the United States. For video data, NATO has adopted the interpretability levels proposed by the Motion Imagery Standards Boards (MISB) in the Recommended Practice (RP) 0901 (Video NIIRS or V-NIIRS).

A common practice for estimating the V-NIIRS level in an objective way is to use the Motion Imagery Quality Equation (MIQE). The MIQE has been derived from a similar equation developed for still images that uses a few technical features such as: ground sampling distance, relative edge response, modulation transfer function, gain and signal to noise ratio. One observation of this study is that the MIQE does not account for video specific parameters such as spatial and temporal encoding.

Furthermore, the study investigates the impact of relevant network parameters (e.g. bandwidth, delay, packet loss, corruption and packet reordering) on the V-NIIRS estimation using MIQE. The experimental setup employed in this analysis includes a video server, a client and a wide area network emulator. Simulated and operational video data was used in the experiment.

The estimated V-NIIRS level of the original stream was compared with that of the stream received on the client side, after transmission through a controlled network channel. The quality of the received video showed visible degradation when the quality of the transmission channel was reduced below certain thresholds (e.g. reduce bandwidth, increase packet loss, etc.). The main artifacts impacting the interpretability level had to do with blocking caused by lossy decompression of video data (e.g. uniform block of 16x16 pixels with no detail information). However, the automated solution for estimating the V-NIIRS, which utilizes MIQE, indicated no significant degradation of the interpretability level for these types of errors.

One parameter in MIQE, which is influenced by network transmission errors, is the Relative Edge Response (RER). The automated calculation of RER includes the selection of the best edge in the frame, which in case of network errors may be incorrectly associated with a blocked region (e.g. of 16x16 pixels). A solution was proposed to correct this inconsistency by removing corrupted regions from the image analysis process. Furthermore, a recommendation was made on how to account for network impairments in the MIQE such that a more realistic interpretability level is estimated.

9016-22, Session 6

Multimodal video quality assessment for talking face analysis

Naty Ould Sidaty, Mohamed-Chaker Larabi, XLIM-SIC (France)

Image and video quality assessment has been for a while an interesting research field. Although a huge amount of work on Image Quality Metrics (IQM) has been done in the past decade. Video Quality Assessment (VQA) is still a challenging issue. Video quality assessment is very important for many video-processing applications (HDTV, video over IP, videoconferencing, video telephony). However, most existing metrics do not rely on perceptual features and do not take into account the multimodal aspect of the video (audio/video). Indeed, the experience of a user is influenced by both medium and for instance a desynchronisation between them may generate a very high annoyance. Because of the importance of heads and especially the talking heads in the video sequence, this paper seeks to introduce an objective solution to better estimate the perceived quality and guarantee a high level of user experience by separating talking and non-talking faces in the video sequence.

It is known that the semantic aspect of the content can attract our eyes. The visual attention has been used in various fields (in video coding for example, coding efficiency can be improved by discarding redundant information out side small fixation region without degradation of perceived quality). The region of attention can be detected based on low level features, detection of moving objects, face detection, etc. Thus, faces and texts are known to be naturally prominent. They

tend to immediately attract our eyes and capture our attention. Trying to evaluate the perceived multimodal video quality based on talking faces, we choose video, with sound, that contains a talking/no talking heads (five videos with different talking scenarios were selected: no talking heads with sound from the outside, one talking head, two talking heads alternatively, two talking heads simultaneously and all faces talk). Because of the importance of faces in many video applications systems, such as video surveillance, TV-content , we have conducted a set of subjective test to determine at what extent viewers are attracted by the talking heads compared with other faces and objects present in the video.

Subjective quality experiment was conducted in a test room with controlled lighting. The viewing conditions conformed to international recommendations. Eye tracker system was used in this experimental test equipped with Tobii Studio, dedicated software, which provides a platform for stimuli presentation, recording, observation, visualization and analysis of eye tracking. Each test session started with eye tracker calibration for the individual subject. A total of 14 people participated in the experiment, 10 males and 4 females, all of them passed Ishihara test and Freiburg visual acuity test. The experimental results and analyze of the data coming from an Eye-tracker show that viewers are well attracted by talking heads than other heads, text and objects. The reported result can be used in various applications especially in video coding to obtain high coding efficiency by using coarse quantization steps or removing high frequency components out side the talking-head region.

Based on the subjective experiment results, we aims to propose an objective audio-visual saliency model to predict overall multimodal video quality by integrating the importance of heads/talking heads and localization of sound emitting region taking into account both temporal and multimodal aspects of a video to assess the overall quality in a similar manner than the end-user judgment.

Because of the absence of audio-visual database, despite a few efforts to create database in this area, a new audio-visual database with different scenarios of talking heads was created in our laboratory. A set of 60 distorted videos were created from these reference videos (5 distorted videos per reference) using H.264 compression with different bit rates.

9016-23, Session 7

Breaking down the problem of blind video quality evaluation

Michele Saad, Intel Corp. (United States) and The Univ. of Texas at Austin (United States); Alan C. Bovik, The Univ. of Texas at Austin (United States)

We break down the problem of perceptual blind video quality assessment (VQA) into components, which we address individually, before proposing a holistic solution. The idea is to tackle the challenges that comprise the blind VQA problem individually in order to gain a better understanding of it.

We address the motion/temporal modeling problem and propose a model of motion characterization that encompasses a characterization of a video's global temporal activity and its local coherency/incoherency of motion trajectories. We characterize motion coherence using a 2D structure tensor model applied to a video's computed motion vectors. The motion coherence tensor summarizes the predominant motion directions over local neighborhoods, as well as the degree to which the local directions of motion flow are coherent. We also propose a model that quantifies the fraction of motion attributed to non-global motion over the global motion computed in a video sequence. We show how this affects human perception of video quality.

We then address the problem of content-dependency in blind VQA. This problem states that, if one knows or restricts the type of content that comprises a video sequence, then one is expected to better predict quality. (We demonstrate this on two video databases). However, how does one address this content-dependency challenge and design blind VQA algorithms that work across a variety of content? We propose an approach that significantly mitigates content-dependency. Our proposed approach relies on computing geometrically pooled spectral ratios derived from a model of natural video statistics (NVS), which we propose. We show that the use of NVS spectral ratios (as opposed to the more usual use of NVS model parameters) serve to mitigate content-dependency and are regularly disturbed with increasing distortion levels making them adequate for blind quality assessment.

We also propose a geometric pooling of features in order to account for ranges of parameters that differ for different content. We show how this pooling strategy improves quality prediction performance.

Additionally, we discuss issues associated with training-based algorithms and provide preliminary results of a blind VQA approach that does away with training completely.

We apply the methods we propose to address each of the above challenges into an approach for blind video quality prediction. There are no existing blind VQA approaches that are non-distortion specific, which makes it difficult to compare our algorithm against other methods. Full-reference and reduced reference approaches have the enormous advantage of access to the reference video or information about it. We do however, compare against them and against the naturalness index NIQE in [Mittal2012], which is a blind IQA approach applied on a frame-by-frame basis to the video.

The median SROCCs (Spearman rank order correlation coefficients) between subjective and predicted scores are computed. The results show that we clearly outperform NIQE, SSIM [Wang2007], PSNR. We show that our approach performs as well as the standardized reduced-reference VQM [Pinson2004] (slightly beating it on the LIVE VQA Database with an SROCC of 0.757), and approaches the prediction performance of the highest performing full reference VQA algorithms, MOVIE [Seshadrinathan2010], and ST-MAD [Vu2011].

We also demonstrate the content-dependency problem on two databases (the LIVE VQA database and the EPFL-PoliMi database) respectively. Finally, on homogeneous content, when the training module was removed, an SROCC correlation of 0.674 was obtained. While it did drop from 0.813 to 0.674, it remained quite high. We discuss training challenges and future directions for training-free algorithms with more preliminary results in the paper.

9016-24, Session 7

Incorporating visual attention models into video quality metrics

Welington Y. L. Akamine, Mylene C. Q. Farias, Univ. de Brasilia (Brazil)

A recent development in the area of image and video quality consists of trying to incorporate aspects of visual attention in the design of visual quality metrics, mostly using the assumption that visual distortions appearing in less salient areas might be less visible and, therefore, less annoying. This research area is still in its infancy and results obtained by different groups are not yet conclusive, as pointed out by Engelke. Some researchers have reported that the incorporation of saliency maps increases the performance of quality metrics, while others have reported no or very little improvement. Among the works that have reported some improvement, most use subjective saliency maps, i.e. saliency maps generated from eye-tracking data obtained experimentally. Although subjective saliency maps are considered as the ground-truth in visual attention, they cannot be used in real-time applications.

To incorporate visual attention aspects into the design of image or video quality metrics, we have to use visual attention computational models to generate objective saliency maps. This raises the question of how the metric performance is affected by the "precision" of the saliency map and the integration model. Another open question is how the content and the distortion type affect the saliency map and, consequently, the metric performance. Very few works so far have tested the incorporation of specific computational attention models into image quality metrics. Moreover, to our knowledge, there are few works that have tested the

incorporation of computational attention models into video quality metrics. In a previous work, we investigated the benefits of incorporating objective and subjective saliency maps into three simple full-reference image quality metrics (MSE, PSNR, and SSIM). In that work, we compared the performance of these original image quality metrics with the performance of quality metrics that incorporated saliency maps using three computational visual attention models (GAFFE, Achanta, and Itti) and subjective saliency maps. Our results showed that the computational models were able to improve the performance of the tested image quality metrics.

In this paper, we investigate the benefits of incorporating visual attention models into video quality metrics. With this goal, we select two popular video quality metrics: the Video Quality Metric (VQM) and the MOtion-based Video Integrity Evaluation (MOVIE). The MOVIE metric generates three quality estimates: a global quality estimate, a spatial quality estimate (Spatial-MOVIE), and a temporal quality estimate (Temporal-MOVIE). The VQM metric is divided into several independent processing stages. We also test two image quality metrics adapted for video: the Structural Similarity Index (SSIM) and the Multi-Scale Structural Similarity Index (MS-SSIM).

We use Itti's computational video attention model, which generates a global saliency map that is a combination of a temporal and a spatial saliency maps. The integration process of the quality metric with the attention information consists of using the gray scale pixel values of the (global) saliency maps as weights for the (global) error maps generated by the quality metrics. For the MOVIE metric, besides combining the global saliency map generated by Itti´s model with the MOVIE error map (temporal + spatial), we also individually combine Itti's temporal and spatial saliency maps with the Temporal-MOVIE and Spatial-MOVIE error maps, respectively. For the VQM metric, we test the incorporation of the saliency map at several stages of the algorithm.

Our results show that the addition of the saliency maps improves the performance of most metrics. It is interesting to notice that the highest improvements in performance correspond to the SSIM and the Spatial-MOVIE metrics. We also modify the original saliency maps with the goal of improving the performance of the metrics. The first modification consists of slightly increasing the focus of attention until it reaches a certain percentage of the frame. The second modification consists of dilating the saliency map with the goal of making the focus of attention region more uniform. Nevertheless, these modifications result in little improvement in performance in comparison with the original saliency maps.

In summary, our results show that saliency maps can improve the performance of video quality metrics, but the improvement of the quality metrics with only spatial error information (SSIM, MS-SSIM and Spatial-MOVIE) is higher than that of the quality metrics with both temporal and spatial information (MOVIE, Temporal-MOVIE and VQM). This result seems to suggest that the temporal saliency map adds information that is lacking to the spatial quality metrics.

9016-25, Session 7

An objective model for audio-visual quality

Helard B. Martinez, Mylene C. Q. Farias, Univ. de Brasilia (Brazil)

There is an ongoing effort to develop video quality metrics that are able to detect impairments and estimate their annoyance as perceived by human viewers. To date, most of the achievements have been made in the development of Full-Reference video quality metrics. In fact, very few metrics have addressed the issue of simultaneously measuring the quality of all media involved (e.g. video, audio, text). Even for the simpler case of audio and video, there are only a few metrics in the literature that estimate the quality of video with audio-visual content. But, in order to design good audio-visual quality metrics it is necessary to understand how audio and video contents are perceived and how the degradations in audio and video affect the overall quality. Previous research shows that video quality influences subjective opinions of audio quality and vice-versa.

The first goal of this paper is to obtain a better understanding of how audio and video components interact and how these interactions affect the overall audio-visual quality. With this goal, we perform three psychophysical experiments and analyze their results. To generate the test sequences for these experiments, we start with original video sequences with both audio and video components. For the first experiment, we consider only the video component of the sequences and compress them using a H.264 codec with different (video) bitrate values. For the second experiment, we consider only the audio component of the sequences and compress them using the MPEG-1 layer 3 codec, with different (audio) bitrate values. Finally, for the third experiment we consider both the video and the audio components of the sequences and compress them independently. In all three experiments, we ask an average of 16 subjects to score the quality of the test sequences.

Results of the experiments allow us to understand how the content of the videos affects the quality perceived by the final user. By comparing the audio content with experiment results, we find that some types of audio are more sensitive to compression than others. The same is true for video, as already observed by other authors. By analyzing the quality scores, we notice that video compression has a higher impact on audio-visual quality than audio compression. Finally, quality scores obtained from the first experiment (only video) are higher than the quality scores obtained from the third experiment (audio and video). This suggests that the audio component might be a distraction during the quality assessment task of audio-visual sequences.

The second goal of this work is to obtain an objective model for audio-visual quality. With this in mind, we test a set of combination models, using the scores of the first (only video) and second (only audio) experiments as basis to obtain the scores of the third (video and audio) experiment. To obtain the audio quality estimates, we use the no-reference audio quality metric SEAM (Single-Ended Assessment Model). To obtain the video quality estimates, we use a full-reference audio quality metric proposed by NTIA – The VQM (Video Quality Metric). Then, we obtain two audio-visual quality metrics by combining these two metrics: the first using a linear model (Pearson correlation = 0.8472) and the second a Minkowski metric (Pearson correlation = 0.8337). The models show a good balance between the audio and video components, with a reasonable correlation coefficient.

In summary, in this work it is observed that the video and audio spatial and temporal characteristics are important while determining the quality scores. Also, sequences with and without audio are scored differently. Preliminary results give us a first version of an audio-visual quality metric. The model presents reasonably good correlation for the tested database. Nevertheless, this model needs to be tested on a more diverse audio and video database.

9016-26, Session 7

Efficient measurement of stereoscopic 3D video content issues

Stefan Winkler, Advanced Digital Sciences Ctr. (Singapore) and Cheetah Technologies (United States)

INTRODUCTION

Many current 3D quality metrics choose a rather simplistic approach of extending 2D quality measurement to 3D by combining quality measurements done separately on left and right views; these are mainly targeted at the evaluation of asymmetric stereo coding. Only recently, more general methods for 3D quality assessment taking into account additional parameters have been proposed. However, little consideration has been given to computational efficiency so far.

In this contribution, we define parameters and metrics that can detect and quantify some common issues with 3D content, namely view mismatch, divergence, disparity range, and (temporal) disparity change. Experimental validation of the metrics is also presented. The important feature of these metrics is high computational efficiency to permit real-time video content analysis.

Note that video distortions such as those introduced by compression are not the main focus of these metrics; instead the aim is to enable checks of 3D-specific content for issues that might make viewers uncomfortable. They could be combined with other 2D quality metrics in order to estimate the overall quality of a stereoscopic 3D presentation.

DISPARITY ESTIMATION

Disparity estimation is a prerequisite for computing depth metrics when the 3D content is represented in separate left and right views. The approach is based on matching the individual scan lines of a frame to compute disparity at the pixel level. The method used here is based on Takaya's and relies on "dynamic time warping" (DTW), a well-known technique to find an optimal alignment between two given (time-dependent) sequences.

Intuitively, the sequences are warped in a nonlinear fashion to match each other. In the context of disparity estimation here, it is applied as follows: The method processes the views scan-line by scan-line. DTW is used to compute the spatially varying shift between the scanlines from the left and right views/images. The result is an estimate of the disparity at each pixel, or in other words a disparity map. Downsampling (or alternatively median filtering) can be applied to the frames before processing in order to reduce noise as well as computation time.

The method has the following benefits:

1. Performance: disparity estimation is the most computationally demanding task in stereo processing; with proper down-sampling, even high-definition (HD) content can be processed within a few milliseconds per frame on a standard PC, which is essential for real-time content monitoring.

2. Resolution: potentially pixel-level precision for disparity (although this has to be traded off with performance and noise).

3. Flexibility: the trade-off between resolution and performance can easily be fine-tuned as necessary.

4. Robustness: disparity estimates are largely correct, without major outliers.

While the resulting disparity maps can be noisy, this is not a big concern for the content metrics proposed below, since we are mainly interested in the overall disparity range and distribution rather than the exact values at every pixel. As will be shown below, the method is sufficiently accurate for the measurements of interest and also robust to various image distortions.

LEFT/RIGHT VIEW (MIS)MATCH

We use histogram correlation to express the magnitude of the mismatch between the two views. It is designed to detect undue variations in color distribution and is commonly used in video scene change detection.

First, the luminance histograms of left and right views are computed, respectively. The correlation between these two histograms is used to quantify how well the views match.

DISPARITY RANGE

Disparity Range measures the range of pixel disparities between the left and right view. It is expressed as the range of disparities of a majority of pixels in a given frame; we use 90% here, but this can be adjusted to trade off robustness to noise with sensitivity. As a guideline, the minimum/maximum disparity should be less than 2-3% of screen width.

For a detailed evaluation of the Disparity Range estimate, we use the New Tsukuba Stereo Dataset. It contains ground truth disparity maps for 1800 frames from a simulated camera fly-through of a computer-generated office environment, featuring a wide variety of content and lighting conditions. Our error statistics data show that the disparity range metric is generally accurate, with errors in an acceptable range.

Since the New Tsukuba Stereo Dataset contains no noise or other types of image distortions, we further evaluate the robustness of the Disparity Range estimate using a sequence created from the IRCCyN/IVC 3D Images Database. It contains six different stereoscopic images, each of which is present in original undistorted form and in 15 distorted versions. Distortions include three different types of processing (JPEG and JPEG2000 compression as well as blurring), which were applied symmetrically to the stereo pairs.

Our disparity range measurements show that the disparity range metric is largely unaffected by image distortions such as compression and blur, some of which reach rather severe levels in this database.

DIVERGENCE

A disparity greater than the inter-ocular distance would force the eyes to diverge and place the object beyond infinity, which is impossible in nature and should be avoided. Therefore, the maximum positive disparity on screen should not exceed the interocular distance of the viewer; in other words, positive (divergent) disparity should not be more than 5-6cm. Naturally, in the image domain, this is screen- and resolution-dependent; for high-definition (HD) video displayed on a 42" screen, this corresponds to roughly 5% of the width of the HD video frame, which is about 100 pixels.

DISPARITY CHANGE

Disparity Change measures the temporal change of disparity distributions between two consecutive frames. Histogram correlation is commonly used for detecting scene changes in video. Therefore, we adapt it here again to use with the disparity maps.

First, a histogram of the disparity map is computed for every video frame. The correlation between the disparity histograms of the current frame and the previous frame is used to quantify the change in disparity.

To evaluate the Disparity Change estimate, we again use the sequence of 96 images created from the IVC 3D Images Database. An additional test on a longer and more realistic compressed video sequence (about 2 minutes of a soccer match) shows the reliability of the metric, with scene cuts being clearly identified and rated according to their severity in terms of disparity change.

CONCLUSIONS

We presented metrics that can detect and quantify some common issues with 3D content, namely view mismatch, divergence, disparity range, and disparity change, all of which may introduce discomfort in viewers if they become too large. The metrics are robust to compression and various other types of image distortions.

The measurement algorithms are suitable for real-time 3D video monitoring applications; even for HD content, all the metrics can be computed within a few milliseconds per frame. The bulk of the computation is due to disparity estimation from the stereo images, which is performed using an efficient method based on dynamic time warping.

The disparity estimation is designed for 3D content stored as separate left and right views; however, the metrics can also be applied to 2D+depth representations.

While we have demonstrated the validity of the metrics in terms of estimating various 3D content parameters, their perceptually acceptable ranges still need to be verified. The 3D content metrics could be easily integrated and combined with other quality metrics (e.g. 2D quality assessment of left and right views) in order to measure viewing comfort or 3D MOS.

9016-28, Session 8

Image characterization of row and column defect correction

Kaushik Atmanadhan, Ramkumar Narayanswamy, Aptina Imaging Corp. (United States)

2D-fuse correction algorithm and has been experimentally introduced in a variety of our sensors such as 12M/13M. Its primary goal is to intelligently replicate an adjacent column or row, once identified, to hide or exclude irreparable or non-tunable single row or column artifacts. This feature allows for an unrestrained row or column fix and impacts final production yield positively. Though useful, new artifacts have been noted under certain lighting conditions and across sensor operating modes. For this reason, this feature is currently disabled in today's production programs and adjustments are necessary to improve the robustness of the algorithm. This paper provides an overview of this feature, its implementation, and namely, characterization of the algorithm

across different snapshot and video modes. We use 12M as the test platform to evaluate Image Quality (IQ) and conclude on its effectiveness with respect to IQ and yield. A performance summary is also provided after varying parameters such as analog gain and frequency. Recommendations and suggestions are also provided and can be used as base for further refinement of the algorithm.

9016-29, Session 8

Analysis of noise power spectrum of gamma rays camera

Hongwei Xie, Faqiang Zhang, Jianhua Zhang, Jinchuan Chen, Institute of Nuclear Physics and Chemistry (China); Linbo Li, Institute of Nuclear Physics and Chemistry (China)

Gamma rays camera is widely used in many studies, including the image diagnostics of the radiation sources, flash photography, and nondestructive assessment (NDA), etc. As a major component of the high sensitivity gamma rays camera, the MCP image intensifier is characterized in the intensified image, tunable shutter time and gain. The gamma rays camera is consisting with rays-fluorescence convertor, the optical imaging system, the MCP image intensifier, CCD and other devices. The performance of the gamma rays camera is mainly dependent on such parameters as the modulation transfer function (MTF), the noise power spectrum (NPS), and the detective quantum efficiency (DQE), etc. All of the parameters are somewhat limited by the noise characteristics of the system. Compared with the standard derivative noise distribution, the NPS, which can reflect the evolution characteristics of the noise of the imaging system with the change of the spatial frequency, could convey more information on the noise distribution in the system. In this paper, theoretical analysis is presented on the major sources of the noise in the gamma rays camera. Based on the analysis, the noise power spectra of the gamma rays camera were calibrated under various radiation dosages respectively with the visible light and gamma rays radiation sources (0.2MeV and 1.25MeV in energy, respectively). As indicated by the experimental results, the noise is majorly induced by the fluctuations of the gain of the MCP image intensifier. And the remarkable noise peak occurs nearby the spatial frequency of about 0.633 Hz/mm. And almost the same phenomena were found with both the 0.2MeV and 1.25MeV radiation energy. Besides, the noise power spectra are in circular symmetrical distribution, whose intensities are rapidly decreased with the increasing spatial frequencies.

9016-30, Session 8

Analysis on relation between Hartmann-Shack wavefront detection error and image restoration quality

Qi Li, Zhihai Xu, Huajun Feng, Yueting Chen, Yuhua Yu, Zhejiang Univ. (China)

The wavefront passed through the atmosphere will produce different degree of distortion, due to atmospheric disturbances, defocus, aberration and etc. Distortion of wavefront can result in image degradation. Conventional methods typically use adaptive optics correction the degradation. Correction system is complex and requires three parts, including wavefront detection, wavefront reconstruction, wavefront correction, and each part requires very precise control. In order to simplify the system structure, we use Hartmann - Shack wavefront sensor to get wavefront information, and then reconstruct the degenerated image using software restoration method.

The paper introduces the background and significance of Hartmann-Shack wavefront sensor, summarizes the foremost application, and put forward an image restoration method based on Hartmann-Shack wavefront sensor. We emphasis on the wave-front reconstruction principle, describes four centroid methods and several image restoration methods. Then we introduce the general model of Optical Transfer Function (OTF) and the way to calculate the OTF of diffraction limited incoherent image system. Take the actual situation into consideration, wave-front distortion is unavoidable, so we deduce the method to calculate OTF with wave-front distortion and several general models with wave-front distortion.

According to OTF calculation method and image restoration method, we analyses the reasons which caused the detection error of Hartmann-Shack wavefront sensor. Based on different wave-front detection error and the image restoration quality, we concluded the allowed maximum detection error of different peak value of wave-front, and found that when the wave front defocus peak value is ?, 2?, 3?, acceptable wavefront error is ablout 15%, 10%, 5%.

9016-31, Session 8

Implementation of an image signal processor for reconfigurable processors

Seung-Hyun Choi, Kwangwoon Univ. (Korea, Republic of); Junguk Cho, Samsung Digital City (Korea, Republic of); Yong-Min Tai, SAMSUNG Electronics Co., Ltd. (Korea, Republic of); Seong-Won Lee, Kwangwoon Univ. (Korea, Republic of)

Recently, the technology progress of smart devices is surprisingly fast. Among many performance improvements and additional special purpose functions, one of the most commercially successful functions is the digital camera. Nowadays almost every smart device has the digital camera feature. Since the environment where the camera with the smart devices is used is sometimes not quite suitable to get good quality pictures. Therefore, Bayer pattern image taken from the CMOS image sensor used in the digital camera should be converted to the RGB image using many complex enhancement functions in the ISP (Image Signal Processor).

The full chain of camera ISP functions for smart devices is presented. The every function of the chain is fully converted to fixed point arithmetic and no special function is used for easy porting to reconfigurable processors.

ISP full chain for vector processors is White Balance(WB), Modified Adaptive Homogeneity-directed Domosaicing(M-AHD), Color Correction(CC), Auto Contrast(AC), Gamma Correction(GC), Modified Bilateral Filter(M-BF), Difference of Gaussian(DoG) based Luminance Transient Improvement(LTI)/Chrominance Transient Improvement(CTI). After performing white balance in the bayer pattern, M-AHD, which is based on AHD, is used as demosaicing algorithm. The AHD consist of three steps; directed interpolation, homogeneity-directed map creation, iteration. The directed interpolation is made of fixed point arithmetic with 2bit extension. The coefficient of homogeneity-directed map creation is converted to fixed point. Also, CIELab that is used as the color space in the original homogeneity-directed map is changed to YCbCr color space. The iteration function is removed for reducing operational loads. After demosaicing, the color correction block finds color features and repairs color artifacts.

Linear stretch method is used in the auto contrast. The linear scale factors in the auto contrast function are calculated in YPbPr color space. A color control function is also used in YPbPr color space to control color saturation and color offset.

Gamma Correction with a reduced LUT(Lookup table) is used and the gamma values between LUT entries are calculated by piecewise linear interpolation method.

M-BF is used as a noise reduction algorithm and based on BF. The original BF has two Gaussian filters. One is for distance weight between pixel locations. The other is for difference weight between pixel intensities. In order to simplify the two Gaussian filters, the Gaussian functions are replaced by fixed point binary threshold functions. The threshold values are determined by pre-calculating the Gaussian filter coefficient for pixel locations and pixel intensities.

For detail enhancement DoG based LTI/CTI is used. The Gaussian mask sizes in the LTI/CTI are 3x3 and 5x5, which are with pre-calculated coefficients.

To verify the performance of the proposed ISP full chain, series of test are performed to check if the quality of result images can pass a commercially arranged test. Test metrics are MTF30, Oversharpening, Edge roughness texture acuity, exposure error, gamma, SNR, Dynamic range, light falloff, color uniformity, SMIA, Delta C00. The experiments are conducted using a CMOS image sensor whose specification is set. There are several test patterns to measure the image quality properly. One pattern is image quality resolution test pattern. It is used to calculate MTF30, over sharpening. Another pattern is color acuity test pattern that is used to evaluate color performance such as Delta C00.

All measured values meet the requirements of the test. Since whole proposed ISP chain is designed only with fixed point addition and multiplication, the proposed ISP chain can be easily ported into any reconfigurable processor.

CGA acceleration result is better than no acceleration for the auto contrast, LTI, CTI function. Acceleration is more than 700%. The results show that the algorithm can be easily converted into the CGA acceleration mode.

Currently the proposed ISP chain is being implemented on a specific reconfigurable processor that has 8way x 16bit core to support VLIW and/or CGA(Coarse-grained Array) units to verify real-time full HD (1920x1080x30) support with commercial quality.

9016-32, Session 9

Noisy images-JPEG compressed: subjective and objective image quality evaluation

Silvia Corchs, Francesca Gasparini, Raimondo Schettini, Univ. degli Studi di Milano-Bicocca (Italy)

Subjective studies within image quality mainly focus on images corrupted by only one distortion. However, consumer images suffer in general of more than one distortion simultaneously. Recently Jayaraman et al. [1] has presented a database of multiply distorted images, where two scenarios are considered: images first blurred and then JPEG compressed, and images first blurred and then corrupted by white Gaussian noise. The authors correlate the psycho-visual scores on this database with a variety of existing Full-Reference (FR) image quality metrics, and a No-Reference (NR) general purpose one [2].

The aim of this work is to study image quality of both single and multiply distorted images. We here address the case of images corrupted by Gaussian noise as single distortion case and images corrupted by Gaussian noise and then JPEG compressed, as multiply distortion case. We are interested in evaluating if and how the quality perception of images corrupted by noise is modified in the presence of JPEG distortion. A subjective study was conducted in two parts to obtain human judgments on the single and multiply distorted images and collect the corresponding psycho-visual data. We study in this work how these subjective data correlate with NR state-of-the-art metrics. In particular we evaluate and compare the performance of single distortion-specific metrics (for measuring both noise and JPEG artifacts) and general purpose ones. Moreover we here investigate proper combining of NR distortion-specific metrics to achieve better correlation performance. The results are analyzed and compared in terms of correlation coefficients.

DATABASES

In this work we consider the IVL database [3]. It consists of 20 reference images of 886x591 pixels (10x15 cm at 150 dpi). Starting from these images we have first generated a database of noisy images consisting of 200 distorted images. The distorted images have been obtained as follows: for each of the 20 reference images we have created 10 corrupted versions with: 1, 2, 3, 4, 5, 6, 8, 10, 12 and 14 gray level of standard deviation on the luminance channel. A second database of multiply distorted images has been generated as follows: each of the 200 noisy images were further corrupted by 3 different levels of JPEG compression (Q factors = 10, 30, and 50, for a total of 600 multiply distorted images).

For the experimental sessions on these two different databases, we have adopted a Single Stimulus method (SS) [4], where all the images are individually shown. We have decided to adopt the SS method to better represent the reality where users of digital photographs do not in general dispose of the reference image (NR image quality assessment). The observers were asked to rate the images within a continuous scale from 0 to 100. Semantic labels Bad, Poor, Fair, Good and Excellent were marked at equal distances along the scale. The experiments were performed following the recommendations in [4]. Full details of the psycho-visual test will be reported in the final version of the paper.

DATA ANALYSIS

The subjective scores, collected in terms of Mean Opinion Scores (MOS) are correlated with several distortion-specific NR metrics (such as [5], [6], and [7]) and general purpose ones [2, 8]. We are here mainly interested in evaluating if single distortion-specific metrics (for measuring both noise and jpeg artifacts) can still predict the subjective scores in case of presence of multiple distortions and if general purpose ones perform better. Moreover, we here propose proper combining of NR distortion-specific metrics and we investigate if they outperform the performance of the single ones, in the case of multiply distorted images.

We adopted as correlation functions both the logistic one and the monotone regression proposed by [9]. As recommended by [10], to evaluate the performance of the regressed metrics we have considered the following statistic coefficients: the Pearson correlation coefficient (CC) and the root-mean-squared error (RMSE), to quantify the prediction accuracy, the Spearman rank-order correlation coefficient (SROCC) to predict the monotonicity of the correlation, and the outlier ratio (OR) to quantify the prediction consistency. To also check whether the numerical differences found in the performance of the analyzed metrics, we also conduct hypothesis testing (T-Test or Wilcoxon Rank-Signed-Test).

The detailed analysis of the correlations of subjective scores with the considered metrics (single distortion-specific and general purpose ones, and proper combining of them), will be deeply presented in the final paper, together with the analysis of the statistical significance of the results.

REFERENCES

[1] Dinesh Jayaraman, Anish Mittal, Anush K. Moorthy and Alan C. Bovik, Objective Quality Assessment of Multiply Distorted Images, Proceedings of Asilomar Conference on Signals, Systems and Computers, 2012.

[2] A. Mittal, A. K. Moorthy and A. C. Bovik, No-Reference Image Quality Assessment in the Spatial Domain, IEEE Transactions on Image Processing 21, 4695-4708, 2012.

[3] S. Corchs, F. Gasparini and R. Schettini, No Reference Image Quality Classification for JPEG Distorted Images, Digital Signal Processing, under revision, 2013.

[4] Recommendation 500-11: Methodology for the subjective assessment of the quality for television pictures, ITU-R Rec. BT.500, 2002.

[5] J. Immerkaer, Fast Noise Variance Estimation, Computer vision and image understanding, vol. 64, 300–302, 1996.

[6] Z. Wang, A. C. Bovik, B. L. Evans, Blind measurement of blocking artifacts in images, in: Proc. International Conference on Image Processing, volume 3, IEEE, 2000, pp. 981-984.

[7] R. Muijs, I. Kirenko, A no-reference blocking artifact measure for adaptive video processing, in: Proceedings of the 13th European Signal Processing Conference 2005.

[8] A. Mittal, R. Soundararajan, A. C. Bovik, Making a completely blind image quality analyzer, IEEE Signal Processing Letters 20 (2013) 209-212.

[9] Y.Han, Y. Cai, Y. Cao, and X. Xu, Monotonic Regression: A New Way for Correlating Subjective and Objective Ratings in Image Quality Research, IEEE Transactions on Image Processing, 21 (4) 2012, 2309-2013.

[10] VQEG, Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II, 2003, http://www.vqeg.org.

9016-33, Session 9

Perceptibility and acceptability of JPEG 2000 compressed images of various scene types

Elizabeth Allen, Sophie Triantaphillidou, Ralph E. Jacobson, Univ. of Westminster (United Kingdom)

The JPEG 2000 image compression standard was developed in the late 1990s to improve upon JPEG in terms of both flexibility and image quality; also to meet the growing demands in image compression of different digital imaging applications, to produce superior rate-distortion at low bit rates to that of existing standards [1]. The lossy version of JPEG 2000 is a transform-based encoder, using a Discrete Wavelet Transform (DWT) and encoding the images by bit plane, which produces various characteristic artefacts, visible below a certain threshold. Part 1 was standardised in 2001 [2,3]. Although not widely adopted in commercial imaging, where baseline JPEG remains the dominant lossy compression algorithm, JPEG 2000 has found application in a number of more specialised sectors of the imaging industries, including forensic imaging, where it is used in the compression of images of fingerprints and footprints, and in certain fields within medical imaging.

The quality benefits of JPEG 2000 over baseline JPEG and other relevant compression algorithms remain an area of research interest, particularly in relation to the compression of images at low bit rates. The authors of this publication have conducted a previous comparison between JPEG and JPEG 2000 across a range of scenes [4], with a further investigation into the scene dependency, analysis and classification of the same sample set of scenes [5]. Recent studies have investigated JPEG 2000 in comparison with JPEG for the compression of dermatological images [6], and the image fidelity of JPEG 2000 compressed images for body CT scans [7].

Image fidelity is a measure of the degree of visible correspondence between a reproduced image and a reference image: in image compression, this is the original uncompressed image. Image fidelity studies provide a measure of the threshold of perceptibility of distortions introduced by lossy image processes and have particular relevance to applications where the integrity of the image is important. Image fidelity may be predicted using objective fidelity metrics, or evaluated using psychophysical studies, aiming to derive a psychometric curve from the observer's responses. The fidelity of an image does not always correlate well with observer preference (i.e. its quality), because for certain images, the distortions introduced by an image processing algorithm at certain levels may not be detrimental, but represent an improvement in image quality [4,5].

Image acceptability studies aim to define the point at which degradation to image quality becomes unacceptable. They can provide a better correlation to perceived image quality than image fidelity within a specific context. In psychophysical studies, acceptability may be evaluated using the same methodology as that used to determine fidelity, but providing observers with a different question.

In many subjective image quality studies the results are found to be scene dependent [8, 9, 10]. Triantaphillidou et al [5] describe different types of scene dependency arising from three main sources, including: observer preferences in relation to image attributes across particular types of scenes (for example a general preference for slight blurring on portraits); the effects of scene content upon the performance of image processing algorithms; and the visibility of artefacts in particular scenes as a result, for example, of masking effects. The correlation between image fidelity and image acceptability for a group of images may provide an indicator of scene dependencies affecting a particular study and can be useful in investigating methods of scene classification.

The investigation consists of two separate but related psychophysical experiments, to determine the fidelity and acceptability of a sample set of images, containing a range of different scene types, and pre-classified objectively using various scene attributes. The images were selected from a larger image database using scene classification methods tested in the previous study by Triantaphillidou et al [5]. The final sample set fell into three broad categories, defined as 'average', 'greater than average', and 'less than average', for a number of different scene attributes and combinations of attributes, including average scene luminance, contrast, colour contrast and busyness.

Observers were presented with pairs of images of the same scene; one compressed using JPEG 2000, the other a reference, uncompressed image on the same display, in an sRGB-viewing environment. They were asked to provide a 'yes' or 'no' answer to the questions: "Do the two images appear to be different?" for the image fidelity investigation, and "is the compressed image acceptable in terms of overall image quality in comparison with the original?" for the image acceptability investigation. The results were used to derive psychometric curves to relate the proportion of observers' 'Yes' responses to the levels of compression for thresholds of perceptibility and acceptability. Results relating the perceptibility and acceptability of individual images and in the context of the original groupings will be presented in this paper. Conclusions will be drawn on the level of JPEG 2000 compression various scene types require to maintain image fidelity, and to maintain an acceptable image quality.

1. Skodras, A.Christopoulos, C. Ebrahimi, T., The JPEG 2000 Still Image Compression Standard. IEEE Sign. Proc. Magazine, pp 36-58 (September 2001)

2. ISO/IEC JTC 1/SC 29/WG 1 (ITU-T SG8) The JPEG 2000 Still Image Compression Standard, M. D., Adams, Dept. of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada V6T 1Z4 (2001)

3. Adams, M. D, Manz, H., Kossentiniy, F., Ebrahimi, T., JPEG 2000: The Next Generation Still Image Compression Standard, Available at the time of writing from www.jpeg.org.

4. Allen, E., Triantaphillidou, S., Jacobson, and R.E., Image quality comparison between JPEG and JPEG 2000 I- Psychophysical Investigation, J. Imag.Sci & Tech, Volume 51, Number 3, May 2007 , pp. 248-258(11)

5. Triantaphillidou, S. Allen, E., Jacobson, and R.E., Image quality comparison between JPEG and JPEG 2000 II- Scene Dependency, Scene Analysis and Classification, J. Imag.Sci & Tech, Volume 51, Number 3, May 2007 , pp. 259-270(12)

6. Guarneri, F., Vaccaro, M., Guarneri, C., and Cannavo, T., JPEG vs. JPEG 2000: benchmarking with dermatological images, Skin Research and Technology, (2013), Early View, online version, John Wiley and Sons.

7. Joong Kim, K., Ho Lee, K., Kang, H., Yeon Kim, S., Hoon Kim, Y., Bohyoung Kim, B., Seo, J., Rafal Mantiuk, R., Objective index of image fidelity for JPEG2000 compressed body CT images, Medical Physics 36, 3218 (2009)

8. Jacobson, R.E., Triantaphillidou,S., Metric approaches to image quality, Colour Image Science: Exploiting Digital Media, ed. MacDonald L.W and Luo M.R, (John Wiley & Sons, England,2002) Ch. 18, p. 371.

9. H. Frieser, and K. Biedermann, Experiments on image quality in relation to the modulation transfer function and graininess of photographs, Photogr. Sci. Eng, 7, 28 (1963).

10. Corey, G.P, Clayton, M.J. and Cuprey, K.N., Scene Dependence of Image Quality, Phot.

Sci. Eng., 1982, 27, 9-13.

9016-34, Session 9

A new image quality assessment database for tiled images

Steven McFadden, Paul A. S. Ward, Univ. of Waterloo (Canada)

Demand for large ultra-high-resolution displays continues to grow as the need to visualize large data sets increases, but the size of an individual display is generally limited by manufacturing constraints. Tiled displays have emerged as a way to overcome these constraints and create arbitrarily large displays by tiling smaller individual displays, each of which shows a portion of the full image. These displays also benefit from

the efficiencies of mass-produced “typical-size” displays.

The two common types of tiled displays used today (LCD and rear-projection) suffer from some common challenges (i.e., color matching and brightness uniformity between tiles) and the worst of these challenges is currently the display gap. When the sub-displays are tiled together, a gap is created between the display area of each tile (a bezel gap for LCD displays; a screen expansion gap for rear-projection displays). This gap manifests as a grid overlaid on top of the image. This grid distortion can be minimized by reducing the size of the gap between each tile’s display area but cannot currently be eliminated entirely (LCDs require at least some bezel for electronics and structural integrity; rear-projection display screens require room to expand due to temperature and humidity changes).

Image quality is commonly measured using objective Image Quality Assessment (IQA) metrics. These metrics are not as reliable as subjective quality measurements but are far cheaper and faster. The performance of any given objective IQA metric is measured by how well it correlates with subjective results in IQA databases. There are currently six commonly used, publicly accessible, IQA databases: LIVE, A57, Toyama, IVC, CSIQ, and TID2008. These include many common image distortion types such as noise, compression artifacts, transmission errors, and blur. None of these databases include any distortions created by tiling displays.

The image distortions created by tiled displays differ significantly from the distortions currently existing in the IQA databases. Unlike the “common” distortions where the entire image is affected more or less equally (subject to the image content), the tiled distortions affect only small portions of the image (i.e. the gaps). Outside of these specific areas, the image quality is perfect with no distortion at all. This difference suggests current IQA metrics may be unsuitable for measuring the quality of tiled displays.

Initial research through the use of an informal subjective user study indicated poor performance of the Structural Similarity Index (SSIM) when applied to tiled images. In this user study, test subjects rated the quality of a set of images corrupted by two different distortions: blur and a tiling grid. The correlation between the SSIM scores and subjective results was significantly lower for grid-distorted images than for blur-distorted images. The blur-distorted images were used as a baseline for the user study and results for these images correlated well with results that use the public databases.

To verify the results of the informal study, a new IQA database for tiled images was created by means of a formal user study. To focus initially on the most significant tiling distortion, this user study only investigated the image quality related to gap distortion (i.e., no color or brightness mismatches were applied to the test images). Early analysis of results from this new database supports the findings of the informal study and further research will include a comparison of current image quality metrics to determine suitable choices for analyzing quality of tiled images.

Conference 9017: Visualization and Data Analysis 2014

Monday - Wednesday 3 –5 February 2014

Part of Proceedings of SPIE Vol. 9017 Visualization and Data Analysis 2014

9017-1, Session 1

FilooT: a visualization tool for exploring genomic data

Mahshid Zeinaly, Mina Soltangheis, Christopher D. Shaw, Simon Fraser Univ. (Canada)

In order to enhance analysis of synthetic health data of the IEEE vast challenge 2010, we introduce an interactive InfoVis tool called FilooT designed as a part of the Interactive Multi- genomic Analysis System (IMAS) project. In this paper we described different interactive views of FilooT: The tabular view for exploring and comparing genetic sequences, the matrix view for sorting sequences according to the values of different characteristics, P-value View for finding the most important mutations across a family of sequences, the Graph View for finding related sequences and the Group View to group them for further investigation. We followed the Nested Process Model framework throughout the design process and the evaluation. To understand the tool's design capabilities for target domain analysts, we conducted a User Experience scenario-based study followed by an informal interview. The findings indicated how analysts employ each of the visualization and interaction designs in their Bioinformatics' task-analysis process. The critical analysis of the results inspired design informing suggestions.

9017-2, Session 1

A framework for analysis of the upper airway from real-time MRI sequences

Samuel de Sousa Silva, António L. J. Teixeira, Univ. de Aveiro (Portugal)

In recent years, real-time Magnetic Resonance Imaging (RT-MRI) has been used to acquire vocal tract data to support articulatory studies.

The large amount of images resulting from these acquisitions needs to be processed and the resulting data analysed to extract articulatory features. This is often performed by linguists and phoneticists and requires not only tools providing a high level exploration of the data, to gather insight over the different aspects of speech, but also a set of features to compare different vocal tract configurations in static and dynamic scenarios.

In order to make the data available in a faster and systematic fashion, without the continuous direct involvement of image processing specialists, a framework is being developed to bridge the gap between the more technical aspects of raw data and the higher level analysis required by speech researchers. In its current state it already includes segmentation of the vocal tract, allows users to explore the different aspects of the acquired data using coordinated views, and provides support for vocal tract configuration comparison. Beyond the traditional method of visual comparison of vocal tract profiles, a regional quantitative method is proposed, considering relevant anatomical features, and analysis supported by an abstract representation of the data both for static and dynamic analysis.

9017-3, Session 2

VAFLE: visual analytics of firewall log events

Mohammad Ghoniem, Ctr. de Recherche Public - Gabriel Lippmann (Luxembourg); Georgiy Shurkhovetskyy, Modern Sciences and Arts Univ. (Egypt); Ahmed Bahey, Nile Univ. (Egypt); Benoît Otjacques, Ctr. de Recherche Public - Gabriel Lippmann (Luxembourg)

In this work, we present VAFLE, an interactive network security visualization prototype for the analysis of firewall log events. Keeping it simple yet effective for analysts, we provide multiple coordinated interactive visualizations augmented with clustering capabilities customized for the forensic analysis of network traffic. We evaluate the usefulness of the prototype in a use case with network traffic datasets from previous VAST Challenges, illustrating its effectiveness at promoting fast and well-informed decisions. We explain how a security analyst may spot suspicious traffic using VAFLE. We further assess its usefulness through a qualitative evaluation involving network security experts, whose feedback is reported and discussed.

9017-4, Session 2

Configurable IP-space maps for large-scale, multi-source network data visual analysis and correlation

Scott B. Miserendino, Corey Maynard, William E. Freeman, Northrop Grumman Corp. (United States)

The need to scale visualization of cyber (IP-space) data sets and analytic results as well as to support a variety of data sources and missions have proved challenging requirements for the development of a cyber common operating picture. Typical methods of visualizing IP-space data require unreliable domain conversions such as IP geolocation, network topology that is difficult to discover, or data sets that can only display one at a time. In this work, we introduce a generalized version of hierarchical network maps called configurable IP-space maps that can simultaneously visualize multiple layers of IP-based data at global scale. IP-space maps allow users to interactively explore the cyber domain from multiple perspectives. A web-based implementation of the concept is described, highlighting a novel repurposing of existing geospatial mapping tools for the cyber domain. Benefits of the configurable IP-space map concept to cyber data set analysis using spatial statistics are discussed. IP-space map structure is found to have a strong effect on data clustering behavior, hinting at the ability to automatically determine concentrations of network events within an organizational hierarchy.

9017-5, Session 3

The CZSaw notes case study

Eric Lee, Ankit Gupta, David Darvill, John C. Dill, Christopher D. Shaw, Robert Woodbury, Simon Fraser Univ. (Canada)

Analysts need to keep track of their analytic findings, observations, ideas, and hypotheses throughout the analysis process. While some visual analytics tools support such note-taking needs, these notes are often represented as objects separate from the data and in a workspace separate from the data visualizations. Representing notes the same way as the data and integrating them with data visualizations can enable analysts to build a more cohesive picture of the analytical process. We created a note-taking functionality called CZNotes within the visual analytics tool CZSaw for analyzing unstructured text documents. CZNotes are designed to use the same model as the data and can thus be visualized in CZSaw's existing data views.

We conducted a preliminary case study to observe the use of CZNotes and observed that CZNotes has the potential to support progressive analysis, to act as a shortcut to the data, and supports creation of new data relationships.

9017-6, Session 3

Linked visual analysis of structured datasets and document collections

Sebastin Kolman, Ekaterina Galkina, Andrew S. Dufilie, Yen Fu Luo, Vivek Gupta, Georges G. Grinstein, Univ. of Massachusetts Lowell (United States)

Analysts use visual analytical systems for exploring and analyzing structured datasets and increasingly require tools to access supporting documents, research papers and news reports. Visual analytical systems for text corpora typically concentrate on techniques for exploring only document collections. We have developed a system for visualizing and analyzing both document collections and structured datasets. We describe a document visualization tool called InfoMaps developed within Weave, an open source framework for data exploration and analysis. Users of Weave analyzing datasets can search for documents from the web and networked repositories, and can use the matched documents as a part of their analysis. Conversely documents in InfoMaps can be used to identify relevant data subsets. In this paper, we discuss InfoMaps, its use and integration with other visual tools of Weave and our approach to the information extraction and integration process.

9017-7, Session 4

A reference web architecture and patterns for real-time visual analytics on large streaming data

Eser Kandogan, Danny Soroker, Steven Rohall, IBM Corp. (United States); Peter Bak, IBM Corp. (Israel); Frank van Ham, IBM Corp. (Netherlands); Jie Lu, IBM Corp. (United States); Harold J. Ship, IBM Corp. (Israel); Chun-Fu Wang, Univ. of California, Davis (United States); Jennifer Lai, IBM Corp. (United States)

Monitoring and analysis of streaming data, such as social media, sensors, and news feeds, has become increasingly important for business and government. Volume and velocity of incoming data are key challenges. To effectively support monitoring and analysis, statistical and visual analytics techniques need to be seamlessly integrated; analytic techniques for a variety of data types (e.g. text, numerical) and scope (e.g. incremental, rolling-window, and global) must be properly accommodated; interaction and coordination among several visualizations must be supported in an efficient manner; and the system should support the use of different analytics techniques in a pluggable and collaborative manner. Especially in web based environments, these requirements pose restrictions on the basic architecture for such systems. In this paper we report on our experience of building a reference web architecture for real-time visual analytics of streaming data, identify and discuss architectural patterns that address these challenges, and report on applying the reference architecture for real-time Twitter monitoring.

9017-8, Session 4

Visualizing confusion matrices for multidimensional signal detection correlational methods

Yue Zhou, Thomas Wischgoll, Wright State Univ. (United States); Leslie M. Blaha, Air Force Research Lab. (United States); Ross A. Smith, Rhonda J. Vickery, Dynamics Research Corp. (United States)

General Recognition Theory (GRT) is a multidimensional signal detection theory framework for capturing sources of perceptual and decisional dependence. The primary type of data for GRT models is an identification-confusion matrix derived in a complete factorial identification task. This confusion matrix plots the responses of study participant for a given signal. The responses may reveal that participants were unable to recognize the signal properly. Such violations of any type of independence in the GRT framework result in response patterns that reflect some form of correlation in the GRT space. While an individual confusion matrix is rather small and relatively easy to visualize, similar studies may be repeated a lot of times resulting in thousands if not millions of confusion matrices that have to be visualized. This paper describes methods that adapt the web-based D3 visualization framework combined with pre-processing tools for the raw data to identify ways of visualization methodologies that enable domain specialists to more easily interpret their data. As the D3 framework utilizes Javascript and scalable vector graphics (SVG) to generate the visualizations it can run readily within the web browser to directly enable deployment of the visualization algorithms by the domain specialists. Parallel coordinate plots and heat maps were developed for the confusion matrix data, and the results were shown to a GRT expert for an informal evaluation of their utility. There is a clear benefit to model interpretation from these visualizations when researchers need to interpret larger amounts of simulated data.

9017-9, Session Key1

What makes cybersecurity visualizations different from other types of visualization *(Keynote Presentation)*

Pak Chung Wong, Pacific Northwest National Lab. (United States)

No Abstract Available

9017-10, Session 5

User-driven sampling strategies in image exploitation

Neal Harvey, Reid B. Porter, Los Alamos National Lab. (United States)

Visual analytics and interactive machine learning both try to leverage the complementary strengths of humans and machines to solve complex data exploitation tasks. These fields overlap most significantly when training is involved: the visualization or machine learning tool improves over time by exploiting observations of the human-computer interaction. This paper focuses on one aspect of the human-computer interaction that we call user driven sampling strategies. Unlike relevance feedback and active learning sampling strategies, where the computer selects which data to label at each iteration, we investigate situations where the user selects which data is to be labeled at each iteration. User driven sampling strategies can emerge in many visual analytics applications but they have not been fully developed in machine learning. User driven sampling strategies suggest new theoretical and practical research questions for both visualization science and machine learning. In this paper we identify and quantify the potential benefits of these strategies in a practical image analysis application. We find user driven sampling strategies can sometimes provide significant performance gains by steering tools towards local minimum that have lower error than tools trained with all of the data. We find these performance gains are particularly pronounced when the user is experienced with the tool and application domain.

9017-11, Session 6

Collaborative data analysis with smart tangible devices

Johannes Fuchs, Roman Rädle, Dominik Sacha, Fabian Fischer, Andreas Stoffel, Univ. Konstanz (Germany)

We present a tangible approach for exploring and comparing multi-dimensional data points collaboratively by combining Sifteo Cubes with glyph visualizations. Various interaction techniques like touching, shaking, moving or rotating the displays support the user in the analysis. Context dependent glyph-like visualization techniques make best use of the available screen space and cube arrangements. As a first proof of concept we apply our approach to real multi-dimensional datasets and show with a coherent use case how our techniques can facilitate the exploration and comparison of data points. Finally, further research directions are shown when combining Sifteo Cubes with glyphs and additional context information provided by multi-touch tables.

9017-12, Session 6

Visualization of off-screen data on tablets using context-providing bar graphs and scatter plots

Peter S. Games, Alark Joshi, Boise State Univ. (United States)

Visualizing data on tablets is challenging due to the relatively small screen size and limited user interaction capabilities. Standard data visualization apps provide support for pinch-and-zoom and scrolling operations, but do not provide context for data that is off-screen. When exploring data on tablets, the user must be able to focus on a region of interest and quickly find interesting patterns in the data. We present visualization techniques that facilitate seamless interaction with the region of interest on a tablet using context-providing bar graphs and scatter plots. Through aggregation, fisheye-style, and overview+detail representations, we provide context to the users as they explore a region of interest. We evaluated the efficacy of our techniques with the standard, interactive bar graph and scatter plot applications on a tablet, and found that one of our bargraph visualizations - Fisheye-style Focus+Context visualization (BG2) resulted in the fewest errors, least frustration and took the least amount of time. Similarly, one of our scatter plot visualizations - User Driven Overview+Detail (SP3) - resulted in the fewest errors, least frustration and took the least amount of time. Overall, users preferred the context-providing techniques over traditional bar graphs and scatter plots, that include pinch-and-zoom and fling-based scrolling capabilities.

9017-13, Session 6

HyFinBall: a two-handed, hybrid 2D/3D desktop VR interface for multi-dimensional visualization

Isaac Cho, Xiaoyu Wang, Zachary J. Wartell, The Univ. of North Carolina at Charlotte (United States)

This paper presents our interactive multi-dimensional visualization and the concept, working prototype and design space of a two-handed, hybrid spatial user interface for desktop VR targeted at users where a minimally immersive, desktop VR system is appropriate. The user interface supports dual button balls (6DOF isotonic controllers with multiple buttons) which automatically switch between 6DOF mode (xyz + yaw,pitch,roll) and planar-3DOF mode (xy + yaw) upon contacting the desktop. The mode switch automatically switches a button ball's visual representation between a 3D cursor and a mouse-like 2D cursor while also switching the available user interaction techniques (ITs) between 3D and 2D ITs. Further, the small form factor of the button ball allows the user to engage in 2D multi-touch or 3D gestures without releasing and re-acquiring the device. We call the device and hybrid interface the HyFinBall interface which is an abbreviation for 'Hybrid Finger Ball.' We describe the user interface (hardware and software), the design space, as well as preliminary results of a formal user study. This is done in the context of a rich, visual analytics interface containing coordinated views with 2D and 3D visualizations and interactions

9017-14, Session 7

Visualizing trends and clusters in ranked time-series data

Michael B. Gousie, John Grady, Melissa Branagan, Wheaton College (United States)

There are many systems that provide visualizations for time-oriented data. Of those, few provide the means of finding patterns in time-series data in which rankings are also important.

Fewer still have the fine granularity necessary to visually follow individual data points through time.

We propose the Ranking Timeline, a novel visualization method for modestly-sized multivariate data sets that include the top ten rankings over time.

The system includes two main visualization components: a ranking over time and a cluster analysis. The ranking visualization, loosely based on line plots, allows the user to track individual data points so as to facilitate comparisons within a given time frame. Glyphs represent additional attributes within the framework of the overall system.

The user has control over many aspects of the visualization, including viewing a subset of the data and/or focusing on a desired time frame.

The cluster analysis tool shows the relative importance of individual items in conjunction with a visualization showing the connection(s) to other, similar items, while maintaining the aforementioned glyphs and user interaction.

The user controls the clustering according to a similarity threshold.

The system has been implemented as a Web application, and has been tested with data showing the top ten actors/actresses from 1929-2010.

The experiments have revealed patterns in the data heretofore not explored.

9017-15, Session 7

Relating interesting quantitative time series patterns with text events and text features

Franz Wanner, Tobias Schreck, Wolfgang Jentner, Lyubka Sharalieva, Daniel A. Keim, Univ. Konstanz (Germany)

In many application areas, the key to successful data analysis is the integrated analysis of heterogeneous data. One example is the financial domain, where time-dependent quantitative data (e.g., trading volume and price information) and textual data (e.g., economic and political news reports) need to be considered jointly. Data analysis tools need to support an integrated analysis, which allows studying the relationships between textual news documents and quantitative properties of the stock market price series. In this paper, we describe a workflow and tool that allows a flexible formation of hypotheses about text features and their combinations, which reflect quantitative phenomena observed in stock data.

To support such an analysis, we combine the analysis steps of quantitative and text-oriented data using an a-priori method. First, based on heuristics we extract interesting intervals and patterns in large time series data. The visual analysis supports the analyst in exploring

parameter combinations and their results. The identified time series patterns are then input for the second analysis step, in which all identified intervals of interest are analyzed for frequent patterns co-occurring with financial news. An a-priori method supports the discovery of such sequential temporal patterns. Then, various text features like the degree of sentence nesting, noun phrase complexity, the vocabulary richness, etc. are extracted from the news to obtain meta patterns. Meta patterns are defined by a specific combination of text features which significantly differ from the text features of the remaining news data. Our approach adequately combines a portfolio of visualization and analysis techniques, including time-, cluster- and sequence visualization and analysis functionality. We provide two case studies, showing the effectiveness of our combined quantitative and textual analysis work flow. The workflow can also be generalized to other application domains such as data analysis of smart grids, cyber physical systems or the security of critical infrastructure, where the data consists of a combination of quantitative and textual time series data.

9017-16, Session Key2

Just-in-time visual analytics and discovery *(Keynote Presentation)*

Eser Kandogan, IBM Almaden Research Ctr. (United States)

No Abstract Available

9017-17, Session 8

Visualization of multidimensional data with collocated paired coordinates and general line coordinates

Boris Kovalerchuk, Central Washington Univ. (United States)

Often multidimensional data are visualized by splitting n-D data to a set of low dimensional data. While it is useful it destroys integrity of n-D data, and leads to a shallow understanding complex n-D data. To mitigate this difficulty an additional and difficult perceptual task of assembling low-dimensional visualized pieces of each record to the whole n-D record must be solved. An alternative way for deeper understanding of n-D data is developing visual representations of n-D data in low dimensions without such data splitting. Methods of Parallel and Radial coordinates are such methods. Developing new methods of this type is a long standing and challenging task that this paper attempts to address by proposing Paired Coordinates that is a new type of n-D data visual representation and by generalizing Parallel and Radial coordinates as a General Line coordinates. The important novelty of the concept of the Paired Coordinates is that uses a single 2-D plot to represent n-D data as an oriented graph based on the idea of collocation of pairs of attributes. The advantage of the General Line Coordinates and Paired Coordinates is that they provide a common framework that includes Parallel and Radial coordinates and generate a large number of new visual representations of multidimensional data without dimension reduction.

9017-23, Session PTues

Evaluation in visualization: some issues and best practices

Beatriz Sousa Santos, Paulo M. Dias, Univ. de Aveiro (Portugal)

Visualization researchers are more and more aware of the importance of evaluation, as it is not only a means of improving techniques and applications, but it can also produce evidence of measurable benefits that will encourage adoption. Yet, evaluating visualization applications or techniques, is not simple. This paper presents some issues that have to be considered while planning an evaluation and a brief description of the methods that the authors have been using to evaluate visualization techniques and applications concerning the users' performance in their tasks when using the visualization, implying the understanding of the phenomenon. A list of guidelines considered as best practices to perform evaluations is also presented and some conclusions are drawn.

(a file contatining the full paper will be uploaded)

9017-24, Session PTues

Interactive word cloud for analyzing reviews

HyunRyong Jung, FactSet Research Systems Inc. (United States)

A five-star quality rating is one of the most widely used systems for evaluating items. However, it has two fundamental limitations: 1) the rating for one item cannot describe crucial information in detail; 2) the rating is not on an absolute scale for comparing to other items. Because of these limitations, users cannot make right decision. In this paper, we introduce our sophisticated approach to extract useful information from user reviews using collapsed dependencies and sentiment analysis. We propose an interactive word cloud that can show grammatical relationships among words, explore reviews efficiently, and display positivity or negativity on a sentence. In addition, we introduce visualization for comparing multiple word clouds and illustrate the usage through test cases.

9017-25, Session PTues

Stars advantages vs parallel coordinates: shape perception as visualization reserve

Vladimir Grishin, View Trends International (United States)

Popular now Parallel Coordinates (PC) (2D Cartesian displays of data vectors) are compared with polar displays (Stars) analytically and experimentally. Advantages of stars vs. PCs by Gestalt Laws are shown. About twice faster feature selection and classification with Stars than PCs are proved by psychological experiments for hyper-tubes structures detection in data space with dimension up to 100-200 and its subspaces, i.e. in range of many tasks of technological and medical diagnostics, decision making, drug design, etc.

1. INTRODUCTION

1.1 Stars vs. parallel coordinates

This research was motivated by controversial results of a few decades discussion about the comparative advantages and shortcomings of different 2D contour representations as star glyphs, parallel coordinates (PCs), bar graphs, etc. for visualization and mining of multi-variant data with dimensions of 10 < D < 100-200 (tens and hundreds of coordinates) which is typical for medical and technical diagnostics, chemical and drug design, many decision making tasks, etc.

On one hand it seemed that this discussion was closed in favor of stars [1,2,3], but recent publications show a much higher popularity of parallel coordinates displays vs. stars[4,5,6]. Apparently quantitative measurement of separate data attributes prevails over qualitative pattern recognition claimed as main visualization goal.

Therefore this work was dedicated to comparative analysis and experimental estimation of these displays capabilities for feature selection and pattern recognition in modeled data structures as hyper-tubes and hyper-spheres in D<=96 data space. Such formal description of data mining results by visualization is really needed for creation of a quantitative theory of visualization.

1.2 Stars vs projections of data space on plane

Additional goal of this work was to demonstrate that human vision using stars can effectively solve data structure analysis for D at least up to 100. This is contrary to current popular methods displaying data

space structures by distances on the 2D plane (projections, principal components, multi dimensional scaling, etc.). They are based on the idea that as far as humans can see only 2D-3D spatial structures, we have to represent multi dimensional (MD) data space structures as 2D-3D spatial structures. Of course human vision can not see data structures directly in the D>3 data space but one can analyze and describe these structures indirectly if the data attributes are mapped into visible image features. Therefore better to use this capability of stars and similar displays to analyze these structures instead or before rejection of the most part of unknown and may be very useful data information by projections.

2. SHAPE PERCEPTION AND GESHTALT LAWS

2.1 Poor usage of shape perception capabilities for data exploration.

Shape perception plays a key role in visual pattern recognition. Although 90% - 95 % of the information in an image is perceived by the human visual system from object shapes this capability is very rarely exploited in current displays. Humans can detect in parallel, compare, and describe many figures by hundreds of local shape features (concave, convex, angle, wave, etc.) with many attributes (size, orientation, location, etc.) and combine them into a multilevel hierarchy (see Fig.1).

Mapping data vectors into simply or multiply connected contours, we can see and describe by means of such features very complicated nonlinear structures in data space. Invariance of visual perception under shape affine transformations radically extends these capabilities. However, modern visualization research usually refers to just a few simplest features on pictures and rarely links it with data structures.

For better understanding and quantitative description of data space structures which leverage visual capabilities we have to use displays that provide analytically simple connections between data attributes and the image features (stars, parallel coordinates, pie- and bar-charts, and other similar displays).

2.2 Contour displays comparison by Gestalt Laws.

About century ago psychologists experimentally have disclosed fundamental laws of figures perception and recognition by human vision, so called Gestalt Laws [7, 8]. Comparison of contour displays by these laws shows the essential perceptual advantages of displays using stars versus parallel coordinates (PC), bar charts, pie charts, etc. Gestalt Laws says that figure will be faster detected in noise; its shape will be more accurate recognized; common pattern of a few figures will be better specified, etc., if figures have next properties:

Closure – stars map the data vectors into a closed holistic figure, PC does not

Symmetry and Similarity – stars show many axial and central symmetries of the image features facilitating detection of their commonalities

Proximity –PC, Bar-graph disrupt proximity due to discontinuity at the ends whereas stars do not

Continuity – stars are closed figures, PCs are open

The following experiments were done to determine stars advantages quantitatively. They were targeted on comparative evaluation of time and accuracy with which subjects can recognize patterns in star plots and parallel coordinates.

3. PSYCHOLOGICAL EXPERIMENTS

3.1 Holistic shapes recognition.

Five modeled data classes were randomly generated as points of Euclidian data space ED, with dimension –D: x={xi}, i=1,..D, where xi –i-th coordinate of x. Random points of each class laid in separate hyper-cone (tube) around its random axes crossing the origin of ED vector rk, where k=1,..5 – class number. These data structures are interesting because if the cone is relatively narrow, stars or PCs representations of its points are figures with visible similarity of shapes having different sizes (see Fig. 2) [9]. Shape variations depend on the distance of points from the tube central line (generatrix).

Of course such “central” tubes are pretty simple data structures but arbitrary tubes with the generatrix crossing any two points r1 and r2 can be transformed in central tubes by the origin shift inside of them. Bent tubes can be approximated in a piece-wise fashion. Besides affine invariance of shape perception allows to detect “balls” of curve tubes. It opens wide possibilities for complex nonlinear structure detection and their interpretation by relationships within different sets of coordinates. Stars allow faster detection and description of the features than PCs...

As far as advantages of stars vs PCs on Fig. 2, especially for D=96, is quite obvious and the time required of volunteers turned out pretty long for PCs analysis, we did only two subjects study to prove and roughly estimate the advantages of the stars representations. Their performance in a grouping task was compared. Subjects were instructed to group a set of randomly mixed stars (30 totals) and the same data visualized by PCs into five groups by the similarity of the whole shape.

5 CLASSES SELECTION

Stars

noise 0/0 0.05/0.1 0.1/0.15

Subject sec errors sec errors sec errors

#1 280 0 585 4 780 5

#4 173 0 312 3 539 4

Parallel Coordinates

noise 0/0 0.05/0.1 0.1/0.15

Subject sec errors sec errors sec errors

#1 985 0 1020 7 2185 11

#4 742 0 823 4 1407 8

In Table 1 the task performance time in seconds and errors amount for each test are given for stars and PCs without noise (first two columns), with moderate noise (columns 3 and 4) and a higher level of noise in last two columns. A random noise vector was generated and added separately to each data vector. The value of noise is given by a statistical mean (first number) and standard deviation (second number) as a share of the maximal possible length of the vector. Performance time and errors increased by double with 10% mean of noise and growth of standard deviation. In this case stars showed a few times better performance than PCs without as well as with noise

3.2 Feature selection

A basic part of visual pattern recognition of shape is feature selection, i.e. detection of shape features common for majority of pictures of one class and not typical for others picture classes.

Therefore in this experiment we compared stars and PCs for selection of a set of local features common for certain sets of figures. To get such figures data vectors having the same combination of different values of 5 to 9 neighbor coordinates in 4-6 location on shape were randomly generated. (see fig. 3)

Between fragments identical for all figures of each class intermediate random fragments different for each figure were placed. It may be considered as “a figurative noise”. No additive noise was in this experiment.

If data vector of data space ED is x(x1, x2, ..xi, ..xD), where xi is value of i-th coordinate, class of vectors represented by above figures can be describe by set of Elementary Conjunctions(EC) of vector coordinates. If all vectors of given class have the same values of coordinates from n-th to (n + p)-th , all of them have identical shape fragment E1 with identical location on figures! Appearance of such fragments can be considered as TRUE value of elementary conjunction (EC), i.e. logical multiplication of logical expression, i.e. “x n =a n ^ x n+1 =a n+1 …^x n+p =a n+p”, where ai are values of indexed coordinates. Length of this fragment is specified by p value.

If a given class has a few common fragments, it is described as: “all vectors of this class have identical fragments E1, E2…En ” . If to take in to account visual perception thresholds for shape discrimination, it describes data space structures having hyper-spheres as projection in subspace of coordinates of its ECs, which combine in complicated nonlinear tubes in these subspaces by affine invariance of shape perception.

Different by shapes and places EC sets (two data classes by 10 data vectors) were generated separately for different dimensions D and different tests A, …,F.

Subjects knew that first class stars (or PC's) placed in first two rows of screen and second class on last two rows.

They were instructed to find all first class fragments not existing in second class and vice versa. Tests were done separately for PCs and stars displays of the same vector classes. Subjects were required to find complete set of coordinates creating each common fragment. Performance time for different D, figures classes and displays for each subject is given in Table. 2 in the same manner as before, i.e. average time of feature selection is the first number and standard deviation of this value is the second number.

Samples A, B, C for dimension 48 had different shapes and placements of the same amounts of fragments, which led to pretty big spread of results for these samples. Because of subjects learning during tests time growth with task dimension was very moderate. This table shows and another factors influencing on performance (subject individuality, a number of tests with him/her, etc.) but despite of them all subjects showed 2-3 times faster informative feature selection with stars than with PCs.

3.3 Classification

Just to roughly estimate of the feature selection time for self learning above data samples of 72 and 96 dimensions (two classes by 10 figures) were represented on separate for star and PC's sheet of paper but with randomly placed pictures. So, subject did not know the class membership of any picture and sorted the plots into 2 sets based on visible similarity. If in similar 3.1 experiment (without feature selection) each picture class had completely identical shapes distorted by noise and with different sizes, then in this experiment pictures of one class had identical set of shape fragments alternated with random fragments, i.e. partial shape similarity.

In this case subjects spent a long time to find the first feature allowing separation of pictures in two classes by 10 and then worked as in the previous experiment searching another features with similar performance time.

For example, search of first feature in sample with D=72 dimension required 112 sec for stars and 274sec for PCs;

Both classes average time of feature detection without noise was:

- 54 sec (mean)/17 sec (standard deviation) for stars and 128/22 sec for PCs if D=72

- 44/23 sec for stars and 92/42 sec for PCs if D=96.

Again, despite of dimension increase, performance enhanced because of subject learning.

When each data vector of sample with D=96 was distorted by 10% additive noise participants spent 284 sec (!) for detection of first informative feature and average 196 sec each feature selection for stars and refused to continue after 27 minutes with PCs when only 7 figures were classified. This provides further evidence confirming the advantages of stars vs. PCs displays up to task solvable with stars but not with PCs.

CONCLUSION

1. Potential advantages of stars vs. parallel coordinates displays for data visualization are argued with Gestalt Laws.

2. About two times increase in feature selection and shape recognition using stars in comparison with parallel coordinates have been shown by experiments with enough representative samples of subjects and data for hyper-tubes and hyper-sphere type structures in data space with D<=96 and its subspaces.

3. It is shown also that by means of stars human vision can effectively select features and recognize patterns of data hyper structures with D up to 100 without dimension reduction.

4. above proved advantages of stars can be extended on higher data dimensions at least up to D <1000 and much bigger data samples (at least thousand and more) by means of:

- multi screen monitors with high resolution or simply many printed sheets for simultaneous display of hundreds data records for comparative visual analysis (technical tools);

- splitting of data records with D>100 between a few stars, searching for interesting subspaces with D< 1000 of data with D>>1000 by means of other methods and further analysis of data structures in it with stars (processing tools);

- combining stars with 2D projections, etc.

Although for other data and experimental tasks value of stars' advantages vs. PCs may be different but at least for data structures mapped in some variation of similar shapes it will be very probable.

Therefore in the most cases it is unreasonable to reject a'priory unknown information by means 2D projections or to overlap it with many PCs on one screen [6], if we can analyze data structures with dimension up to 1000.

REFERENCES

[1] Grishin, V., [Pictorial Analysis of Experimental Data], Nauka Publishing, Moscow, 237 (1982).

[2] Grishin, V., "CAM operator functions and pictorial representation of information," Proc. IFAC/IFIP/Int. Conference on Analysis, Design & Evaluation of Man–Machine Systems, (1988).

[3] Grishin, V., Sula, A. and Ulieru, M., "Pictorial analysis: a multi-resolution data visualization approach for monitoring and diagnosis of complex systems", Intl. Journal of Information Sciences. Papers 152, 1-24. (2003).

[4] Bertini, E., Tatu, A. and Keim, D., "Quality metrics in high-dimensional data visualization: An overview and systematization," IEEE Trans. on Visualization and Computer Graphics. Papers 17(12), 2203-2212 (2011).

[5] Lee, M.D., Reilly, R.E. and Butavicus, M.A., "An empirical evaluation of Chernoff faces, star glyphs, and spatial visualizations for binary data," APVis '03 Proc. of the Asia-Pacific symposium on Information visualization, Volume 24, 1-10 (2003).

[6] Ware, C., [Information Visualization], Morgan Kaufmann Publishers, 1- 486 (2000).

[7] Ward, M., Grinstein, G. and Keim, D., [Interactive Data Visualization: foundations, techniques, and applications], A.K Peters, Ltd. Natick, MA, 1- 496 (2010).

[8] Klippel, A., Hardisty, F. and Weaver, C, "Star plots: How shape characteristics influence classification tasks," Cartography and Geographic Information Science. Papers 36(2), 149-163 (2009).

9017-26, Session PTues

Possibility space for GIS suitability analysis

Wutthigrai Boonsuk, Eastern Illinois Univ. (United States); Chris Harding, Univ. of Iowa (United States)

In Geographic Information System (GIS), suitability analysis is used to model the spatial distribution of suitability within a region of interest with regard to a planning goal. This analysis is based on the combination of multiple geospatial source datasets, which spatially overlap and each encode a factor that contributes with a certain weight to the overall suitability. "Possibility space" refers to an event space that represents all possible outcomes of the suitability analysis. This paper proposed an interactive possibility space for real-time visualization and exploration with a goal to help understand meaningful relationships between variable combinations and the suitability outcomes. A case study for siting windfarm locations in North-west Iowa is presented to demonstrate the practical application and usefulness of the possibility space.

9017-27, Session PTues

Improving chemical mapping algorithm and visualization in full-field hard x-ray spectroscopic imaging

Cheng Chang, Wei Xu, Yu-Chen K. Chen-Wiegart, Jun Wang, Dantong Yu, Brookhaven National Lab. (United States)

In the big data era, high performance computing (HPC) becomes indispensible to cope with the un-precedent data intensive and computing intensive processing requirements and thereby attracts attention from domain scientists. By dividing processing tasks into sub-tasks according to their intrinsic parallelism and mapping each sub-task to individual core or processing unit in state-of-the-art high performance computing servers, domain scientists do not have to suffer the time-consuming data analysis and as a result, can discover new science in a much faster fashion. In this paper, we expedite one critical data analysis routing for the Transmission X-ray Microscope Beamline of NSLS and NSLS-II facilities to identify the correlation between two-dimensional morphology images and chemical state distribution at nanoscale, and thereby understand the mechanism of microstructural evolution and chemical reaction associated with functional materials. We focus on a chemical composition identification problem based on 2D spectroscopic images, in particular, to fit the spectrum of mixed material with standard spectra of individual chemical elements. We implement and compare two fitting approaches: (i) a brute force enumeration method, and (ii) the constrained least square minimization algorithm proposed by us. The precision of the constrained least square minimization algorithm is about 2.5% higher than the brute force one. To handle the increased computational cost of the new approach, we implement a parallel version of the software program in a PBS-based HPC cluster and accelerate the computing time about 37 times faster than the sequential one. In summary, HPC and parallelization provide a viable approach to analyze the more complex time-resolved X-ray images of 3-dimension or higher.

9017-28, Session PTues

Progressively consolidating historical visual explorations for new discoveries

Kaiyu Zhao, Matthew O. Ward, Elke A. Rundensteiner, Huong N. Higgins, Worcester Polytechnic Institute (United States)

A significant task within data mining is to identify data models of interest. While facilitating the exploration tasks, most visualization systems do not make use of all the data models that are generated during the exploration. In this paper, we introduce a system that allows the user to gain insights from the data space progressively by forming data models and consolidating the generated models on the fly. Each model can be a a computationally extracted or user-defined subset that contains a certain degree of interest and might lead to some discoveries. When the user generates more and more data models, the degree of interest of some portion of some models will either grow (indicating higher occurrence) or will fluctuate or decrease (corresponding to lower occurrence). Our system maintains a collection of such models and accumulates the interestingness of each model into a consolidated model. In order to consolidate the models, the system summarizes the associations between the models in the collection and identifies support (models reinforce each other), complementary (models complement each other), and overlap of the models. The accumulated interestingness keeps track of historical exploration and helps the user summarize their findings which can lead to new discoveries. This mechanism for integrating results from multiple models can be applied to a wide range of decision support systems. We demonstrate our system in a case study involving the financial status of US companies.

9017-29, Session PTues

Comparative case study between D3 and highcharts on lustre data visualization

Omar M. ElTayeby, Clark Atlanta Univ. (United States); Dwayne John, The National Institute for Computational Sciences (United States) and The Univ. of Tennessee Knoxville (United States); Pragnesh Patel, Scott Simmerman, The National Institute for Computational Sciences (United States)

One of the challenging tasks in visual analytics is to target clustered time-series data sets, since it is important for data analysts to discover patterns changing over time while keeping their focus on particular subsets. In order to leverage the human's ability to quickly visually perceive these patterns, multivariate features should be implemented according to the attributes available. However, a comparative case study has been done using JavaScript libraries to demonstrate the differences in capabilities of using them. A web-based application to monitor the Lustre file system for system administrators and operation teams has been developed using D3 and Highcharts. Lustre file systems are responsible of managing Remote Procedure Calls (RPCs) which include input output (I/O) requests between clients and Object Storage Targets (OSTs). The objective of this application is to provide time-series visuals of these calls and storage patterns of users on Kraken, a University of Tennessee High Performance Computing (HPC) resource in Oak Ridge National Laboratory (ORNL).

9017-30, Session PTues

Spatial partitioning algorithms for data visualization

Raghuveer Devulapalli, Mikael Quist, John G. Carlsson, Univ. of Minnesota (United States)

Spatial partitions of an information space are frequently used for data visualization. Weighted Voronoi diagrams are among the most popular ways of dividing a space into partitions. However, the problem of computing such a partition efficiently can be challenging. For example, a natural objective is to select the weights so as to force each Voronoi region to take on a pre-defined area, which might represent the relevance or market share of an informational object. In this paper, we present an easy and fast algorithm to compute these weights of the Voronoi diagrams. Unlike previous approaches whose convergence properties are not well-understood, we give a formulation to the problem based on convex optimization with excellent performance guarantees in theory and practice. We also show how our technique can be used to control the shape of these partitions. More specifically we show how to convert undesirable skinny and long regions into fat regions while maintaining the areas of the partitions. As an application, we use these to visualize the amount of website traffic for the top 101 websites.

9017-31, Session PTues

Visualization of probabilistic relationships in shape-maturity data for lunar craters

Prasun Mahanti, Mark S. Robinson, Arizona State Univ. (United States)

Probabilistic modeling and visualization of crater shape-maturity relationships is explored in context to the remote sensing data acquired from Apollo, Clementine and Lunar Reconnaissance Orbiter lunar missions. Unlike any earlier attempt in understanding relationships between lunar crater features (depth and diameter), relative age of crater formation (Pre-Nectarian to Copernican) and optical maturity of the lunar surface (OMAT values), the joint probability of these variables is modeled in this work. The proposed model is strongly dependent on data density and is not based on deterministic equations as in earlier works. Once developed,a joint probability model can accommodate additional factors through conditional probability weights in a Bayesian network architecture. It is expected that probabilistic modeling will facilitate visualization of relationships between experimental variables and eventually help gain additional insight into lunar cratering mechanisms and linkages between crater morphology, spectral properties and crater degradation mechanisms. The described simple Bayesian network in this work is by no means complete, but illustrates the potential of the proposed novel method in the context of ever-increasing high resolution lunar data in recent times.

9017-32, Session PTues

SocialMood: an information visualization tool to measure the mood of the people in social networks

Guilherme Oliveira de Amorim, Roberto Yuri Silva Franco, Rodolfo Barros Moraes, Bruno Nascimento Figuereido, João Paulo Miranda dos Santos, José Alfredo Lira Dobrões, Ricardo Alexandre Afonso, Bianchi Serique Meiguins, Information Visualization Group (Brazil)

This paper produces a guideline for the development of information visualization tools for social networks using it for this a case study on the psychological behavior of people in social networks. At the end of the paper are presented results of evaluations of usability and performance, compared to other visualization tools for social networking.

9017-33, Session PTues

Technique and cue selection for graphical presentation of generic hyperdimensional data

Lee M. Howard, Robert P. Burton, Brigham Young Univ. (United States)

Several presentation techniques have been created for visualization of data with more than three variables. Packages have been written, each of which implements a subset of these techniques. However, these packages generally fail to provide all the features needed by the user during the visualization process. Further, packages generally limit support for presentation techniques to a few techniques. A new package called Petrichor accommodates all necessary and useful features together in one system. Any presentation technique may be added easily through an extensible plugin system. Features are supported by a user interface that allows easy interaction with the data. Annotations allow users to mark up visualizations and share information with others. By providing a hyperdimensional graphics package that easily accommodates presentation techniques and includes a complete set of features, including those that are rarely or never supported elsewhere, the user is provided with a tool that facilitates improved interaction with multivariate data to extract and disseminate information.

9017-34, Session PTues

Evaluation of stream surfaces using error quantification metrics

Ayan Biswas, Han-Wei Shen, The Ohio State Univ. (United States)

Visualizing stream surfaces in three-dimensional flow fields is a popular flow visualization method for its ability to depict flow structures with better depth cues compared to simply rendering a large number of streamlines. Computing stream surfaces accurately, however, is non-trivial since the result can be sensitive to multiple factors such as the accuracy of numerical integration, placement of sampling seeds, and tessellation of sample points to generate high quality polygonal meshes. To date, there exist multiple stream surface generation algorithms but verification and evaluation of the quality of the stream surfaces remain an open area of research. In this paper we address this issue, propose different stream surface evaluation metrics and study different aspects of stream surface generation process like choice of algorithms, seeding curve placement, initial seeding curve density, choice of algorithm parameters with four verification metrics to reach meaningful conclusions.

9017-18, Session 9

Visual abstraction of complex motion patterns

Halldor Janetzko, Dominik Jäckle, Oliver Deussen, Daniel A. Keim, Univ. Konstanz (Germany)

Today's tracking devices allow high spatial and temporal resolutions and due to their decreasing size also an ever increasing number of application scenarios. However, understanding motion over time is quite difficult as soon as the resulting trajectories are getting complex. Simply plotting the data may obscure important patterns since trajectories over long time periods often include many revisits of the same place which creates a high degree of over-plotting. Furthermore, important details are often hidden due to a combination of large-scale transitions with local and small-scale movement patterns. We present a visualization and abstraction technique for such complex motion data. By analyzing the motion patterns and displaying them with visual abstraction techniques a synergy of aggregation and simplification is reached. The capabilities of the method are shown in real-world applications for tracked animals and discussed with experts from biology. Our proposed abstraction techniques reduce visual clutter and help analysts to understand the movement patterns that are hidden in raw spatiotemporal data.

9017-19, Session 9

Abstract rendering: out-of-core rendering for information visualization

Joseph A. Cottam, Andrew Lumsdaine, Indiana Univ. (United States); Peter Wang, Continuum Analytics (United States)

A fundamental premise of visualization is that a useful correspondence between pixels and data can be built.

However, visualization programs rarely operate at the pixel level.

Instead the most common visualization models work with an abstract canvas and geometric shapes.

Conversion to individual pixels is often only tacitly acknowledged in consideration of visual or computational optimizations.

This paper introduces Abstract Rendering, which augments existing geometry-based models with a step in which source data is associated directly with pixels.

Explicitly acknowledging this connection provides opportunities to more directly express certain types of visualizations (such as density maps formed with alpha composition) and provides efficient execution via data parallel formulations of common visualizations.

This paper defines Abstract Rendering and discusses implementation considerations learned from our current system.

9017-20, Session 10

GlyphSea: Visualizing Vector Fields

Emmett McQuinn, IBM Almaden Research Ctr. (United States); Amit Chourasia, San Diego Supercomputer Ctr. (United States) and Univ. of California, San Diego (United States); Jürgen P. Schulze, California Institute for Telecommunications and Information Technology (United States) and Univ. of California, San Diego (United States); Jean-Bernard Minster, Scripps Institution of Oceanography (United States) and Univ. of California, San Diego (United States)

Understanding of vector fields is important in many science and engineering domains. Traditionally, glyphs have been used to represent vector data as cones, arrows, ellipsoids and other geometric shapes.

However, these glyphs have drawbacks of being view dependent, orientation ambiguous, and sometimes-even requiring specific geometry resolution. We propose a straightforward new method of procedural dipole texturing of glyph shapes, which overcomes these drawbacks and could even enhance existing methods. We demonstrate our method with an interactive application (GlyphSea), which incorporates additional features such as screen space ambient occlusion, displacements, lattice, halos and other contextual visual cues. We also discuss the results and informal feedback from scientists on insights gained by exploring time varying vector datasets in astrophysics and seismology.

9017-21, Session 10

Simulation and visualization of velocity fields in simple electrokinetic devices

Prasun Mahanti, Thomas Taylor, Douglas Cochran, Michael Keebaugh, Mark Hayes, Arizona State Univ. (United States)

Capillary electrophoresis and similar techniques which use an electrified contracting-flow interface (gradient elution moving boundary electrophoresis, electrophoretic exclusion, for examples) are widely used and have several applications, but the detailed flow dynamics and local electric field effects within this zone have only recently been quantitatively investigated. The motivating force behind this work is establishing particle flow based visualization tools enabling advances for arbitrary interfacial designs beyond this traditional flow/electric field interface. These tools work with pre-computed 2-dimensional fundamental interacting fields which govern particle and(or) fluid flow and can now be obtained from various computational fluid dynamics (CFD) software packages. The particle-flow visualization calculations implemented in the tool and are built upon a solid foundation in fluid dynamics. The module developed in here provides a simulated video particle observation tool which generates a fast check for legitimacy. Further, estimating the accuracy and precision of full 2-D and 3-D simulation is notoriously difficult and a centerline estimation is used to quickly and easily quantitate behaviors in support of decision points. This tool and the recent quantitative assessment of particle behavior within the interfacial area have set the stage for new designs which can emphasize advantageous behaviors not offered by the traditional configuration.

9017-22, Session 10

Streamline similarity analysis using bag-of-features

Yifei Li, Chaoli Wang, Ching-Kuang Shene, Michigan Technological Univ. (United States)

Streamline similarity comparison has become an active research topic recently. In this paper, we present a novel streamline similarity comparison method inspired by the bag-of-features idea from computer vision. Our approach computes a feature vector, spatially sensitive bag-of-features, for each streamline as its signature. This feature vector not only encodes the statistical distribution of combined features (e.g., curvature and torsion), it also contains the information on the spatial relationship among different features. This allows us to measure the similarity between two streamlines in an efficient and accurate way: the similarity between two streamlines is defined as the weighted Manhattan distance between their feature vectors. Compared with previous distribution based streamline similarity metrics, our method is easier to understand and implement, yet producing even better results. We demonstrate the utility of our approach by considering two common tasks in flow field exploration: streamline similarity query and streamline clustering.

Conference 9018: Measuring, Modeling, and Reproducing Material Appearance

Monday - Tuesday 3 –4 February 2014

Part of Proceedings of SPIE Vol. 9018 Measuring, Modeling, and Reproducing Material Appearance

9018-1, Session Key1

Perceiving, measuring, and modeling 3D material appearance *(Keynote Presentation)*

Christopher W. Tyler, Smith-Kettlewell Eye Research Institute (United States)

Objects by their nature are 3D and have to be understood in 3D in order to be effectively perceived and manipulated from many angles. Since objects are perceived through their surfaces, the 3D representation is the natural operating domain of the appearance of surface materials. I will consider a variety of aspects of the role of 3D representations in surface appearance, including the history of 3D surface representation, the role of different forms of lighting, and the process of integration of multiple perceptual cues to provide our assessment of surface structure and appearance.

9018-2, Session 1

Rapid determination of the photometric bidirectional scatter distribution function by use of a near-field goniophotometer

Frédéric B. Leloup, Katholieke Univ. Leuven (Belgium); Ward De Ketelaere, ETAP NV (Belgium); Peter Hanselaer, Peter Hanselaer, Katholieke Univ. Leuven (Belgium)

The bidirectional scatter distribution function (BSDF) characterizes the scattering properties of a material for any angle of illumination or viewing, and offers as such a complete description of the spatial optical characteristics. An accurate determination of the BSDF is important in many scientific domains, such as computer graphics, architectural and lighting design, and the field of material appearance (e.g. the characterization of color and gloss).

Many BSDF measuring instruments have been reported in the literature. The majority of these instruments are goniometric measurement devices, by use of which the BSDF is determined by scanning all incoming and outgoing light flux directions in sequence. For this, the sample, detector, and/or source perform relative individual movements. With this type of instruments a broad angular coverage may be realized both in and out of the plane of incidence, and both in reflectance as well as in transmittance mode. However, the major restriction constitutes the measurement time, which may run to the order of several hours depending on the accuracy (angular resolution) and complexity (spectral coverage, absolute measurement capability, etc.) of the reported measurement data.

To speed up acquisition, alternative measurement devices have been proposed which detect multiple angles simultaneously. This can e.g. be achieved by use of a camera in combination with optical accessories, such as a mirrored hemisphere which is used as a projection surface, or by use of a specimen holder with known surface curvature (e.g. a spherical material sample). The major drawback of these type of devices is however that, due to the fact that they have been optimized for time efficiency, the other features are generally restricted (e.g. the ability to perform measurements in reflectance or transmittance mode only).

In this paper an alternative goniometric measurement system is presented, enabling to acquire the photometric BSDF in a full three-dimensional space, with a high angular resolution (0.1°) in a time efficient way (about 30 minutes). A near field goniophotometer (type RiGO 801 by TechnoTeam), which is originally designed to measure luminance intensity distributions and luminous fluxes of lamps and luminaires, has been converted for this purpose. Within the near field goniophotometer, a tungsten halogen illumination source has been installed, designed based on the principle of Köhler illumination, and providing a uniform illumination spot of 2.6 cm diameter at the sample position. Photometric measurements are made by use of a photometer which is moved on a spherical surface around the measurement sample. Furthermore, the goniophotometer is equipped with a colorimetric camera, which is used to check the adjustable sample position (variable light incidence angle), but which can also be used to acquire colorimetric data if required. Measurements are made relative to a white Ceram reference sample, for which the spectral bidirectional reflectance distribution function (BRDF) has been characterized in another home-built measurement setup, capable of performing absolute spectral BSDF measurements.

Besides discussing the measurement procedure, test sample measurements will be presented to illustrate the versatility of the device. These include BRDF measurements of gloss samples, and bidirectional transmittance distribution function (BTDF) measurements of lighting diffusers.

9018-3, Session 1

Multidimensional reflectometry for industry (xD-Reflect) an European research project

Andreas Höpe, Physikalisch-Technische Bundesanstalt (Germany); Annette Koo, Measurement Standards Lab. (New Zealand); Francisco Martinez Verdú, Univ. de Alicante (Spain); Frédéric B. Leloup, Katholieke Univ. Leuven (Belgium); Gaël Obein, Conservatoire National des Arts et Metiers (France); Gerd Wübbeler, Physikalisch-Technische Bundesanstalt (Germany); Joaquin Campos Acosta, Consejo Superior de Investigaciones Científicas (Spain); Paola Iacomussi, Istituto Elettrotecnico Nazionale Galileo Ferraris (Italy); Priit Jaanson, MIKES Mittatekniikan keskus (Finland); Stefan Källberg, SP Technical Research Institute of Sweden (Sweden); Marek ?míd, Czech Metrology Institute (Czech Republic)

The European Metrology Research Programme (EMRP) is a metrology-focused programme of coordinated R&D funded by the European Commission. It supports research collaboration between the National Metrology Institutes (NMIs) of Europe.

The general objective of xD-Reflect is to meet the demands from industry to describe the overall macroscopic appearances of modern surfaces by developing and improving methods for their measurement which rightly correlates with the visual sensation. In specific, the project deals with the “Goniochromatism”, “Gloss” and “Fluorescence” properties of dedicated artefacts, which will be investigated in three main work packages (WP). Two additional transversal WP reinforce the structure. “Modelling and Data Analysis” with the objective to give an irreducible set of calibration schemes and handling methods and “Visual Perception”, which will produce perception scales for the different visual attributes.

Multidimensional Reflectometry means the enhancement of spectral and spatial resolution of reference gonioreflectometers for BRDF measurements, using modern detectors, conoscopic optical designs, CCD cameras, line scan cameras, and modern light sources in order to describe new effects like sparkle, graininess/coarseness.

Participating partners within the project are 8 NMIs: LNE-CNAM (France), PTB (Germany), MIKES (Finland), INRIM (Italy), CMI (Czech Republic), SP (Sweden), CSIC (Spain), MSL (New Zealand) and the universities KU Leuven (Belgium) and Alicante (Spain).

The specific work package of “Goniochromatism” has five tasks where as a major activity a round robin comparison of the goniometric scales of

the participants will be arranged. Other tasks deal with the identification of a basic set of parameters for goniometric effects and data handling recommendations for goniochromatic materials.

A general aim is to find an irreducible set of scaling parameters for goniometric effects like lightness flop, colour-flop or sparkle/graininess. In recent years, sparkling effects of pigments have strongly increased. The sparkling of effect pigments can be recognized as many tiny but very intense light spots, like bright stars twinkling at the night sky. Sparkle is an obvious effect to human observers, but cannot be measured with current spectrophotometers because of its small length scale. Furthermore, the same effect pigments, when they are measured and perceived in diffuse illumination, causes a new visual texture effect named graininess or coarseness.

For the measurement of these visual appearance-related quantities the Physikalisch-Technische Bundesanstalt (PTB) is operating two robot-based gonioreflectometers. The first one is the German national measurement standard for calibrations in directed/directed diffuse reflection geometries. The second one is a research set-up denoted with the acronym ARGon3, which stands for “3D appearance robot-based gonioreflectometer”. Compared to other gonioreflectometers, there are two new features within this setup. First, a photometric luminance camera with a spatial resolution of 28 µm at the device under test (DUT) enables spatially high-resolved measurements of luminance and colour coordinates. Second, a line-scan CCD-camera mounted to a spectrometer provides measurements of the BRDF, in full V(?)-range (360 nm – 830 nm) with arbitrary angles of irradiation and detection relative to the surface normal. As examples, goniometric BRDF measurements of different interference effect pigments in 3D-space together with subsequent colorimetric representation of the obtained data will be presented.

9018-4, Session 1

Evaluation of the shape of the specular peak for high glossy surfaces *(Invited Paper)*

Gaël Obein, Shiraz Ouarets, Guillaume Ged, Conservatoire National des Arts et Metiers (France)

Gloss is the second most relevant visual attribute of a surface beside its colour. Where the colour originates from the wavelength repartition of the reflected light, gloss originates from the angular repartition of the reflected light. When we ask to an observer to evaluate the gloss of a surface, the observer always put its eyes is the specular direction and then tilts a little the sample. It means that gloss is located in and around the specular direction, in a peak that is called the specular peak. This peak is flat and large on matte surfaces and narrow and peaky on high gloss surfaces. For the late ones, typically, the gloss of car paintings, the FWHM of the specular peak is less than 2° and starts to be difficult to measure. We developed a dedicated facility capable to measure specular peaks with a FWHM up to 0,1°. We measured the evolution of the peak according to the angle of illumination and according to the specular gloss of the sample in the restricted field of very glossy surface. The facility and peaks measured will presented in the paper. The next step will be to identify the correlations between the peak and the roughness of the samples.

9018-5, Session 1

Analysis of aberrations and pixel information in goniometric multispectral imaging

Julie Klein, Georg Schmücking, RWTH Aachen (Germany)

Goniometric acquisitions aim at characterizing completely the remission properties of an object by measuring it under different viewing angles and with different illumination angles. It can be performed with a spectrometer, a line-scan CCD or a camera. Over the last years, cameras with a better spectral accuracy than RGB cameras have been utilized [Ber12, Hoe12, Rum10]. In this work, we analyze the aberrations and distortions appearing in goniometric multispectral imaging.

We utilize a multispectral camera featuring a filter wheel with 19 bandpass filters whose central wavelengths are spread from 380nm to 780nm and whose bandwidths are 10 to 20nm. Using such a camera in a goniometric measurement gives access to information about the acquired object for different viewing and illumination angles with a high spectral and spatial resolution. We perform here in-plane measurements with a light source at angle 45° and the measurement device positioned between 30° and -70°.

Aberrations cannot be avoided in filter wheel multispectral cameras. They are caused by the optical filters in the filter wheel and have to be corrected to avoid errors like color fringes in the multispectral images, where the color channels are assembled. We deal with the transversal aberrations, i.e., the aberrations appearing along the sensor plane. We first analyze the aberrations over the image plane, for each color filter and each acquisition angle. They can be modeled with the addition of two terms: a position-dependent displacement and a global displacement [Bra11]. These displacements can be measured in each color channel separately. They can also be measured simultaneously in all the channels, since they only depend on the angles of the incoming rays. This allows the parameters of the model to be optimized and outliers to be ignored. We compare both approaches (separate and simultaneous measurement) and the results of the correction.

Another important effect in goniometric imaging is that the images from the different viewing angles are deformed geometrically: one object point is imaged on a different position in each image. The images are thus rectified to look as if they were acquired from the same position. In this way, the measured information from one object point is on the same position in all the images from the different viewing angles. The rectification can either be a homography [Kou03] or a more sophisticated transformation [Vav13]. We analyze the relations between the correction of transversal aberrations and this rectification.

Since a camera captures a whole area and not only an infinitesimal region of the object, the effective viewing angles for the different pixels in one given image, i.e., the angles between the perpendicular to object surface and the rays that actually come from the object points, are slightly different. This means, for instance, that the information captured in the image for a viewing angle of 0° does not correspond to remission angles of 0° within the whole image plane, but rather to angles reaching a few degrees on the sides of the image. This additional angular information could allow getting more angular information from fewer acquisitions. The acquisition system is modeled with a pinhole camera to first simulate these effects. Real acquisitions of a colored object give similar results.

References

[Ber12]: Roy S. Berns, Tongbo Chen, David R. Wyble and Lin Chen. “Practical total appearance imaging of paintings”. In IS&T Archiving Conference, pages 162-167, June 2012.

[Bra11]: Johannes Brauers and Til Aach, “Geometric calibration of lens and filter distortions for multispectral filter-wheel cameras”. In IEEE Transactions on Image Processing, Vol. 20, No. 2, pages 496-505, February 2011.

[Hoe12]: A. Höpe, T. Atamas, D. Hünerhoff, S. Teichert and K.-O. Hauer. “ARGon?: “3D appearance robot-based gonioreflectometer” at PTB”. In Review of Scientific Instruments, Vol. 83, pages 045102-1--045102-8, 2012.

[Kim06]: Akira Kimachi, Norihiro Tanaka, and Shoji Tominaga. “A goniometric system for measuring surface spectral reflection using two robot arms”. In Proc. IS&Ts 3rd European Conference on Colour in Graphics, Imaging, and Vision (CGIV), pages 378-381, Leeds, UK, June 2006.

[Kou03]: Melissa Koudelka, Sebastian Magda, Peter N. Belhumeur, and David J. Kriegman. “Acquisition, compression, and synthesis of bidirectional texture functions”. In 3rd International Workshop on Texture Analysis and Synthesis, 2003.

[Rum10]: Martin Rump, Ralf Sarlette and Reinhard Klein. “Groundtruth data for multispectral bidirectional texture functions”. In Proc. IS&Ts 5th

European Conference on Colour in Graphics, Imaging, and Vision (CGIV), pages 326-331, June 2010.

[Vav13]: Radomir Vavra and Jiri Filip. “Registration of multi-view images of planar surfaces”. In KyoungMu Lee, Yasuyuki Matsushita, JamesM. Rehg, and Zhanyi Hu, editors, Computer Vision - ACCV 2012, volume 7727 of Lecture Notes in Computer Science, pages 497-509. Springer Berlin Heidelberg, 2013.

9018-6, Session 1

A line scan camera based stereo method for high resolution 3D image reconstruction

Pengchang Zhang, Yuji Ogane, Jay Arre O. Toque, Yusuke Murayama, Ari Ide-Ektessabi, Kyoto Univ. (Japan)

3D geometrical shape acquisition based on stereo vision has been attracting increasing interests and demands in the realm of cultural heritage digitization. Although it has found many successful applications, high resolution still remains a challenging problem for researchers. Previous studies mostly focused on area camera-based stereo vision for imaging medium or large size objects such as architecture, monuments, and sculptures.?One point in the reconstructed 3D image represented a physical size in the magnitude of centimeter or even larger. Here we define resolution as Dots Per Inch (DPI). Accordingly, traditional methods demonstrated a reconstructed resolution less than 3DPI. In this paper, focus was given to increase the resolving ability of reconstructed points to the magnitude of sub-millimeter (greater than 30DPI) by stereo vision using line scan camera. This technique is capable of delivering high resolution images as the basis of extracting finer 3D spatial information and color information at the same time. Such high a resolution offers us greater details of the geometry, texture and material information which are quite beneficial to the scientific documentation, restoration, research and visualization of cultural heritages.

There are two critical problems that need to be addressed: camera calibration and correspondence.

Area camera obtains the whole image at one shot with its optical center fixed, while, line scan camera captures an images by shifting the optical center along one direction perpendicular to its line image sensor. In this way, line scan camera has a distinct geometric image formation model compared with area camera. In this paper, the mathematical equation describing the relationship between the coordinates of a pixel in the image and its corresponding 3D spatial coordinates was analyzed and derived as the basis for camera calibration

Camera calibration was further improved by introducing the correction parameters for both camera position and lens distortion. For line scan cameras, the optical axis should be perpendicular to the object surface and the line sensor to the direction of the optical center movement in order to avoid geometric distortion and keep the aspect ratio constant. On the other hand, lens itself also causes distortion in images. Combined with these distortion factors, an improved mathematical description was derived for the camera calibration in order to get more accurate camera parameters. Camera calibration was implemented by using a self-designed calibration chart.?It was composed of two pieces of flat paper with certain degrees apart. On each paper an array of black/white squares were used with known 2D dimension. The degree between the two pieces of paper was chosen to obtain the 3D coordinates of the features on chart with high accuracy.

Image rectification is required for area camera stereo to simplify the process for correspondence. However, the imaging mechanism of line scan camera itself makes it unnecessary. Correspondence is a critical issue when recovering 3D shapes in stereo vision. For high resolution reconstruction, dense correspondence was an extremely challenging task. In this paper, a Phase-Only Correlation (POC)-based high accuracy correspondence search was employed with a hierarchical strategy to speed up the search progress. In POC-based correspondence, several techniques were employed to optimize the peak value search and improve the robustness over image noise. The effect of window size used for matching on reconstructed quality was also examined.

The proposed line scan camera stereo method was experimentally verified and evaluated by designing and implementing the idea through a semi-cylindrical shaped 3D scanner. In this scanner, there was a fixed stage for housing the object to be scanned, and the camera can be rotated along the semi-circular frame in both sides of the central axis to capture images from different viewpoints. Some 3D artworks were used as the subjects for experiment. A laser scanner was used as reference for comparison to evaluate the accuracy of the reconstructed result. The effect of resolution on reconstructed accuracy was also investigated. The result showed that this method demonstrated great effectiveness and efficiency, and can be used to reconstruct 3D shape of cultural heritage with color information captured simultaneously at a resolution greater than 50DPI

9018-7, Session 1

An investigation into the micro surface of artworks using alternative lighting techniques

Flavia Tisato, Univ. degli Studi di Ferrara (Italy); Carinna E. Parraman, Univ. of the West of England (United Kingdom)

The technique of raking light consists of illuminating objects at an oblique angle from a light source that is very close to the surface. By using raking light, the details of brushstrokes and the surface texture of a painting is accentuated, both by the increased illumination of surfaces facing the light source, and the exaggerated shadows of non-illuminated surfaces. This sort of method provides useful information about the surface topography and relief of the artefact, and that is why it is widely used in the examination of works of art. Furthermore, raking light can help to emphasize various painting techniques, and can for example, assist conservators through non-invasive examination of artworks. It also may be useful in order to monitor the effects of conservation interventions.

In the context of recent developments in additive layer manufacturing and projects investigating the application of contemporary texture to surfaces or construction of surface topology, paintings can be considered as a three-dimensional landscape, although composition and colour are key components of an artwork, texture is also vital, as it provides a unique ‘fingerprint’ of an artwork’s surface) [1].

In this paper, an alternative approach is presented, starting with an illuminated dome that is attached to a stereo-microscope, which is able to gain - at the same time - both colour and texture features of the sample. By using a stereo-microscope, a more detailed observation is possible. Important elements including colour, texture, the morphology of the sample surface, in terms of specular and diffused components of reflected light, can be summarized/condensed in the measure of Bidirectional Texture Function [2].

The whole apparatus consists of an enclosed rotating dome in which is placed the sample, which must be of reduced size or, however, easy to handle (for example, a sheet of paper). A closed hemispherical covering, approximately 20 cm high, is placed above the sample, in order to guarantee a “micro-environment” with a constant geometry of illumination and observation of the sample. On the external shell of the dome a hole is present, through which a light source is inserted, a small lambertian bright 0.5W LED lamp (6000k) illuminates the sample surface at a fixed angle. The dome is rotated around the sample so that the sample is illuminated from different directions. By rotating the dome, it is possible to obtain both single pictures and a 360degree rotating view.

In order to provide a better understanding of the relationship between ink or paint on the different substrates, the samples are examined under a stereo-microscope. The samples are captured using a Nikon SMZ800 stereo zoom microscope with a P-ED Plan 0.5x objective, attached to a G-US2 universal table stand so that a range of samples can be examined. Nikon Elements software can be used as both a visualisation and measuring tool. Images obtained will emphasize the three-dimensionality of every smallest detail, of a single brushstroke, making it possible to digitally reconstruct the surface.

This method would be useful in order to better characterize the sample with its colorimetric and texture features, then it would allow an easier digital reconstruction of the sample and, thus, a more simple reproducibility by printing. Furthermore, for restorers involved in a conservative intervention, it could be a useful way to monitor their work and to provide a feedback, that is, a scientific basis for the subsequent steps.

References

[1] http://www.whiteboardmag.com/3d-scan-van-gogh-perfect-replica-formart

[2] J. C. Ashbaugh, R. S. Berns, B. A. Darling, L. A. Tamplin, Artist Material BRDF Database for Computer Graphics Rendering, in: 17th Color Imaging Conference Final Program and Proceedings, Society for Imaging Science and technology, 2009, pp. 62 - 68

9018-8, Session 1

Topographical scanning and reproduction of near-planar surfaces of paintings

Willemijn Elkhuizen, Tim Zaman, Technische Univ. Delft (Netherlands); Wim Verhofstad, Oce Technologies B.V. (Netherlands); Pieter P. Jonker, Joris Dik, Delft University of Technology (Netherlands); Jo M.P. Geraedts, Delft University of Technology (Netherlands) and Oce Technologies B.V. (Netherlands)

Paintings are versatile near-planar objects with material characteristics that vary widely. The fact that paint has a material presence is often overlooked, mostly because of the fact that we encounter many of these artworks through two-dimensional reproductions. The capture of paintings in the third dimension is not only interesting for study, restoration and conservation, but it also facilitates making three dimensional reproductions through novel 2.5D printing methods.

The varying material characteristics of paintings are first investigated, after which an overview is given of the feasible imaging methods that can capture a painting's color and topography. Because no imaging method is ideally suited for this task, a hybrid solution between fringe projection and stereo imaging is proposed involving two cameras and a projector. Fringe projection is aided by sparse stereo matching to serve as an image encoder. These encoded images processed by the stereo cameras then help solve the correspondence problem in stereo matching, leading to a dense and accurate topographical map, while simultaneously capturing its color.

Through high-end cameras, special lenses and filters we capture a surface area of 170 cm2 with an in-plane effective resolution of 50 ?m and a depth precision of 9.2 ?m. Semi-automated positioning of the system and data stitching consequently allows for the capture of surfaces up to 2 m2.

The topographical map and color data are used to make hardcopy reproductions, using a specially developed printing system.

Several (Dutch) paintings (by Rembrandt and Van Gogh) are scanned and reproduced using this technique. These 2.5D printed reproductions are evaluated, individually and in a side-by-side comparison with the original.

9018-9, Session 1

State of the art of 3D scanning systems and inspection of textile surfaces

Miguel Montilla, Sergio A. Orjuela Vargas, Univ. Antonio Nariño (Colombia); Wilfried Philips, Univ. Gent (Belgium)

The rapid development of hardware and software in the digital image processing field has boosted research in computer vision for applications in industry. The development of new electronic devices and the tendency to decrease their prices makes possible new developments that few decades ago were possible only in the imagination. This is the case of 3D imaging technology which permits to detect failures in industrial products by inspecting aspects on their 3D surface.

Particularly, the use of 3D imaging technology in the textile industry poses a challenge to deal with irregularities that are exhibited at relative small scales. The modern textile industry aims to produce textiles with as little defects as possible since the presence of defects can decrease the final price of products from 45% to 65%. The quality of a textile surface is monitored by statistical processes. High levels of production require acceptance sampling plans consisting in assessing the quality of piles or groups checking a certain portion of samples. Detection of defects requires localizing the defective regions, which are categorized into degrees of importance such as major, minor and critical flaws. Surface properties in textiles such as pilling, hairiness and roughness, produce effects like occlusions or shadows that may affect the 3D scanning.

Quality evaluation of appearance in textile materials is conducted during end-process inspection. Changes in surface appearance of textile materials are defined by diverse visual characteristics related to appearance such as losing or deformation of shape or fit, degradation of surface and changing in color, handle or pilling among others. Several studies have been performed in search of objectives AR grading assessment systems using Automated Visual Inspection. Most of the research has been conducted using algorithms on grey-scale images based on intensity, like photographs. Besides, most investigations report that appearance changes perceived by humans can be represented with image features using logarithmic representations. A big limitation of using grey-scale images for detecting surface appearance changes is that they do not capture well the 3D structure of the textile surfaces. Therefore, to better distinguish surface textiles deviations from the original, some researchers explore the use of 3D topographic measures of depth obtained using non destructive acquisition methods. In this type of systems, the depth information is digitized into an image where the pixels represent depth instead of colors, permitting to evaluate structural changes in materials while being less affected by color and lighting changes. 3D scanning methods are reported to be more accurate than imaging methods. 3D information can be used to quantify features related to appearance characteristics such as texture, hairiness, structure, thickness, smoothness or roughness, wrinkle, shape changes, seam puckering, fuzziness, snagging, and pilling. Considering that information related to colors and patterns can be obtained from intensity images, a complete description of appearance changes in textile surfaces can be then achieved by combining intensity, colors and 3D information. Despite the international effort, few automated commercial systems for inspecting appearance of textiles exist. In search of an optimal solution for scanning textiles we present in this paper a review of existing techniques for digitizing 3D surfaces. Topographic details of textiles can be obtained by digitizing surfaces using laser line triangulation, phase shifting optical triangulation, projected-light, stereovision systems and silhouette analysis. Although we are focused on methods that have been used in the textile industry, we also consider potential mechanisms used for other applications. We discuss the advantages and disadvantages of the evaluated methods and state a summary of potential implementations for the textile industry.

9018-10, Session 2

Colour printing 7.0: goals, challenges, and accomplishments *(Invited Paper)*

Jon Yngve Hardeberg, Gjøvik Univ. College (Norway)

No Abstract Available

9018-11, Session 2

Estimating Neugebauer primaries for multi-channel spectral printing modeling

Radovan Slavuj, Ludovic G. Coppel, Gjøvik Univ. College (Norway); Melissa Olen, University of the West of England (United Kingdom); Jon Yngve Hardeberg, Gjøvik Univ. College (Norway)

The idea of spectral printing is to match input spectra and to minimize metamerism. Multichannel printing systems provide an enlarged set of possible color combinations and it can produce better spectral match to the input spectra than conventional CMYK systems. The most used empirical spectral models are the Yule-Nielsen modified spectral Neugebauer (YNSN) model and its cellular extension. These models make the prediction with an essentially linear combination of the spectral reflectances of the Neugebauer primaries (NP) that act as nodes in the multidimensional space.

The majority of printing substrates are not capable of receiving more than around 280% of inks amount per patch without ink bleeding, preventing the creation of the NP training set for those substrates. Instead physical color mixture prediction models can be used to simulate NP's that are not printable. Chen et al [1] used weighted linear regression [2] to estimate what is referred to as non-printable primaries for the cellular YNSN model and Abebe et al. [3] used the single constant Kubelka-Munk (KM) theory to perform the same on the four channel printing system. In some cases YNSN performs better when NP are estimated [3] where for many substrates this is necessary as a set of NP is not physically realizable.

For this work we compared two continuous tone color mixing models and their performance in NP estimation: KM and general radiative transfer theory with the DORT2002 [4] simulation tool. As opposed to KM, DORT2002 simulates anisotropic reflectance, which has been shown to significantly affect the measurement of radiance factor, especially for highly absorbing media [5]. The substrate that is used for the experimental evaluation was artist cotton paper (300 g/m2) that has great absorbance capabilities and can accommodate up to 700% of ink coverage. For printing we employed 7-channel Canon iPF 8000 where each channel was individually controlled through Caldera RIP. For this stage of the project we have used 45/0 measurement geometry and we have determined the absorption coefficient and scattering coefficient K/S unit directly from the reflectance. Preliminary results show that for the selected patches, average model precision errors range from 0.86 CIE ?E*00 units for 2-color mixing to 4.4 unit for 7-colour mixing with KM. Using DORT2002, CIE ?E*00 is about 20% lower and the same trend is observed for the spectral root mean square (sRMS) difference.

For additional substrates, where ink absorbance is significantly lower, we will perform the same estimation. To evaluate the outcome of the estimation we will apply YNSN model on all selected substrates. Also, as the Neugebauer training set is oversampled in the dark region, the accuracy of the measurement instrument might influence performance of the model. In this case we will compare two measurement geometries and performance of the estimation for both 45/0 and d/8. The advantage of using DORT2002 (besides the moderate improvement in determining K/S over KM model) is that it can simulate different measurement geometries. This would enable simulation of the otherwise tedious measurement procedure with d/8 instrument.

References

[1] Yongda Chen, Roy S. Berns, Lawrence A. Taplin, 2004. Extending Printing Color Gamut by Optimizing the Spectral Reflectanceof Inks, Proc. IS&T/SID 12th Color Imaging Conference. pp. 163-169

[2] Balasubramanian, R. (1999), Optimization of the spectral Neugebauer model for printer characterization, J. Electron. Imaging 8 (2), 156–166, (1999).

[3] Abebe, M., Gerhardt, J., and Hardeberg, J. Y. (2011), Kubelka-Munk theory for efficient spectral printer modeling. Proc. SPIE 7866, Color Imaging XVI: Displaying, Processing, Hardcopy, and Applications article id. 786614, 15 pp.

[4] DORT2002 v3.0 manual, Mid Sweden University, www.miun.se

[5] Neuman, M., and Edström, P., Anisotropic reflectance from turbid media. I. theory, J. Opt. Soc. Am. A 27, 1032–1039 (2010).

9018-12, Session 2

An interactive tool for gamut masking

Ying Song, Cheryl Lau, Sabine Süsstrunk, Ecole Polytechnique Fédérale de Lausanne (Switzerland)

Artists often want to change the colors of an image to achieve a particular aesthetic goal. For example, they might limit colors to a warm or cool color scheme to create an image with a certain mood or feeling. Gamut masking is a technique that artists use to limit the set of colors they can paint with. They draw a mask over a color wheel and are only allowed to use the hues within the mask. However, creating the color palette from the mask and applying the colors to the image requires skill. We propose an interactive tool for gamut masking that allows amateur artists to create an image with a desired mood or feeling. Given an image, the tool provides an initial mapping to the desired gamut. Then, the user can experiment with changing the colors of selected local areas in the image.

Our tool supports any gamut mask shape, giving the user the flexibility to create a mask that will sufficiently represent the desired mood. The user draws a 2D gamut mask over a color wheel either by hand or by selecting a preset shape. We extract a 3D gamut in CIE LUV space from the 2D user-drawn mask by extruding it over all lightness values and intersecting it with the display gamut. Since we represent our gamut in a voxel grid, we are able to handle any shape gamut the user may draw, including gamuts with extreme concavities, multiple components, and holes.

Once a gamut is created, the image is mapped to this gamut. The user can choose to locally change the colors in the image, redraw the gamut mask, or remap the image to the gamut. To represent the image, we cluster the image pixels by their original colors and spatial locations, similar to Lau et al. [2011]. By representing the image using clusters, the user can click on a single image pixel and change the color of the pixel's cluster in order to change the color of the entire local area in a blended fashion. To automatically map the image colors to the gamut, we represent the clusters as spheres and map them to their closest points within the gamut such that their spheres are fully within the gamut. While our gamut mapping algorithm is similar to the method of Lau et al., a difference is that their method maps an image to a single output according to an optimization function. Instead, we allow the user to explore different outputs by letting the user change the colors of different local areas in the image according to the user's artistic intent. Our interactive tool for gamut masking allows the user to impose a gamut mask on an image and refine the colors locally to achieve a desired mood.

C. Lau, W. Heidrich, R. Mantiuk. 2011. Cluster-Based Color Space Optimizations. Proc. IEEE. International Conference on Computer Vision.

9018-13, Session 2

A new connection space for low-dimensional spectral color management

Steven Le Moan, Technische Univ. Darmstadt (Germany); Philipp Urban, Fraunhofer-Institut für Graphische Datenverarbeitung (Germany)

Spectral color management consists of extending the traditional color reproduction techniques to being independent from viewing conditions (illuminant, standard observer). Multi- and hyper-spectral pixels are typically described by several dozens of reflectance values, allowing to analyze their color renderings under various lluminants with great accuracy. Nevertheless, not only does a 30-dimensional image creates huge memory requirements, raw spectral data often lack of perceptual meaning. Indeed, color sensation can only be estimated in low-

dimensional spaces such as CIELAB or CIECAM02, which require to specify certain viewing conditions. In other words, what we can refer to as the spectral space contains information aplenty, but is memory-costly and does not allow for direct perceptual analysis, whereas colorimetric spaces are perceptually meaningful, computationally efficient, but limited to, e.g., one particular illuminant. The whole challenge when it comes to spectral color management lies in finding a tradeoff between these two conditions. In recent years, there have been several attempts at creating low-dimensional Interim Connection Spaces (ICS), to represent the most important features of multi- or hyper-spectral pixels in only a few dimensions. A good ICS should have no more than a few components (typically 6), it should span an entropy large enough to allow for an accurate spectral reconstruction, but small enough for limited memory requirements (e.g. for building look-up tables). Also, distances within the space should have a good correlation with perceptual distances for various viewing conditions. Finally, a spectral reproduction should not only be better than a colorimetric reproduction on average over several illuminants, it should be at least as good under one common light such as CIED50. Therefore, a good ICS should as well allow to be competitive with colorimetric workflows under some specific viewing conditions.

Derhak and Rosen [1] introduced LabPQR, an ICS with three colorimetric dimensions (Lab), optimized for a certain illuminant, as well as 3 spectral dimensions (PQR), which convey the dominant structure of the difference between original and reconstructed spectra from colorimetric values (also referred to as metameric black). Although it has been reported that LabPQR can be used successfully in several applications, it struggles to meet all of the aforementioned criteria in that its spectral dimensions lack of perceptual meaning. Moreover, it has been shown more recently by Zhang et al. [2] that a multiple-XYZ space can outperform LabPQR when it comes to spectral reconstruction accuracy. The authors proposed indeed to concatenate several colorimetric spaces (XYZ) from different illuminants, the latter being computed as the principal components from a set of dozens of real lluminants. Such approach however introduces redundancy, due to the fact that some perceptual attributes do not vary much from one illuminant to another. Moreover, the muliple-XYZ ICS described by the authors lacks of an actual strategy to weight the various illuminants into consideration, based on their relative importance in a given application.

Based on these remarks, we propose an alternative ICS which has the advantage of being simple, meaningful, and with only five dimensions. As in LabPQR, we propose to use actual CIELAB values from a common illuminant (CIED50) as the first three dimensions. That way, we make sure to meet the last of the aforementioned criteria, about competitiveness with metameric reproductions. Only, in our case, we used the hue-linear LAB2000HL [3] space for a better perceptual uniformity. The remaining two dimensions are then built to represent chromatic features only, as we observed that lightness components from various illuminants are usually very correlated. We use a set of real illuminants which we de-correlate from CIED50 and then extract the first principal component from the result. The thusly obtained synthetic illuminant is then used to render the spectral image in LAB2000HL and only a00HL and b00HL are eventually used to finish building the proposed ICS. We performed experiments on a database of 16 multispectral images of natural scenes and 74 illuminants. Results indicate that, although the multiple-XYZ strategy permits a slightly better spectral reconstruction accuracy on average, the loss of accuracy engendered by our ICS is still within the range of the Just-Noticeable Distance and therefore does not counterweight the advantages of using 5 dimensions rather than 6. Moreover, our strategy yields better results than multiple-XYZ for its main illuminant (CIED50) and better results than LabPQR on average.

[1] Derhak, M. and Rosen, M., “Spectral colorimetry using labpqr: an interim connection space,” Journal of Imaging Science and Technology 50(1), 53–63 (2006).

[2] Zhang, X., Wang, Q., Li, J., Yang, P., and Yu, J., “The interim connection space based on human color vision for spectral color reproduction,” JOSA A 29(6), 1027–1034 (2012).

[3] Lissner, I. and Urban, P., “Toward a unified color space for perception-based image processing,” IEEE Transactions on Image Processing 21(3), 1153–1168 (2012).

9018-14, Session 2

Extension of Murray-Davies tone reproduction model by adding edge effect of halftone dots

G. M. Atiqur Rahaman, Ole L. Norberg, Per Edström, Mid Sweden Univ. (Sweden)

Halftone technique is used in different industrial sectors to reproduce color in paper, ceramic tiles etc. The color reproduction is modeled by classical Murray Davies (MD) formula, linearly combining the spectral reflectance of full tone ink and the substrate, scaled by the fractional colorant coverage. Here, the assumption is the color of both full tone ink and substrates are uniform and constant. In an ideal situation, MD model works well as it is insightful that the overall reflectance is composed of all the light reflected from different components off the print. But due to the dot gain phenomenon the fundamental assumption rarely applies. Hence, the MD formula has been modified in various ways to remove the limitation. Most important improvement was made by introducing Yule-Nielsen (Y-N) n-factor which models non-linear relationship due to lateral light scattering in paper. Practically n-factor needs to be fitted by nonlinear optimization and accepted as a useful tool for printer modeling and characterization. So the basic technique for most available regression based color prediction models is to use the measured average reflectance of an area of halftone dots and of unprinted paper to estimate the model parameters through optimization. Thus it emulates the system behavior rather than modeling physics of the process. However, in this study, we propose an expansion of basic MD formula to directly account the reflectance from optically gained area and adjusting the other parameters accordingl

In a recent study, we show that by microscale haltone image analysis, it is possible to calculate actual fractional coverage of full tone ink (a_i), optically gained (a_o) and paper between dots (a_p). So instead of scaling the measured reflectance of 100% ink (R_(?,i)) by average effective coverage, we scale it by actual a_i. Moreover, we add the reflectance of optically gained area (R_(?,o)) scaled by a_o. As a result, the proposed expansion of basic MD formula is: [R ?_?=a_i * R_(?,i)+a_o * R_(?,o)+a_p * R_(?,P)], where, R_(?,P) is the measured reflectance of white paper, and R ?_? is the predicted reflectance. As a preliminary approach, we replace R ?_? by measured reflectance and estimate R_(?,o) from a set of color patches with varying reference coverage. Then using the estimated R_(?,o), the spectra for any ink coverage is predicted by the expanded formula.

Compared to simple MD with and without DG, the proposed expansion gives closer match to the measured spectra. The preliminary results show that the average RMS error for cyan, magenta and yellow inks in uncoat paper are 1.9%, 0.8% and 0.7%, where the errors of MD with dot gain 2.4%, 2.7% and 2.7% and without dot gain 4.2%, 3.3% and 3.2%.

The full article covers in-depth result analysis for a large set of print samples covering a wide range of variations in paper, ink and halftoning method as well as print technology. It also compares accuracy with Y-N modified MD model and addresses the issue of estimation or measurement of R_(?,o). The advantage can be a closer relation between print result and physical properties of paper and ink. It can also be beneficial in product development of paper grades and optimization of the printing process.

9018-15, Session 3

Mathematical limitations when choosing psychophysical methods: geometric versus linear grey scales

Niels Dekker, Akzo Nobel Coating B.V. (Netherlands); Marcel P. Lucassen, LUCASSEN Colour Research (Netherlands); Eric J. J. Kirchner, Akzo Nobel Coating B.V. (Netherlands); Philipp Urban,

Fraunhofer-Institut für Graphische Datenverarbeitung (Germany); Rafael Huertas Roa, Univ. de Granada (Spain)

A common psychophysical method to investigate perceived differences in material appearance is the Grey scale method. For example, color difference equations can be derived from visual tests in which observers are asked to compare the perceived color difference between a pair of samples with the perceived color differences for so-called grey scales. The grey scales consist of a series of achromatic (grey) pairs with known color differences.

Two designs for grey scales have been proposed in the past. Linear grey scales show color differences that increase linearly in terms of the CIELAB color difference dEab. Alternatively, geometric grey scales are characterized by color differences that increase exponentially. Geometric gray scales are used in e.g. ISO standards and standard procedures of the American Association of Textile Chemists and Colorists and by the Fastness Tests Coordinating Committee.

We compared using linear versus geometric grey scales by carrying out psychophysical experiments in which we varied the type of grey scale. In this concise test, we imitated the design of a hypothetical larger study on color differences by investigating one color center recommended by the CIE (L*=44, a*=37, b*=23). Colors were displayed as uniform color patches on a color calibrated EIZO CG221 monitor inside a dark room.

Sample sets consisted of 30 sample pairs of colors close to the color center, with color differences within each pair ranging from dEab=0.13 to 2.50. Ten observers of normal color vision were asked to score the perceived color difference for each sample pair by comparison to the grey scales.

In one session, this grey scale consisted of a linear scale of six achromatic pairs with color differences of dEab=?dL*=0.0, 0.6, 1.2, 1.8, 2.4 and 3.0. In a second session, sample pairs were assessed with a geometric grey scale. In order to be able to assess the effect of only the type of grey scale, the geometric grey scale also consisted of six achromatic pairs. We chose color differences of dEab=dL*=0.0, 0.4, 0.8, 1.6, 3.2 and 6.4.

The results from these experiments show that with the geometric scale, the scores 1 and 2 (corresponding to dEab=dL*=0.0 and 0.4) become cluttered. Additionally, the scores 4, 5 and 6 (corresponding to dEab=dL*=1.6, 3.2 and 6.4) become cluttered as well, if scored at all. Therefore, the six-point geometric scale effectively is only a four-point scale.

By analyzing several other designs for the geometric grey scales we find that for investigating small color differences it is mathematically impossible to construct a geometric gray scale that avoids the problems found here. This may explain why in the past, observer variability has been found to be larger for geometric grey scales than for linear grey scales [1]. Another potential explanation is that when using a geometric scale, observers may find it more difficult to give intermediate scores, as this would require a mental nonlinear interpolation.

[1] L.M. Cárdenas, R. Shamey and D. Hinks, “Development of a novel linear gray scale for visual assessment of small color differences”, AATCC Review 9 (2009) 42-47

9018-16, Session 3

The visual appearance and surface texture of materials according to the old masters

Carinna E. Parraman, Univ. of the West of England (United Kingdom)

Primary components of colour reproduction of textured materials, are firstly, the accurate rendering of the appearance of texture [1], and secondly, the ability to print a surface topology that moves towards 2.5D texture printing [2]. The line of enquiry is based on the author’s interest in the relationship between the ‘direct manipulated’ mark [3] that is digitally generated (mouse, drawing tablet’s, iPads); compared to the analogue mark (produced by a brush, a charcoal smudge, an etched or hand drawn line) and its surface tactility or gonio characteristics [4], [5]; compared to its printed reproduction (for example an inkjet, electrostatic, four-colour separation).

There is an emerging area for accurate rendering and synthetic application of texture, for example surface rendering in 2D and 3D CAD; and due to recent developments in novel materials and decorative printing inks, textures and embellishments are being incorporated to enhance the surface qualities of packaging. However, convincing naturalistic rendering and texture has proven to be more difficult. Where the human visual system is more forgiving in the perception of halftoned images, texture is problematic, as our visual system is able to discriminate the difference between natural and patterned texture. A natural texture appears homogeneous, but remains random - each element is similar but remains unique. However a patterned texture, although homogeneous is composed of the same repeatable and recognisable elements. Furthermore to render surfaces with no discernable pattern structure that comprises unlimited variations can result in large file sizes.

As demonstrated in the recent art exhibition by painter, photographer and printmaker David Hockey at the Royal Academy, London (21 January - 9 April 2012) [6] comparing his paintings and drawings to his iPad works, the surface qualities of the inkjet prints from the iPad drawings were disappointing and devoid of surface characteristics. http://www.youtube.com/watch?v=0jabJKtqK0k. Whereas Hockney’s thickly applied paint onto canvas, the paint has a multi-dimensional quality, the varying translucency and opacity of the brush strokes can be seen, as can gloss and matte differential between oil on canvas and watercolour on paper.

The difference between Hockney’s approach to drawing using paint on an iPad and drawing on paper can be loosely described as the difference between digital (graphical user interface, pixels, colour picker tools, vector, raster) and analogue (autographic, pigments, brushes, fluid dynamics, materials, texture) and the need to develop methods that are a verisimilitude of real materials [7].

The evolving question is, what are the elements of paintings produced by old masters that capture the qualities, texture, grain, reflection, translucency and absorption of a material, that through the application of coloured brush marks, were able to create a convincing likeness for the material qualities of wood, metal, glass and fabric?

There is a difference here between the photoreal methods developed by artists working in the 20th century who were interested in the creation of hyper-real images. These methods and paintings emulated photographic images and therefore artists attempted to remove any brush marks and surface characteristic to obtain a photoreal quality. However, the painters in the 16th and 17th century who were interested in creating a convincing representation of the attributes of a material, these paintings on close inspection demonstrate a gestural almost abstracted version of the material and surface. The paper suggests that in order to create both a convincing visual appearance, a high level of detail is not necessary, that too much information possibly hinders the appearance. It suggests that by using a more gestural approach, whereby the relationship of mark and colour, and by modulating the fluid dynamics of a mark through a textured surface, a more convincing rendering of texture can be achieved.

Artists have been long aware of the psychological aspects of the juxtaposition of colour in exploiting the optical qualities and arranging visual effects in artworks. The artists, such as Velázquez, Goya, Holbein, Raphael, Raimondo de Madrazo, Gainsborough, Reynolds [8] demonstrated their mastery of texture by juxtaposing velvet with fur, satin alongside stiff silver embroidery. In order to better understand the convincingness of the visual appearance of texture, in this instance, the study has concentrated on the accurate rendering of garments. What material qualities were they able to convey to the viewer through the medium of paint, that a diverse range of materials - satin, silk, wool, velvet, woven, tapestry, brocade, metal could be demonstrated to the viewer.

[1] P. Campisi, A. Neri, & G. Scarano, ‘Reduced Complexity Modelling and Reproduction of Colored Textures’. IEEE Transactions on Image Processing, 9, 2000, 510-518.

[2] C. Parraman (2012), Special Session Dark Side of Color V, ‘Dark Texture in Artworks’, Proc. IS&T/SPIE Electronic Imaging, San Francisco,

California USA, 23-27 January 2012, Vol.8292-0H.

[3] Shneiderman, B. Leonardo's Laptop: Human Needs and the New Computing Technologies, MIT. 2003

[4] N. Pauler, O. Norberg & P. Edström, 'Mechanisms involved in the optical interaction between ink and substrate'. Advances in Printing and Media Technology, 2009, 36.

[5] J. C. Ashbaugh, R. S. Berns, B. A. Darling, L. A. Tamplin, 'Artist Material BRDF Database for Computer Graphics Rendering', in: 17th Color Imaging Conference Final Program and Proceedings, Society for Imaging Science and technology, 2009, pp. 62 - 68).

[6] M. Gayford, 'David Hockney: A Bigger Picture', The Infinity of Nature Royal Academy of Arts Magazine. London: Royal Academy 2011.

[7] A. Blatner, J. Ferwerda, B. Darling, 'TangiPaint: A Tangible Digital Painting System', 19th Color and Imaging Conference Final Program, November 7-11, 2011 San Jose, California. Society for Imaging Science and Technology and Society for Information Display, 2001, 102-107.

[8] Examples of Images

RAPHAEL, "Portrait of a Cardinal" (1510-11) http://www.museodelprado.es/imagen/alta_resolucion/P00299.jpg

RAPHAEL, "Portrait of Pope Julius II" (1511),

http://www.nationalgallery.org.uk/cid-classification/classification/picture/raphael,-portrait-of-pope-julius-ii/260682/*/moduleId/ZoomTool/x/-132/y/-437/z/4

HOLBEIN, "The Ambassadors" (1533),

http://www.nationalgallery.org.uk/cid-classification/classification/picture/hans-holbein-the-younger,-the-ambassadors/271993/*/moduleId/ZoomTool/x/152/y/0/z/1

National Gallery, London

9018-17, Session 3

On pictures and stuff: image quality and material appearance

James A. Ferwerda, Rochester Institute of Technology (United States)

Would be nice to be able to submit figures (imaging conference... pictures worth 1000 words... sigh...)

Images have always presented a puzzle for perceptual scientists because they serve as visual representations of objects while also being objects themselves. Much effort has gone toward understanding how images represent the three-dimensional properties of objects, and there is now considerable knowledge about the relations between the geometric projections used in image rendering and the perceived shapes of depicted objects. Considerably less effort has focused on how images convey other important object properties such as materials and textures, and although there is a vast literature on image quality that purports to speak to these issues, the premise of this paper is that much this work conflates the signal properties of images with the visual messages they convey.

Efforts to increase image quality typically focus on improving the signal coding capabilities of the medium (resolution, frame rate, dynamic range, color gamut, etc.), with little regard for the messages (visual information) that the images will be used to communicate. The faith is that if the image signal is ideal then the message will be conveyed with high fidelity. This approach seems logical and has mathematical support from the fields of signal processing and information theory, however the danger of focusing exclusively on the signal properties of images is that we may miss insights and opportunities that come from distinguishing between the imaging medium and the visual messages conveyed by that medium.

Figures 1 illustrates the distinctions can be drawn between the signal and message properties of images. The left panel shows a grayscale photograph of a black sports car parked on a concrete pad. Both the car and the pad show distinct reflections of the surrounding environment that suggest that the car is glossy and the concrete pad is wet. The grayscale levels and contrasts in the image also suggest that the car is black (or a dark color) and the concrete pad has a lighter shade. The right panel shows the same scene represented by a halftoned image created to simulate the contrast and sharpness of a typical newspaper print. This image is clearly different than the one on the left, and in conventional terms one would say that its quality is low. However, as a visual representation of the scene, this image is largely equivalent to the one on the left in that we can still perceive important properties of the depicted objects such as the shape, reflectance, and gloss of the black car, and the reflectance and gloss/wetness of the concrete pad.

We have conducted a series of experiments to investigate the visual system's ability to "see through" image distortions such as the ones shown in Figure 1 to perceive the object and scene properties the image depicts. In a series of experiments we are investigating the ability of conventionally low quality (low contrast, blurry, disordered) images to faithfully represent the material properties of objects. Our approach is to perform gloss scaling experiments using images of glossy objects and to compare the scales produced by high and low quality images. Figure 2 shows examples of some of the images being used in the experiments.

Contrary to the predictions of standard image quality metrics, we are finding that the ability to discriminate objects with different gloss properties is not reduced as much as would be expected by these distortions. It is as if observers are able to see through the distortions to perceive the material properties of the depicted objects. On the basis of these experiments we are developing new image quality metrics that take into account recent findings on the role of light reflection statistics in material perception, and analyze interactions between the statistics of light structuring by materials and the statistics of image coding distortions.

The focus of this work is on understanding the relationships between the characteristics of image signals and the fidelity of the visual messages images convey. Our goals are to learn more about how images work as visual representations, to develop more meaningful image quality metrics that better predict how well images with different signal properties serve as visual representations of the objects they depict.

9018-18, Session 4

Modeling cloth at micron resolution *(Invited Paper)*

Kavita Bala, Cornell Univ. (United States)

Fabrics are ubiquitous and play an important role in our daily lives. Human beings are able to assess subtle distinctions in appearance when judging different fabrics like silk versus nylon. Accurately capturing fabric appearance to the exacting standards that humans expect has been a long standing challenge. While various models have been proposed in graphics for cloth, until recently they were not adequate in terms of quality. This is because fabrics fundamentally have a very complex structure. Fibers are twisted to produce yarns, and differ greatly depending on their source: e.g., cotton from cotton plants, wool from sheep, silk from silk worms. The resulting yarns are woven into patterns. The combination and structure of the weave plays a major role in the appearance of the final fabric. We will particularly focus on woven fabrics, though other challenges exist in modeling fabrics beyond wovens.

To accurately represent the appearance of fabrics it is important to model the complex structure of fibers, yarns and weaves. Until recently models either were: surface-based models that did not correctly capture how light interacts with fabrics; or programmable models of fabrics that required intensive programming effort to create and still missed the irregularities that are typical for real fabric. We have introduced a new pipeline to address this challenge of modeling fabrics.

Our approach uses CT modeling to capture the complex 3D structure of materials representing both the macro-scale and mesoscale structure of fibers and yarns.

This modality of modeling appearance is new to graphics and has proven to be crucial in developing appearance models that can correctly

predict the subtle differences between silk vs. wool, cotton vs. nylon, or between different types of silk. While this approach shows great promise in bringing micron resolution detail into fabric models it poses a few fundamental challenges. (1) CT data includes no optical information; that information has to be inferred. (2) The captured region scanned is small, even with the highest resolution devices it is possible to get micron resolution detail only for about a 1/2 sq. cm. area of material. But true fabrics can be meters long with complex designs and patterns. Scaling the data to that size is hard. (3) The complexity of the corresponding models is a challenge.

We discuss our solution to these three challenges. We introduce a new modeling approach that combines CT scans with a single photograph to build a full volumetric appearance model for fabrics like silk satin and velvet. We show how volume synthesis can be used to create user-specified designs and weave patterns using small swatches of CT scans to create arbitrarily large fabric models. Finally, we select the appropriate scale of the model that provides maximum visual fidelity but also manages the complexity of the volume models.

Together our approach can accurately and efficiently visualize micron resolution models of woven fabrics made of silk, cotton and wool.

9018-19, Session 4

Towards a better understanding of the color shift of effect coatings by densely sampled spectral BRDF measurement

Alejandro Ferrero, Berta Bernad, Joaquin Campos Acosta, Consejo Superior de Investigaciones Científicas (Spain); Francisco Javier M. Martinez-Verdu, Esther Perales, Univ. de Alicante (Spain); Ivo van der Lans, Akzo Nobel N.V. (Netherlands); Eric J. J. Kirchner, Akzo Nobel Coating B.V. (Netherlands)

Color of effect coatings depend strongly on the illumination and viewing angles, giving them a very appealing appearance. As a consequence, these coatings have become very popular in the automotive industry and in other application such as cosmetics and security inks. Effect coatings consist of a transparent medium containing traditional absorption pigments, and flake-shaped effect pigments.

To completely characterize the color of these coatings under any illuminant and for any illumination/viewing geometry, the spectral Bidirectional Reflectance Distribution Function (BRDF) should be measured for a large number of measurement geometries, thus providing all information required to characterize the color shift.

GEFE, the gonio-spectrophotometer developed at Instituto de Óptica in CSIC (IO-CSIC), allows the spectral BRDF to be measured at any geometry, including out-of-plane and retro-reflection angles. Twenty-four effect coating samples were prepared by AkzoNobel and characterized by using this instrument. These samples contain metallic and/or interference pigments. The samples can be classified into several groups:

(1) Metallic coatings. The texture was varied from very fine to very coarse in order to assess the dependence of the color travel on the texture.

(2) Interference coatings showing a varying degree of color travel as depending on illumination/viewing angle. These samples included for example Iriodin®, Xirallic®, Colorstream® and Meoxal® pigments from Merck and Chromaflair® pigments from JDSU.

The spectral BRDF was measured by GEFE following a normalized procedure. The polar angles (with respect to the sample normal) were chosen from 0o to 70o in steps of 10o, both for illumination (?i) and viewing (?s) directions. The azimuth angle of the viewing directions (?s) was varied from 0o to 180o in steps of 30o, assuming symmetry with respect to the incident angle. The azimuth angle of the illumination (?i) was not varied, since preliminary measurements showed that anisotropy for this angle was negligible. The overall number of measurement geometries was 384 (neglecting possible double counting due to symmetries).

The results are presented in CIELAB diagrams to show the color travel. For the sake of clarity, only in-plane and non-specular measurements are shown, as usually. The points in the plots were grouped using two different types of connecting lines: (1) The well-known interference lines of constant aspecular angle (corresponding to an almost constant orientation angle of the flakes), and (2) a recently defined line of almost constant incident angle with respect to the flakes' normal, which is an important alternative to the commonly used aspecular lines of constant illumination angle. Conclusions based on these two types of lines are quantitatively drawn from the measurement data, and interpreted in view of the absorption pigments and metallic/interference pigments.

9018-20, Session 4

Lateral light propagation and angular variation of the reflectance of paper

Ludovic G. Coppel, Gjøvik Univ. College (Norway)

The appearance of translucent materials is strongly affected by bulk (or sub surface) scattering. Bulk scattering does not only control the shade of the material but also leads to the blurring of surface features such as topography and coloured spots. In computer rendering, it has been recognised that subsurface scattering must be taken into account in order to accurately capture the appearance of e.g. skin, orange juice or snow [1]. Keeping angular variation into account, this require to model the bidirectional surface scattering distribution function (BSSRDF), which relates the lateral propagation and reflection angle of light rays hitting a surface, as opposed to the bidirectional distribution function (BRDF), which assumes that all reflected rays exit at the entrance position on the surface. For paper and carton board, which can be fluorescing, lateral light propagation and angle-resolved reflection have been studied extensively but treated separately. Lateral light propagation is known to make printed dots appear larger than their physical size while several authors have studied the BRDF of paper and prints to predict gloss and angular varia-tion of colour. Recent studies have shown that the bulk reflection of all turbid media is anisotropic [2, 3] while the fluorescence component is nearly Lambertian [4] and that directional inhomogeneity due to the planar fibrous structure requires the use of non rotationally invariant single scattering phase function to render the relative large extent of lateral propagation observed in uncoated papers [5].

The present work aims at modelling the BSSRDF of turbid media in order to study the angular variation of the reflectance as function of the lateral propagation within the medium. Using general radiative transfer theory, which has been shown to describe accurately the light propagation in paper, the BSSRDF can be simulated, although its direct measurement remains impractical. The Open PaperOpt simulation tool [6] is modified to perform Monte Carlo simulations of the spatial- and angle-resolved reflectance of turbid media for different scattering and absorption coefficients, phase functions and surface topographies representative for several paper grades. The model used treats surface and bulk scattering separately and traces the path of a large number of wave packets interacting with the medium. A 1 mm x 1 mm area is illuminated at different incident angles and the scattered angle of the reflected wave packets is then recorded as function of their lateral propagation.

The simulated average (or standard) BRDFs show a specular reflectance peak, but also increase with in-creasing polar angle, from 0? normal to the paper surface to 90? parallel to the surface. The simulations are in accordance with observed reflectance anisotropy from paper [7]. The BSSRDF simulations show that the bulk reflection is anisotropic and that the anisotropy decreases with propagation distance.

The angle-resolved reflectance of turbid media is thus function of the lateral light propagation within the substrate. This may impact on the appearance at different angles and make measurements of the lateral light propagation dependent on the instrument geometry. Since the model used can handle topographical surfaces and ink layers, future work includes to model the BSSRDF of 2.5 prints.

References

[1] H. W. Jensen, S. R. Marschner, M. Levoy, and P. Hanrahan, "A practical model for subsurface light transport," in "Proceedings of the 28th annual conference on Computer graphics and interactive techniques," (ACM, New York, NY, USA, 2001), SIGGRAPH '01, pp. 511–518.

[2] M. Neuman and P. Edström, "Anisotropic reflectance from turbid media. I. theory," J. Opt. Soc. Am. A 27, 1032–1039 (2010).

[3] M. Neuman, L. G. Coppel, and P. Edström, "A partial explanation of the dependence between light scattering and light absorption in the Kubelka-Munk model," Nord. Pulp Pap. Res. J. 27, 426–430 (2012).

[4] N. Johansson and M. Andersson, "Angular variations of reflectance and fluorescence from paper - the influence of fluorescent whiten-ing agents and fillers," in "20th Color and Imaging Conference," (Society for Imaging Science and Technology, 2012), pp. 236–241.

[5] T. Linder, T. Löfqvist, L. G. Coppel, M. Neuman, and P. Edström, "Lateral light scattering in fibrous media," Opt. Express 21, 7835–7840 (2013).

[6] L. G. Coppel, P. Edström, and M. Lindquister, "Open source Monte Carlo simulation platform for particle level simulation of light scattering from generated paper structures," in "Paper making research symposium," , E. Madetoja, H. Niskanen, and J. Hämäläinen, eds. (Kuopio, 2009), p. 97.

[7] N. Johansson, M. Neuman, M. Andersson, and P. Edström, "Separation of surface and bulk reflectance by absorption of bulk scattered light," Appl. Opt. 52, 4749–4754 (2013).

9018-21, Session 4

Printing gloss effects in a 2.5D system

Teun Baar, Océ Print Logic Technologies (France) and Télécom ParisTech (France); Sepideh Samadzadegan, Technische Univ. Darmstadt (Germany); Maria V. Ortiz Segovia, Océ Print Logic Technologies (France); Philipp Urban, Fraunhofer-Institut für Graphische Datenverarbeitung (Germany); Hans Brettel, Télécom ParisTech (France)

An important aspect of digital prints is the perceived gloss level that is seen to be dependent on various parameters of the print process. Currently, not many studies have investigated the relation between the print controls and the gloss output, but it is seen that the paper substrate, the type of inks and the print method all influence the gloss level. In these cases there is a strong correlation between the surface roughness of the printout and the amount of specular reflection. Getting benefit of 2.5D printing, controlling the micro surface texture/ roughness and hereby influencing the gloss appearance is doable. Therefore, in order to control the spatially varying reflection properties of the printout, an understanding between print controls and gloss appearance output is important. The aim of this paper is to understand the relationship between the Delayed/Drying Time, as one of the printer control parameters in our 2.5D printing process, and the gloss level of the printout.

We have used two prototype printers that have the ability to superimpose layers of ink, a wet-on-wet and a wet-on-dry system. We have observed a strong relationship between print parameters and the gloss level of the printout. It has been seen that the gloss level depends on different aspects of the print system, such as the substrate and the inks, the parameters for the UV exposure, the temperature of the heating element for ink crystallisation, the levels of varnish that are applied and the 2.5D surface texture. Moreover, when superimposing different layers of ink, the time in between the placement of the inks showed to be an important parameter for the obtained gloss level. We performed several experiments where colour patches, using cyan (C), magenta (M), and yellow (Y) with different coverages of 0%, 50%, and 100%, were printed on top of two separate passes of solid white ink. All the patches were created in a 3-pass print mode where the time between the passes was varied to study the influence of time delay on surface roughness and thereby glossiness appearance.

The gloss level and surface roughness of the obtained samples was measured using "BYK micro-TRI-gloss" gloss meter. This device conforms to these standards: ASTM D 523, D 2457, DIN 67530, ISO 2813, ISO 7668, JIS Z 8741 [1]. The measurements are conducted for 20°, 60°, and 85°.

This data are useful in order to estimate the relationship between the time delay between layers and the roughness of the print surface. Such information can later be used as a mechanism to create print modes with specific gloss levels or to intentionally introduce spatial gloss variations to the prints. The achieved results represent the general decreasing trend of glossiness appearance as the time delay between the second deposited white layer and the top CMY layer varies between 1 to 4 seconds. While, no noticeable decreasing or increasing order is visible as the time goes on from 4 to 10 seconds. Physical observations of the generated patches also confirm the measurements.

Finding the limits of the gloss levels that can be printed will lead to a gloss gamut which together with a colour gamut can be used as a basis to which an input image containing both spatial colour and gloss information can be mapped. For this application it is also important to consider the effect of the print controls on the colour shifts with different gloss levels applied. During one of the observations we noticed that having multiple ICC profiles corresponding to different gloss levels might help to achieve colour consistency between the same colours with spatially varying gloss levels. Psychophysical experiments need to be conducted to explore the relationships between physical gloss measurements and perceived gloss levels as well as the accuracy of colour in our printouts.

[1] http://www.karg-industrietechnik.de/english/products/Gloss_Meter_Ed1_Rev4_EN_EVLR.pdf

9018-22, Session 5

Measured materials for rendering and visualization: applications, existing approaches, and open challenges *(Invited Paper)*

Jan Meseth, RTT AG (Germany)

The use of measured materials for rendering and visualization has come a long way. Although initial ideas based on BRDFs have been published many years ago, for a long time they have only been adopted in special applications like optics simulations. Due to the recent popularity of physics based rendering technology, which calls for and supports physically plausible material representations, measured materials are finally at the edge of becoming widely used in product design, marketing, movie production and even computer games.

This talk is organized in three parts. First, an overview of applications of measured materials in various fields of computer graphics is given and the specific requirements of these use cases are outlined. It is shown that in some cases measurements are the method of choice due to their high accuracy while in other cases they are used as they capture the visual richness of materials much more easily than even a trained user could model in a reasonable amount of time.

Second, existing approaches for digitizing materials are categorized with respect to the kinds of materials they can measure adequately, and reviewed concerning their fitness for the industrial use cases. This part also establishes a link between devices used in rather traditional color measurement and components of complex materials that can be measured with them.

Third, remaining challenges for material measurement techniques are presented, ranging from limited accuracy and cost issues of digitization technologies over prohibitive storage requirements to missing intuitive editing capabilities. These challenges are intended as starting points for future research activities.

9018-23, Session 5

Image ghosting reduction in lenticular relief prints

Teun Baar, Océ Print Logic Technologies (France); Marjan Shahpaski, Ecole Polytechnique Fédérale de Lausanne (Switzerland); Maria V. Ortiz Segovia, Océ Print Logic Technologies (France)

Commonly known lenticular prints use a lens-like system superimposed on a standard 2D print to control the light directed into each direction. By controlling the light directed to the prints, different images can be observed just by changing the viewing direction. These lenticular systems are often seen in artwork and publicity campaigns.

Thanks to our 2.5D or relief printing system, we are capable of creating a lenticular effect embedded directly on the prints that does not require the use of a system of lenses. Relief printing techniques allow the reproduction and control of surface characteristics such as reflectance and texture. Through the combination of various surface parameters many special effects can be reproduced in the prints. One of the applications is the creation of lenticular prints, on which we combine two (or more) images that are meant to be viewed from different directions or with different illumination conditions. In the case of having two source images, the lenticular effect in the prints is achieved by using a zigzag-shaped surface composed of a chain of continuous small triangles. Each of the two source images is entirely printed either on the left or on the right sides of the triangular structures, where each side corresponds to one of the two intended views. The spatially varying height, width, and angles of the triangular structures depend on the desired appearance of the print for the distinct viewing or illumination directions.

A problem that is often encountered in lenticular prints is the crosstalk or ghosting effect between different views. Ghosting occurs when some parts of one image remain visible for the illumination or viewing direction corresponding to the other image. Methods for reducing these artefacts have been proposed not only for lenticular applications [1] but also for related crosstalk effects in stereoscopic displays [2]. Our goal is to design a compensation method for ghosting reduction in lenticular prints reproduced by a 2.5D printing system.

Given the fact that ghosting is strongly related to the content of the images, we propose to use an image content driven technique. Our algorithm identifies the regions in the source images that are prone to cause ghosting for a given set of viewing angles. For that purpose, a model of the appearance of the ghosting effect in our lenticular prints is implemented. The model uses as parameters the source images, a set of viewing angles, and the height, width and angles of the triangular structure used as the support for the lenticular system. Such parameters along with the information about problematic regions can be optimized and tuned in order to achieve the best approximation to the intended views.

In preliminary experiments, improvements have been observed with respect to the quality of the lenticular effect for different viewing directions but the impact of the compensation on the quality of the source images still needs to be evaluated. Psychophysical experiments will be conducted to estimate the actual improvements of the algorithm. This study is also applicable for future work on printing reflectance distribution functions combining 2.5D surface texture and surface controllable reflection, which will also require compensation of ghosting effects from different views.

[1] Stephen Gulick, Jr. et al, Lenticular image product presenting a flip image(s) where ghosting is minimized, PATENT NO. US 6,405,464. JUNE 2002

[2] Janusz Konrad, Senior Member, , Bertrand Lacotte and Eric Dubois, Cancellation of image crosstalk in time-sequential displays of stereoscopic video IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 9, NO. 5, PP. 897–908, MAY 2000

9018-24, Session 5

Continuous tone printing in silicone from CNC milled matrices

Stephen Hoskins, Peter MacCallion, Univ. of the West of England (United Kingdom)

The Centre for Fine Print Research (CFPR) has been reappraising 19th Century print processes for the last fifteen years and Hoskins has published on the continuous tone properties of Woodburytype and its relation to other 19th processes such as photo-ceramics in previous papers to the IS&T PICT1 and EI in 20032. Since the invention of Walter Bentley Woodbury's 'Woodburytype process in 1865 3, it has been possible to autographically print a true continuous tone black and white photographic image. New research at CFPR builds upon previous research to combining 19th Century Photomechanical techniques with digital technology to reappraise the potential of these processes

CNC milled relief printing blocks from which casts are taken will enable a physical and tactile surface quality that more closely resembles the autographic surface of the Woodburytype and the three colour carbon printing processes. Therefore this latest research aims to explore the potential of creating coloured pictorial imagery from a continuous tone relief surface. The specific context for the research is to create a surface quality and visual appearance that retains the character of the earlier autographic processes, such as letterpress and screen printing, which are undergoing a revival, This is happening for two reasons, one, for their hand crafted content, and secondly, because they create a physical and tactile surface that has a subjective quality, hard to define using standard testing methodologies.

To prove the potential the research team have been using CNC milled images where the height of the relief image is dictated by creating a tone curve and then milling this curve into a series of relief blocks from which the image is cast in a silicone ink. A translucent image is cast from each of the colour matrices and each colour is assembled - one on top of another - resulting is a colour continuous tone print, where colour tone is created by physical depth of colour.

A number of factors have come together to enable a reappraisal of continuous tone printing. Software is now readily available that will split a photographic image into a bump map image where the Z-axis corresponds to a set of height steps on a tonal step wedge from black to white. This software is easily transcribed to either a DXF for processing in G Code to CNC (Computer Numeric control) in order mill an image. In addition new silicone materials have become more readily available that can be precisely coloured.

The aim is to create a digitally initiated, physical tactile surface that retains the qualities of arlier autographic process which is capable of being transferred to a range of substrates and demonstrates the potential for future research into new methods of continuous tone printing. This research is part of a larger parallel research project to investigate the potential of creating a physically printed relief surface that better represents the autographic mark making requirements of visual artists and designers.

1 Thirkell, Paul, Hoskins, Stephen, 2003, A reassessment of past colour collotype printing achievements as a model for current digital, archival printing practice. IS&T SPIE PICS Digital Photography Conference, Rochester New York

2 Hoskins, Stephen, Thirkell Paul, 2003 The relevance of 19th century continuous tone photomechanical printing techniques to digitally generated imagery IS&T SPIE Electronic Imaging

Conference, Santa Clara, California

3 Nadeau, Luis, Encyclopedia of printing, photographic and photomechanical processes, 1989-1990 Atelier Luis Nadeau, Canada. ISBN 0969084153

9018-25, Session 5

Characterization of relief printing

Xing Liu, Purdue Univ. (United States); Lin Chen, Rochester Institute of Technology (United States); Maria V. Ortiz Segovia, Océ Print Logic Technologies (France); James A. Ferwerda, Rochester Institute of Technology (United States); Jan P. Allebach, Purdue Univ. (United States)

Conventional printers are constrained to reproduce image contents in a two-dimensional space as they are not capable of using any height or surface information that may exist in a third dimension. Relief printing technology developed by Océ allows the superposition of several layers of ink on different types of media which creates a variation of the surface height defined by the input to the printer. Evaluating the reproduction accuracy of distinct surface characteristics is of great importance to the application of the relief printing system. Therefore, it is necessary to develop quality metrics that are specially designed to evaluate the quality of the relief process.

In this paper, we attempt to objectively evaluate the relief printing quality in the third dimension, i.e. height information. To achieve this goal, we define metrics and develop models that aim to evaluate relief prints in two aspects: overall fidelity and surface finish. For the overall fidelity, three metrics are calculated: MTF, distance and RMSE between the input height map and scanned height map, and print surface angle accuracy. To characterize the surface finish property, we measured the surface roughness and raggedness, generated surface normal maps and developed a light reflection model that serves as a simulation of the difference that may be perceived by human observers between ideal prints and real prints. Three sets of test targets were designed and printed by the Océ relief printer prototype for the calculation of the above metrics: (i) twisted target, (ii) sinusoidal wave target, and (iii) ramp target. The twisted target and ramp target were designed to evaluate both the overall fidelity and the surface finish. We choose suitable metrics or models for the two targets respectively according to their surface pattern features. The sinusoidal wave target was used only for the MTF calculation as a probe to the overall fidelity. Furthermore, a zone plane pattern target was used to validate the performance in different orientations of the relief printers. Additionally, we investigated the space connecting the mathematical models and the printer device to better understand the performance of the relief printer.

In general, the calculated metrics demonstrate that the height of relief prints is reproduced accurately with respect to the input design. However, there is a dependence on the printing direction. For the ramp target with a symmetric shape along printing direction, the left side is always steeper than the right side. As expected, the MTF analysis shows that there is a strong relationship between the frequency of fine details and the accuracy of the reproduction. The MTF decreases as the frequency increases. The surface finish analysis led to the conclusion that the surface roughness and raggedness have a dependency on the geometry of the relief prints. As the structure of the relief prints gets steeper, the printed surface becomes rougher. The light reflection model provides a heuristic visualization of the surface characteristics for the ideal prints and the real prints under the same light environment setting. From the simulation we see that the visual experience depends on the direction of the incident illumination and viewing angle. That is to say, the surface exhibits different extent of roughness when we change the direction of incoming light or rotate the relief prints. Finally, the investigation of the printer performance shows that the relationship between different settings and prints thickness is nonlinear. This work provides the beginnings of an approach to characterizing relief printers to provide high quality reproduction of surfaces with complex color and 3D texture properties.

9018-26, Session 5

Method for visualization and presentation of priceless old prints based on precise 3D scan

Eryk Bunsch, Wilanow Palace Museum (Poland); Robert Sitnik, Warsaw Univ. of Technology (Poland)

Graphic prints and manuscripts constitute main part of the cultural heritage objects created by the most of the known civilizations. Their presentation was always a problem due to their high sensitivity to light and changes of external conditions (temperature, humidity). Today it is possible to use an advanced digitalization techniques for documentation and visualization of mentioned objects. In the situation when presentation of the original heritage object is impossible, there is a need to develop a method allowing documentation and then present to the audience all the esthetical features of the object. During the course of the project scans of several pages of one of the most valuable books in collection of Museum of Warsaw Archdiocese were performed (6 pages already scanned, another 6 planned to be scanned till the end of the year 2013). The book in question is „Ornatus cum Apocalipsi. Liber Vulgarium devote contempla comnis pulcheris figures", also known as "Great Dürer Trilogy". The book dates back to beginning of XVI century and consists of three series of wood engravings by the Albrecht Dürer: "The life of Mary", "Great Passion" and "Book of Revelation". There are 46 pages made of vat paper (320 mm x 465 mm) adorned with graphic prints of German master.

For a few years now the Warsaw University of Technology, Mechatronics Faculty in cooperation with Wilanow Palace Museum has been developing technology for works of art documentation based on precise three-dimensional measurements. Structured light illumination scanning was chosen as registering technique.

Used measurement system consists of a custom designed, structured light-based, high-resolution measurement head with automated digitization system based on the industrial robot. This device was custom built to meet conservators' requirements, especially no ultraviolet nor infrared radiation emission in the direction of measured object. Only low intensity visible light (400-700 nm) is allowed and the emission occurs only during measurements, it is blocked (light sources are turned off) during positioning and pauses. Such measurement head is built on the basis of commercially available multimedia projector (Casio XJ-A255) with custom lens, allowing for close range focusing (250 mm from the device) and small image size (50 mm x 50 mm x 16 mm). The new optics block and no heat or UV radiation towards image require completely changed cooling system, which was carefully designed to ensure also a low level of vibrations. After the modifications, series of tests have been performed. Whole measurement sequence is comparable to 2 hours of light exposure of 50 lx. The detector used in this measurement head is a standard high-quality DSLR, namely Canon EOS 60D.

Documentation of one page from the book requires about 380 directional measurements which constitute about 3 billion sample points. The distance between the points in the cloud is 20 µm.

The 3D measurement technique used in this study is structured light illumination, employing the temporal phase shifting method combined with hierarchical phase unwrapping. Temporal phase shifting is realized by projection of a set of shifted in phase sinusoidal modulated patterns on the object surface. These patterns, deformed on the object surface, are acquired by the detector. After the projection of sine patterns a Gray code sequence is displayed. It is used for binary enumeration of the periods of the sine pattern. From the whole acquired sequence a phase map is calculated.

Most of the old types of paper, similarly as the prints made today using traditional methods, is not flat at all. Provided that the measurement with MSD of 2500 points makes it possible to show to the publicity the spatial structure of this graphics print.

An important aspect is the complexity of software environment created for data processing, in which massive data sets can be automatically processed and visualized.

Very important advantage of software which is using directly clouds of points (no need to create triangle mesh) is possibility to use freely virtual light source. Additionally with full freedom to operate the measurement data (the angle of view and proximity can be chosen freely) it gives the possibility to observe all the esthetic nuances of the heritage object using digital data without the necessity to physically access the object.

ACKNOWLEDGMENTS

This work has been partially supported by the Ministry of Culture and National Heritage (Poland) by KULTURA+ framework (2011-2015).

Conference 9019: Image Processing: Algorithms and Systems XII

Monday - Wednesday 3 –5 February 2014

Part of Proceedings of SPIE Vol. 9019 Image Processing: Algorithms and Systems XII

9019-1, Session 1

On the pass band of light-field displays *(Invited Paper)*

Atanas R. Boev, Robert Bregovic, Tampere Univ. of Technology (Finland)

Light-field (LF) displays aim at recreating a visual replica of the 3D scene of interest. They do so by recreating the light-field of the scene as it is cast towards different directions. Depending on the way of doing this, displays can be broadly distinguished as multiview displays, which recreate limited amount of scene images, each seen from different perspective; ray-space displays, which generate a large number of rays with different color, intensity and direction; and holographic displays, which reconstructs the wavefront of a scene LF by illuminating an interference pattern with a coherent light.

Neither of these methods is able to recreate a visually indistinguishable replica of the scene. The perceptual differences between the replica and the real scene are interpreted by the human visual system as artifacts. In order to quantitatively assess the visibility of such artifacts, it is essential to characterize the display in terms of visual throughput.

The visibility of artifacts depends on the interaction between the spectrum of the visualized content and the display's transfer function. This interaction can be conveniently expressed in frequency domain. Therefore, in order to assess the visibility of artifacts, one can study the performance of the display in frequency domain through a quantity which we call a display passband.

In this paper, we explain the concept of a display passband, and apply it for analysis of a range of LF displays. We demonstrate how the passband can be used as a formal representation of the ability of the display to create distortion-free scene representations.

9019-2, Session 1

A novel method of filtration by the discrete heap transforms

Artyom M. Grigoryan, Mehdi Hajinoroozi, The Univ. of Texas at San Antonio (United States)

In this paper, we describe the method of filtering the frequency components of the 1-D and 2-D signals, by using the discrete signal-induced heap transforms (DsiHT), which are composed by elementary rotations or Givens transformations. The transforms are fast, because of a simple form of decomposition of their matrices, and they can be applied for signals of any length. Fast algorithms of calculation of the direct and inverse heap transforms do not depend on the length of the processed signals. Due to construction of the heap transform, if the input signal contains an additive component which is similar to the generator, this component will be eliminated in the transform of this signal. The energy of this component will be preserved in the first point, only. In particular case, when such component is the wave of a given frequency, this wave will be eliminated in the heap transform. Different examples of the filtration over signals and images are described and compared with the known method of the Fourier transform. In this paper, we consider a subclass of discrete unitary transforms, which we call the heap transforms, or transforms which are generated by input signals. The complete systems of basic functions of heap transforms are referred to as waves generated by input signals, the waves with their specific motion in the space of functions. These generators play the role of keys without which it is not possible to reconstruct the initial signals or images. We consider only the heap transforms which are defined by a simple system of rotations. The heap transform is described by the unique set of angles, which represents the angular representation of the transform, or the signal-generator. This set of angles allows for moving the energy of processed components of the signal-generator to one point, and such heap is located at the first component of the transform. When the same system of rotations is performed over a signal which has the component that is similar to the generator, this component will be eliminated at each point except the first one where the energy of the component is moved. In particular case, when the heap transform is generated by the sine wave, this transform removes the component of the signal with the frequency of this wave. In other words, the sine wave, or its frequency is represented by the set of angles, and a signal rotated by this set of angles in the heap transform is losing this wave. Therefore the DsiHT can work as the linear filter, and our preliminary experimental results show that the filtration by the heap transform is comparative with the filtration by the Fourier transform. The heap transform has fast algorithms for any length of processed signals and uses less operations of multiplication than the DFT. It should be mentioned that the heap transformation transfers the signal into another signal and it may filter the frequency during this transformation. Therefore, the calculation of the inverse heap transform is not required when filtering the signal, and the method of DFT requires the calculation of the inverse transform.

The described above heap transformations represent a class of discrete unitary signal induced transformations which are defined by systems of moving functions. The movement of the basic functions is accomplished with rotation and angular representation is defined for the signal. Unlike the theory of discrete wavelets, these wave-functions move in the field associated with the generator, and during this movement they interact. The transforms are fast, because of a simple form of decomposition of their matrices, and they can be applied for signals of any length. The one of the most interesting properties of the heap transform can be effectively used for signal and image filtration. If the input signal contains an additive component which is similar to the generator, this component will be eliminated in the transform of this signal. When such component is the wave of a given frequency, this wave will be eliminated in the heap transform. The method of heap transform is simple and requires less operations than the method of the Fourier transform. The heap transformation transfers the signal into another signal and it may filter the frequency during this transformation. Therefore, the calculation of the inverse heap transform is not required when filtering the signals and the method of DFT requires the calculation of the inverse transform. Our preliminary results show that the heap transform-based method of filtering can be effectively used for filtering images. Different examples of filtering images by the heap transform generated by different frequencies are described.

9019-3, Session 1

Alpha-rooting method of color image enhancement by discrete quaternion Fourier transform

Artyom M. Grigoryan, Sos S. Agaian, The Univ. of Texas at San Antonio (United States)

The goal of image enhancement techniques is to improve a characteristic or quality of an image, such that the resulting image is better than the original, when compared against specific criteria. History of digital image enhancement has been more than 50 years. It has remained one of the most active research areas in image processing and computer vision. Recently, several image enhancement algorithms, and many applicable systems have been exploited. Current research in image enhancement covers such wide topics as algorithms based on the human visual system, histograms with hue preservation, JPEG-based enhancement

for the visually impaired, and histogram modification techniques. Two major classifications of image enhancement techniques can be defined: spatial domain enhancement and transform domain enhancement. Digital image enhancement is a powerful tool for many image-processing applications when the critical details are not seen clearly enough. The purpose of image enhancement is to improve a digital image quality and to support the human perception. For instance in medical imaging, such as a computer tomography and magnetic resonance, three-dimensional images (or stack of 2-D images) of different organs and tissues are produced. There are many sources of interference in the production of medical images, such as the movement of a patient, insufficient performance and noise of imaging devices. When the quality of an image is poor in its contrast, to improve the quality of image, enhance edges, and reduce the noise for diagnostic purposes, methods of enhancement can be used.

Existing methods for image enhancement focus mainly on properties of the processed image while excluding any consideration of the observer characteristics. With their specific nature, various enhancement methods are required for various types of images and applications. The Fourier transform plays important role in image processing and is used in different stages of processing such as filtering, coding, recognition, and restoration analysis. This transform was generalized for application of the Hartley, Hadamard, cosine, and other transforms. Transform-based methods of image enhancement are based on manipulation with all or part of spectral components of the transform. We focus on the well-known method of alpha-rooting, although other methods, including the log-alpha-rooting, modified unsharp masking, and methods based on wavelet transforms are also used in image enhancement. Many traditional methods of image enhancement were applied for processing the color images. The Fourier transform-based method is one of these methods, which in many cases is applied to each color plane-component of the color image separately. In other words, the color image is considered as a triplet of separate 2-D gray scale images and each of these images represents red, green, or blue component of the color. Quaternion numbers of Hamilton's was used in Ell's works, and after that time much attention was given to the transformation of the color components to the imaginary subspace of the quaternion numbers, ``imaginary part" of which consists of three components.

In this paper, new methods of image enhancement are proposed for enhancement color images, which are based on application of the concept of the two- dimensional discrete quaternion Fourier transform (DQFT) together with the well-known method of the alpha-rooting. The application of the alpha-rooting method which is based on the traditional two-dimensional discrete Fourier transform (2-D DFT) results in high quality images, when comparing with other transforms, such as Hadamard, Hartley, and cosine transforms. The image enhancement technique of alpha rooting can be used for enhancing high contrast edge information and sharp features in images, as well as for enhancing even low contrast images. Therefore the use of the 2-D DQFT in alpha-rooting adapted to the case of color images is much promised. The color images are considered in RGB format, and three color components (r,g,b) are placing into the tree imaginary parts (i,j,k) of the quaternion number space. Since in the 2-D DQFT, any given spatial variation of a color component (r,g,b) is separated into different real-and-imaginary parts of the spectral point, this transform may separate the information of colors in the spectral domain. Our preliminary results show that the application of the 2-D DQFT plus alpha-rooting method can be effectively used for enhancing color images. Examples of application of the proposed method on different color images and comparison with the traditional method when the 2-D DFT based alpha-rooting is applied for each color plane separately are given. To estimate the quality of the color image enhancement, we generalize the known EME measure of image enhancement and applied it for images enhanced by both 2-D DQFT and DFT.

9019-5, Session 2

(JEI Invited) Multiple description discrete cosine transform-based image coding using DC coefficient relocation and AC coefficient interpolation *(Invited Paper)*

Nafees Mansoor, A. K. M. M. Islam, Univ. Teknologi Malaysia (Malaysia); M. A. Razzak, Independent Univ., Bangladesh (Bangladesh)

The development of an effective and dependable multiple description coding scheme over a lossy communication system is described. A general framework of an effective multiple description robust communication system with two-channel and four-channel cases is presented with a proposed block-based DC coefficient relocation and AC coefficient interpolation approach. A benefit of such system is that, if all the channels work properly, a very good quality, probably lossless, reconstruction can be obtained from the received descriptions. On the other hand, if some of the channels do not work properly, a lower but still quite satisfactory quality of reconstruction can be obtained. The performance of the proposed scheme is mea- sured with mean squared error, peak signal-to-noise ratio, entropy, and bit rate to analyze the reconstruction quality, and the computed results are matched with that of the other schemes. The system complexity is also computed and compared. Comparing the simulation results, it is observed that the proposed scheme, which uses a DC coefficient relocation and AC coefficient interpolation scheme, gives comprehensive enhancements over the other recently developed schemes.

9019-6, Session 2

Edge preserving motion estimation with occlusions correction for assisted 2D to 3D conversion

Petr Pohl, Michael Sirotenko, Victor Bucha, Ekaterina Tolstaya, Samsung R&D Institute Rus (Russian Federation)

CONTEXT: In this article we propose high quality motion estimation based on variational optical flow formulation with non-local regularization term and RANSAC motion clustering for occlusion motion correction.

OBJECTIVE: Our motion estimation method is intended as a part of 2D to 3D conversion tool for background spatio-temporal inpainting, automatic adaptive key frame detection and key points tracking. The goal of background spatio-temporal inpainting is to restore background image behind objects using video data where background can be seen during processed video sequence and spatial inpainting methods for remaining parts. Here motion estimation and background motion approximation allow to connect video data from different frames. The goal of adaptive key frame detection is to automatically estimate temporal complexity of processed scene and suggest key frames for human interaction further apart in simple cases and more densly in case of complex changes of the scene. Analysis of dense motion estimation result gives very good information about temporally local complexity of the scene. Both these applications require sharp edges of motion around moving objects.

METHOD: We adapted efficient primal-dual optimization algorithm proposed by Chambolle and Pock [1], which is suitable for GPU implementation. The main drawbacks of basic total variation optical flow algorithm with L1 norm (TV-¬L1) are incorrect smoothing around motion edges and unpredictable behavior in occlusion areas. We extended the algorithm to use color information and replaced TV-L1 regularization by local neighborhood weighting known as non-local smoothness term proposed by Werlberger, Pock and Bischof [2] or Sun, Roth and Black [3]. We tested two practically useful updates of optical flow estimation. First one was a smoothness coefficient variability during pyramid computation of optical flow. The idea is to allow less smooth motion

result on coarse levels of pyramid not to miss larger motion of important parts of frames because of smoothing. The second update was a non-local smoothness term heuristic, that allows to half the computation time with acceptable decrease of quality. Our occlusion correction method is a post-processing step, after computing optical flow motion vectors for every frame in both directions. Occlusion correction comprises from these steps: occlusion detection, joint clustering of forward and backward motion, clusters inpainting and occlusion motion inpainting. The pixels in forward (resp. backward) occlusion areas are not seen on next (resp. previous) frame, but their consistency with motion cluster can usually be determined from previous (resp. next) frame. So we used 3 frames scheme for joint forward-backward occlusion aware motion clustering. The clustering is done by slightly adapted RANSAC algorithm first proposed by Fischler, Bolles [5] with similarity motion model. We choose similarity model (shift, scaling and rotation), because the results were more stable than results achieved with more general affine or homography transform motion model. The main idea of our occlusion correction is to use motion of a consistent motion cluster and inpaint it in occlusion area. Areas considered occlusions in both directions were inpainted by local image similarity cross-bilateral cluster in-painting with Gaussian weights.

RESULTS: Our motion estimation method is placed around 20th place on Middlebury optical flow dataset, but only one other method reports better processing time on Urban sequence. We also tested the proposed motion estimation on a wide set of TV-quality (960x540) and Full-HD (1920x1080) videos. The motion estimation results were usually stable and reasonably reliable. Computation times of CUDA GPU implementation were around 2 seconds per frame (960x540). Main problems were coming from larger motion, rapid change of light conditions and bad conditioned results because of aperture problem. On well textured areas without motion edges the achieved accuracy is about 0.2 pixel, the average error on Middlebury optical flow benchmark is dependent on sequence and ranges from 0.1 to 0.6 pixel.

9019-7, Session 2

Exemplar-based inpainting using local binary patterns

Viacheslav V. Voronin, Vladimir I. Marchuk, Nikolay V. Gapon, Roman A. Sizyakin, Don State Technical Univ. (Russian Federation); Karen O. Egiazarian, Tampere Univ. of Technology (Finland); Aleksandr I. Sherstobitov, Don State Technical Univ. (Russian Federation)

This paper focuses on a novel image reconstruction method based on a modified exemplar-based technique. The basic idea in our approach is to find example (i.e. patch) from the image using local binary patterns and replace the lost data with it. It is proposed to use a multiple criterion for searching a similar patch because in many cases the exemplar-based method produced visually poor results. The criterion for searching the best match uses several terms, such as the Euclidean metric for pixel brightness and Chi-squared histogram matching distance for local binary patterns. The use of textural geometric characteristics together with a color informational allows getting a more useful description of the patches. For patches restoring we use algorithm as in texture synthesis method proposed by Efros and Freeman instead of simple copy-and-paste procedure. Such approach allows to optimize overlap region between patches using minimum error boundary cut. We obtained good quality results using our approach compared with original exemplar-based method. Using our technic, we demonstrate the performance via several examples, showing the effectiveness in removal large objects as well as recover small regions on the test images.

9019-8, Session 2

Local feature descriptor on base indexing 2D kernel of local polynomial approximation

Aleksandr I. Sherstobitov, Vladimir I. Marchuk, Don State Technical Univ. (Russian Federation); Karen O. Egiazarian, Tampere Univ. of Technology (Finland); Viacheslav V. Voronin, Dmitry Timofeev, Don State Technical Univ. (Russian Federation)

This paper present a novel texture descriptor on base kernel of local polynomial approximation and design 2D histogram of indexes. In the first part of article present design of 2D kernel of local polynomial approximation (k-LPAp) for arbitrary degree p. The basic idea in our approach is to split initial texture image on local non overlap square blocks, transform neighbored pixels into set of vectors, convolute kernel of local polynomial approximation with each vectors from designed set and to find indexes, for which of L1 norm between central pixel of local area and vector of neighbor pixels is minimal. In results using proposed technique is design 2D descriptor texture with fist dimension by count of neighbored pixels in the local area and second dimension by count of using kernels k-LPAp. The criterion of best match is uses the correlation coefficients for compare descriptors different texture images. The accuracy of texture classification are research on the base texture images different classes: Brodatz album, Outex, KTH-TIPS2 and Columbia-Utrecht (CUReT). We obtained accuracy results using our approach compared with modern methods texture analysis and classification. Using Fourier transform of designed descriptor are allow get robust and rotate invariant properties

9019-9, Session 3

Metric performance in similar blocks search and their use in collaborative 3D filtering of grayscale images

Aleksey Rubel, Vladimir V. Lukin, National Aerospace Univ. (Ukraine); Karen O. Egiazarian, Tampere Univ. of Technology (Finland)

Similar blocks (patches) search plays an important role in image processing. However, there are many factors making this search problematic and leading to errors. Noise in images that arises due to bad acquisition conditions or other sources is one of the main factors. Performance of similar patch search might make worse dramatically if noise level is high and/or if noise is not additive, white and Gaussian. In this paper, we consider the influence of similarity metrics (distances) on search performance. We demonstrate that robustness of similarity metrics is a crucial issue for performance of similarity search. Two models of additive noise are used: AWGN and spatially correlated noise with a wide set of noise standard deviations. To investigate metric performance, five test images are used for artificially inserted group of identical blocks. Metric effectiveness evaluation is carried out for nine different metric (including several unconventional ones) in three domains (one spatial and two spectral). It is shown that conventional Euclidian metric might be not the best choice which depends upon noise properties and data processing domain. After establishing the best metrics, they are exploited within non-local image denoising, namely the BM3D filter. This filter is applied to intensity images of the database TID2008. It is demonstrated that the use of more robust metrics instead of classical ones (Euclidean) in BM3D filter allows improving similar block search and, as a result, provides better results of image denoising for the case of spatially correlated noise.

9019-11, Session 3

Tensor transform based adaptive window size estimation

Emiliano Morales, Artyom M. Grigoryan, Sos S. Agaian, The Univ. of Texas at San Antonio (United States)

The ability to define boundaries and shapes in a moving window across digital images has numerous advantages over utilizing fixed window sizes in image processing. For example, it has been shown that the use adaptive window sizing, and most notably in shape adaptive sizing, results in superior filtering in contrast to applying filters to fixed windows or over entire images. Additionally, the ability to spatially define regions and shapes is of great aid in edge detection and object/target detection missions, such as the automated cancerous region discrimination in medical images and target identification. Various adaptive window size determination methods have been proposed and are of popular interest in research. Methods to locally determine adaptive varying scales in different directions, to extract shapes and regions, include the anisotropic Local Polynomial Approximation based on the Intersection of Confidence Interval (ICI) rule. The authors of this novel approach used their LPA-ICI in conjunction with the Shape Adaptive Discreet Cosine Transform (SA-DCT) to define the shape of the transform's support in a point-wise adaptive manner to filter locally and reconstruct the image. Thanks to the shape adaptive nature of this method, reconstructed edges were preserved.

In this paper, we propose a different and novel approach to determine adaptive varying scales utilizing high energy direction image components derived from Tensor Transforms. It has been demonstrated by Grigoryan, 2011, that an NxN image can be uniquely decomposed to direction images based on the concept of tensor representation and its advanced form, the paired representation. The tensor transform has been considered for images where N is prime, power of two, and a power of odd primes. The tensor and paired representations in frequency and time domain define an image as a set of 1-D signals, called splitting signals. It has also been shown that splitting an image into 1-D signals can be a fast method of calculating a 2-D DFT. Another advantage of splitting signals that is of interest is that when each splitting signal is calculated as the sum of the image along parallel lines, defining the direction image as a component of the original image is made possible. This valuable property can be used for image reconstruction from projections, when splitting-signals or their direction images are calculated direction from projection data. In other words, the sum of directional images can reconstruct an image by principle of superposition, with negligible error. The approach proposed stems from the fact that high energy directional images, such as those in 0, 45, and 90 degrees, contain the necessary information to derive the size of varying scales in the splitting signal's direction. By noting that the varying vector sizes are indicative of the rate of increase in magnitude in a specific direction, contours generated can be used to identify specific regions in an image and capture the varying directional scale. Preliminary results show that this method can be effectively used for determining the size of directional scales for varying angles, including in the presence of noise. Examples of the application of the proposed method on different images and comparison with the traditional methods are given. To estimate the quality of the image enhancement, we generalize the known EME measure of image enhancement.

9019-12, Session 3

Generalized non local means filtering for image denoising

Sudipto Dolui, Univ. of Pennsylvania (United States); Iván Camilo Salgado Patarroyo, Oleg Michailovich, Ivan Camilo Salgado Patarroyo, Univ. of Waterloo (Canada)

Image denoising, either as an independent application or an initial stage for subsequent processing tasks, is known to be one of the most fundamental problems in image processing. As a result, substantial efforts have been extended to design effective denoising tools, which are capable of ridding digital images of background noises while preserving their essential features and content. A specific type of denoising method, known as neighbourhood filtering, estimates each pixel intensity of an original image as a function of its observed noisy counterpart and its neighbouring pixel intensities. The recently proposed neighbourhood filtering technique known as non-local means (NLM) filtering has been proven to outperform a large number of alternative denoising schemes. In its original form, however, NLM filtering is optimized for suppression of additive white Gaussian noise (AWGN) and is not suitable for heteroscedastic noises. Consequently, several efforts have been made by the research community to extend NLM filtering to a wider range of noise statistics. It appears, however, that many of these formulations fail to consistently perform favourably under varying noise statistics, as situations are frequent in which a NLM method optimized for one type of noise yields unacceptable filtering artifacts for another.

One determining factor in the design of any NLM filter is the definition of its associated similarity measure (SM), which quantifies the degree of similarity between different image neighbourhoods along with their respective centroids. In general, such a quantity is not known a priori and hence has to be estimated from the noisy image intensities. Consequently, the SM should be tuned according to the specific type of noise scenario at hand, and hence, the key to the unification of the different NLM schemes lies in the definition of a universal SM which is guaranteed to work irrespective of the noise statistics. Accordingly, the main contribution of the present work is in providing a consistent and rigorous method for deriving such an SM.

From a theoretical viewpoint, an ideal SM should possess a series of desirable properties irrespective of the noise statistics at hand. In particular, an ideal SM should be symmetric, bounded, and should attain its maximum value when the two image neighbourhoods being compared are equal. In addition, the maximum value attained by such SM should be independent of the specific intensities associated with the equal neighbourhoods under comparison. Although intuitive, these requirements are not met by a number of existing SMs found in the literature.

In this work, we provide two SM definitions which offer the above mentioned characteristics under varying noise statistics. The derivation of these definitions is based on a rigorous Bayesian approach which aims at calculating the posterior probability that any two pixels have equal underlying noiseless intensities, given their observed noisy counterparts. Formally, denoting X1 and X2 as the original image intensities and Y1 and Y2 as their corresponding noisy versions, one may express the above-mentioned posterior probability as $P(X1=X2|Y1,Y2)$. It is important to note that in the case when X1 and X2 are assumed to be continuous random variables, $P(X1=X2|Y1,Y2)=0$ for all possible values of the intensities involved, thus rendering this notion of equality inconsequential. Two possible alternatives of defining this equality is to define the random variables $U=X1-X2$ and $V=X1/X2$ and evaluating their respective conditional densities at $U=0$ and at $V=1$, respectively, conditioned on Y1 and Y2. Subsequent normalization of these quantities results in the proposed SM definitions, which we have referred to as correlative subtractive SM (CSSM) and correlative rational SM (CRSM), respectively. Both of these SMs satisfy all the favourable characteristics of an ideal SM listed above, independently of the noise statistics. It should be noted that although derived using similar philosophy, the CSSM and CRSM often result in different expressions of SMs.

In addition to the general form of the CSSM and the CRSM, the closed form expressions of these SMs for a number of important noise scenarios are provided. In particular, we consider the multiplicative Rayleigh, multiplicative Gamma, Poisson and Rician noise contamination models, whose practical relevance lies in their prevalence in the contexts of ultrasound, optical, astronomical, microscopy, and magnetic resonance imaging. None of these noise types is additive or homoscedastic, and all of them depend on the underlying noise-free intensities in a nontrivial manner. As such, one expects the original NLM filter designed under the assumption of AWGN model to offer suboptimal performance in these settings. In addition to highlighting the theoretical aspects of the proposed SMs for each noise contamination model, their explicit

expressions are compared to different SMs available in the literature for the same types of noise. Furthermore, relevant theoretical connections are made. For example, it has been shown that the proposed SM formula corresponding to Rician noise agrees with the original SM for AWGN contamination in the case of high signal to noise ratio (SNR). This observation agrees with the fact that Rician noise tends to AWGN for high levels of SNR.

Finally, the practical utility of the proposed SMs have been demonstrated through numerical experiments. It should be noted that the main contribution of the current work has been to present a unified theoretical framework for the derivation of various SMs for different types of noise and, wherever relevant, to connect the obtained results to various SMs which have been already proposed in the literature. Consequently, in many cases, exhaustive comparisons of different SMs and their associated denoising methods do not seem to be necessary, since some of such results have already been reported in earlier works. Nevertheless, to make the present work self-contained, some principal experimental results have been demonstrated. In addition, in order to support the proposed formulations as well as to prove the superiority of NLM filtering, the results have also been compared with those produced by a number of state of the art filtering techniques, which do not necessarily fall in the NLM category of filters. Along with visual comparisons, two quantitative performance metrics, viz. normalized mean square error and the structural similarity index (SSIM), have been used in order to evaluate the quality of the denoising methods. The results obtained using the proposed NLM formulation have been found to outperform the alternative techniques, both visually and in terms of the above-mentioned performance metrics.

9019-13, Session 4

Calibration of a dual-PTZ-camera system for stereo vision based on parallel particle swarm optimization method

Yau-Zen Chang, Huai-Ming Wang, Chang Gung Univ. (Taiwan); Shih-Tseng Lee M.D., Chieh-Tsai Wu M.D., Chang Gung Memorial Hospital (Taiwan); Ming-Hsi Hsu, Chang Gung Univ. (Taiwan)

This work investigates the calibration of a stereo vision system composed of two PTZ (Pan-Tilt-Zoom) cameras. As accuracy of the system highly depends not only on optical properties of the cameras, but also intrinsic and extrinsic parameters of the configuration, where accurate geometric relationships between rotation axes of the cameras are the major concern of this study.

We derived a complete geometric model of the dual-PTZ-camera system and proposed a calibration procedure for the parameters of the model. The proposed calibration method is based on Zhang's approach, together with a parallel particle swarm optimization (PSO) method, which is an improved version of the original algorithm. In the calibration procedure, an augmented checkerboard composed of eight small checkerboards was used.

The experimental system is composed of two Sony EVI-D70 PTZ cameras. By implementing the proposed calibration procedure, the root-mean-square errors (RMSE) in the horizontal and vertical direction are 0.192 mm and 0.115 mm, respectively when the target is within 1.5 m, and the RMSE of overlapped points between the small checkerboards is 1.3958 mm.

Besides, 3D reconstruction of several real objects, including a 3D painting and a human body is demonstrated. Practical implementation results show that the performance of the system is satisfactory.

9019-14, Session 4

Probabilistic person identification in TV news programs using image web database

Marco Leo, Federica Battisti, Marco Carli, Alessandro Neri, Univ. degli Studi di Roma Tre (Italy)

The automatic labelling of faces in programs of broadcast TV is still a challenging problem. The high variability in view points, the facial expressions, the general appearance and lighting conditions, as well as occlusions, rapid shot changes, and cameras motions, produce significant variations in image appearance. For many automatic applications, such as multimedia content monitoring, face recognition is a key feature to provide an acceptable level of performance in term of percentage of correct identifications.

Our investigation deals with the problem of automatic face recognition in TV broadcast programs. This is a challenging task due to the presence of fast variations in the video recording and to the amount of potential poses that a human face can assume during the TV broadcast. The target of the proposed method is to identify the presence of a specific person in a video. To this aim a coarse to fine approach is proposed. In the first phase, a guess of the possible candidates is computed by extracting face features from the video under test. The identification is performed by refining the coarse estimation by matching these features with the ones extracted by the set of web images, previously annotated.

Face recognition techniques could simplify the work of the human operator supervising semantic annotation tools based on identification of people in the video stream. The application of automatic tools for face recognition in multimedia sequences is not yet fully established and the human intervention is still needed. An example of an approach based on the human supervision is provided by the portal Amazon Mechanical Turk [1] that is a virtual place for activities where, at present, almost exclusively men are used for the realization of systems of video annotation. In that system, the crowdsourcing model is adopted for exploiting human intelligence in refining the results of video annotation tools.

As well known, the performances of face recognition techniques can degrade due to lighting conditions, changes in skin colour, or to the orientation of the subjects in the frame in the video sequence. In the scenario of monitoring multimedia contents, person identification can have applications in the generation of meta-data for indexing and fine-grained retrieval of specific scenes.

State-of-the art [2, 3] approaches the problem mainly considering face tracks; however there are many situations in which faces to be recognized are among unknown others (i.e. in the news typically there is an anchor-man in foreground and a secondary face on a screen in the background or a man speaking in a crowd).

Many methods have been proposed for Automatic Face Recognition (AFR) such as the ones based on Bayesian eigenfaces [4] or Fisherfaces [5]. These methods present a good accuracy on a small number of controlled test sets.

In [6], a probabilistic method for identifying characters in TV series or movies is described. Here, a similar model is designed for detect faces in news programs: the main task is the identification of different persons showing up in TV programs, i.e., for isolating faces repetitively appearing (anchor-men/women) with respect to the interviewed or to the people to be identified.

The proposed approach can be divided into three steps: in a first step, the newscast is segmented in scenes and, as secondary step, the face recognition on the face tracks is performed and, eventually, the speaker identification is carried out. Finally, the outputs of the recognition process are filtered for performing video annotation for person tracks.

A frontal face detector [7] can be exploited on each frame of the video and, in order to achieve a low false positive rate, a conservative threshold on detection confidence can be adopted. The use of frontal face detector can restrict the video content that can be annotated but it has to be robust to the heterogeneity of the image database acquired from the Web.

References

[1] Amazon Mechanical Turk https://www.mturk.com/mturk/welcome

[2] M. Everingham, J. Sivic, and A. Zisserman, '"Hello! My name is..." Buffy – Automatic Naming of Characters in TV Video', Procs. of the Workshop of British Machine Vision Association, 2006.

[3] J. Sivic, M. Everingham, and A. Zisserman, '"Who are you?" – Learning person specific classifiers from video', Procs. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009.

[4] B. Moghaddam, W. Wahid, and A. Pentland, 'Beyond eigenfaces - probabilistic matching for face recognition', Procs. of IEEE Conference on Automatic Face and Gesture Recognition, 1998.

[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, 'Eigenfaces vs. fisherfaces: Recognition using class specific linear projection', Procs. of IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7):711–720, July 1997.

[6] M. Tapaswi, M. Bauml, and R. Stiefelhagen, '"Knock! Knock! Who is it?" Probabilistic Person Identification in TV-Series', Procs. of IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[7] P. Viola and M. Jones, 'Rapid object detection using a boosted cascade of simple features' , Procs. of IEEE Conference on Computer Vision and Pattern Recognition, 2001.

9019-15, Session 4

Spatial - temporal features of thermal images for Carpal Tunnel Syndrome detection

Kevin Estupiñan Roldan, Marco A. Ortega P., Hernan D. Benitez Restrepo, Pontificia Univ. Javeriana, Cali (Colombia)

Carpal Tunnel Syndrome (CTS) is a median entrapment neuropathy, that causes paresthesia, pain, numbness, and other symptoms in the distribution of the median nerve due to its compression at the wrist in the carpal tunnel [1]. Physical assessment test for CTS are: Phalen's test and Tinel's test. In the first, the patient flex both wrists for a minute. This test is 80% sensitive to the CTS diagnose. In Tinel's test, the doctor taps over the median nerve to produce a sensation of tickling. This test is the most specific test for CTS diagnose with a specificity of 60%. When the above tests show positive results, the diagnosis is confirmed by electromyography. Electrodiagnostic tests analyze the electric waves of nerves and muscles 1. These tests can help detect median nerve compression in the carpal tunnel. These methods are invasive and sometimes painful. Infrared Thermography (IT) is an alternative method of examination, useful in the field of medicine for its safety, lack of pain and invasiveness, and low running costs. Previous work [2] mainly use temperature and its variation in time as features to represent IR (Infrared) images of hands in healthy and ill subjects. Spatial variation of temperature is neglected obtaining fair results in classification errors. In this work, we propose a set of spatial-temporal features extracted from IR images taken in healthy and ill patients with CTS. These features are based on the time variation of temperature spatial gradient on regions of interest in subject's hands. We validate these features by classification with linear SVC (Support Vector Classifiers) and LOO (Leave One Out) error. The results of the proposed approach show linear separability and lower validation errors when compared to features used in previous works that do not account for temperature spatial variability.

9019-16, Session 4

A speed-optimized RGB-Z capture system with improved denoising capabilities

Aleksandra Chuchvara, Atanas P. Gotchev, Tampere Univ. of Technology (Finland)

We propose an end-to-end system for 3D scene sensing which combines a conventional high-resolution RGB camera with a low-resolution Time-of-Flight (ToF) range sensor. The system comprises modules for range data de-noising, data re-projection and non-uniform to uniform up-sampling and aims at composing high-resolution 3D video output for driving auto-stereoscopic 3D displays in real-time. In our approach, the ToF sensor is set to work with short integration time with the aim to increase the capture speed and decrease the amount of motion artifacts. However, reduced integration time leads to noisy range images. We specifically address the noise reduction problem by performing a modification of the non-local means filtering in spatio-temporal domain. Time-consecutive range image frames are optimally utilized not only for efficient denoising but also for accurate non-uniform to uniform upsampling on the high-resolution RGB grid. Use is made of the reflectance signal of the ToF sensor for providing a confidence-type of feedback to the denosing and data fusion modules. We demonstrate a real-time performance of the system working in low-power regime.

9019-17, Session 5

On a mathematical characterization of tri- and tetra-chromatic metamerism

Alfredo Restrepo, Univ. de los Andes (Colombia)

Two light beams that are seen as having the same colour but that have different spectra are said to be metameric. In photography or in computer vision, "same color" is read to mean same readings of the photodetector components R, G, B or R, G, B, UV, etc.

We provide basis for the kernel of the linear transformation I^R ? R^3, where I^R is the set of spectra.

The colour of a light beam is based on the reading of several photodetectors with di?erent spectral responses and metamerism results when a set of photodetectors is unable to resolve two spectra. The spectra are then said to be metameric.

We are mainly interested in exploring the concept of metamerism in the tetrachromatic case.

Applications are in computer vision, computational photography and satellite imaginery, for example.

9019-18, Session 5

Refractory neural nets and vision

Thomas C. Fall, Kalyx Associates (United States)

A prominent example of biologic inspiration is artificial neural nets (ANN); these have been developed over the decades into useful tools and are based on the biologic understanding that nerve connections can be trained. The ANN has a set of nodes (artificial neurons) connected in an acyclic network with weighted links where the weights can be modified through a learning process. These aspects of neuron behavior have been well exploited, but the question that arises is: "are there other aspects of neural behavior that could be exploited?" to which we answer "Yes!" In this paper, we will explore an aspect of the neuron that does not seem to have been examined, namely the neuron's refractory period - the period of time after it fires when it is recovering its ability to fire. We will see that this is a mechanism that enables short term memory. Arrays of these refractory neurons can be overlaid on an image and the neurons fire stochastically depending on the brightness of the pixel underneath that neuron. This is the retinal neural array. If we then move that neuron array, the short term memory due to refractory period will distinguish where there were differences between the two views. That is, if the neuron was over a bright area and gets moved to an area similarly bright, then there would be no change in expected stochastic behavior – similarly for dark to dark. However, if the neuron starts out over a dark area, then there is a low likelihood of it having fired and so has a high likelihood of being in fire-able state. If that neuron now moves over a bright area, there is a

high likelihood of firing. Returning to the underlying biology of the vision system, there is indeed the fact that the retinal array does get moved – this mechanism is the ocular microtremor (OMT). The OMT is around 80 Hz, so there is a settling period of around 12 milliseconds between movements of the retinal array. Since the absolute refractory period is 1 millisecond, this gives the retinal array 12 rounds to stabilize over the new view before the next OMT moves the array again.

If we had two regions which touched along a boundary with different levels of brightness, then upon an OMT, we would expect to see a flash if neurons over dark pixels moved to be over light pixels. On the converse, moving from light to dark, we would expect fewer neurons to fire than would be expected at equilibrium. Because this is stochastic, we need to keep the results for a given OMT movement in its own aggregation array that is indexed according to the index of the image pixel. Left-right OMT flash snapshots would be in one aggregation array, up-down in another. The OMT will not discover boundaries that are parallel to the direction of its movement, so we need to combine results from different directions. Interestingly, these different OMT movement results can also characterize different textures. Though it is tempting to say this might be an actual portion of the biological vision information processing, this paper will concentrate only on the computational aspects. To summarize, the refractory neural net (RNN) consists of an array of refractory, stochastic neurons that stabilize over the image and are then moved by an OMT. They effectively bring the memory of the previous view, overlaying it on the new view. If a neuron moves from a dark pixel (and thus unlikely to have recently fired) to a bright one, we will likely see a firing event. The event upon movement is recorded in an aggregational array indexed by the pixel's index.

Because RNN is stochastic, it is not broken by noise. It just takes longer to get enough movement events to sufficiently amplify the signal so as to be discernible. Similarly if the signal is weak (that is, low contrast) the RNN will need to run longer to get enough amplification to discern the image.

From a strictly computational perspective, this approach allows for highly localized computation. The OMT moves cells in parallel from pixels to near neighbor pixels. The resulting firings are moved to appropriate aggregational arrays. Except for the refractory comparison to nearby neighbors to determine the OMT flash, there are no long range correlation effects. This makes it plausible that the RNN could be implemented suitably for distributed processing

This looks promising from computational perspective. But what is important is, does it have any value for developing vision information? To assess this, some simple images were built up to test the approach. Two images were built up to do a feasibility test of RNN capabilities. One was for testing boundary detection and had two regions. The left portion is shaded from black at the top to grey at the bottom whereas the right portion is shaded from grey at the top to white at the bottom. The grey at the bottom of the left has the same grey value as the top of the right. If we only used thresholding, this shade of grey would end up in the wrong segment. But locally, the boundary between the two regions has about the same difference of grey values all along the boundary. Running the RNN on that for 40 rounds of left-right OMT produced an aggregation array with significantly enhanced pixels along the border.

The other image, built on the same grid, was to test texture issues – could RNN distinguish between different textures? The left hand region was a black/white checkerboard. The right hand side had vertical rectangles of alternating black/white. Vertical OMTs of 1 pixel showed very significant activity on the left, moderate on the right. OMT of two pixels showed no increase of activity on the left, but very significant increase on the right.

In summary, this paper discusses how the biological aspect of neurons, their refractory period, and the biological aspect of the eye, the ocular microtremor, can be used to perform fairly deep analyses of visual fields. In particular, arrays of these neurons can be used to do local comparisons of image regions producing region boundaries and segmenting regions of different textures. Because the operations are local, they are eminently suitable for distributed processing techniques such as Hadoop map-reduce, making these candidates for cloud computing.

9019-19, Session 5

Statistical shape analysis for image understanding and object recognition

Peter F. Stiller, Texas A&M Univ. (United States)

In order to analyze the effects of noise on certain recognition and reconstruction algorithms, including the sensitivity of the so-called object/image equations and object/image metrics, one needs to study probability and statistics on shape spaces. Work along these lines was pioneered by Kendall and has been the subject of many papers over the last twenty years (see Mardia and Dryden). Most that work, especially the early work, focused on 2D shape for point features under similarity (rotation, translation, and scale). What remains missing from all the prior work is the projection from one dimension to another. For example, if our object feature points are subject to some uncertainty, how is this reflected in the distribution of image shapes we will get? Conversely, errors and noise in image feature point locations should translate into some distribution on the object shapes capable of producing that image distribution. There is no clear cut way to do this currently. One difficulty is that there is usually no uniform distribution on the (usually) non-compact space of all projections. One could pick a set of object feature points in 3D based on some distribution, but then it makes no sense to "pick a random projection to 2D" to see how the images are distributed. However when the shape spaces are compact there is an approach to questions of this sort. In this case it is possible to describe a uniform distribution on the set of shapes that can be produced by a given object shape. Using this fact we can transfer the given distribution on object shapes to one in image shape space that reflects the probability of a given image shape occurring. We investigate these distributions in various cases for both weak perspective and scaled orthographic projection when the object feature points are chosen from iid spherical normals. The resulting distributions in image space (say in the affine case) peak along certain Schubert varieties in the Grassmann manifold that parameterizes all image shapes.

9019-20, Session 5

A blind detection system for image contrast modification

Federica Battisti, Marco Carli, Antonino Laudani, Francesco Riganti Fulginei, Univ. degli Studi di Roma Tre (Italy)

The number of digital image collections available to users of the Internet is growing at an incredible rate. As a result, in recent years the role of images in the communication process has gained larger importance. If in the past the accurate tampering of an analog photo was not an easy task, with the advent of the digital world, thousands of easy-to-use tools are available for reshaping, color/content modification, resizing, image patching, and so on.

The aim of this work is the design of an effective system for blind evaluation of contrast modification in color images. Among others, the contrast feature has been selected since its modification can severely impact on human perception. Recent studies demonstrate that it can modify the perceived quality and the general feeling and emotion conveyed by an image.

Here we present an approach for detecting the presence of contrast modifications that does not require the original image. The blind estimation of these modifications, even if less effective and reliable than a full-reference one, is the only assessment system that can be employed in real situations in which the original data in unknown.

The proposed approach is based on the creation of a database containing images whose contrast is modified. Then, an automatic classification of contrast variation is used. The method validation is performed by means of subjective tests.

In a previous work [1], an approach based on a SVM (Support Vector

Machine) classifier, with Radial Basis Function kernel, using 200 images per class for the training procedure, and 100 images per class for the test procedure, was employed. In particular, the effectiveness of the overall system was evaluated by varying the contrast of the test images using Adobe Photoshop with values of ±35 (heavy positive/negative) and ±20 (light positive/negative). The objective was, given an image, to estimate its positive and negative adjustments in contrast.

In this contribution the same database as in [1] is used in order to be able to perform a comparison and evaluate the improvements with respect to [1]. 11 features are extracted from the histogram of the green color component of each test image: the maximum and minimum luminance values and the corresponding minimum occurrences cardinality, the number of elements corresponding to 0 and to 255 values, the mean, the variance, the skewness, the kurtosis, the entropy, and the Tenengrad value. After the features extraction procedure, a Neural System (NS) is used for classifying the 5 output contrast levels by using a Final Decision Maker (FDM). The NS is made of 5 different neural sub-systems dedicated to the 5 classes (CLs) of contrast level.

Each CLi is referred to the related class (1, 2, 3 , 4 or 5) of contrast level and consists of a neural sub-system composed by five Neural Networks (NNs) able to decrease the error probability of the single NN generalizations. With the aim of improving the performances of the whole NS, the procedure described in [2] has been utilized in the training of each NN.

In the final paper we will present the results of the contrast detector and we will analyze the performances with respect to the method in [1]. Furthermore a subjective experiment will be performed for assessing the visibility of the contrast modification.

REFERENCES

[1] P. Zontone, M. Carli, G. Boato, F.G.B. De Natale, “Impact of contrast modification on human feeling: an objective and subjective assessment”, Image Processing (ICIP), 2010 17th IEEE International Conference on, 1757-1760.

[2] A. Laudani, M. Parodi, F. Riganti Fulginei, and A. Salvini, “Automatic and Parallel Optimized Learning for Neural Networks performing MIMO Applications”, Advances in Electrical and Computer Engineering, 2013, vol 1.

9019-21, Session 5

2D-fractal-based algorithms for nanoparticles characterisation

Giuseppe Bonifazi, Silvia Serranti, Roberta Palmieri, Univ. degli Studi di Roma La Sapienza (Italy)

Fractal geometry concerns the study of non-Euclidean geometrical figures generated by a recursive sequence of mathematical operations. The proposed 2D-fractal approach was applied to characterise the image structure and texture generated by fine and ultra-fine particles when impacting on a flat surface. The work was developed with reference to particles usually produced by ultra-fine milling addressed to generate nano-particles population. In order to generate different particle populations to utilize in the study, specific milling actions have been thus performed adopting different milling actions and utilising different materials, both in terms of original size class distribution and chemical-physical attributes. The aim of the work was to develop a simple, reliable and low cost analytical set of procedures with the ability to establish correlations between particles detected by fractal characteristics and their milled-induced-properties (i.e. size class distribution, shape, surface properties, etc.). Such logic should constitute the core of a control engine addressed to realize a full monitoring of the milling process as well as to establish correlation between operative parameters, fed and resulting products characteristics.

9019-22, Session 6

Non-stationary noise estimation using dictionary learning and Gaussian mixture models

James M. Hughes, LGS Innovations Inc. (United States); Daniel N. Rockmore, Dartmouth College (United States); Yang Wang, Michigan State Univ. (United States)

Stationarity of the noise distribution is a common assumption in image processing. This assumption greatly simplifies denoising estimators and other model parameters and consequently assuming stationarity is often a matter of convenience rather than an accurate model of noise characteristics. The problematic nature of this assumption is exacerbated in real-world contexts, where noise is often highly non-stationary and can possess time- and space-varying characteristics. Regardless of model complexity, estimating the parameters of noise distributions in digital images is a difficult task, and estimates are often based on heuristic assumptions. Recently, sparse Bayesian dictionary learning methods were shown to produce accurate estimates of the level of additive white Gaussian noise in images with minimal assumptions. We show that a similar model is capable of accurately modeling certain kinds of non-stationary noise processes, allowing for space-varying noise in images to be estimated, detected, and removed. We apply this modeling concept to several types of non-stationary noise and demonstrate the model’s effectiveness on real-world problems, including denoising and segmentation of images according to noise characteristics, which has applications in image forensics.

9019-23, Session 6

Weighted denoising for phase unwrapping

Satoshi Tomioka, Shusuke Nishiyama, Hokkaido Univ. (Japan)

In order to measure an optical distance of an object, two well-known methods were developed: one is the Fourier transform method (FTM), and the other is the phase shifting method. From the view of the property of an object to be measured, the differences between them are the number of interferograms; the FTM requires only a single interferogram, while the phase shifting method requires at least three interferograms. When we wish to observe an object that changes rapidly over time, the FTM is desirable since the interferogram can be recorded by a digital camera with short exposure time before the change of the object does not appear. However, the short exposure time causes decrease of the signal intensity. In this case, the noise by the camera results in a significant error. If the noise level is low, the optical distance that is obtained by two steps: the wrapped phase bounded in a certain range is evaluated by the FTM, and the unwrapped phase that is a phase in a continuous range is evaluated by phase unwrapping algorithms. The noise causes singular points at which the unwrapped phase depends on paths to unwrap and the unwrapped phase is not unique. There are several phase unwrapping algorithms that can remove the noise by regularization of the wrapped phase; e.g. the least-square method with a discrete cosine transform, the singularity spreading phase unwrapping and the phase unwrapping using localized compensator. However, the noise level is too high, the denoising by the phase unwrapping algorithms cannot remove the noise sufficiently. In such a case, the denoising process is required before the phase unwrapping.

Since a source of the noise by the camera is the thermal noise, the noise has a uniform distribution. In contrast, the signal of interferogram depends on a profile of incident wave of an interferometer, which has a spatial distribution such as a Gaussian beam. As a result, the signal to noise ratio has the spatial distribution; however, in the past denoising algorithms the distribution of the signal intensity was not considered. The phase at pixels with sufficiently large signal intensity is reliable. This means that the denoising for these points should not be applied so much. In this paper, the weighted denoising method in which the weight

depends on the intensity is shown.

Furthermore, two cost function is demonstrated in this paper; one is a complex-valued cost function that was proposed in past, and the other is a real-valued cost function. The basic idea of the denoising is averaging of the wrapped phase. Since the wrapped phase has a discontinuity, a direct averaging of the wrapped phase includes error. To solve this problem, the complex-valued representation was proposed, in which the phase is represented as an argument of a complex exponential function. We also employ the complex-valued representation. In one of the past methods, the complex-valued cost function is defined as summation of two terms; one is a square difference between an estimation of the complex representation and that of the neighboring points, the other is a square difference of the estimation of the complex representation and that of the observed phase. The square was not a square of the absolute value in the complex-valued cost function. The stationary point of the complex-valued cost function satisfies the equation that the partial derivative is zero. In the past method, the Newton method that requires the second order partial derivative were applied. In this paper, the solution can be obtained by a simple iterative process. Moreover, it is proved that the convergence of the process to find the stationary point is ensured. However, the stationary point of the complex-valued cost function cannot be ensured the condition that both the real part and the imaginary part are minimized at once. In contrast, in the real-valued cost function, the squares in the complex-valued cost function are replaced by squares of the absolute values of the complex differences. It can be ensured that the stationary point is the minimal point. However, this method has a drawback. It cannot be proved whether the equation to minimize the real-valued cost function converges at a unique solution any time.

In order to demonstrate the applicability of the two cost functions and the validity of the weighed denoising, numerical simulations are examined. Since the wrapped phase that is input of the denoising algorithm is prepared from the unwrapped phase is known, the error of the denoising algorithm can be evaluated after the phase unwrapping process by comparing with the original known unwrapped phase. The known unwrapped phase is generated by a simulation of the interferometry system; i.e., the distributions of an incident wave profile, the phase modulation by a object, the spatial carrier by a FTM measurement, and the noise by a digital camera are simulated.

After the numerical simulations, the following conclusions are demonstrated. The weighted cost function is valid for effective denoising. The choice of weights between the two terms in the cost functions that are the term for averaging operation and the term for reliability of the observed phase is still important. They are a trade-off between denoising and avoiding the increase of local error. There are no significant differences between the types of two cost functions in the numerical simulation, although it is not a sufficient condition that the complex-valued cost function is always applicable.

9019-24, Session 6

A sliding-window transform-domain technique for denoising of DSPI phase maps

Asen Shulev, Bulgarian Academy of Sciences (Bulgaria); Atanas P. Gotchev, Tampere Univ. of Technology (Finland)

We have developed a technique for denoising speckle pattern fringes, which makes use of an overcomplete expansion in transform domain combined with suitable thresholding of transform-domain coefficients related with the speckle size. In this paper, we modify the technique to work on noisy phase maps obtained by Digital Speckle Pattern Interferometry (DSPI). The modified version utilizes a complex-valued representation for the phase maps and consequently employs Discrete Fourier transform in sliding window mode for obtaining the sought overcomplete expansion. We discuss issues related of the window size and local threshold value selection. We compare the two approaches by simulating the processes of speckle pattern fringe and phase map formation. Furthermore, we demonstrate the performance of both techniques for denoising of real phase maps, obtained through a phase shifting DSPI in an out-of-plane sensitive set-up.

9019-25, Session 6

A fast method of optimal directions dictionary learning algorithm for sparse decomposition with its application to image denoising

Hossein Rabbani, Isfahan Univ. of Medical Sciences (Iran, Islamic Republic of)

In this paper, we address the problem of dictionary learning for sparse decomposition. In other words, we seek to form basis that sparsely represent the signals. One of the methods proposed for this aim is method of optimal directions (MOD). In this article, we propose a variation of this method that is much faster in which sparse representations and dictionary atoms are simultaneously obtained. We use approximate sparse solution in each step of sparse coding and iteratively close approximate sparse solution to exact sparse solution. We experimentally show that our algorithm have the same characteristics with its counterpart. The algorithm is also applied for the purpose of image denoising which leads to a fast dictionary learning image denoising algorithm with favorable results.

9019-29, Session PWed

Shape-dependent image processing tools for improved and automated response of adhering medulloblastoma cancer cell edges in a unique imaging system

Frederick C. Weissbach, Brent M. Nowak, Sos S. Agaian, The Univ. of Texas at San Antonio (United States)

Researchers at the University of Texas at San Antonio Health Science Center (UTSAHSC) have developed a unique device for imaging the adhesion process of a metastasizing medulloblastoma cell on the leptomenginges in the cerebral spinal fluid. The device is part of an emph{in vitro} system used to observe live-cell reactions with a fluid flow. Large quantities of video data consisting of the adhesion of fluorescing medulloblastoma exists to be processed with numerous concerns. First, properly detecting the boundaries of adhered cells with emphasis on locality, uniformity, and noise resistance will improve upon further biological research and help provide quantifiable results. Second, removing the human factor will provide both autonomy for a system to help with large amounts of data as well as preventing errors from human estimation in a normally difficult environment for specificity. This was achieved by modifying the popular Canny edge detection algorithm. We exchanged the Gaussian smoothing function with the bilateral filter, included the Otsu threshold determination method, and customized shape-dependent gradient operators for enhanced edge response. The process was applied to real adhering cancer cell images taken from videos provided by the UTSAHSC team and compared to traditional Canny methods.

9019-30, Session PWed

Fibonacci thresholding: signal representation and morphological filters

Artyom M. Grigoryan, Sos S. Agaian, The Univ. of Texas at San Antonio (United States)

One of the simplest operations used in several standard image segmentation techniques, which include the global thresholding,

semi-thresholding, multilevel thresholding and variable thresholding, is the thresholding. This operation is widely used in signal and image processing, especially in nonlinear filtering. The cross sections or threshold sets describing one- and multi-dimensional real-valued functions, signals and images, play very important role in nonlinear filtering and mathematical morphology, as a powerful set-theoretical method of analysis and synthesis various nonlinear operations, filters, and systems of signal and image processing. This notion allows for transferring the basic operations of set algebra, such as Minkowski's addition and subtraction of sets, and the dilation and erosion of sets on the function algebra. The cross-sections are typically are horizontal. A more general concept of ga-cross sections by arbitrary homothetic curves ga that generalizes the traditional horizontal cross sections were proposed by Grigoryan (SPIE 2003). On the base of these kinds of cross sections, the corresponding set representations in the form of umbrae, as well as the function processing transformations such as dilation and erosion, the median and order statistic filters were given.

In this paper, we consider the traditional horizontal cross sections and describe a general concept of weighted thresholding, which can be used for a new set-theoretical representation of signals and images. This representation also can be used for obtaining new signals which contain a large number of key features of the original signal, as well as for the design of new morphological filters. The property and applications of the weighted thresholding are described and the decompositions of signals and images, which are based on such thresholding with different types of level sets, are analyzed. The weighted thresholding-based representation maps many operations of non binary signal and image processing to the union of the simple operations over the binary signals and images. We consider one of the most interesting cases, when the level set of thresholding is defined by the Fibonacci series. The weighted thresholding can be described by direct formulas of calculation. The implementation of the gray scale transforms for the weighted thresholding is simple and the look in table method can be used in calculations. The examples of the Fibonacci thresholding are given and advantages of using such thresholding in comparison with the standard morphological methods of segmentation are given. The implementation of the Fibonacci levels for fast calculation of the median filters is also described.

The preliminary experimental results and applications of the weighted thresholding concept, as a new tool for signal and image processing, are described. These experimental results show that the Fibonacci thresholding is much promised and can be used for many applications in signal and image processing. As an example, the advantages of the Fibonacci thresholding in the segmentation (for enhancement and edge detection) are described and compared with the standard morphological methods of segmentation. The main advantage of such representation is in an opportunity to implement new nonlinear operations by weighted threshold operation and enhance a large number of geometrical features that are presented in the original signals and images, manipulating with weighted coefficients. The weighted thresholding is invariant under the morphological transforms, including the basic ones, erosion and dilation, opening and closing. On the base of the weighted thresholding, we can develop new algorithms for efficient calculation and representation of nonlinear filters (for instance the median type filters), for image enhancement and segmentation.

(see for more detail the attached file)

9019-31, Session PWed

Novel image darkness and brightness measures

Sos S. Agaian, Mehdi Roopaei, Wuxia Chen, The Univ. of Texas at San Antonio (United States)

Novel discriminant measure for image darkness and brightness is introduced in the current article. The new measure is applicable on the gray scale and color images. In the proposed measure, intensity distribution of pixels is calculated for a specific span and integrated by the whole range intensity. The best intensity interval is achieved by implementing statistical criteria. To show the effectiveness of the represented measure, darkness and brightness of different image categories like infrared and thermal images are examined.

9019-32, Session PWed

Image de-noising through symmetric, bell-shaped, and centered weighted median filters based sub-band decomposition

Sos S. Agaian, Sirajul Salekin, The Univ. of Texas at San Antonio (United States)

This paper presents a new method for removing combination of Gaussian, Speckle and Salt and Pepper type noise from an image by using several nonlinear technique based on the median transform. The benefit of sub-band decomposition using median transform over the wavelet decomposition method is that the nonlinear filters are not subject to Gibb's phenomenon which causes the ringing effects associated with the linear subband methods and they can be computed with low computational complexity. It has been experimentally observed that noisy coefficients have a higher value at the first scale of multiresolution analysis and later the value decreases in the subsequent scale and so on. So, we have proposed to apply subsequently decreasing threshold on every single step of the multiresolution coefficients so that the de-noising method filters out the noise while preserving good image quality. The two-band filter bank shown Figure 1 is the basis of the image denoising procedure we used to evaluate the merits of various weighted median filters. The 2-dimensional extension to this basic filter bank is shown in Figure 2. Here LL, LH, HL and HH are the resulting four subbands after one transform stage. For coding efficiency, LL is further transformed by yet another stage. Figure 3 shows the multiresolution based on median transform that consists of a series of smoothing of the input image, with successively broader kernels. Each successive smoothing provides a new resolution scale. The multiresolution coefficient values constructed by differencing images at successive resolution scales are not necessarily zero mean, and so the potential artifact-creation difficulties related to this aspect of wavelet transforms do not arise. A number of noisy images contaminated with different combinations of the Gaussian, Speckle and Salt and pepper noises are denoised by the approach proposed in figure 3 and compared using SNR measure with different wavelet denoising algorithms. The experimental results validate that the proposed algorithm outperforms the traditional wavelet decomposition method for noise removal. More of the denoising tests will be carried out using the weighted median filter based subband decomposition given in figure 1 and 2. Various support regions will be explored to determine the optimum structure for use in image denoising and outcome will be combined here before final presentation.

9019-33, Session PWed

Parametric rational unsharp masking for image enhancement

Changzhe Yin, Yicong Zhou, Univ. of Macau (Macao, China); Sos S. Agaian, The Univ. of Texas at San Antonio (United States); C. L. Philip Chen, Univ. of Macau (Macao, China)

Digital images have great impacts in most important fields such as medical, military, meteorology, security, and entertainment. The visual quality of images may not be always desired because of many factors such as out of focus, and various environment conditions (frog, water, and windstorm). And sometimes we need more details in digital images and demand high visual quality. Unsharp masking (UM) as one of simple and well-known methods of image enhancement is an effective tool to improve the visual quality of fine details in images. UM enhances an

image by subtracting a lowpass filtered image from its original, or by adding a scaled high-frequency part of the image to its original. It can enhance edges and fine details images. However, it also has side effects such as (1) high sensitivity to noise, (2) out of range problem, and (3) halo effect.

To overcome these problems, many algorithms have been developed to improve UM. One method replaced the highpass filter with the nonlinear Teager filtering and applies one mapping function to reduce the noise sensitivity. Cubic unsharp masking uses a quadratic function of local gradients as a modulation component. Intensity-based gain adaptive UM focuses on the gain factor instead of the filter. It adopts a Gaussian kernel to the gain to overcome over-range problem. Generalized unsharp masking algorithm makes use of the log-ratio approach to avoid the problems about out-of-range and halo effect. Rational unsharp masking (RUM) is a powerful UM technique using a control term to limit noise amplification and halo effect. The control term works as the thresholding approach. Through adjusting parameter g_0, we can choose which region of gradient to be amplified the most. Although RUM performs better than the classical UM in suppressing noise and halo effect, the improvement is limited and out of range is still a problem to RUM.

In this paper, we proposed a new parametric rational unsharp masking scheme (PRUM). The PRUM contains two parts: detail extraction using a nonlinear filter, and the nonlinear mapping. Detail extraction uses two gain parameters to extract the horizontal and vertical details independently. The nonlinear mapping is designed to reduce the background noise which usually has small values of high frequency components while true details are large. After the mapping process, small values in the image will be eliminated and rest values decrease linearly at the same time to avoid the steep change situation. Finally, a log-like addation is adopted in order to limit the output in a porper range.

The results show the PRUM's excellent enhancement performance.

9019-34, Session PWed

Sparse presentation-based classification with position-weighted block dictionary

Jun He, Tian Zuo, Bo Sun, Xuewen Wu, Lejun Yu, Fengxiang Ge, Beijing Normal Univ. (China); Chao Chen, Naval Academy of Armament (China)

This paper is aiming at applying sparse representation based classification (SRC) on general objects of a certain scale. Authors analyze the characteristics of general object recognition and propose a position-weighted block dictionary (PWBD) based on sparse presentation and design a framework of SRC with it (PWBD-SRC). Principle and implementation of PWBD-SRC have been introduced in the article, and experiments on car models have been given in the article. From experimental results, it can be seen that with position-weighted block dictionary (PWBD) not only the dictionary scale can be effectively reduced, but also roles of image blocks taking in representing a whole image can be embodied to a certain extent. In reorganization application, an image only containing partial objects can be identified with PWBD-SRC. Besides, rotation and perspective robustness can be achieved. Finally, a brief description on some remaining problems has been proposed in the article.

9019-26, Session 7

Alternating direction optimization for image segmentation using hidden Markov measure field models *(Invited Paper)*

José M. Bioucas-Dias, Filipe Condessa, Univ. Técnica de Lisboa (Portugal); Jelena Kovacevic, Carnegie Mellon Univ. (United States)

Image segmentation is fundamentally a discrete problem. It consists in finding a set of image regions, that is, defining a partition of the image domain, such that the pixels in each region exhibit some kind of similarity. Image segmentation is very often formulated as the minimization of an energy containing a data misfit term (the negative logarithm of the likelihood in Bayesian terms) built on image features informative to the segmentation goal and a regularization term (the negative logarithm of the prior in Bayesian terms) promoting some form of spacial regularity, such as piecewise smoothness on the segmentation result.

The above formulation ends up being an integer optimization problem that, apart from a few exceptions, is NP-hard and thus impossible to solve exactly. This roadblock has stimulated active research aimed at computing "good" approximations to the solutions of those integer optimization problems. Relevant lines of attackhave focused on the representation of the regions in terms of functions, instead of subsets, and on convex relaxations which can be solved in polynomial time.

In this paper, inspired by the "hidden Markov measure fields, introduced by Marroquin et al. in 2003, we sidestep the discrete nature of image segmentation by formulating the problem in the Bayesian framework and introducing a hidden set of continuous random fields determining the probability of a given segmentation. Armed with this model, the original discrete optimization is converted into a convex program, provided that the negative logarithm of the prior for the hidden fields is convex. The hidden fields are inferred efficiently with the Constrained Split Augmented Lagrangian Shrinkage Algorithm (SALSA).

In addition to the supervised scenario, in which the model parameters are assumed to be known, we also introduce an expectation maximization (EM) algorithm to be applied to semi-supervised and unsupervised scenarios, implementing an M-step similar to the algorithm designed for the supervised case. The effectiveness of the proposed methodology is illustrated with simulated and real hyperspectral and medical images.

9019-27, Session 8

Multispectral imaging and image processing *(Invited Paper)*

Julie Klein, RWTH Aachen (Germany)

No Abstract Available

9019-28, Session 8

On the performance of multi-rate filter banks *(Invited Paper)*

Robert Bregovic, Atanas P. Gotchev, Tampere Univ. of Technology (Finland)

No Abstract Available

Conference 9020: Computational Imaging XII

Wednesday - Thursday 5 –6 February 2014

Part of Proceedings of SPIE Vol. 9020 Computational Imaging XII

9020-31, Session PWed

Medical MR image compressed sensing reconstruction based on wavelet and l_1 norm optimization

Xiaoming Huang, Beijing Jiaotong Univ. (China); Ivan Jambor, Harri Merisaari, Marko Pesola, Chunlei Han, H. J. Aronen, Univ. of Turku (Finland); Gangrong Qu, Beijing Jiaotong Univ. (China)

Purpose:

Compressed sensing (CS) has many applications in medical image processing as well as other areas of signal processing. This study was designated to analyze the procedure of using wavelet and l_1 norm based CS method in medical MR image reconstruction. Moreover, we have proposed a method of using prior known structure information in wavelet transform coefficients to improve the speed of CS MR image reconstruction.

Methods:

Since MR data acquisition is performed in a transform domain (k-space) instead of the image domain, the application of CS is particularly useful. On the other hand, most of the medical images usually do not possess truly sparse representations in any domain as most of the transform coefficients are of non-zero values. Fortunately, many images are compressible in the sense that information is strongly captured by a small number of these coefficients. Parallel imaging could be combined with CS to further increase the data under-sampling rate resulting into increased data acquisition and shorter imaging times. CS has three major components: (1) sparse representation in transform domain (k-space); (2) incoherent measurements (random sampling), and (3) no-linear CS reconstruction.

We considered the following sparse reconstruction algorithms in CS:

min_omiga||Psi(x)||_1 s.t. ||Phi(x)-Phi(y_n)||_1 < epsilon

Where x is the image to reconstruct, y_n = y + n is the under-sampled k-space data, y is the real data, n is a noise process such as Additive White Gaussian Noise (AWGN) found on each of the quadrature measurement channels in MRI, Psi is the sparse transform (e.g. wavelet or discrete cosine transform), Phi is the multi-coil acquisition operator (under-sampled Fourier transform + coil sensitivities), epsilon is a statistic describing the magnitude of the error, omiga is the domain of the image.

We used the MR image BRAIN as the original data to complete the whole CS reconstruction process. The images were reconstructed with the above model solved by the iteration with the same stopping criterion. Since the wavelet coefficients of real MR images often tend to have a quad-tree structure to some extent, this property of wavelet transform can be combined into MR image reconstruction by CS methods resulting into faster MR image reconstruction.

Results:

Although CS has proved to be a very effective method for MR image reconstruction in our study, combination of other useful information with CS method can further improve the results.

Conclusion:

Combination of CS with wavelet transform can effectively be used for the MR image reconstruction. The use of some prior known sparse structure produces improved reconstruction speed, which is of utmost importance in medical MR clinical practice. Prior known sparse structure pattern has a potential to be useful in other specific MR applications.

9020-32, Session PWed

Texture mapping 3D models of indoor environments with noisy camera poses

Peter Cheng, Michael Anderson, Stewart He, Avideh Zakhor, Univ. of California, Berkeley (United States)

Automated 3D modeling of building interiors is useful in applications such as virtual reality and environment mapping. Applying textures to these models results in useful photorealistic visualizations of indoor environments. To generate such textures, modeling systems must capture images of their surrounding environment and compute a camera pose for each image in order to project the image onto surfaces reconstructed from captured point clouds. The accuracy of these camera poses is innately dependent on the ability of the modeling system to estimate its location and orientation over the period of its data collection, a process typically referred to as “localization.” With accurate localization and camera pose information, neighboring images projected onto a surface for texturing reveal minimal visual artifacts at their borders, barring the existence of occlusion or parallax effects.

In this paper, we capture 3D modeling data using a backpack system, transported by an ambulating human operator. This platform offers unique advantages over commonly-used wheeled systems in terms of agility and portability, but generally produces noisier localization and camera poses, as it exhibits 6 degrees of freedom of motion, instead of 3 in wheeled systems. As a result, projecting captured images onto surfaces results in highly visible misalignments. Existing approaches to texture mapping adjust camera poses and image projections in an attempt to reduce these misalignments. As shown in this paper, such approaches are not always resilient against error accumulation, and often struggle when images have varying resolution, or contain featureless or repeating textures.

We propose a method for generating accurate and seamless textures on all manner of surfaces, despite the presence of erroneous camera poses. This method begins by partitioning the recovered environment geometry into a set of approximately planar regions, each to be textured separately. Depending on the orientations of the cameras containing imagery for a region, one of two different approaches is used to generate a texture for that region.

For both approaches, we perform pose refinements, in order to better align images to geometry as well as to each other. First, straight lines are detected in images, and transformations are calculated to maximize their alignment to straight lines in the environment geometry. Second, SIFT matches between all pairs of overlapping images are calculated, and a least squares problem containing all SIFT matches as well as geometry alignment transforms is solved in order to optimize image locations. Our first approach for texturing handles regions where camera poses are at arbitrary angles to surfaces, and as a result, their resolution can vary spatially. For this approach we break up the region into numerous tiles, and select the best image to texture each tile, accounting for camera angle and distance as well as selections made by neighboring tiles. Our second approach is tailored towards situations where source images are taken at headings mostly perpendicular to large, near-planar areas to be textured. In such a situation, we find that image projections are spatially uniform, and using entire images reduces visible seams without sacrificing texture resolution. We then select images for texturing by solving a shortest-path problem which qualitatively reduces discontinuities in the final texture. For our particular backpack system, we apply the former approach to poorly-imaged surfaces, such as floors, ceilings, and small environmental features, while the latter approach is used for long, planar walls.

We demonstrate the effectiveness of the two approaches on a variety of 3D models generated from point clouds captured by the backpack system.

9020-33, Session PWed

Spatial adaptive blending method for robust ultrasound image deconvolution

Sung-chan Park, Jooyoung Kang, Yun-Tae Kim, Kyuhong Kim, Jung-Ho Kim, Jong Keun Song, Samsung Advanced Institute of Technology (Korea, Republic of)

1. Background, Motivation, and Objective After receiving ultrasound signal with imaging device, we obtain the low resolution image due to the device and tissue's spatial response. We can denote this response as the point spread function (PSF). If you know the PSF previously and assume that the input image has the convolution model between the PSF and restored image, we can obtain the high resolution image by the deconvolution process. The recent papers have reported that the classification errors of the tissue characterization are reduced with the help of the deconvolution methods. But, when the ultrasound wave propagates the human body, its velocity factors change, which make the PSF shape different at each region. To solve the PSF estimation problem, many algorithms have been developed to find its magnitude and phase information. Some algorithms simply eliminate the phase component and estimate the magnitude with the low pass filter on the cepstrum domain. Other algorithms use the phase unwrapping methods which need high computational complexities. They are rarely error free because of the ill-posed problem and produce the PSF estimation errors and make the image overblurred by the blur artifacts. For the commercialization of the ultrasound deconvolution method, the robustness of the image deconvolution without artifacts is essential.

2. Methods To solve this problem, we present a new spatial varying adaptive blending scheme based on the best linear unbiased estimator and derive the equation based on Gauss-Markov model. With a stochastic image blending of the deconvolution images, we obtain high resolution results which suppress the blurring artifacts enough although the input deconvolution images have restoration errors. We analyze the deconvolution artifacts based on the ultrasound velocity. We build the two types of deconvolution images which are used for the blending. One is the PSF which has symmetric shape. The other is the PSF which has non-symmetric shape. We observed that the PSFs' shape has variations under this range. After brightness alignment between the images, we blend them with the optimal weight of the best linear unbiased estimator. In this case, we measure its noise statistics on the sample space and compute the covariance matrix. Although there exist cases which have the poor PSF estimation, we can suppress the artifacts efficiently by our blending scheme.

3. Results and Discussion We verify our algorithm on the real and simulation data. In all the cases, we can observe that the artifacts which are produced by the PSF estimation error are suppressed and shows the highest resolution among the input candidate images. We compare it with the other deconvolution algorithms and show only ours can suppress the artifacts. As we add the more image candidates, we can obtain the higher accuracy of image restoration. But, it pushes to increase the computational complexity.

9020-34, Session PWed

Reconstruction of compressively sampled ray space by using DCT basis and statistically-weighted L1 norm optimization

Qiang Yao, Keita Takahashi, Toshiaki Fujii, Nagoya Univ. (Japan)

In recent years, ray space (or light field in other literatures) photography has gained a great popularity in the area of computer vision and mathematical theory. Especially, the free viewpoint image generation based on ray space and the application of light field camera have convinced people that the study of ray space holds a great significance in areas of both academic research and practical engineering.

Generally speaking, ray space can represent all the information of a 3-D virtual space and it can be constructed from images that are captured by a camera array horizontally arranged in front of the objects. However, according to the plenoptic sampling theory, the composition of a ray space requires a large number of images, which poses a great challenge to data acquisition, storage and transmission.

Several researchers have investigated compressed sensing methods for ray space or light filed data acquisition. Several hardware designs for conducting compressively sampling have also been proposed, and L0-norm or L1-norm optimization schemes were employed for the reconstruction of complete data from the incomplete samples. However, there is still an open question that how to fully exploit the unique structure of ray space or light field data in compressed sensing framework. One of the most successful approaches is based on dictionary learning, however, the main drawback of this method is that dictionary learning method loses statistical property of sparse representation which can be used to direct data reconstruction and the performance heavily depends on similarity between the dictionary and the data that we want to acquire.

In this paper, we propose to exploit the unique structure of ray space data in DCT domain and to integrate this structure into L1-norm based optimization. A statistical weighted matrix based on the unique structure of ray space data is proposed to promote sparsity of solution and we name it statistically-weighted L1 norm optimization. The support of sparse solution can be identified more precisely and the quality of reconstructed ray space can be enhanced greatly.

The key point of this paper is to integrate the structure of ray space data into the process of reconstruction. Actually, L1 optimization only minimizes the sum of absolute value of each non-zero element in sparse solution. Therefore, the elements with smaller amplitude have higher probability to be chosen. According to the special structure of ray space data in DCT domain, the high frequency components are always getting smaller amplitude while low frequency components tend to have larger amplitude, thus high frequency components are more likely to be chosen as the elements in the sparse solution. However, this is totally conflict with the fact that low frequency components of ray space data are much more significant than high frequency components in sparse representation.

Therefore, corresponding weights are designed to encourage low frequency components and to constrain high frequency components in optimization procedure. It is natural to adopt the inverse value of each non-zero element as the weight, however, the value is only available after the optimization problem is solved. Thus, it becomes a "chicken-egg" dilemma.

In order to break this dilemma, we resort to an estimation strategy. Actually, in compressed sensing of ray space, we find that original data and the measurements share the same sparse representation. Since the unique structure of ray space data in DCT domain can be obtained, this structure can be integrated to design the corresponding weighed matrix in the reconstruction of ray space from incomplete measurements. In addition, the weighted matrix is a square matrix with only non-zero elements along main diagonal and it is possible to reuse the available methods for solving the structure-based L1 optimization, such as linear programming.

In our preliminary experiment, the proposed method could identify much more precise support of sparse solution and achieve much better quality than previous plain-L1 optimization. Especially in extremely low sensing ratio, about 10 percent, the proposed method could still recover the low frequency support which is of more significance for ray space data representation in frequency domain. In addition, since proper structure is exploited and integrated in optimization, the convergent speed of optimization is also accelerated so that the proposed method also reduces the running time of reconstruction.

9020-35, Session PWed

Image matching in Bayer raw domain to de-noise low-light still images, optimized for real-time implementation

Ilya V. Romanenko, Apical (United Kingdom); Eran Edirisinghe, Loughborough Univ. (United Kingdom)

Images temporal accumulation is a well-known approach to improve signal to noise ratios of still images taken in a low light conditions. However the complexity of known algorithms often lead to high hardware resource usage, memory bandwidth and increased computational complexity, making their practical use impossible. In our research we attempt to solve this problem with an implementation of a practical spatial-temporal de-noising algorithm. Image matching and spatial-temporal filtering was performed in Bayer RAW data space, which allowed us to benefit from predictable sensor noise characteristics. Proposed algorithm accurately compensates for global and local motion and efficiently removes different kinds of noise in noisy images taken in low light conditions. In our algorithm we were able to perform global and local motion compensation in Bayer RAW data space, while preserving the resolution and effectively improving signal to noise ratios of moving objects. Proposed algorithm is suitable for implementation in commercial grade FPGA's and capable of processing 12MP images at capturing rate (10 frames per second).

The main challenge for still images matching is the compromise between the quality of the motion prediction and the complexity of the algorithm and required memory bandwidth. Still images taken in a burst sequence must be aligned to compensate for objects movements in a scene and for camera shake. High resolutions of still images as well as significant time between successive frames produce significant displacements of the parts of an image and create additional difficulty for image matching algorithms. In photo applications it is very important that the noise is efficiently removed in both static backgrounds and moving objects and the resolution of the image is maintained. In our proposed algorithm we solved the issue of matching current image with accumulated image data in Bayer RAW data space in order to efficiently perform spatio-temporal noise reduction. In our approach we propose to decompose the image using Laplacian pyramid and perform motion estimation and compensation by calculating optical flow on low resolution image scales, while images matching on a pixel and sub-pixel level is performed by the spatio-temporal block matching algorithm which we proposed in our previous research for video de-noising applications. The final stage of image de-noising is the images accumulation in Bayer RAW domain. The images accumulation is based on a Gaussian background modeling method, which allowed us further reduce memory bandwidth requirements. The implementation of a background modeling in Bayer RAW data space allowed us to evaluate temporal difference against predicted noise levels. The advantage of this approach is the ability to build a noise reduction algorithm with a very reliable motion prediction and a good amount of signal to noise improvement for static and moving parts of an image, while keeping the resource usage and memory bandwidth low. Performing noise reduction in Bayer RAW data space giving us an important advantage of greatly improved signal to noise ratios for the other blocks in image processing pipeline, reside after spatio-temporal noise reduction, such as: dynamic range compression, de-mosaic, sharpening, color correction etc.

Taking into account the achievable improvement in noise levels(on the level of the best known noise reduction techniques) and low algorithmic complexity, enabling its practical use in commercial applications, the results of our research can be very valuable.

9020-36, Session PWed

Real-time focal stack compositing for handheld mobile cameras

Mashhour Solh, Texas Instruments Inc. (United States)

Extending the depth of field using a single lens camera on a mobile device can be achieved by capturing a set of images each focused at a different depth or focal stack then combine these samples of the focal stack to form a single all-in-focus image or an image refocused at a desired depth of field. Focal stack composting in real time for a handheld mobile camera has many challenges including capturing, processing power, handshaking, rolling shutter artifacts, occlusion, and lens zoom effect. In this paper, we describe these challenges and present a system for a real time focal stack compositing system for handheld mobile camera and the alignment and compositing algorithms. We will also show all-in-focus images captured and processed by a cell phone camera running on Android OS.

9020-37, Session PWed

Image deblurring using the direction dependence of camera resolution

Yukio Hirai, Hiroyasu Yoshikawa, Masayoshi Shimizu, Fujitsu Labs., Ltd. (Japan)

The blurring that occurs in the lens of a camera has a tendency to get further degraded in areas away from the on-axis of the image. In addition, the degradation of the blurred image in an off-axis area exhibits directional dependence. The blurred image can be modeled as an image, from which the point spread function (PSF) is convolved with the object.

Conventional methods have been known to use the Wiener filter or the Richardson–Lucy algorithm. These methods use the already-known PSF in the restoration process, thereby preventing an increase in the noise elements. The non-uniform degradation, which depends on the direction, is not improved even though the edges are emphasized by these conventional methods.

In this paper, we analyze the directional dependence of the resolution based on the modeling of an optical system using a blurred image. We have proposed a novel image deblurring method that employs a reverse filter obtained by optimizing the coefficient of directional dependence of the regularization term in the maximum a posterior probability (MAP) algorithm.

At first, we explain the directional dependence of the resolution. When we define the model of an optical system, the shape of the lens aperture becomes a circle in the central portion. However, it becomes an oval in the area away from the on-axis, which is induced by a phenomenon called vignetting. The oval axis can be defined as the circumferential and radial directions of the concentric circle drawn on the optical axis. We can quantify the directional dependence of the resolution by measuring the modulation transfer function (MTF) along these two directions. It is shown that the radial resolution shows a larger comparative decrease.

Subsequently, we describe our image deblurring method. We consider the angular information of these two directions to determine the regularization term of the MAP algorithm. In addition, we define the weight coefficients not only for the decrease in noise but also for the direction. The definition of the weight coefficients is important when dealing with degradation in resolution. In addition to the gain of the reverse filter, these aforementioned variables are optimized such that the difference between the MTFs obtained for the radial and circumferential directions may decrease. Finally, we can obtain a reverse filter by means of these optimization processes.

Using the proposed method, we deblurred the area of an image away from the on-axis and estimated the MTFs in the radial and circumferential directions. Therefore, as a result, we confirmed the improvement in the difference between these MTFs. We demonstrated that the proposed method is effective against non-uniform degradation, which depends on the direction.

9020-38, Session PWed

Illumination modeling and optimization for indoor video surveillance environment

Krishna Reddy Konda, Nicola Conci, Univ. degli Studi di Trento (Italy)

Introduction. The decrease in the cost of image sensors along with the increasing availability of different types of cameras, made the deployment of camera networks an attractive area of research. However, when planning a camera network, most of the attention is usually directed towards camera positioning so as to satisfy specific coverage requirements, disregarding the impact that the illumination of the scene might have on the captured video. Although planning of illumination and camera position should be carried out in parallel, in this paper we demonstrate that better coverage of the environment can be achieved by optimizing the illumination, given an infrastructured scenario for video surveillance applications. Contrast and brightness in an image or video are direct result of the illumination conditions and most computer vision algorithms are highly sensitive to this type of variations, which should fall within the dynamic range of cameras in order to be acceptable. In a real time deployment scenario it is impossible to evaluate all possible light source location and its effect on the quality of the video, because of the sheer size of the solution space. Furthermore, the solution space is highly non-linear, owing to the non-linear nature of the radiometric properties of the surfaces.

Proposed Solution. We propose to approach this problem by modelling a given scenario in a virtual domain, thus taking into account all possible parameters of the environment. Simulating the camera view in the synthetic environment will enable us to simulate various views for different light source configurations. In order to select the best configuration, we need to define a quality metric, so as to compare the various cases. However, the evaluation of the obtained configuration may vary depending on the computer vision task, for which the system is deployed. Hence we propose to use an entropy metric, based on the distribution of the histogram bins of a given image [1]. Intuitively, an even distribution of the bin intensity across various grey levels of the image implies higher entropy, while the concentration towards higher (lower) grey levels indicates image sensor saturation. In order to simulate the virtual views of the given environment, we use the software POVRAY [2], a ray- tracer which is based on light ray reflection model; it allows us to simulate light sources and other objects which are typical of indoor surveillance environments with minute details like reflectance, irradiance, attenuation etc. Considering that regular optimization methods cannot be used to solve such a problem, we use particle swarm optimization PSO [3].

References

[1] Zujun Hou and Wei-Yun Yau, Visible entropy: A measure for image visibility," in Pattern Recognition (ICPR), 2010 20th International Conference on. IEEE, 2010, pp.4448-4451.

[2] Victoria Australia Persistence of Vision Pty. Ltd., Williamstown, Persistence of vision (tm) raytracer,"2004

[3] J. Kennedy and R. C. Eberhart, Particle swarm optimization," in IEEE Int. Neural. Networks Conf., 1995, vol. 4, pp. 1942-1948.

9020-39, Session PWed

Nonlinear and non-Gaussian Bayesian-based handwriting beautification

Cao Shi, Jianguo Xiao, Peking Univ. (China); Canhui Xu, Wenhua Jia, Peking Univ (China)

Handwriting has been partly replaced by typing in daily life, due to the increasing usage of personal computer. Interestingly, the evolution progress of personal computer results in the popularity of handwriting on tablet computer. People have become familiar with typing even more than handwriting skills, especially in the office. And writing letters on touch screen is significantly different with handwriting on paper. To improve user experience and realize handwriting using computer system, it is urgent to assist user in enhancing visual effect of handwriting in computer system.

A natural methodology to imitate handwriting depends on pen trajectory, through analyzing influence over handwriting from physical properties of writing tools, exploring geometric deformation of writing tools under external force, finally, achieving imitation of handwriting process and painting process.

To reduce high computational complexity, spline curves with mechanical parameter are employed to compose glyph. Moreover, prior knowledge, such as inherent geometric features of sample characters, and machine learning algorithm are exploited to improve visual beauty of glyph.

However, few consider making use of computer font to beautify handwriting. Unlike previous work, in this paper, a handwriting beautification strategy is proposed by exploiting Bayesian dynamic model to adjust typeface so as to gain the best nonlinear and non-Gaussian approximation of handwriting.

The preliminary step to beautify handwriting is handwriting capture and image normalization. Handwriting image is captured, of which the size and the format are normalized according to algorithm configuration. After obtaining handwriting image, the second step is to recognize the character and a typeface. In practice, research on character recognition has been developed for several decades and remains as a challenging task. In this paper, character recognition is skipped, and the recognition result is a manual parameter. And then a typeface, designed similarly to handwriting, is used to beautify handwriting. The third step is called Parameter Estimation, in which Bayesian dynamic model is employed. The Measurement Equation calculates the likelihood between normalized Handwriting image and the typeface, and the State Equation estimates the affine transformation parameters to transform typeface, resulting in handwriting beautification.

Experimental results demonstrate the proposed framework provides a creative handwriting beautification methodology to improve visual acceptance. The beautification result is able to reduce bad visual effects caused by bad handwriting habits, such as the size of letter is too small, a stroke is too short or long to keep the balance, or the pen trajectory is unsmooth, etc.

Furthermore, a big difference between handwriting and some kinds of typeface is the serif in typeface. The original usage of typeface is for publishing. The serif is adopted to avoid the lack of ink at the end of stroke in the printing process. And then the serif is kept for monitor display to provide vivid visual effect for handwriting beautification result.

Reversely, the other kinds of typeface are designed according to handwriting. The random variation of curvature on these kinds of typeface let the beautification result obtain the randomicity from handwriting.

9020-40, Session PWed

Multi-exposure image fusion with compensation for multi-dimensional camera shake and foreground object movement

Eran A. Edirisinghe, Manal Al-Rawahi, Sara Saravi, Loughborough Univ. (United Kingdom)

High Dynamic Range (HDR) images can be obtained by capturing images at different exposures and then fusing them to reproduce a HDR image. However, in order to produce ghosting artefact free images, the fusion technique has to be able to compensate for motion due to both camera shake and object movement.

In our previous work [1] we proposed the use of Scale Invariant Feature Transforms (SIFT) and Coherent Point Drift (CPD) to respectively identify and align key feature points in two differently exposed images prior to their fusion in Contourlet Transform Domain. Although this approach

allowed compensation against multi-dimensional camera shake, it did not compensate for object movement within a scene.

In [2] authors proposed a single image-based ghost free high dynamic range image fusion approach that uses local histogram stretching and spatially-adaptive de-noising. Although this approach can compensate for foreground object movement, it has the disadvantage of being based on a single exposure image, resulting in entirely over or under exposed regions being irrecoverable. However the histogram stretching approach presented can be used to create a set of multi-exposure images from dual exposure images.

In [3] an approach to perform spatio-temporal noise reduction in images, which relies on the idea of matching, blending and recursive accumulation of images into a frame buffer, to improve signal to noise ratio, was proposed. In this approach errors due to motion were handled by the noise reduction system, which removed potential image artefacts. Therefore it is possible to consider the application of this approach in matching images taken at different exposures. However this method has the limitation that object displacement due to in-plane camera rotation and camera movement in a direction perpendicular to the image plane, cannot be compensated.

The application of image alignment approach of [1] to remove multi-dimensional camera shake prior to prior to using the spatio-temporal noise reduction approach of [3] in removing artefacts due to local foreground object movement in the images being fused provides a practical solution to deal with shortcomings of utilising the two approaches independently in multi-exposure image fusion. In addition the histogram stretching approach of [2] can be used to generate a set of multi-exposure images from two exposure images.

We therefore present a novel approach that combines ideas from [1], [2] and [3] to solve the problem of multi-exposure image fusion in the presence of multi-dimensional camera shake and foreground object movement. We provide experimental results to evaluate the subjective and objective performance of the proposed. We demonstrate the significantly improved performance obtainable as compared to state-of-the-art algorithms.

1) Lluis-Gomez A., Saravi.S., Edirisinghe E.A., Subjectively optimised multi-exposure and multifocus image fusion with compensation for camera shake, SPIE 8436, 2012.

2) J. Im, J.Jeon, H.Hayes, J.Paik, 'Single image-based ghost-free high dynamic range imaging using local histogram stretching and spatially-adaptive denoising', IEEE Transactions on Consumer Electronics, Vol 57, No 4, 2011.

3) I.V.Romanenko, E.A.Edirisinghe, D.Larkin, 'Block matching noise reduction method for photographic images applied in Bayer RAW domain and optimised for real-time implementation', SPIE 8437, 2012.

9020-41, Session PWed

LCAV-31: a dataset for light field object recognition

Alireza Ghasemi, Nelly J. Afonso, Martin Vetterli, Ecole Polytechnique Fédérale de Lausanne (Switzerland)

We present LCAV-31, a multi-view object recognition dataset designed specifically for benchmarking light field image analysis tasks. The principal distinctive factor of LCAV-31 compared to similar datasets is its design goals and availability of novel visual information for more accurate recognition (i.e. light field information).

Light field imaging has received increased attention in recent years mostly due to availability of consumer light field cameras such as Lytro and Raytrix. In this emerging imaging technology rays from multiple views are captured from a scene. The improved sensing capability and the extra information captured by a light field camera enables novel lines of research in order to significantly improve current vision tasks such as object recognition, reconstruction depth mapping.

Many light field processing algorithms have been proposed in recent year due to the increased attention toward this area. The need to evaluate different aspects of these algorithms has led to various light field image sets proposed so far. The structure and associated ground truth information of any of these datasets have been optimized for a specific task such as reconstruction or depth estimation.

On the other side, many object recognition benchmarks are already known among the computer vision community. There is a huge diversity regarding the objectives, scope, capturing setup and other properties between these datasets. However, they mostly consist of traditional color or grayscale images. This restricts their use in evaluation of algorithms which utilize data of modern sensors such as depth or light fields. This issue and other drawbacks of current object recognition datasets such as bias and lack of object diversity bold the need to construct a novel dataset with more information available per object than traditional color images.

LCAV-31 is the first light field object recognition dataset which provides a unified framework for benchmarking light field retrieval, object recognitions and tracking systems. It provides the possibility to evaluate effectiveness of different sensed information such as color and light field in object recognition accuracy.

We first study important properties of different state-of-the-art datasets. We analyze various types of bias in dataset creation and their effect on evaluation results. We also analyze the diversity and robustness of current datasets. Then we explain how we avoid the biases and achieve a high level of diversity in LCAV-31.

The dataset is composed of 31 object categories captured from ordinary household objects. We captured the color and light field images using the recently popularized Lytro consumer camera. Different views of each object have been provided as well as various poses and illumination conditions. We explain all the details of different capture parameters and acquisition procedure such that one can easily study the effect of different factors on the performance of algorithms executed on LCAV-31.

Finally we apply a set of basic object recognition algorithms on LCAV-31. The results of these experiments can be used as a baseline for further development of novel algorithms.

9020-42, Session PWed

Scale-invariant representation of light field images for object recognition and tracking

Alireza Ghasemi, Martin Vetterli, Ecole Polytechnique Fédérale de Lausanne (Switzerland)

We propose a scale-invariant feature descriptor for representation of light-field images. The proposed descriptor can significantly improve tasks such as object recognition and tracking on images taken with recently popularized light field cameras.

A desirable feature descriptor should be invariant to certain transformations such as scaling and rotation. However, achieving true invariance requires certain information of the scene which are mostly lost during the image formation process in modern traditional cameras.

An image is usually formed by projecting the three-dimensional real-world into a two dimensional image plane. This process is obviously not reversible and therefore certain informations such as depth and occlusion are lost during the image formation. Without having these valuable informations, current algorithms rely mostly on color and texture information for description. This is error-prone and leads to a significant amount of wrong matches.

Light field imaging is an emergent technology which has gained popularity in recent years. In light field imaging, rays from multiple views are captured from a scene. Utilizing the extra information content of the rays can significantly improve computer vision algorithms. However, we first need to adapt current algorithms to exploit light field information or develop completely new algorithms.

A crucially important part of many computer vision tasks is robust extraction and description of a set of features. To achieve the goal of perfect scale-invariant light field feature description, we first study the effect of scaling on the internal structure of light fields. We consider horizontal and vertical slices of the light field signal called epipolar planes. We show that each scene point corresponds to a line in an epipolar plane. Further we show that gradients of the epipolar lines are proportional to the depths of their corresponding scene points. Therefore, scaling the scene can be interpreted as shearing the epipolar plane.

Consequently, we attack the problem of scale-invariance by a shear-covariant integral transform approach. The novel transform we propose is

begin{center}

begin{equation}

M(lambda,mu)=intint f(x,y)delta (lambda y + lambdaln lambda x - mu) dxdy

end{equation}
end{center}

which maps the original epipolar planes $f(x,y)$ to a new $lambda-mu$ space in which each point $(lambda_0,mu_0)$ corresponds to a line in the epipolar plane.

The key property of the $lambda-mu$ space is that we can easily isolate and cancel the effect of uniform shearing in it. A simple computationally efficient way for doing so is by using the following transform:

begin{center}

begin{equation}

g(s)=intint M(lambda,mu)delta(mu-slambda) dlambda dmu

end{equation}
end{center}

The discretized values of $g(s)$ are then invariant to shearing

the source epipolar plane. As we showed, unlike some other approaches in computer vision our proposed approach has a solid theoretical justification which proves its scale-invariance.

We test our proposed representation using various light field images of different types, both synthetic and real. Our experiments show very promising results in terms of retaining invariance under various scaling transformations.

9020-43, Session PWed

Image indexing based on a circularity features

Ana M. Herrera-Navarro, Univ. Autonoma de Queretaro (Mexico); Hugo Jiménez-Hernández, Ctr. de Ingenieria y Desarrollo Industrial (Mexico); Ivan Terol-Villalobos, CIDETEQ (Mexico)

Image features represent most common used primitives to perform higher-level methods based on image processing. Good selection of these features helps to provide expressivity and representatively of information belonging in images. However, toward to the search of general feature descriptor task, represent a difficult path. This is, feature images are directly dependable of image semantic and the type of analysis to be performed. In this situation, several descriptors have been created and have been applied in controlled environment for particular task. Speaking in formal aspect, each image feature is described as a formal metric, which warrant certain similitude behavior into set of image applied; or in other hand, use a set of local heuristic based on previously knowledge of problem. Some of the most representatives include Euler number, Harris descriptors, Gabor filter and more recently SIFT features. Last one has received certain acceptance because try to use a set of filter that be invariant to scale rotation and certain degree of perspective. However, the use of more robust features have the disvantage of an increasing computer complexity.

In this sense, for a particular application, as image indexing, a set of features is needed to identify the information content on an image. In recent years, several geometrical metrics are developed to characterize basic geometric shapes. One useful geometric shape is the circle. In this sense in the literature there are several works done around the measurement of circularity. All circularity measures have low computational complexity, which make suitable to use for analyzing big chunks of data, in a such a way, circularity measures are used as topological descriptors of an image. The advantage to use circularity measures is that these are invariant to certain image deformation as rotation, scale or border width.

A proposed circularity measure expressed in last work, deal with the problem of partial information of border of circle, or occlusions. This fact becomes reliable to discover hidden geometries in partial information. Then, this measure is defined as follows:

CircularityIndex=sum f(x) from x –k to +k

Where f(x) is probability density function (pdf) of inner radius expressed by the border evidence, and k is a confidence index related with the probability density behavior.

This work presents a novel procedure to index automatically images based on a circularity measure into a set of images. The circle is measured from all connected border, which can be estimated. In literature, there are several ways to compute borders. However, they depend of external factors as size, noise, and orientation. To solvent this problem, we introduce a border a general detector, which is based on a derivative Gaussian and Lk distance.

The variation of Gaussian parameter to discover the border and Lk distance allows to be more/less sensitive to the borders. After, connecting all maxima zones from derivatives a path connected are located. All connected (but not necessarily closed) border are used to be characterized by a circularity measure. As it is appreciated all shapes does not correspond to circles, but the proposal only take a circle measure as reference to measure the similarity expressed to respect to parametric geometry.

All given measured are used to estimate a pdf of circularity measures. Then, from different images the pdf of circularity measure are estimated. Using the behavior of two pdf's the images are matched. Next, to develop a good similarity criterion a cumulative pdf (cpdf) are estimated starting from both pdf's. Now, to match both images the cumulative are discretized into m uniform parts. Each one cpdf is expressed as a 1 x m vector. The distance of this vector is estimated using Lk norm as follows

d(cpdf1,cpdf2)=(sum(|cpdf1(xi)-cpdf2(xi)|^k))^(1/k)

where xi is th component of vector.

This criterion under apply to a set of image can be discover dominant geometrical structures referenced to the ideal circles. Computationally this measure is cheaper and can be applied to big chunk of data. In experimental process, this method is applied to index different images and to detect in video sequences, the changes of scene.

9020-44, Session PWed

Comparative analysis of the speed performance of texture analysis algorithms on a graphic processing unit (GPU)

Jennifer C. Triana, Sergio A. Orjuela Vargas, Univ. Antonio Nariño (Colombia); Wilfried Philips, Univ. Gent (Belgium)

with standards. This type of evaluation is made from a visual inspection task conducted by human experts. Visual examination is conducted for determining surface quality. It includes examining texture, surface

characteristics, dye shade variations, design details, weave patterns, construction particulars, pilling assessment and wrinkling evaluation among other applications. Generally, it is divided in two types: defect detection and appearance deviations from standard samples or photographs. The main drawback of such evaluations is that it is subjected to human subjectivity which affects accuracy and repeatability, decreasing the reliability. Therefore, evaluation of products appearance for quality assurance purposes has improved by analyzing visual characteristics of the products using an automatized computer vision system. Computer vision is based on image analysis, which consists on extract relevant information from images. The primitive image descriptors are shape, texture and color. Color features are among the most widely used visual features. They are invariant against geometric transformations. Optimal color features must be highly discriminative and stable under varying viewing conditions. Shape features are used to characterize geometric details in an image. Texture refers to innate properties of objects related to their structural surface organization. Particularly, the visual perception of texture plays an important role in the biological visual system for recognizing and differentiating regions in a scene. Texture analysis in computer vision is inspired on biological visual perception and is one of the most difficult computer vision processes.. It used to automatically differentiate regions or identify deviations from patterns. In this context, texture can be defined as variation in data at scales smaller than the scale of interest. Therefore, texture analysis of images consists of characterizing local intensity changes on images. Inspection systems based on computer vision are known as Automated Visual Inspection (AVI) systems. Currently, industry manages big quantities of production. Therefore, a challenge in developing AVI systems is to implement low computational costs algorithms capable of conducting inspection in real time. Recently, Graphic Process Units (GPU) has been used to accelerate the execution of image analysis algorithms thanks to the use of a vast number of simple, data-parallel, deeply multithreaded cores and high memory bandwidths. These main features make the use of the Graphic Process Units (GPU) suitable for processing large-scale data-parallel load of high-density computing, turning it into an attractive area for research and development. The device's programming is possible thanks to the use of the Compute Unified Device Architecture (CUDA) programing, only supported by NVidia Graphics Cards. In this approach we implemented a set of classic texture image algorithms. We first implement the algorithms sequentially, making use of the OpenCv Libraries. Then, we compare the speed performance of the same algorithms for evaluation of texture by using CUDA programing in a Windows 7x64 system with 2.90 GHz core processor, 4GB RAM and a GeForce GT 525M. We also compare the accuracy of the algorithms by using a dataset of images exhibiting transitional textures. These dataset is composed of photographs of samples of carpets that have been subjected to progressive degrees of degradation by the use of mechanical devices. The comparison is conducted following experimental designs. Results show a comparison of the texture algorithms executed on a CPU and the GPU. Comparison shows Improvements in computation speed were achieved over traditional CPU-based algorithms. As the image size increase, some texture analysis Algorithm computations can get a more than 10X speedup, and others Algorithm computations like Co-occurrence Matrix can get an about 40X speedup. This may facilitate the selection of algorithms for further applications.

9020-1, Session 1

Video colorization based on optical flow and edge-oriented color propagation

Mayu Otani, Nara Institute of Science and Technology (Japan); Hirohisa Hioki, Kyoto Univ. (Japan)

We propose a novel method for video colorization. Colorization is a process of adding color to monochrome images or videos. Typical applications of colorization includes restoring colors of old films and giving special color effect for photos. Since a video has a huge number of frames, an automatic colorization algorithm is required.

Several methods have been proposed for video colorization. In these methods, it is assumed that key frames are selected out in advance and are partially colorized by providing initial color strokes manually. Levin et al. defined a cost function under assumption that neighboring pixels of similar intensities have similar colors. Colors are propagated from initial strokes to remaining pixels of all the frames spatiotemporally by minimizing the cost function. Motions between frames are estimated by computing dense optical flow. Jun Hee Heu et al. proposed a method where color propagation in a frame is controlled by a pair of accuracy and priority measures defined for each pixel. A block-matching algorithm is employed to estimate motions and to propagate colors and accuracy values between frames. Even when we carefully give initial color strokes, these methods occasionally spread colors inappropriately, which is considered to be caused by wrong motion estimation and carelessness about edges in images.

To overcome these problems, we propose a new colorization method based on sparse optical flow and edge-oriented color propagation. On propagating colors between frames, our method computes optical flow only for feature points to reduce the influence of wrong motion estimation. Frames are thus once partially colorized. The remaining grayscale pixels are colorized by propagating colors of feature points with considering edges. Note that key frames can be colorized by giving initial strokes and propagating them.

Assume that we have a pair of colorized key frames and an in-between grayscale frame to be colorized. At the first stage of our method, feature points are extracted from both key frames. The colors around feature points in the key frames are then copied to the grayscale frame by computing optical flow. We avoid copying colors for featureless points at this stage, since optical flow for such points often become unreliable. The colors of feature points are propagated to the the rest of the frame at the second stage of our method. For propagating colors appropriately, a pair of priority and accuracy measures is defined for each pixel. In order to prevent inappropriate colors from spreading across edges, nine different sets of neighboring pixels are provided for each pixel and one of them is adaptively selected on propagating color. In this way, all grayscale frames between the two key frames are colorized.

We have carried out experiments of still image colorization and video colorization. The still image experimental results show our propagation algorithm works well. Video experimental results demonstrate that our method can colorize videos with good quality and thus our method is useful. We also show that our method enables us to easily modify colors in colored video streams.

9020-3, Session 1

Image enhancement with blurred and noisy image pairs using dual edge-preserving filtering technique

Yuushi Toyoda, Hiroyasu Yoshikawa, Masayoshi Shimizu, Fujitsu Labs., Ltd. (Japan)

The general electronic image stabilization function of digital cameras is realized by combining multiple images captured with short exposure times. Each short-exposed image is less affected by motion blur but is noisy. Therefore, we need to combine short-exposed images to obtain a less-blurry and less-noisy image. However, the performance of image stabilization carried out using this method is theoretically dependent on the number of images that need to be combined. For instance to estimate the performance for three stops by a shutter-speed conversion, it is necessary to combine eight images. Consequently, a significantly large image memory and a considerable amount of processing time are required.

In this paper, we propose a two-image approach that adaptively combines a short-exposed image and a long-exposed image. The short-exposed image is less affected by motion blur, whereas the long-exposed image is less affected by noise. The combined method combines only the good properties of the two images to form a less-blurry and less-noisy

image.

The basic concept of the proposed method is the generation of an image in which the edge part is mainly the combined short-exposed image and the smooth part is mainly the combined long-exposed image. We use the absolute difference between the two images for distinguishing the edge part from the smooth part. Because after position alignment and photometric calibration between the images, the absolute difference consists of motion blur and noise. If the absolute difference is low, it seems to be in the smooth part, and vice versa. Further, in the above-mentioned basic concept, the motion blur can be controlled. However, the edge of the combined image is directly outputted from the short-exposed image's noise.

To reduce the noise in the edge part, we have developed a filtering technique that combines the long-exposed image at the edge part to the utmost extent. In the edge part, the long-exposed pixel of the same position has a significantly large absolute difference because of the motion blur. However, there will be compoundable long-exposed pixels, if we search not for a pixel having the same position but for a pixel that is located a little distance away from the edge part. Thus, we have developed a dual edge-preserving filtering technique, which searches and compounds similar pixels in the spatial and the intensity domains from both images. The developed filtering technique can be summarized in the following steps:

1. Set each pixel of the short-exposed image as the target pixel for filtering.

2. Select the pixels of the short-exposed image and of the long-exposed image that are located near the target pixel and have similar pixel intensity.

3. Calculate the average of the selected pixels to obtain an output pixel of the target pixel.

As described above, the proposed technique seeks out similar pixels to combine from both images not only in the smooth part but also from the edge area. We have confirmed the effectiveness of the proposed approach by using a number of images taken by a handheld camera and using vibratory apparatus.

9020-4, Session 1

Computational efficiency improvements for image colorization

Chao Yu, Gaurav Sharma, Univ. of Rochester (United States); Hussein Aly, Military Technical College (Egypt)

Colorization refers to the problem of adding colors to monochrome images and videos. Colorizing is of considerable interest for the large collection of black-and-white movies, television shows, and photographs that were produced before color capture became widely prevalent. Colorization is traditionally an extremely time-consuming task that requires a skilled artist to manually add and iteratively adjust the colors and shading in an image. Recently, promising results have been obtained with optimization based colorization algorithms that require only a sparse set of colors to be manually specified from which colors are automatically propagated throughout the image while preserving natural smoothness constraints.

We propose an efficient algorithm for colorization of greyscale images. As in prior work, colorization is posed as an optimization problem: a user specifies the color for a few scribbles drawn on the greyscale image and the color image is obtained by propagating color information from the scribbles to surrounding regions. In this formulation, colorization is obtained by solving a large sparse linear system, which normally requires substantial computation and memory resources. Our algorithm improves the computational performance through three innovations over prior colorization implementations. First, the linear system is solved iteratively without explicitly constructing the sparse matrix, which significantly reduces the required memory. Second, we formulate each iteration in terms of integral images obtained by dynamic programming, reducing repetitive computation. Third, we use a coarse-to-fine framework, where a lower resolution subsampled image is first colorized and this low resolution color image is upsampled to initialize the colorization process for the fine level. The improvements we develop provide significant speed-up and memory savings compared to the conventional approach of solving the linear system directly using off-the-shelf sparse solvers, and allow us to colorize images with typical sizes encountered in realistic applications on typical commodity computing platforms.

The algorithm we propose also has the advantage that the memory and computation cost are (largely) independent of the texture content in the image and of the number of regions for which the user specifies the approximate color via scribbles. In contrast, for prior algorithms developed for this problem, the computation/memory requirements increase with user inputs or amount of image texture and are therefore less predictable.

According to our experimental study, the run time requirements for the conventional approach are approximately 2.5 times those for the proposed method, consistently across the different image sizes. The reduction in memory requirement is even more significant and, more importantly, grows with increase in the size of the image being colorized. For a 2560 by 2048 image, the proposed algorithm uses 2071 Megabytes (MB) as opposed to 12502 MB required for the conventional optimization algorithm.

9020-5, Session 2

Architectures and algorithms for x-ray diffraction imaging

David A. Castañón, Boston Univ (United States); Ke Chen, Boston Univ. (United States)

X-Ray imaging is the predominant modality used in luggage inspection systems for explosives detection. Conventional or dual energy X-ray computed tomography (CT) imaging reconstructs the x-ray absorption characteristics of luggage contents at the different energies; however, material characterization based on absorption characteristics at these energies is often ambiguous. X-ray diffraction imaging (XDI) measures coherently scattered X-rays to construct diffraction profiles of materials that can provide additional molecular signature information to improve the identification of specific materials. In this paper, we review our recent work on developing XDI algorithms for different architectures, which include limited angle tomography and the use of coded aperture masks. We illustrate the performance of different approaches using Monte Carlo propagation simulations through 3-D media.

9020-6, Session 2

Joint metal artifact reduction and material discrimination in x-ray CT using a learning-based graph-cut method

Limor Martin, Ahmet Tuysuzoglu, Prakash Ishwar, William C. Karl, Boston Univ. (United States)

No Abstract Available

9020-7, Session 2

Accurate estimation of noise correlations in a CT image

Frédéric Noo, Marta Heilbrun, The Univ. of Utah (United States)

No Abstract Available

9020-8, Session 2

Linear discriminant analysis (LDA) for rapid deconvolution of photon counting measurements by digital filtering

Shane Z. Sullivan, Paul D. Schmitt, Ryan D. Muir, Emma L. DeWalt, Garth J. Simpson, Purdue Univ. (United States)

No Abstract Available

9020-9, Session 2

Magnified neutron radiography with coded sources

Philip R. Bingham, Hector J. Santos-Villalobos, Nickolay V. Lavrik, Oak Ridge National Lab. (United States); Jens Gregor, The Univ. of Tennessee (United States); Hassina Bilheux, Oak Ridge National Lab. (United States)

A number of neutron facilities worldwide have developed radiography instruments. These instruments provide transmission images following the same exponential attenuation seen in x-ray radiography. However, the unique cross sections for neutrons and x-rays make these two modalities complementary such that one modality will be able to make a measurement for a specific application while the other would not. X-ray radiography has pushed resolution limits down to the micron level using focused sources and magnified imaging systems. Unfortunately, neutron source intensity is limited and optics for neutrons are inefficient making the use of a single focused spot for magnified imaging infeasible. As a result, neutron radiography is performed with a parallel beam and is limited in resolution by the detectors. Scintillator based detectors can reach 25-50um resolution and micro channel plate detectors claim 15um resolution. A team at the Oak Ridge National Laboratory (ORNL) has been developing a magnified neutron radiography capability using coded aperture techniques. In this effort, a coded aperture mask is placed between the neutron and the object to be radiographed to produce a coded array of point sources. This allows the neutron beam to provide the high resolution associated with the point sources in the mask along with the higher neutron flux resulting from multiple sources. Magnified imaging using coded sources has been performed at the CG-1D imaging beam line at the High Flux Isotope Reactor using multiple mask resolutions (200, 100, 50, 20, and 10um) with magnifications up to 25x. Challenges of achieving these resolutions and moving to higher resolution are mask quality, neutron beam non-uniformity, aperture alignment, and gravity. This paper discusses the achievable mask pattern geometries for the coded neutron masks, presents experimental results using these masks for magnified neutron radiography using convolution based reconstruction with aperture/anti-aperture pairs, and presents initial model based SIRT reconstruction results that include both beam shape and gravity in the model.

9020-10, Session 3

A super-resolution algorithm for enhancement of flash lidar data: flight test results

Alexander Bulyshev, Analytical Mechanics Associates, Inc. (United States); Farzin Amzajerdian, Vincent E. Roback, Robert A. Reisse, NASA Langley Research Ctr. (United States)

The results of using 3D super-resolution (SR) processing method for Flash Lidar range data obtained during a helicopter flight test campaign are presented. The flight test was conducted in frame of NASA Autonomous Landing and Hazard Avoidance Technology (ALHAT) project. ALHAT project objective is to develop the necessary technologies for autonomous safe landing of on the surface of celestial bodies such as the Moon, Mars, asteroids. The flights took place in NASA Kennedy Space Center (KSC) in December 2012. One of the test objectives was to verify the ability of 3D SR procedure to generate low-noise, high resolution digital elevation model (DEM) and to reproduce a time resolved relative positions and orientations of the instrument. 3D SR method was developed earlier and tested in computational modeling, and laboratory experiments, and dynamically from a moving vehicle at the NASA Langley's lidar test range facility.

The flight test was conducted over a simulated lunar terrain facility constructed specifically for demonstrating the landing sensors and guidance technologies by ALHAT project at NASA KSC. The simulated lunar terrain is 100 m X 100 m in dimensions and includes hundreds of rocks and craters with different size and shapes. The Flash Lidar was built by NASA Langley using 128X128 pixels camera from Advance Scientific Concept. The camera consists of a InGaAs Avalanche Photodiode (APD) array, an Integrated Readout Circuit (ROIC), and associated controller and data/command interfaces. A 5 degree field-of-view lens was used with the lidar for data collection. The flights were conducted over different incident angles from 20 to 50 degrees relative to the ground from over 1000 m slant range to the target field. Collected data were processed by the SR algorithm to generate high resolution DEMs and the platform navigation state vector (i.e., position and attitude angles). The generated DEMs of the target area were compared with a truth elevation map provided by a ground-based surveying instrument. Each individual high resolution DEM with the size 1024X1024 pixels was constructed using 20 Flash Lidar frames. The results of comparisons show that the method allows constructing high quality DEM with visible ability to identify hazards like rocks and craters as small as 50cm in diameter and 30cm in height/depth from over 700 m away. The 6 degree of freedom state vector of the platform as function of time obtained from the SR algorithm were compared with the data from a high-grade Inertial Measurement Unit and GPS. Comparisons of the trajectory data also confirm the SR algorithm's ability for relative navigation based on 3D video flow approach. The trajectory error is estimated as 30cm in position and 0.01 degree in pointing angle.

9020-11, Session 3

Automatic image assessment from facial attributes

Ray Ptucha, David Kloosterman, Brian Mittelstaedt, Alexander Loui, Eastman Kodak Co. (United States)

Personal consumer photography collections often contain photos captured by smart phones, tablets, digital cameras, and video recording devices. These collections are stored on multiple devices, may span multiple family members, and often include contacts via online social networks. The task of gathering, organizing, and assembling still and video assets in preparation for sharing with others can be quite difficult. For example, after returning from a family vacation, the family may decide to create a slideshow or photobook. The collection size can be daunting, often containing a myriad of quality levels and redundant material. Current commercial photobook applications are mostly manual-based requiring significant user interactions for selecting the right images. To assist the consumer in organizing these assets, we propose an automatic method to assign a fitness score to each asset, whereby the top scoring assets can be suggested for product creation. Our method uses cues extracted from analyzing pixel data, metadata embedded in the file, as well as ancillary tags or online comments.

Although evidence suggests less than 30% of photos contain faces, over 75% of images used in consumer photobooks contain faces. Furthermore, when a face occurs in an image, its features have a dominating influence on both aesthetic and compositional properties of the displayed image. As such, this paper will emphasize the contributions faces have on affecting the overall fitness score of an image. With

such fitness scores available, only the highest scoring assets are recommended for initial inclusion into product. Similarly, when assets are deemed to be similar enough, only the asset with the highest score of the group of similar assets is used in the product. Further, there needs to be a way to determine which assets are superior to others such that they may be displayed more prominently (larger, more centrally located, or more duration).

Our facial attribute assessment algorithm utilizes rules that are based upon head size, location, expression, pose, eye blink, and eye gaze. We introduce a concept of facial impact, such that faces that are larger and more centrally located have more influence on the final fitness score. To understand consumer preference, we conducted a psychophysical study that spanned 27 judges, 5,598 faces, and 2,550 images. Preferences on a per-face and per-image basis were independently gathered to train our classifiers.

After face detection, automatically localized eye and mouth corner points define an affine warp to a normalized canonical face representation. Once in this normalized state, linear extension of graph embedding manifold learning enables facial understanding classification. We describe how to merge differing attributes such as pose and gaze into a single classifier, and then how to weight the impact of each on a face by face basis. Individual face attributes are used to train classifier frameworks that determine a single face's preference from a consumer perspective. Holistic image attributes are used to train classifier frameworks that determine a preferred composition when multiple faces are in a single frame or image. Aside from face detection and facial feature point localization, the overhead of our proposed approach is negligible. We believe this is the largest psychophysical study on consumer face preference to date. Our novel methods of facial weighting, fusion of facial attributes, and dimensionality reduction produce state-of-the-art results suitable for commercial applications.

9020-12, Session 3

Closely spaced object discrimination computation using quantum annealing model

John J. Tran, Information Sciences Institute (United States); Robert F. Lucas, The Univ. of Southern California (United States); Kevin J. Scully, Darren L. Semmen, The Aerospace Corp. (United States)

One of the challenges of automated target recognition and tracking on a two-dimensional infrared (IR) focal plane is the ability to detect and discriminate closely spaced objects (CSO).

To date, one of the best CSO approximation algorithm involves first subdividing a cluster of image pixels into equally spaced grid points; then it conjecture that N targets are located at the centers of those sub-pixels and calculate the irradiance values of those conjectured targets that minimizes the sum of squares of the residuals.

Computing the initial starting point is a linear least-squares fit and the remaining approximation calculation employs a non-linear fitting algorithm (e.g. Levenberg-Marquardt, Nelder-Mead, trust-region, expectation-maximization, etc.). As the result, the overall time complexity is exponential and thus cannot be reasonably solved in polynomial time.

Although numerous strides have been made over the years, vis-a-vis heuristic optimization techniques, the CSO discrimination problem remains largely intractable, due to its combinatoric nature.

We propose a novel approach to address this computational obstacle by employing the adiabatic quantum annealing (AQA) optimization model.

First, we frame the CSO problem as an Ising spin glass problem, where each non-linear combination of the grid points maps to a Hamiltonian state of the AQA.

Second, we program the Hamiltonian configurations into the D-Wave Two machine and using its quantum annealing property to determine minimal energy configurations. Although the result can be stochastic, a cumulative average should give high confidence to the correct optimal Euclidian positions for hypothesized CSO objects.

Finally, we conduct analytical comparisons and timing experiments of the AQA model against several known CSO optimization heuristics and other hardware-based optimization models, such as the massively parallel graphics processors, e.g. the CUDA platform.

9020-13, Session 3

3D quantitative microwave imaging from sparsely measured data with Huber regularization

Funing Bai, Aleksandra Pizurica, Univ. Gent (Belgium)

Quantitative microwave imaging aims at reconstructing the exact permittivity profile of an unknown scattering object by illuminating the object with microwaves and by measuring the scattered field. Different regularization approaches have been proposed for solving this nonlinear ill-posed inverse problem, including Multiplicative Smoothing (MS), Value Picking (VP) and Total Variation (TV). MS applies Tikhonov regularization in a multiplicative fashion and hence it tends to oversmooth the results, while VP produces sharp results but is applicable only to piece-wise constant profiles with few distinct permittivities values. TV is also effective for piecewise constant profiles. We introduced previously a new regularization approach based on the Huber function, which is suitable not only for piecewise constant, but also for more general permittivity profiles (that are of interest e.g. in biomedical imaging). We demonstrated already very encouraging first results on simulated data. In this paper, we perform thorough analysis of this regularizer, studying the influence of the parameters under different levels of noise. Moreover, we evaluate the whole approach not only on simulated data but also on real 3D electromagnetic measurements. Our focus is on reconstructions from relatively few measurements (sparse measurements), which are of interest to speed up the reconstruction process.

We consider in this paper scattering measurements of inhomogeneous targets from the Institute Fresnel, that are available in the so-called Fresnel database, and that are commonly used to test new algorithms in the inverse scattering community. Validating inversion algorithms on experimental data is much more reliable than using simulations only, which are prone to inverse crime. Reconstructions from the experimental data are quite challenging due to measurement noise and discretization noise as well as mismatch between the actual incident fields and their simulation in the forward solvers.

Furthermore, it is interesting to explore potentials for speeding up the reconstruction process. In quantitative microwave imaging, using a smaller number of data points (i.e. performing a sparse reconstruction), is especially of interest because long computation times are currently limiting practical use of this imaging modality. We show that using our approach reconstruction times for 3D objects can be drastically reduced (e.g., from several hours to several minutes on the same computer architecure, without affecting significantly the quality of the reconstructed results).

Our results on experimental data from the 3D Fresnel database motivate strongly the use of Huber regularization. The advantages over realted regularization methods, like MS and VP that were demonstrated previously on simulated data are now confirmed with real experimental data. Moreover, thorough analysis of the influence of different parameters presented in this paper gives new insights in the behavior of the Huber regularizer in quantitative microwave imaging and provides useful guidelines for its practical use in different scenarios (e.g. different levels of measurement noise).

9020-14, Session 4

Novel tensor transform-based method of image reconstruction from limited-angle projection data

Artyom M. Grigoryan, Nan Du, The Univ. of Texas at San Antonio (United States)

The solution of the problem of image reconstruction from a finite number of projections and with a limited angular range in the projection data is very important in many practical applications, such as computed tomography and electron microscopy. The exact solution of image reconstruction from a finite number of projections does not exist. The inverse Radon transform-based method or the method of convolution uses a complete and infinite number of projections and all line-integrals within each projection. The implementation of this method for a finite number of projections requires the approximation of the transform; the quality of image reconstruction is highly dependent upon the number of projections. The limited angular range of available projections makes the problem of image reconstruction difficult. Different algorithms were developed to solve this problem; the iterative image-Fourier space revision methods, iterative image-projection space methods, and backprojection methods with the deconvolution of the point spread function along each projection. Many methods work under the ideal assumption that 2-D objects in images have structures which can be represented by very sparse matrices.

In this paper, we apply a novel approach for reconstructing the discrete image f(xn,ym) on the Cartesian lattice from a finite number of projections of the image f(x,y). This approach is based on the tensor transform and an idea of transferring the geometry of integrals from the image space to the Cartesian lattice. This transformation means the line-integrals (or ray-integrals) over the image f(x,y) can be used for exactly calculating all line-sums (or ray-sums) over the Cartesian lattice, which are required to calculate the tensor transform of f(xn,ym). In the tensor representation, a discrete image is considered as the sum of direction images; each one carries the spectral information of the image at frequencies of different sets covering the Cartesian lattice. Therefore, the discrete image can be reconstructed by these direction images through the 2-D DFT or directly by performing the inverse tensor transform. We consider the parallel beam scanning scheme for projections. The reconstruction of the image on the lattice is exact. The proposed approach is presented for the continuous model, when the image f(x,y) is on the unit square and consists of cells (image elements) divided by the Cartesian lattice; the intensity of the image on each such cell is constant. We analyze the problem of image reconstruction from limited angle data and compare with the known algorithms, including the projection onto convex sets. Preliminary results show very good results of image reconstruction when the angular range scanned is 26° and down to 10°. The experimental results of image reconstruction from a finite number of projections with a limited range of angles are illustrated. The proposed method is also analyzed for the case with the noise in projection data. The results with a noisy projection data show that the proposed method of reconstruction is robust relative to an additive signal-independent noise. The reconstruction SNR for the images obtained from the noisy projection data are given and compared with other methods of image reconstruction.

9020-15, Session 4

Statistical x-ray computed tomography from photon-starved measurements

Zhiqian Chang, University of Notre Dame (United States); Jean-Baptiste Thibault, GE Healthcare (United States); Ken Sauer, University of Notre Dame (United States); Charles Bouman, Purdue University (United States)

No Abstract Available

9020-16, Session 4

Model-based iterative tomographic reconstruction with adaptive sparsifying transforms

Luke Pfister, Yoram Bresler, Univ. of Illinois at Urbana-Champaign (United States)

There is a growing concern about the public health risk posed by the radiation dose delivered by x-ray CT. X-ray dose reduction has therefore taken on increased importance. Unlike standard linear filtered backprojection reconstruction, iterative reconstruction algorithms can produce high-quality images from low-dose data by incorporating detailed models of the image being reconstructed, and of the data acquisition process and noise statistics. Recently, signal models in which the data is assumed to have a sparse representation in some form have shown enormous promise.

The performance of a sparse signal model strongly depends on how accurately the model represents the data. Recent work has explored techniques to adaptively learn such a signal model from the data itself using the synthesis sparsity model. These adaptive sparse representations incorporated directly into the tomographic reconstruction process. Jointly reconstructing an image from tomographic data while learning a synthesis dictionary has shown improvement over methods such as total-variation (TV) regularization, especially at representing complex structures.

Unfortunately, synthesis sparsity methods scale poorly with problem size and significantly increase image reconstruction time, limiting their utility for practical imaging applications.

Recently, Ravishankar and Bresler proposed an alternative type of sparse modeling known as transform sparsity. A matrix, called a sparsifying transform, is learned, which, when applied to data, causes the result to be nearly sparse. The nearest sparse vector to the transformed signal is known as a transform sparse code. Unlike the synthesis sparsity case, sparsifying transforms can be learned and the transform sparse codes can be found using computationally efficient algorithms.

We propose an image reconstruction algorithm that combines statistical iterative reconstruction with an adaptive sparsifying transform penalty. Our algorithm begins with a filtered backprojection reconstruction and an initial sparsifying transform, then alternates between updating the sparsifying transform, calculating the transform sparse codes, and reconstructing the image, using the sparsifying transform as a regularizer.

The sparsifying transform and transform sparse code updates are performed using closed-form solutions at low computational cost.

To capture the statistical properties of low-dose tomographic data, we utilize a quadratic approximation to the negative log-likelihood of the Poisson distribution. The Alternating Direction Method of Multipliers (ADMM) is used to separate the statistical weighting from the forward projection operator, allowing for the use of efficient Fourier-based preconditioners in the image reconstruction step.

We evaluate the algorithm through numerical experiments using the FORBILD mathematical phantom and 2D fan-beam projection data generated from clinical CT scans. Dose reduction is achieved by reduction in photon flux and/or by reduction in the number of views. Our algorithm is compared to ADMM-based TV and synthesis dictionary learning regularized reconstruction algorithms. Results indicate that adaptive sparsifying transform regularization retain the structure-preserving performance of synthesis sparsity methods at a speed comparable to total-variation regularization.

9020-17, Session 4

Structured illumination for compressive x-ray diffraction tomography

Joel A. Greenberg, Mehadi Hassan, Kalyani Krishnamurthy, David Brady, Duke Univ (United States)

Coherent x-ray scatter (also know as x-ray diffraction) has long been used to non-destructively investigate the molecular structure of materials for industrial, medical, security, and fundamental purposes. Unfortunately, molecular tomography based on coherent scatter typically requires long scan times and/or large incident fluxes, which has limited the practical applicability of such schemes. One can overcome the conventional challenges by employing compressive sensing theory to optimize the information obtained per incident photon. We accomplish this in two primary ways: we use a coded aperture to structure the incident illumination and realize massive measurement parallelization and use photon-counting, energy-sensitive detection to recover maximal information from each detected photon. We motivate and discuss here the general imaging principles, investigate different coding and sampling strategies, and provide results from experimental and theoretical studies for our structured illumination scheme. We find that this approach enables real-time molecular tomography of bulk objects without a loss in imaging performance.

9020-18, Session 4

Generalized Huber functions for model-based reconstruction from anomalous data

Singanallur Venkatakrishnan, Purdue Univ. (United States); Lawrence F. Drummy, Air Force Research Lab. (United States); Marc De Graef, Carnegie Mellon Univ. (United States); Jeff Simmons, Air Force Research Lab. (United States); Charles A. Bouman, Purdue Univ. (United States)

No Abstract Available

9020-19, Session 5

Effects of powder microstructure on CT number estimates

Jeffrey S. Kallman, Sabrina dePiero, Stephen G. Azevedo, Harry E. Martz, Lawrence Livermore National Lab. (United States)

No Abstract Available

9020-20, Session 5

Coded aperture x-ray scatter tomography

Andrew Holmgren, Kenneth P. MacCabe, Martin P. Tornai, David J. Brady, Duke Univ (United States)

We present a system for X-ray tomography using a coded aperture. A fan beam illuminates a 2D section of an object and the scatter signal creates projections of the coded aperture onto the detector plane. From these projections we estimate the scattering density. Our system produces a tomographic image from each snapshot; as such, we can either reconstruct a static object scanned over time or an x-ray video of a non-static object. We demonstrate results of both cases: we show a toy figurine, and an X-ray video of a clock hand as it rotates in the plane of the beam. We also discuss our measurement models and aperture design. In addition to reconstructing scattering objects, our reconstruction algorithm estimates the scattered radiance of each object -- a step toward materials imaging and identification.

9020-21, Session 5

Marked point process models for microscope images of materials

Huixi Zhao, Mary L. Comer, Purdue Univ. (United States)

No Abstract Available

9020-22, Session 5

Model-based, one-sided, time-of-flight terahertz image reconstruction

Stephen M. Schmitt, Jeffrey A. Fessler, Univ. of Michigan (United States); Greg D. Fichter, David A. Zimdars, Picometrix, LLC (United States)

In the last decade, terahertz-mode imaging has received increased attention for non-destructive testing applications due to its ability to penetrate many materials while maintaining a small wavelength. We propose a model-based reconstruction algorithm that is able to detect defects in the spray-on foam insulation (SOFI) used in aerospace applications that has been sprayed on a reflective metal hull. In this situation, X-ray based imaging is infeasible since only one side of hull is accessible in flight.

Prior work in model-based THz reconstruction has modeled the object of interest as a grid of point reflectors and matched the reflection strength to the received pulse. In our method, we model the object as a grid of materials, each section of which has a constant index of refraction. The delay between the transmission and reception of a THz pulse is related to the integral of the index of refraction along the pulse's path, and so we adapt methods for statistical reconstruction of computed tomography (CT) images to reconstruct our image. Specifically, we minimize a cost function consisting of a data-fit penalty and an edge-preserving regularizer that penalizes the first differences of neighboring pixels. Because the SOFI being imaged with our method is on the outside of a metal hull, we must also take into account that the path of a THz pulse reflects off of the hull. We do this by considering a virtual object reflected over the hull; the path of the pulse is then approximately a straight line through the object and then its virtual counterpart.

We present the results of our reconstruction method using real data of the timing of THz pulses passing through a block of SOFI with holes of a known location and radius layered on a reflective backplane. The holes (defects) are clearly visible in the reconstructed image of the refraction index map of the block.

9020-23, Session 6

Resolution enhancement and noise reduction using an array of cameras

Ibrahim E. Pekkucuksen, Umit Batur, Texas Instruments Inc. (United States)

Multi-camera arrays have been investigated for some time for different applications [1]. In this paper, we propose a multi-camera array algorithm that is designed to combine images from a set of closely packed set of cameras to obtain an output image with higher resolution and lower noise than any of the input images.

Our algorithm consists of several parts, each of which are crucial to get the desired output. The first step of the algorithm is camera calibration. We assume that the cameras in the array are parallel to one another

without any intentional toe-in angle. However, perfect physical calibration of multiple cameras is very difficult to achieve and maintain. So, there will always be some unintentional misalignment between the cameras that needs to be corrected. At the same time, not all the mismatch we observe between the camera views is a result of camera misalignment. Some of this mismatch is the disparity which is a function of the scene depth and the baseline between the cameras. We decouple the disparity from the misalignment using a modified perspective model, and initially only remove the camera misalignment while preserving the disparity between the views. The correspondence between the cameras is established by feature detection and matching without use of any special charts.

The second step of the algorithm is to locally estimate the disparity present between the views. We define one of the cameras as the reference camera and bring all the pixels from other views onto its grid by correcting for the estimated disparity. This is done by multi-level block search along the baseline direction, and then sub-pixel disparity estimation by curve fitting to block matching errors. By the end of this step, we have all the pixels from all the cameras lying on the reference grid.

The third and the final step of the algorithm is to interpolate a higher resolution output image from the nonuniform samples obtained from the cameras. These samples can be directly used for interpolation or they can be preprocessed to improve their fidelity by using other samples (from other cameras) that fell close to them. Resolution enhancement is achieved by sub-pixel placements of multiple samples coming from different cameras. We observed that Lanczos filtering is a good candidate for this task. Since many samples from a small area are combined to produce a single output pixel, we also observed a nice denoising effect without resorting to additional processing.

We simulated a camera array by capturing a still scene with a single camera from slightly different locations. We also tested our algorithm on several lightfield and super-resolution datasets with promising results.

[1] Stanford Multi-camera Array, http://lightfield.stanford.edu/acq.html

9020-24, Session 6

Fast edge-preserving image denoising via group coordinate descent on the GPU

Madison G. McGaffin, Jeffrey A. Fessler, Univ. of Michigan (United States)

Besides the ever-increasing pixel count on consumer electronics and increasingly large scientific image datasets, pure image denoising problems have appeared as components in medical image reconstruction algorithms, presenting problems with essentially hundreds of megapixels. The denoising operation in these applications is often are expressed as a constrained convex problem, and the “denoised” image is produced by running an iterative optimization algorithm. This process is computationally complex, and efficient and hardware-aware algorithms are essential.

In this work, we consider the penalized weighted least squares denoising problem with an edge-preserving regularizer and box constraints on the pixel values. The edge-preserving regularizer penalizes the differences between adjacent pixels through a possibly non-smooth penalty function. This formulation sufficiently general to handle Tikhonov regularization, anisotropic total variation (TV), and “corner-rounded” TV-like regularizers like the Huber or Fair potentials. We target the general-purpose graphics processing unit (GPGPU), which offers significant acceleration for algorithms performing large numbers of independent and identical calculations with highly structured memory access patterns.

Our algorithm divides the image into four strided groups of pixels, similar to a four-colored checkerboard. In inner iterations, we optimize the cost function over each of these groups sequentially. This is a group coordinate descent (GCD) algorithm, which is known to be convergent. By arranging the image pixels in memory so each group is contiguous (as opposed to interleaved in the natural ordering) we make the requisite memory accesses favorable for GPGPU implementation (“coalesced,” in NVIDIA’s terminology). This group structure causes each inner group optimization to further decompose into an independent 1D constrained optimization problem for each pixel in the group.

These subproblems are tackled with either a majorize-minimize approach in the case of a differentiable potential function or an inner iterative split-Bregman-like algorithm for nondifferentiable potential functions (e.g., the absolute value). In both cases the box constraints are trivial to apply.

In a preliminary experiment, we corrupted an approximately 32-megapixel subset of the Large Magellanic Cloud survey from NASA’s Swift project with white Gaussian noise. We used the differentiable Fair potential to penalize the difference between each pixel and its eight neighbors. We generated a reference image by running 1500 iterations of preconditioned conjugate gradients using a fast preconditioner. We computed per-iteration and -time convergence curves for conjugate gradients, preconditioned conjugate gradients, a separable quadratic surrogates-based descent algorithm, and the proposed group coordinate descent algorithm. No box constraints were enforced to allow a fair comparison with the second-order methods. All algorithms where implemented in OpenCL on an NVIDIA GTX 480. The proposed GCD algorithm reached within 2 gray levels (of 256) in RMSD to the converged solution in 0.957 seconds; the next fastest reached that level of convergence after nearly a minute. The propsed GCD algorithm exhibits acceleration beyond what could be easily accounted for by differences implementation efficiency, and is particularly well-suited to take advantage of massively-parallel architectures like GPGPUs.

9020-25, Session 6

Signal processing to improve imagery from compressive sensing cameras under unstable illumination

Donna Hewitt, Justin Fritz, Tyler Weston, James Tidman, Matt Herman, Lenore McMackin, InView Technology Corp. (United States)

Compressive sensing (CS) is a well-known technique for acquiring images from fewer measurements than typical Nyquist sampling requirements. There is a vast literature surrounding the mathematics and applications of CS, including studies on performance limitations. However, there have been no exhaustive investigations on the effects of noise on compressive imaging from a practical standpoint. As realized by the “single-pixel” architecture, CS is very sensitive to measurement noise. The signal for the CS single-pixel camera is not the average light level resulting from the projection of the scene onto the detector, as attained by traditional cameras, but the small variations between measurement patterns. Since these variations are a tiny fraction of the average values, noise on the measurements masks the signal in many instantiations.

In the application of shortwave infrared (SWIR) imaging, InView Technology has developed cameras utilizing the single-pixel architecture. Previously, we described our efforts to design and build low noise yet high speed detection, amplification and digitization circuitry for the camera. There is still a need to process the raw data. In this paper we describe the raw data processing that we have developed to enhance the imaging performance of the camera in the presence of measurement noise.

Measurement noise can occur in several modalities in the single-pixel camera. One such modality is due to instability in the illumination. CS camera data are acquired as a series of measurements over time, with the implicit assumption that neither the scene nor the illumination varies during the measurement time. Light source intensity changes corrupt the measurements with noise that is commonly an order of magnitude greater than the signal. To filter out these irregularities, we require knowledge of the temporal frequency of the fluctuations.

For light changes that occur slowly when compared to the CS measurement frequency, light fluctuations can be estimated and

removed using calibration frames. Calibration frames are reference frames that measure the scene illumination at various points in time, independent of the CS measurements. Using least squares, a best fit can be approximated for the rate of change of illumination across these calibration frames, or for slow enough changes, simple linear interpolation works well. Another method we discuss for low frequency light drift is a rolling mean calibration where successive groups of measurements are averaged to determine the mean light level from which deviations can then be found. For light level fluctuations of a specific frequency—such as 120 Hz incandescent lighting, the measurement fluctuations can be least-squares fit to a sinusoid or other pre-defined curve. Unknown lighting situations that may have a combination of light sources fluctuating over a wider frequency band require further characterization. In these cases, least mean square adaptive filtering techniques have shown much promise. We demonstrate our CS imaging results under a variety of light sources and discuss how the filtering processes are implemented in the single-pixel camera.

9020-27, Session 7

(JEI Invited) Adaptive compressive sensing algorithm for video acquisition using a single-pixel camera

Imama Noor, Univ. of Memphis (United States)

We propose a method to acquire compressed measurements for efficient video reconstruction using a single-pixel camera. The method is suitable for implementation using a single-pixel detector, along with a digital micromirror device or other types of spatial light modulators. Conventional implementations of single-pixel cameras are able to spatially compress the signal, but the compressed measurements make it difficult to exploit temporal redundancies directly. Moreover, a single-pixel camera needs to make measurements in a sequential manner before the scene changes, making it inefficient for video imaging. We discuss a measurement scheme that exploits sparsity along the time axis for video imaging. After acquiring all measurements required for the first frame, measurements are acquired only from the areas that change in subsequent frames. We segment the first frame, detect the magnitude and direction of change for each segment, and acquire compressed measurements for the changing segments in the predicted direction. Next, we compare the reconstruction results for a few test sequences with existing techniques and demonstrate the practical utility of the scheme.

9020-28, Session 7

Light field panorama by a plenoptic camera

Zhou Xue, Loïc Baboulaz, Paolo Prandoni, Martin Vetterli, Ecole Polytechnique Fédérale de Lausanne (Switzerland)

Plenoptic cameras, also known as light-field cameras, can capture both color and geometric information of a scene in a single shot. This is achieved by recording the intensity and direction of each light rays. These cameras enable many new imaging applications, including depth estimation, digital refocusing and perspective shift. Light-field acquisitions also provide new perspectives for many of the standard problems in image processing such as denoising, super-resolution and image stitching. The first consumer-grade plenoptic camera, the Lytro, has gathered a lot of interest since it became available in the market in 2011. It provides a compact and portable solution for the acquisition of the light field by adding a microlens array to a conventional camera. Its major limitation, however, is the low resolution of the rendered images: the camera's sensor is indeed used to record both the spatial and the angular information of the scene resulting in a rendered image with a resolution equal to that of the microlens array. In the angular domain, the range of the captured light directions is limited by the main lens settings.

To increase both the spatial and angular resolutions, we propose to use the plenoptic camera as an image scanner in order to align and stitch together multiple light field images; the novelty of our solution involves alignment and stitching directly in the light field domain. We begin by showing the sampling pattern of the plenoptic camera in the light field with the help of a simplified camera model which includes a thin lens and a moving pinhole camera behind the lens in continuous space. Given the complete light field of a scene, we derive which part such camera is able to capture in a shot and we describe the sampling pattern at the camera plane. We further characterize the bandwidth of the light field sample in continuous space when the plenoptic camera captures the light field of a smooth surface "painted" with band-limited signals then we define the transformations between the light field samples considering different intrinsic camera parameters such as focal length, focal plane, sensor size and lens size. We also derive how the light field is transformed when the plenoptic camera is translated or rotated.

Based on the above analysis we finally show that classic image alignment and stitching is a sub-problem of the more general stitching problem of light fields. We propose a feasible stitching framework to create a panorama light field image using multiple translations of a plenoptic camera along a plane parallel to the sensor plane. We apply a re-sampling operation on the aligned light field data and generate a set of panorama images from virtual cameras behind the original camera plane with different perspective shifts. These images compose a new light field image with both increased spatial and angular information using virtual intrinsic and extrinsic camera parameters. Simulation results on synthetic data from a virtual plenoptic camera also support our model and analysis.

9020-29, Session 7

Efficient volumetric estimation from plenoptic data

Paul Anglin, Stanley J. Reeves, Brian S Thurow, Auburn Univ (United States)

The commercial release of the Lytro camera, and greater availability of plenoptic imaging systems in general, have given the image processing community cost-effective tools for light-field imaging. While this data is most commonly used to generate planar images at arbitrary focal depths, reconstruction of volumetric fields is also possible. Similarly, deconvolution is a technique that is conventionally used in planar image reconstruction, or deblurring, algorithms. However, when leveraged with the ability of a light-field camera to quickly reproduce multiple focal planes within an imaged volume, deconvolution offers a computationally efficient method of volumetric reconstruction. Related research has shown than light-field imaging systems in conjunction with tomographic reconstruction techniques are also capable of estimating the imaged volume, and have been successfully applied to particle image velocimetry (PIV). However, while tomographic volumetric estimation through algorithms such as multiplicative algebraic reconstruction techniques (MART) have proven to be highly accurate, they are computationally intensive. In this paper, the reconstruction problem is shown to be solvable by deconvolution. Deconvolution offers significant improvement in computational efficiency through the use of fast Fourier transforms (FFTs) when compared to other tomographic methods. This work describes a deconvolution algorithm designed to reconstruct a 3-D particle field from simulated plenoptic data. A 3-D extension of existing 2-D FFT-based refocusing techniques is presented to further improve efficiency when computing object focal stacks and system point spread functions (PSF). Reconstruction artifacts are identified; their underlying source and methods of mitigation are explored where possible, and reconstructions of simulated particle fields are provided.

9020-30, Session 7

Computationally efficient background subtraction in the light field domain

Alireza Ghasemi, Mahdad Hosseinikamal, Martin Vetterli, Ecole Polytechnique Fédérale de Lausanne (Switzerland)

In this paper we present a novel approach for depth estimation and background subtraction in light field images. our approach exploits the regularity and the internal structure of the light field signal in order to extract an initial depth map of the captured scene and uses the extracted depth map as the input to a final segmentation algorithm which finely isolates the background in the image.

Light field or plenoptic imaging has attracted increased attention in both research and consumer community in recent years. This exponentially growing popularity stems mostly from the introduction of the consumer-targeted light field cameras Lytro and Raytrix. In light field imaging multiple views of the scene of interest are captured using the directional information of the light rays incident on the sensing device. In color imaging in contrast, only intensities of light rays are captured on the sensor cells.

The advanced sensing capability of a light field camera enables capturing much more information from a scene than traditional pinhole ones. For example information such as depth map and occlusion ordering can be extracted using the information inherent in the light field signal. Therefore, we can think of light field imaging as a method to recover the information which are lost during the color image formation process in pinhole cameras which involves projecting the real-world three-dimensional scene into a two-dimensional image plane.

The extra information content of a light field image can be very efficiently exploited in order to improve many of the current computer vision algorithms. Computer vision tasks such as segmentation, depth map estimation and three-dimensional reconstruction can highly benefit light field information for performance improvement. This requires adaptation of current algorithms to work in the light field space or develop new methods from scratch.

Background subtraction is natural application of the light field information since it is highly involved with depth information and segmentation. However many of the approaches proposed so far are not optimized specifically for background subtraction and are highly computationally expensive. Here we propose an approach based on a modified version of the well-known Radon Transform and not involving massive matrix calculations. It is therefore computationally very efficient and appropriate for real-time use.

Our approach exploits the structured nature of the light field signal and the information inherent in the plenoptic space in order to extract an initial depth map and background model of the captured scene. We apply a modified the Radon transform and the gradient operator to horizontal slices of the light field signal to infer the initial depth map. The initial depth estimated are further refined to a precise background using a series of depth thresholding and segmentation in ambiguous areas.

We test on method on various types real and synthetic of light field images. Scenes with different levels of clutter and also various foreground object depth have been considered in the experiments. The results of our experiments show much better computational complexity while retaining comparable performance to similar more complex methods.

Conference 9021: Document Recognition and Retrieval XXI

Wednesday - Thursday 5 –6 February 2014

Part of Proceedings of SPIE Vol. 9021 Document Recognition and Retrieval XXI

9021-24, Session PWed

Two-stage approach to keyword spotting in handwritten documents

Mehdi Haji, IMDS Software (Canada) and Concordia Univ. (Canada); Mohammad R. Ameri, Concordia University (Canada); Tien D. Bui, Ching Y. Suen, Concordia Univ. (Canada); Dominique Ponson, IMDS Software (Canada)

Separation of keywords from non-keywords is the main problem in keyword spotting systems which has traditionally been approached by simplistic methods, such as thresholding of recognition scores. In this paper, we analyze this problem from a machine learning perspective, and we study several standard machine learning algorithms specifically in the context of non-keyword rejection. We propose a two-stage approach to keyword spotting and provide a theoretical analysis of the performance of the system which gives insights on how to design the classifier in order to maximize the overall performance in terms of F-measure.

9021-25, Session PWed

Extraction and labeling high-resolution images from PDF documents

Suchet K. Chachra, Zhiyun Xue, Sameer Antani, Dina Demner-Fushman, George R. Thoma, National Library of Medicine (United States)

Accuracy of content-based image retrieval is affected by image resolution among other factors. Higher resolution images enable extraction of image features that more accurately represent the image content. In order to improve the relevance of search results for our biomedical image search engine, Open-I, we have developed techniques to extract and label high-resolution versions of figures from biomedical articles supplied in the PDF format. Open-I uses the open-access subset of biomedical articles from PubMed Central repository hosted by the National Library of Medicine. Articles are available in XML and in publisher supplied PDF formats. As these PDF documents contain little or no meta-data to identify the embedded images, the task includes labeling images according to their figure number in the article after they have been successfully extracted. For this purpose we use the labeled small size images provided with the XML web version of the article. This paper describes the image extraction process and two alternative approaches to perform image labeling that measure the similarity between two images based upon the image intensity projection on the coordinate axes and similarity based upon the normalized cross-correlation between the intensities of two images. Using image identification based on image intensity projection, we were able to achieve a precision of 92.84% and a recall of 82.18% in labeling of the extracted images.

9021-26, Session PWed

Structure analysis for plane geometry figures

Tianxiao Feng, Xiaoqing Lu, Lu Liu, Keqiang Li, Zhi Tang, Peking Univ. (China)

As there are increasing numbers of digital documents for education purpose, we realize that there is not a retrieval application for mathematic plane geometry images. In this paper, we propose a method for retrieving plane geometry figures (PGFs), which often appear in geometry books and digital documents. First, detecting algorithms are applied to detect common basic geometry shapes from PGF image. Based on all basic shapes, we analyze the structural relationships between two basic shapes and combine some of them to a compound shape to build the PGF descriptor. Afterwards, we apply matching function to retrieve candidate PGF images with ranking. The great contribution of the paper is that we propose a structure analysis method to better describe the spatial relationships in such image composed of many overlapped shapes. Experimental results demonstrate that our analysis method and shape descriptor can obtain good retrieval results with relatively high effectiveness and efficiency.

9021-27, Session PWed

On-line signature verification method by Laplacian spectral analysis and dynamic time warping

Changting Li, Liangrui Peng, Changsong Liu, Xiaoqing Ding, Tsinghua Univ. (China)

As smartphones and touch screens are more and more popular, on-line signature verification technology can be used as one of personal identification means for mobile computing. In this paper, a novel Laplacian Spectral Analysis (LSA) based on-line signature verification method is presented and an integration framework of LSA and Dynamic Time Warping (DTW) based methods for practical application is proposed. In LSA based method, a Laplacian matrix is constructed by regarding the on-line signature as a graph. The signature's writing speed information is utilized in the Laplacian matrix of the graph. The eigenvalue spectrum of the Laplacian matrix is analyzed and used for signature verification. The framework to integrate LSA and DTW methods is further proposed. DTW is integrated at two stages. First, it is used to provide stroke matching results for the LSA method to construct the corresponding graph better. Second, the on-line signature verification results by DTW are fused with that of the LSA method. Experimental results on public signature database and practical signature data on mobile phones proved the effectiveness of the proposed method.

9021-28, Session PWed

A slant removal technique for document page

Ergina Kavallieratou, Univ. of the Aegean (Greece)

The slant removal is a necessary preprocessing task in many document image processing systems. In this paper, we describe a technique for removing the slant from the entire page, avoiding the segmentation procedure. The presented technique could be combined with the most existed slant removal algorithms. Experimental results are presented on two databases.

9021-30, Session PWed

Robust binarization of degraded document images using heuristics

Jon Parker, Georgetown Univ. (United States) and Johns Hopkins Univ. (United States); Ophir Frieder, Gideon Frieder, Georgetown Univ. (United States)

Historically significant documents are often discovered with defects that make them difficult to read and analyze. This fact is particularly troublesome if the defects prevent software from performing an automated analysis. Image enhancement methods are used to remove

or minimize document defects, improve software performance, and generally make images more legible. We describe an automated, image enhancement method that is input page independent and requires no training data. The approach applies to color or greyscale images with hand written script, typewritten text, images, and mixtures thereof. We evaluated the image enhancement method against the test images provided by the 2011 Document Image Binarization Contest (DIBCO). Our method outperforms all 2011 DIBCO entrants in terms of average F1 measure – doing so with a significantly lower variance than top contest entrants. The capability of the proposed method is also illustrated using select images from a collection of historic documents stored at Yad Vashem Holocaust Memorial in Israel.

9021-31, Session PWed

A machine learning based lecture video segmentation and indexing algorithm

Di Ma, Bingqing Xie, Gady Agam, Illinois Institute of Technology (United States)

Video segmentation and indexing are important steps in multimedia document understanding and information retrieval. This paper presents a novel machine learning based approach for automatic structuring and indexing of lecture videos. By indexing video content, we can support both topic indexing and semantic querying of multimedia documents. In this paper, our proposed approach extracts features from video images and then uses these features to construct a model to label video frames. Using this model, we are able to segment and indexing videos with accuracy of 95% on our test collection.

9021-1, Session 1

Writer identification on historical Glagolitic documents

Stefan Fiel, Fabian Hollaus, Melanie Gau, Robert Sablatnig, Technische Univ. Wien (Austria)

This work aims at automatically identifying scribes of historical Slavonic manuscripts. The quality of the ancient documents is partially degraded by faded-out ink or varying background. The writer identification method used is based on textual features, which are described with Scale Invariant Feature Transform (SIFT) features. A visual vocabulary is used for the description of handwriting characteristics, whereby the features are clustered using a Gaussian Mixture Model and employing the Fisher kernel. The writer identification approach is originally designed for grayscale images of modern handwritings. But contrary to modern documents, the historical manuscripts are partially corrupted by background clutter and water stains. As a result, SIFT features are also found on the background. Since the method shows also good results on binarized images of modern handwritings, the approach was additionally applied on binarized images of the ancient writings. Experiments show that this preprocessing step leads to a significant performance increase: The identification rate on binarized images is 98.9%, compared to an identification rate of 87.6% gained on grayscale images.

9021-2, Session 1

Probabilistic modeling of children's handwriting

Mukta Puri, Sargur N. Srihari, Univ. at Buffalo (United States); Lisa Hanson, Bureau of Criminal Apprehension Laboratory (United States)

There is little work done in the analysis of children's handwriting, which can be useful in developing automatic evaluation systems and in quantifying handwriting individuality. We consider the statistical analysis of children's handwriting in early grades.

Samples of handwriting of children in Grades 2-4 who were taught the Zaner-Bloser style were considered. The commonly occurring word "and" written in cursive style as well as hand-print were extracted from extended writing.The samples were assigned feature values by human examiners using a truthing tool.

Changes were measured using a feature space distance measure. Results indicate that the handwriting develops towards more conformity with Zaner-Bloser. Bayesian networks were also learnt from the data to enable answering various probabilistic queries, such as determining grade based on handwriting.

9021-3, Session 1

Variational dynamic background model for keyword spotting in handwritten documents

Gaurav Kumar, Univ. at Buffalo (United States); Safwan Wshah, Xerox Corp. (United States); Venu Govindaraju, Univ. at Buffalo (United States)

We propose a bayesian framework for keyword spotting in handwritten documents. This work is an extension to our previous work where we proposed dynamic background model, DBM for keyword spotting that takes into account the local character level scores and global word level scores to learn a logistic regression classifier to separate keywords from non-keywords. In this work, we add a bayesian layer on top of the DBM called the variational dynamic background model, VDBM. The logistic regression classifier uses the sigmoid function to separate keywords from non-keywords. The sigmoid function being neither convex nor concave, exact inference of VDBM becomes intractable. An expectation maximization step is proposed to do approximate inference. The advantage of VDBM over the DBM is multi-fold. Firstly, being bayesian, it prevents over-fitting of data. Secondly, it provides better modeling of data and an improved prediction of unseen data. VDBM is evaluated on the IAM dataset and the results prove that it outperforms our prior work and other state of the art line based word spotting system.

9021-4, Session 1

Boosting bonsai trees for handwritten/printed text discrimination

Yann Ricquebourg, Christian Raymond, Baptiste Poirriez, Aurélie Lemaitre, Bertrand Coüasnon, IRISA / INRIA Rennes (France)

This article deals with the use of bonzaiboost, implementing small trees boosting, in image classification process- ing. This algorithm proved its efficiency in Natural Language Processing essentially with symbolic features, and its good properties could be of great interest in the numeric world of pixel images. We designed a first system in the context of handwritten/printed text discrimination. Then, we conducted experiments to compare it to usual SVM-based classification revealing convincing results with very close performance, but with faster predictions and behaving far less as a black-box. Those promising results tend to make use of this classifier in more complex recognition tasks like multiclass problems.

9021-5, Session 2

Form similarity via Levenshtein distance between ortho-filtered logarithmic ruling-gap ratios

George Nagy, Rensselaer Polytechnic Institute (United States); Daniel P. Lopresti, Lehigh Univ. (United States)

Geometric invariants are combined with edit distance to compare the ruling configuration of noisy filled-out forms. Gap-ratios used as features capture most of the ruling information of even low-resolution and poorly scanned form images The Hough Transform on edgelets, combined with ortho-normal filtering, was used for for detecting horizontal and vertical rulings in tables or forms. The radial components of the resulting line coordinates are invariant to rotation. The ratios of distances between parallel rulings are invariant to translation and scaling, and can be calculated efficiently from the radius-angle representation provided by the Hough Transform. The N-2 ratios of N-1 adjacent intervals offer a complete representation of the relative positions of N parallel lines. Logarithmic scaling of the ratios allows uniform quantization for the conversion of the numeric values of the ratios to strings of symbols. String comparison, one of the few effective methods for classifying objects with a variable number of features, was chosen to provide a similarity measure for pairs of forms. The edit distance is tolerant of missed and spurious rulings. Since the method was developed for an application that precludes public dissemination of the data, it is illustrated on public-domain death certificates.

9021-6, Session 2

Form classification and retrieval using bag of words with shape features of line structures

Florian Kleber, Markus Diem, Robert Sablatnig, Technische Univ. Wien (Austria)

In this paper a document form classification and retrieval method using Bag of Words and local shape features of form lines is proposed. In a preprocessing step the document is binarized and the form lines (solid and dotted) are detected. Shape features based on the line information describe local line structures, e.g. line endings, crossings, boxes. The dominant line structures build a vocabulary for each form class. According to the vocabulary an occurence histogram of structures of form documents can be calculated for the classification and retrieval. The proposed method has been tested on a set of 489 documents and 9 different form classes.

9021-7, Session 3

OCR for Google Books *(Invited Paper)*

Ray Smith, Ashok Popat, Google (United States)

No Abstract Available

9021-8, Session 4

Utilizing web data in identification and correction of OCR errors

Kazem Taghva, Shivam Agarwal, Univ. of Nevada, Las Vegas (United States)

In this paper, we report on our experiments for detection and correction of OCR errors with web data. More specifically, we utilize Google search to access the big data resources available to identify possible candidates for correction. We then use a combination of the Longest Common Subsequences (LCS) and Bayesian estimates to

automatically pick the proper candidate.

Our experimental results on a small set of historical newspaper data show a recall and precision of 51% and 100%, respectively. The work in this paper further provides a detailed classification and analysis of all errors. In particular, we point out the shortcomings of our approach in its ability to suggest proper candidates to correct the remaining errors.

9021-9, Session 4

How well does multiple OCR error correction generalize?

William B. Lund, Eric K. Ringger, Brigham Young Univ. (United States); Daniel D. Walker, Microsoft Corp. (United States)

As the digitization of historical documents, such as newspapers, becomes more common, the archive patron's need for accurate digital text from those documents increases. The contributions of this paper are 1. in demonstrating the applicability of novel methods for improving optical character recognition (OCR) on disparate data sets, including a new synthetic training set, 2. enhancing the correction algorithm with novel features, and 3. assessing the data requirements of the correction learning method. First, we correct errors using conditional random fields (CRF) trained on synthetic training data sets in order to demonstrate the applicability of the methodology to unrelated test sets. Second, we show the strength of lexical features from the training sets on two unrelated test sets, yielding an absolute reduction in word error rate on the test sets of 6.52%. New features capture the recurrence of hypothesis tokens and yield an additional absolute reduction in WER of 2.30%. Further, we show that only 2.0% of the full training corpus of over 500,000 feature cases is needed to achieve correction results comparable to those using the entire training corpus, effectively reducing both the complexity of the training process and the learned correction model.

9021-10, Session 4

Video text localization using wavelet and shearlet transforms

Purnendu Banerjee, Bidyut B. Chaudhuri, Indian Statistical Institute (India)

Text in video is useful and important in indexing and retrieving the video documents efficiently and accurately. In this paper, we present a new method of text detection using a combined dictionary consisting of wavelets and a recently introduced transform called shearlets. Wavelets provide optimally sparse expansion for point-like structures and shearlets provide optimally sparse expansions for curve-like structures. By combining these two features we have computed a high frequency sub-band to brighten the text part. Then K-means clustering is used for obtaining text pixels from the Standard Deviation (SD) of combined coefficient of wavelets and shearlets as well as the union of wavelets and shearlets features. Text parts are obtained by grouping neighboring regions based on geometric properties of the classified output frame of unsupervised K-means classification. The proposed method tested on a standard as well as newly collected database shows to be superior to some existing methods.

9021-11, Session 5

A Markov chain based line segmentation framework for handwritten character recognition

Yue Wu, Shengxin Zha, Huaigu Cao, Daben Liu, Raytheon BBN Technologies (United States); Premkumar Natarajan, The Univ. of Southern California (United States)

In this paper, we present a novel text line segmentation framework following the divide-and-conquer paradigm: we iteratively identify and re-process regions of ambiguous line segmentation from an input document image until there is no ambiguity. Two complimentary methods of line detection, referred to as the underline and highlight line detectors, are implemented. Both line detectors are modeled as Markov Chains, and thus good candidates can be identified via the Viterbi algorithm. Regions of ambiguous line segmentation are located by examining the pattern of lines detected by the two types of line detectors. As a result, we can easily identify already good line segmentations, and largely simplify the original line segmentation problem by only reprocessing ambiguous regions. We evaluated our line segmentation algorithm using the ICDAR 2009 handwritten document dataset showed that the performance of the proposed method is close to top-performing systems submitted to the competition. Our results show that our algorithm is robust against skewness, noise, variable line heights and touching characters. Our idea can also be applied to other text analysis tasks such as word segmentation and page layout analysis.

9021-12, Session 5

Handwritten text segmentation using blurred image

Aurélie Lemaitre, Jean Camillerapp, IRISA / INRIA Rennes (France); Bertrand Coüasnon, IRISA - INSA (France)

In this paper, we present our new method for the segmentation of handwritten text pages into lines, which has been submitted to ICDAR'2013 handwritten segmentation competition. This method is based on two levels of perception of the image: a rough perception based on a blurred image, and a precise perception based on the presence of connected components.

The combination of those two levels of perception enables to deal with the difficulties of handwritten text segmentation: curvature, irregular slope and overlapping strokes. Thus, the analysis of the blurred image is efficient in images with high density of text, whereas the use of connected components enables to connect the text lines in the pages with low text density. The combination of those two kinds of data is implemented with a grammatical description, which enables to externalize the knowledge linked to the page model. The page model contains a strategy of analysis that can be associated to an applicative goal. Indeed, the text line segmentation is linked to the kind of data that is analysed: homogeneous text pages, separated text blocks or unconstrained text. This method obtained a recognition rate of more than 98% on last ICDAR'2013 competition.

9021-13, Session 6

What makes big visual data hard? *(Invited Paper)*

Alexei (Alyosha) Efros, Univ. of California, Berkeley (United States)

No Abstract Available

9021-14, Session 7

Optical music recognition on the International Music Score Library Project

Christopher S. Raphael, Rong Jin, Indiana Univ. (United States)

A system is presented for optical recognition of music scores. The system processes a document page in three main phases.

First it performs a hierarchical decomposition of the page, identifying systems, staves and measures. The second phase, which forms the heart of the system, interprets each measure found in the previous phase as a collection of non-overlapping symbols including both primitive symbols (clefs, rests, etc.) with fixed templates, and composite symbols (chords, beamed groups, etc.) constructed through grammatical composition of primitives (note heads, ledger lines, beams, etc.). This phase proceeds by first building separate top-down recognizers for the symbols of interest. Then, it resolves the inevitable overlap between the recognized symbols by exploring the possible assignment of overlapping regions, seeking globally optimal and grammatically consistent explanations.

The third phase interprets the recognized symbols in terms of pitch and rhythm, focusing on the main challenge of rhythm. We present results that compare our system to the leading commercial OMR system using MIDI ground truth for piano music.

9021-15, Session 7

Document flow segmentation for business applications

Daher Hani, Abdel Belaïd, LORIA (France) and Univ. de Lorraine (France)

The aim of this paper is to propose a document flow supervised segmentation approach applied to real world heterogeneous documents. Our algorithm treats the flow of documents as couples of consecutive pages, and studies the relationship that exists between them. At first, sets of features are extracted from the pages where we propose an approach to model the couple of pages into a single feature vector representation. This representation will be provided to a binary classifier which classifies the relationship as either segmentation or continuity. In case of segmentation, we consider that we have a complete document and the analysis of the flow continues by starting a new document. In case of continuity, the couple of pages are assimilated to the same document and the analysis continues on the flow. If there is an uncertainty on whether the relationship between the couple of pages should be classified as a continuity or segmentation, a rejection is decided and the pages analyzed until this point are considered as a "fragment". The first classification already provides good results approaching 90% on certain documents, which is high at this level of the system.

9021-16, Session 7

LearnPos: a new tool for interactive learning positioning

Cérès Carton, Aurélie Lemaitre, Bertrand Coüasnon, IRISA / INRIA Rennes (France)

The analysis of 2D structured documents often requires localizing data inside of a document during the recognition process. In this paper we present LearnPos a new generic tool, independent of any document recognition system. LearnPos models and evaluates positioning from a learning set of documents. Thanks to LearnPos, the user is helped to define the physical structure of the document. He then can concentrate his efforts on the definition of the logical structure of the documents. LearnPos is able to furnish spatial information for both absolute and

relative spatial relations, in interaction with the user. Our method can handle spatial relations compose of distinct zones and is able to furnish appropriate order and point of view to minimize errors. We prove that resulting models can be successfully used for structured document recognition, while reducing the manual exploration of the data set of documents.

9021-17, Session 7

Document page structure learning for fixed-layout e-books using conditional random fields

Xin Tao, Zhi Tang, Canhui Xu, Peking Univ. (China)

In this paper, a CRF based model is proposed to learn document page structure by combining support vector machines(SVMs) and conditional random fields(CRFs). Features related to each logical label and their dependencies are extracted from various PDF original attributes. Both local evidence and contextual dependencies are integrated in the proposed model so as to achieve better logical labeling performance. With the merits of SVM as local discriminative classifier and CRF modeling contextual correlations of adjacent fragments, it is capable of resolving the ambiguities of semantic labels. The experimental results show that CRF based models with both tree and chain graph structures outperform the SVM model with an increase of macro-averaged F1 by about 10%.

9021-18, Session 7

Automatic comic page image understanding based on edge segment analysis

Dong Liu, Yongtao Wang, Zhi Tang, Luyuan Li, Liangcai Gao, Peking Univ. (China)

Comic page image understanding aims to analyse the layout of the comic page images by detecting the storyboards and identifying the reading orders automatically. It is the key technique to produce the digital comic documents suitable for reading on mobile devices. In this paper, we propose a novel comic page image understanding method based on edge segment analysis. First, we propose an efficient edge point chaining method to extract Canny edge segments (i.e., contiguous chains of Canny edge points) from the input comic page image; second, we propose a top-down scheme to detect line segments within each obtained edge segment; third, we develop a novel method to detect the storyboards by selecting the border lines and identify the reading the orders. The proposed method is performed on a data set consisting of 2000 comic page images from ten printed comic series. The experimental results demonstrate that the proposed method achieves satisfactory results on different comics and outperforms the existing methods.

9021-19, Session 8

Scalable ranked retrieval using document images

Rajiv Jain, Douglas Oard, David Doermann, Univ. of Maryland, College Park (United States)

Despite the explosion of text on the Internet, hard copy documents that have been scanned as images still play a significant role for some tasks. The best method to perform ranked retrieval on a large corpus of document images, however, remains an open research question. The most common approach has been to perform text retrieval using terms generated by optical character recognition. This paper, by contrast, examines whether a scalable segmentation-free image retrieval algorithm, which matches sub-images containing text or graphical objects, can provide additional benefit in satisfying a user's information needs on a large, real world dataset. Results on 7 million scanned pages from the CDIP v1.0 test collection show that content based image retrieval finds a substantial number of documents that text retrieval misses, and that when used as a basis for relevance feedback can yield improvements in retrieval effectiveness.

9021-20, Session 8

A contour-based shape descriptor for biomedical image classification and retrieval

Daekeun You, Sameer Antani, Dina Demner-Fushman, George R. Thoma, National Library of Medicine (United States)

Contours, object blobs, and specific feature points are utilized to represent object shapes and extract shape descriptors that can then be used for object detection or image classification. In this research we develop a shape descriptor for biomedical image type (or, modality) classification. We adapt a feature extraction method used in optical character recognition (OCR) for character shape representation, and apply various image preprocessing methods to successfully adapt the method to our application. The proposed shape descriptor is applied to radiology images (e.g., MRI, CT, ultrasound, X-ray, etc.) to assess its usefulness for modality classification. In our experiment we compare our method with other visual descriptors such as CEDD, CLD, Tamura, and PHOG that extract color, texture, or shape information from images. The proposed method achieved the highest classification accuracy of 74.1% among all other individual descriptors in the test, and when combined with CSD (color structure descriptor) showed better performance (78.9%) than using the shape descriptor alone.

9021-21, Session 8

Semi-automated document image clustering and retrieval

Markus Diem, Florian Kleber, Stefan Fiel, Robert Sablatnig, Technische Univ. Wien (Austria)

In this paper a semi-automated document image clustering and retrieval is presented to create links between different documents based on their content. Ideally the initial bundling of shuffled document images can be reproduced to explore large document databases. Structural and textural features, which describe the visual similarity, are extracted and used by experts (e.g. registrars) to interactively cluster the documents with a manually defined feature subset (e.g. checked paper, handwritten). The methods presented allow for the analysis of heterogeneous documents that contain printed and handwritten text and allow for a hierarchically clustering with different feature subsets in different layers.

9021-22, Session 8

Fast structural matching for document image retrieval through spatial databases

Hongxing Gao, Maçal Rusiñol, Dimosthenis Karatzas, Josep Lladós, Univ. Autònoma de Barcelona (Spain)

The structure of document images plays a significant role in document analysis thus overflowing efforts have been made such as layout analysis, which is computationally expensive. In this paper, we first employ Distance Transform based MSER (DTMSER) to efficiently extract the document structure in terms of a dendrogram of semantic key-regions (letters, words, paragraphs). Then a fast structural matching method is proposed to query the structure of document (dendrogram) based on

spatial database which carries advanced techniques. The experiment demonstrate remarkable improvement of our proposed method over BoW and pyramidal spatial BoW.

9021-23, Session 9

The Lehigh Steel Collection: a new open dataset for document recognition research

Barri Bruno, Daniel P. Lopresti, Lehigh Univ. (United States)

Document image analysis is a data-driven discipline. For a number of years, research was focused on small, homogeneous datasets such as the University of Washington corpus of scanned journal pages. More recently, library digitization efforts have raised many interesting problems with respect to historical documents and their recognition. In this paper, we present the Lehigh Steel Collection (LSC), a new open dataset we are currently assembling which will be, in many ways, unique to the field. LSC is an extremely large, heterogeneous set of documents dating from the 1960's through the 1990's relating to the wide-ranging research activities of Bethlehem Steel, a now-bankrupt company that was once the second-largest steel producer and the largest shipbuilder in the United States. As a result of the bankruptcy process and the disposition of the company's assets, an enormous quantity of documents (we estimate hundreds of thousands of pages) were left abandoned in buildings recently acquired by Lehigh University. Rather than see this history destroyed, we stepped in to preserve a portion of the collection via digitization. Here we provide an overview of LSC, including our efforts to collect and scan the documents, a preliminary characterization of what the collection contains, and our plans to make this data available to the research community.

Conference 9022: Image Sensors and Imaging Systems 2014

Wednesday - Thursday 5 –6 February 2014

Part of Proceedings of SPIE Vol. 9022 Image Sensors and Imaging Systems 2014

9022-27, Session PWed

(JEI Invited) Optical imaging of high frequency focused ultrasonic field using a Nomarski interferometer coupled with multichannel lock-in detection *(Invited Paper)*

Smain Femmam, Univ. de Haute Alsace (France)

In this paper we present a non-destructive optical technique for quantitatively imaging focused ultrasonic field in continuous-wave mode within clear media. We use a Nomarski interferometer configuration coupled with a CCD array and multichannel lock-in detection. 2-D full field image of pressure amplitude can be obtained in only one acquisition. This method was experimentally demonstrated with a high frequency ultrasonic transducer, and experiment measurements of acoustic pressure which have been obtained at 27 MHz. This technique is very interesting because it doesn't introduce any disturbance in the ultrasonic field and doesn't need any scanning mechanisms to acquire images. It can be useful for probing high frequency acoustic fields and for calibrating HF-piezo-electric transducers.

9022-28, Session PWed

Iterative compressive sampling for hyperspectral images via source separation

Simeon Kamdem Kuiteing, Mauro Barni, Univ. degli Studi di Siena (Italy)

Compressive Sensing (CS) is gaining more interest as a way to lower measurement time, storage volume and compression requirements for on-board acquisition of remote-sensing images. In the case of HyperSpectral Images (HSI) the huge amount of data acquired by conventional sensors creates significant handling problems on satellites or aircrafts because the compression process must usually be performed in a computationally constrained environment. In literature, different approaches to reconstruct HSI based on CS theory have been investigated. The main difficulty of these methods is that the complexity of the reconstruction stage is relatively high for practical applications, as it is cubic in the number of samples. This problem becomes computationally intractable when the CS approach has to be applied to multidimensional signals.

To overcome this dilemma, Vandergheynst et al.* proposed to exploit the spatial correlation typical of HSI to develop a CS reconstruction scheme based on source separation (SCS) which relies on a linear mixture model to generate the data cube as the product of independent sources images with their corresponding spectral vector, where the spectral vectors are supposed to be known. Instead of recovering the whole data cube, this scheme directly applies the CS framework on each source image separately allowing to decrease considerably the number of measurements to be sent to the decoder and consequently its computational complexity. Despite its good performance in terms of overall computation time, the SCS method achieves an MSE that is not small enough for many HSI because it performs 2D spatial CS reconstruction on each source independently, so failing to exploit the correlation along the spectral dimension.

In order to simultaneously exploit the dependency among all the three dimensions of the data cube and to reduce the computational complexity, we propose to use a prediction-based iterative CS reconstruction algorithm which employs a linear correlation model to describe inter-band correlation, and iteratively apply this model to the output of the CS reconstruction algorithm applied independently to each band. At each iteration, each band f_i of the whole datacube reconstructed by the SCS scheme is predicted by applying a prediction operator π to reconstructed channels f_(i-1) and f_(i+1). The measurement of the predicted band acquired with the same sensing matrix used to acquire that band, is subtracted from the measurement of the band itself and the CS reconstruction is only applied to the measurement of the prediction error which we use to reconstruct the signal as the sum of the predicted band with the CS reconstruction of the prediction error. This process is performed on all bands, and is iterated until convergence. As a starting point of the iterative procedure we take the HSI reconstructed by applying the SCS scheme by Vandergheynst et al.

In this way, we combine the advantages of the SCS algorithm (in term of low computational complexity) and the iterative procedure (in terms of MSE reduction and improved reconstruction quality). A number of simulations show that the proposed approach allows achieving good performance in terms of MSE values while decreasing significantly the computational complexity with respect to conventional reconstruction methods.

* M. Golbabaee, S. Arberet, and P. Vandergheynst, "Multichannel compressed sensing via source separation for hyperspectral images," Proc. in Eusipco, 2010.

9022-29, Session PWed

Multiple object tracking and behavior analysis method for video surveillance applications

Jie Su, Harbin Univ. of Science and Technology (China)

In order to realize efficient multiple object detection and tracking, an efficient visual detection and tracking method is proposed, which meets the requirements of unsupervised working conditions and is based on a novel Bayesian tracking model by using the ability of managing multi-modal distributions without explicitly computing the association between tracked objects and detections. Behavior analysis method is proposed and can be used as the direct of objects moving trend. The proposed algorithm can be proved to be robust to erroneous, distorted and missing detections. Its superior performance1 can be proved compared with the formal work.

9022-30, Session PWed

Hyperspectral imaging applied to end-of-life concrete recycling

Silvia Serranti, Giuseppe Bonifazi, Univ. degli Studi di Roma La Sapienza (Italy)

The recycling of end-of-life concrete into new concrete is one of the most interesting options for reducing worldwide natural resources use and emissions associated with the building materials sector. The possibility to realize a larger re-use of aggregates from old concrete, can strongly contribute to reduce environmental impact (lower exploitation of natural resources, reduction of CO2 emissions, airborne dust production, etc.).

In this paper a new technology, based on HyperSpectral Imaging (HSI) sensors, and related detection architectures, is investigated in order to develop suitable and low cost strategies addressed to: i) preliminary detection and characterization of the composition, in terms of constituting material characteristics and assessment, of the structure to dismantle and ii) definition and implementation of innovative smart detection engines for sorting and/or demolition waste flow stream

quality control, that is to perform recovered materials and/or products certification.

The proposed sensing architecture is fast, accurate, affordable and it can strongly contribute to bring down the economic threshold above which recycling is cost efficient.

Investigations have been carried out utilizing two different HSI devices working in different wavelength ranges: i) NIR Spectral Camera™, embedding an ImSpector™ N17E (SPECIM Ltd, Finland) acting in the range 1000-1700 nm, ii) SisuCHEMA XL? Chemical Imaging workstation (SPECIM Ltd, Finland), a complete and high speed HSI system operating in the SWIR region (1000-2500 nm).

Spectral data analysis was carried out utilizing the PLS_Toolbox (Version 6.5.1, Eigenvector Research, Inc.) running inside Matlab® (Version 7.11.1, The Mathworks, Inc.), applying different chemometric techniques, selected depending on the materials under investigation.

The developed procedure allows to assess the characteristics, in terms of materials identification, such as recycled aggregates and related contaminants, as resulting from end-of-life concrete processing. A good classification of the different classes of material was obtained, being the model able to distinguish aggregates from other materials (i.e. glass, plastic, tiles, paper, cardboard, wood, brick, gypsum, etc.).

Results showed as the proposed technology is particularly suitable to analyse quality and characteristics of concrete before/after dismantling and of the products obtained by the recycling process. Furthermore, results also demonstrated the great potentiality of the HSI as an automated tool for recognition/classification of different classes of DW materials. They are of particular interest especially with reference to secondary raw materials, where expensive and sophisticated control architectures cannot be often adopted both for technical (e.g. particles of different size, shape and composition) and economic reasons.

9022-31, Session PWed

A indirect time-of-flight measurement technique for sub-mm range resolution using impulse photocurrent response

Takahiro Usui, Keita Yasutomi, Sangman Han, Taishi Takasawa, Keiichiro Kagawa, Shoji Kawahito, Shizuoka Univ. (Japan)

To apply Time-of-Flight (TOF) range imagers for new applications such as 3D scanner, sub-millimeter or higher range resolution is required. Conventional TOF range imagers have only a few centimeter of range resolution. For higher range resolution, it is necessary to shorten the pulse width of light sources down to several hundred of picosecond. However, such a light source usually has a large distortion, causing the big error in the range calculation. In this paper, we propose a new type of TOF measurement method using an impulse photocurrent response.

In the proposed technique, a short pulse laser with a pulse width of approximately a hundred picosecond is used. The pulsed light is emitted to a target object and the reflected light generates charge at a photodiode. These charges are modulated by a lock-in pixel. For the range calculation, at least two signals with different time windows are used. Since the pulse width of light source is very short, which can be regarded as an impulse input, the range calculation is determined only by the photocurrent response of the lock-in pixel.

The higher speed photocurrent response provides the higher range resolution. Accordingly, we used the draining only modulation (DOM) pixel, which consists of a photodiode, a draining gate and floating diffusion. The lateral electric field is created in the channel between the PD and FD. The draining gate is formed along the channel. When a draining gate is opened, all of generated charges at photodiode are drained out. While the draining gate is closed, generated charges are transferred into a FD. Since the DOM pixel does not have any transfer gate in signal path, high speed charge modulation can be achieved. In the 3D device simulation, photocurrent response of less than 50ps is achieved.

To proof the concept of the proposed technique, a test chip is fabricated in a 0.11um CIS technology. In the measurement, a 445nm laser with pulse width of 100 ps is used for the light source. The emission trigger of the laser is given by the sensor board via a digital delay generator (DDG) which can accurately control the delay time of laser trigger. The change of trigger delay is equivalent to change a time of flight or target distance. In this measurement, the number of repetitive accumulation is set to 40,000, and the repetition frequency is 7.5 MHz. The measurable range is measured to be 50mm within nonlinear error of 5%. The average range resolution of 210um is achieved.

9022-32, Session PWed

Theoretical study of an efficient bracketing camera system's architecture

Amine Besrour, Univ. of Carthage (Tunisia); Hichem Snoussi, Univ. de Technologie Troyes (France); Mohamed Siala, Univ. of Carthage (Tunisia); Fatma Abdelkefi, Univ of Carthage (Tunisia)

No Abstract Available

9022-1, Session 1

A time-resolved image sensor for tubeless streak cameras

Keita Yasutomi, Sangman Han, Min-Woong Seo, Taishi Takasawa, Keiichiro Kagawa, Shoji Kawahito, Shizuoka Univ. (Japan)

Recently, much attention has been paid to time-resolved methods for biological imaging. One of the time-resolved imaging devices is a streak camera. Streak cameras acquire an image in which one of the axes corresponds to a temporal resolution instead of a spatial resolution. Although the conventional streak camera has high time resolution (up to 300 fs), the device requires high voltage and bulky system due to the structure with vacuum tube.?This paper presents a time-resolved CMOS image sensor with draining only modulation(DOM) pixels for tube-less streak cameras.

The proposed system of streak camera consists of cylindrical lens and the time-resolved imager. By using the cylindrical lens, an incident light is expanded along the only horizontal axis, i.e. every pixel in a horizontal line receives the same light. Since each pixels in a horizontal line have a different time window in the proposed time-resolved imager, a streak image can be taken.

The proposed imager has the DOM pixels with storage diodes and delay-based pulse generator to create a short time window. The DOM pixels consists of a pinned photodiode(PPD), a pinned storage diode(PSD), a draining gate(TD) and floating diffusion(FD). The lateral electric field is created in the channel between the PD and SD. The draining gate is formed along the channel. When a draining gate is opened, all of generated charges at the photodiode are drained out. While the draining gate is closed, generated charges are transferred into a SD. The stored signal in the SD is read out in the same manner as 4T pixels, thereby the reset noise is cancelled. Since the DOM pixel does not have any transfer gate in signal path, high speed charge modulation and loss-free charge accumulation can be achieved.

In the streak camera, a control signal with a short pulse of several hundred picoseconds is required in order to generate the time window. To do this, the delay-based pulse generator is used, in which the pulse width is determined by the amount of an logic delay. Because of parasitic components in control signal lines, it is difficult to provide the very short pulse directly into the pixel array. In the proposed imager, the control signal is reproduced by an simple logic in a pixel. Since the number of transistors in the in-pixel logic is small, reduction of the fill factor is also small.

A prototype time-resolved CMOS image sensor with the proposed pixel is designed and implemented using 0.11um CMOS image sensor technology. The image array has 30(Vertical) x 128(Memory length) pixels with the pixel pitch of 22.4um. In the simulation results, the time resolution of 250 ps can be achieved.

9022-2, Session 1

Pixel structure with 10 nsec fully charge transfer time for the 20M frame per second burst CMOS image sensor

Ken Miyauchi, Tohru Takeda, Katsuhiko Hanzawa, Yasuhisa Tochigi, Tohoku Univ. (Japan); Shin Sakai, TOHOKU UNIVERSITY (Japan); Rihito Kuroda, Tohoku Univ. (Japan); Hideki Tominaga, Ryuta Hirose, Kenji Takubo, Yasushi Kondo, Shimadzu Corp. (Japan); Shigetoshi Sugawa, Tohoku Univ. (Japan)

Recently, ultra-high speed (UHS) image sensors have been utilized to analyze UHS phenomena such as combustion, materials fracture, and electric discharge [1-2]. In this paper, we demonstrate the technologies related to the pixel structure achieving the fully charge transfer time of less than 10 nsec for the 20M frame per second (fps) burst CMOS image sensor.

In this image sensor, each pixel has multiple on-chip memories corresponding to the number of burst video operation frames, and the pixel array region and the on-chip memory array region are spatially separated [3]. This design makes it possible to enlarge the size of the fully-depleted pinned photodiode (PD) and to increase the number of burst video operation frames simultaneously. The PD size is 30.0 umH X 21.3 umV in the 32.0 umH X 32.0 umV pixel. The pixel structure achieving the fully charge transfer time of less than 10 nsec is necessary for 20M fps burst video operation.

The longest path from the PD edge to the floating diffusion (FD) of this image sensor is about 25 um. To achieve 10 nsec fully charge transfer time, about an electric field greater than average of 400 V/cm toward the FD is needed in the entire PD region. In this image sensor, the FD and the transfer-gate-electrode (TG) are placed at the bottom center of the PD. The n-layer for the PD consists of the semicircular regions centered on the FD and the sector-shaped portions extending the edges of the semicircular regions. In each part of the sector-shaped portions, the n-layer width becomes narrower from the proximal-end to the distal-end to generate a constant electric field toward the FD direction. By using the three-dimensional field effect, this achieves the comparable effect by graded-dopant concentration. In addition, the sector-shaped portions are densely arranged to satisfy the charge collection efficiency. Under these concepts, we designed the PD structure, which included the n-layer shape and the PD dopant profile with the condition of three times n implantation, and the TG structure. The n-layer concentration has three levels, and the concentration level of the near-FD region is the highest. Based on the device simulation, the photoelectron generated in the longest path is transferred to the FD within 7 nsec.

An UHS CMOS image sensor with the abovementioned pixel structure has been fabricated. The photoelectron was transferred to the FD within 10 nsec through the experiment changing the pulse timings of the 635 nm laser-diode exposure and turning on the TG. Secondly, we carried out the experiment exposing the pixels to the array spot lights displaced slightly in the pixel area. From the experiment, all the pixels of which PD area were exposed had equivalent amount of signal levels. Therefore, we confirmed that the entire PD area had sensitivity. Thirdly, image lag was below the measurement limit. The image sensor achieved 20M fps burst video operation and the maximum power consumption was 12 W.

[1] T .G. Etoh, et al., “An Image Sensor Which Captures 100 Consecutive Frames at 1 000 000 Frames/s,” IEEE Tans. Elec. Dev, Vol. 50, No. 1, pp. 144-151, 2003.

[2] J. Crooks, et al., “Kirana: a solid-state megapixel uCMOS image sensor for ultra-high speed imaging,” Proc. IS&T/SPIE Electronic Imaging, Sensors, Cameras, and Systems for Industrial and Scientific Applications XIV, pp. 865903-1-865903-14, Jan. 2013.

[3] Y. Tochigi, et al., “A Global-Shutter CMOS Image Sensor With Readout Speed of 1-Tpixel/s Burst and 780-Mpixel/s Continuous,” IEEE J. Solid-State Circuits, Vol. 48, No. 1, pp. 329-338, 2013.

9022-3, Session 1

Novel CMOS time-delay summation using single-photon counting for high-speed industrial and aerospace applications

Munir M. El-Desouki, King Abdulaziz City for Science and Technology (Saudi Arabia)

Time-delay integration (TDI) is a popular imaging technique that is used in many applications such as machine vision, dental scanning and satellite earth observation. One of the main advantages of using TDI imagers is the increased effective integration time that is achieved while maintaining high frame-rates. Another use for TDI imagers is with moving objects, such as the earth’s surface or industrial machine vision applications, where integration time is limited in order to avoid motion blurs. Such technique may even find its way in mobile and consumer based imaging applications where the reduction in pixel size can limit the performance during low-light and high speed applications. Until recently, TDI was only used with charge-coupled devices (CCDs) mainly due to their charge transfer characteristics. CCDs however, are power consuming and slow when compared to CMOS technology. Additionally, CCDs are no longer favorable for mobile applications that are currently dominated by CMOS technology. In the past couple of years, there have been some attempts to implement TDI imagers using CMOS technology and the interest is slowly growing.

In this work, we report on novel (USPTO patented) architectures that use single-photon counting based TDI techniques that are implemented in standard CMOS technology allowing for complete camera-on-a-chip solutions. The presentation will start with reviewing some of our work to achieve single-photon detection in standard CMOS technologies using a 130 nm mainstream digital CMOS technology from IBM. A novel time-domain technique will also be shown that allows for 3 to 4 orders of magnitude higher dynamic range than conventional wide dynamic range imagers, while simultaneously offering high sensitivity and high speed operation at low cost. And finally, a single-photon counting based imager, configured for use with TDI, will be presented that was fabricated in a standard CMOS 150 nm process from LFoundry.

9022-4, Session 1

Ultra-high speed video capturing of time dependent dielectric breakdown of metal-oxide-silicon capacitor up to 10M frame per second

Fan Shao, Daiki Kimoto, Kiichi Furukawa, Hidetake Sugo, Tohru Takeda, Ken Miyauchi, Yasuhisa Tochigi, Rihito Kuroda, Shigetoshi Sugawa, Tohoku Univ. (Japan)

Many papers have described the model of time dependent dielectric breakdown (TDDB) of Metal-Oxide-Silicon (MOS) capacitor [1-2]. However, the transient analysis of TDDB with the ultra-high speed (UHS) video capturing with 10 usec order period or less has never been reported. In this paper, the UHS video capturing results of TDDB of MOS capacitor are reported using the UHS camera with flexibly variable trigger timing types, a wide spectral sensitivity from visible light to near-infrared light, the maximum frame rate of 20M frame per second (fps) and the

number of record length of 256 frames [3].

In the measurement system, the Si-wafer with MOS capacitors was chucked on the stage of a prober. The backside voltage of the MOS capacitor was applied to 0 V, and the gate voltage of the MOS capacitor was boosted from 0 V to 100 V instantaneously. The UHS camera with an increased lens was set above the MOS capacitor. A metal halide lamp was used as the imaging light source. It is widely known that the times to breakdown by Fowler-Nordheim current have a variation. In addition, the time span of the breakdown of a thick-oxide MOS capacitor is known to be very short. In order to capture the breakdown, we set a trigger circuit which detects the rapid current increase through the MOS capacitor at the breakdown, which supplies the trigger pulse to the UHS camera. The details of the MOS capacitor for this experiment were as follows; the capacitor area was 500 x 500 um2, the dopant concentration of the n-Si was 1.0?10^16 cm-3, the gate insulator film was thermal grown SiO2 film, the thickness of the oxide was 100 nm and the gate electrode was made of aluminum.

Some movies have succeeded to capture the intermittent emissions of light at some points on the gate during the breakdown. From the movie taken at 50 Kfps, we confirmed the total time of the intermittent emissions of light was about 1 msec. From the movie taken at 1 Mfps, the order of the time interval of the emissions of light was about 10 usec and simultaneous emission of light at multiple points was not observed. From the movie taken at 10 Mfps, the time span of the emission of light was less than 100 nsec. We consider that the local current caused by the breakdown of the insulator generates heat into the gate electrode and leads the emission of light. These movies have succeeded to analyze TDDB both transiently and spatially. We will evaluate some capacitors with different gate electrode materials and so on. We will show some movies at the presentation. It is expected that this video capturing technique is to be utilized in a wide range of scientific fields such as material science, electronic engineering and nanotechnology.

9022-5, Session 2

Low data rate architecture for smart image sensor

Amani Darwish, Gilles Sicard, Laurent Fesquet, TIMA Lab. (France)

Nowadays, the power management of microelectronic devices is crucial, especially for mobile applications. Therefore, designers have developed strategies for managing and reducing circuit power consumption, while maintaining high performance capabilities in term of speed or computation. However, in the field of image sensors there is almost no industrial CMOS image sensors optimized to operate at very low power consumption as in the mean time they are massively employed in mobile devices.

Previous work on asynchronous image sensor architectures has been presented in order to reduce the excessive data flow in high-resolution CMOS image sensors. Nevertheless, the proposed sensor's pixel suffers from a very high number of transistors per pixel [2]. Furthermore, this architecture requires an arbiter - for managing an asynchronous sequential protocol - that proportionally increases its complexity with the size of the matrix [3].

Our work targets an innovative image sensor architecture (≈ 20-25 transistors) based on non-uniform sampling [1] and event-driven asynchronous, computing able to drastically decrease the image sensor data flow. This data reduction will contribute to significantly reduce the power in such devices. The proposed low-power architecture has been designed in order to only extract the relevant image information of a CMOS image sensor.

To reach such a goal, we completely rethink the sensor architecture including the design of the analog and digital parts: pixel and digital read-out system respectively :

• The first idea is to remove the temporal pixel color redundancies [4] between two consecutive frames. Indeed, the pixel is able to check the color redundancy with the previous frame. In case of redundancy, the pixel does not send any signal. On the other side, if the pixel color is different, the pixel sends a request signal in order to capture the new pixel value. Then the pixel waits for an acknowledgment signal, which will reset the pixel. This means that each pixel is able to behave asynchronously from the others.

• The second idea is to remove the spatial color redundancies, thanks to a specific read-out system. The read-out system records the addresses of pixels - that send request along with the instant of these and acknowledges each pixel. As the pixel requests sent during a specific time interval are referring to the same level of illumination, the read-out system is able to remove the spatial redundancies.

Simulations using both Matlab© and Modelsim© software tools were perform to validate the behavior of the pixel and of the read-out protocol. The Matlab simulations targeting the removing of spatial redundancies have shown that we reached a very low reading rate compared to a conventional reading, even for fine textured images. This has been supplemented by VHDL simulations showing the correct behavior of the asynchronous reading protocol of the system. These simulation results meet our expectations and validate the proposed ideas.

In conclusion, these proposed ideas lead to design an innovative smart image sensor architecture intended to work at low power consumption in mobile devices because it only produces relevant information. This architecture is able to drastically reduce the data stream by suppressing the spatial and temporal redundancies that exist in videos.

References:

[1] E. Allier, G. Sicard, L. Fesquet, and M. Renaudin, "Asynchronous level crossing analog to digital converters," Measurement, vol. 37, no. 4, pp. 296–309, Jun. 2005.

[2] J. Kramer, "An on/off transient imager with event-driven, asynchronous read-out," in 2002 IEEE International Symposium on Circuits and Systems. Proceedings (Cat. No.02CH37353), 2002, pp. II–165–II–168.

[3] A. Myat, T. Linn, D. A. Tuan, C. Shoushun, and Y. K. Seng, "Adaptive priority toggle asynchronous tree arbiter for AER-based image sensor," 2011 IEEE/IFIP 19th International Conference on VLSI and System-on-Chip, vol. 2, pp. 66–71, Oct. 2011.

[4] H. Amhaz, H. Abbass, H. Zimouche, and G. Sicard, "An improved smart readout technique based on temporal redundancies suppression designed for logarithmic CMOS image sensor," 2011 18th IEEE International Conference on Electronics, Circuits, and Systems, pp. 472–475, Dec. 2011.

[5] X. Guo, X. Qi, and J. Harris, "A time-to-first-spike CMOS image sensor," Sensors Journal, IEEE, vol. 7, no. 8, pp. 1165–1175, 2007.

9022-6, Session 2

Frameless, time domain continuous image capture

Henry G. Dietz, Univ. of Kentucky (United States)

In 1994, Bishop, Fuchs, McMillan, and Zagier published a paper observing that frameless rendering by randomizing pixel update timing leaves fewer visible artifacts than traditional double-buffering for high-speed graphical output.

The current paper suggests that imaging sensors can be constructed to use a type of frameless capture in order to achieve similar but even greater benefits.

Ideally, each sensel is controlled by a nanocontroller literally fabricated under the sensel. In the proposed design, each sensel's value is read by software executed on the corresponding nanocontroller that times how long it takes to reach a charge threshold and then converts sequences of such samples into a smooth waveform describing how Ev changes over time for that sensel. Although the sampling is asynchronous and can be at a very high rate, Ev generally varies slowly over time or changes to follow the waveform of a nearby sensel, so very high compression of the

waveforms theoretically could be achieved before they leave the sensor's massively-parallel nanocontroller array.

From a frameless image data stream, still images can be computed as the average value of each sensel's waveform over the desired interval. Thus, the interval represented by an image can be determined, and changed, after the capture. This makes it possible to losslessly render video at any desired frame rate -- for example, using the same frameless data stream to create videos rendered at 24FPS, PAL, and NTSC. It also is possible to "nudge" a still image forward or backward in time to capture the perfect moment. Knowing the waveform for each pixel's Ev also means that the time interval represented by an extracted image frame is indepedent of the photon integration period; for example, an image summarizing 1/60s can contain high-quality data for both pixels that would have saturated and those that would have been below the noise floor with a real 1/60s integration period. Thus, all image data is HDR (high dynamic range) with a low and directly controllable Ev noise level. This also implies that consecutive individual frames in a video can have apparent integration times, independent of lighting, that ensure smooth and complete motion blur rather than the "jumping telephone poles" artifacts so commonly seen in pan sequences.

This paper describes our work toward building a frameless imaging sensor using nanocontrollers, basic processing of time domain continuous image data, and the expected benefits and problems. The nanocontroller architecture has been developed over the past decade and is described in earlier publications; thus, emphasis is placed on the complete system design and abstract properties of this approach.

9022-7, Session 2

Digital vision sensor for collision avoidance and navigation

Joseph H. Lin, Peter Grossmann, Daniel R. Schuette, MIT Lincoln Lab. (United States)

Autonomous vehicles, particularly unmanned air vehicles (UAVs), are playing an increasingly prominent role in military and civilian applications. Small UAVs (0.25 – 2 m wingspan) capable of safely operating at low altitudes in close proximity to structures and terrain offer exciting new opportunities. Critical to the success of these small UAVs is the ability to sense and react to their environment, specifically to detect and avoid obstacles. MIT Lincoln Laboratory is developing a unique image sensor and new UAV flight-control algorithms to fill this capability gap.

We describe the architecture of the image sensor, which integrates silicon Geiger-mode avalanche photodiodes with custom designed readout integrated circuitry. High-frame rate images (~1000fps) feed into on-chip data processing circuits that compute apparent velocities of objects in the scene (optical flow). Combined with the UAV's velocity, these apparent velocities provide a coarse depth map that can be used for collision avoidance and local navigation. This is similar to early visual processing that birds and insects use to fly through cluttered environments.

The optical flow algorithm is derived from the classic Horn and Schunck algorithm. This algorithm assumes that scene reflectance remains constant from frame to frame, thus image motion can be computed from spatio-temporal gradients. The large linear system of equations that results is typically solved using an iterative method such as Gauss-Seidel. However, we unroll this iteration over multiple frames taking advantage of the high frame rate of our imager. Due to the high frame rate, pixel velocities do not exceed 1 pixel/frame thus the optical flow is continually updated from successive frames. We show MATLAB results using simulated and real images that compare the accuracy of this algorithm with other popular optical flow algorithms such as the Lucas and Kanade algorithm.

We also show that a pipelined column-parallel architecture is more computationally and energy efficient than a processor-memory architecture. Energy efficiency of the architecture is important, as the small UAVs will have a limited power budget. We estimate that our 256x256 image sensor will consume less than 250mW.

The APD detectors are fabricated in MIT Lincoln Laboratory's micro-electronics fabrication facility and the readout electronics and optical flow circuitry will be fabricated in a 90nm standard CMOS process. We plan to tape out the read-out chip by the end of this year.

9022-8, Session 2

Smart imaging for power-efficient extraction of Viola-Jones local descriptors

Jorge Fernández-Berni, Ricardo A. Carmona-Galán, IMSE-CNM (Spain) and Univ. de Sevilla (Spain); Rocío del Río Fernández, Instituto de Microelectrónica de Sevilla (Spain) and Univ. de Sevilla (Spain); Juan A. Leñero-Bardallo, IMSE-CNM (Spain) and Univ. de Sevilla (Spain); Manuel Suarez-Cambre, Univ. de Santiago de Compostela (Spain); Ángel B. Rodríguez-Vázquez, IMSE-CNM (Spain) and Univ. de Sevilla (Spain)

In computer vision, local descriptors permit to summarize relevant visual cues through so-called feature vectors. These vectors constitute inputs for trained classifiers which in turn enable different high-level vision tasks: object recognition, tracking, surveillance etc. While local descriptors certainly alleviate the computation load of subsequent processing stages by preventing them from handling raw images, they still have to deal with individual pixels. Feature vector extraction can thus become a major limitation for low-power embedded vision hardware implementing conventional processing architectures due to the unavoidable data serialization and consequent memory bottleneck. In this paper, we present a power-efficient smart imager conceived to provide, among other processing capabilities, the computation of integral images at different scales. These images are intermediate representations that speed up feature extraction. In particular, the chip operation is tailored for extraction of Haar-like features. These features, that encode differences in average intensities between rectangular regions, feed the cascade of classifiers at the core of the Viola-Jones framework, one of the best approaches reported to achieve real-time object recognition. The prototype imager has been designed for the standard UMC 0.18um 1P6M CMOS process. It includes a reconfigurable QVGA SIMD focal-plane mixed-signal sensing-processing array featuring low-power massively parallel operation, peripheral circuitry to carry out such reconfiguration, random pixel access readout circuitry and four full-custom SAR A/D converters providing a throughput of 4MSa/s at frame rates around 30fps. In addition to integral image computation, the array can be reprogrammed to deliver other early vision tasks, namely concurrent rectangular area sum, block-wise HDR imaging, Gaussian pyramid generation and image pre-warping for subsequent reduced kernel filtering. Finally, we describe the FPGA-based system designed to interface the chip and test it.

9022-9, Session 3

Time-to-digital converter based on analog time expansion for 3D time-of-flight cameras

Muhammad Tanveer, Luleå Univ. of Technology (Sweden); Ilkka Nissinen, Jan Nissinen, Juha T. Kostamovaara, Univ. of Oulu (Finland); Johan Borg, Jonny Johansson, Luleå Univ. of Technology (Sweden)

This paper describes an architecture and achievable performance of a time-to-digital converter by cascading a time stretcher and a gated ring oscillator based time-to-digital converter (GRO-TDC). The analogue time expansion, where the time interval to be measured is stretched by a factor k, is realized by charging a capacitor with a current I followed by discharging the capacitor by a current I/k . The currents are created by wide swing cascode current source/sinks with a current ratio of k. The time stretching method involves two conversions: time to charge and

then charge to time. Whereas these are performed in each individual pixel, the final time to digital conversion is performed by the global GRO-TDC, where a multiphase gated ring oscillator is used to measure the stretched time interval by counting full clock cycles and storing the states of the ring oscillator within the clock period to obtain increased resolution. A block diagram of the proposed structure is shown in

The nine phase single ended gated ring oscillator operates as an interpolator and as a clock during charge to time conversion. The layout of the gated inverters in the oscillator is designed by placing them in an order to compensate for the parasitic loading of capacitances for each stage and to minimize the e?ffects of process variations. The clock of the 8-bit ripple counter is enabled at the start of the charge to time conversion by the Start-Stretch signal. The result of the counter will be ready after the clock of the counter is disabled to the counter by the Stop-Stretch timing mark from the comparator. Special attention is paid to the design of the hysteresis based comparator using positive feedback. The comparator is designed to achieve acceptable robustness against transistor mismatch, small power dissipation, off?set voltage, linearity, speed, small area and good noise immunity. The digital flip-flops functioning as an interpolator will store the time interval measurement responses with in the clock cycle. By selecting an appropriate stretch factor and suitable clock frequency for the gated ring oscillator, measurement error of few cm in distance is achievable. To ensure reliable recording of the timing signals, the counter is synchronized by using a dual edge synchronization scheme where one of the two flip-flops always has enough delay between data change and clock edge to avoid metastability problems.

9022-10, Session 3

Experiment on digital CDS with 33-M pixel 120-fps super hi-vision image sensor

Jun Yonai, Toshio Yasue, Kazuya Kitamura, Tetsuya Hayashida, Toshihisa Watabe, Hiroshi Shimamoto, Japan Broadcasting Corp. (Japan); Shoji Kawahito, Shizuoka Univ. (Japan)

We have developed a CMOS image sensor with 33 million pixels and 120 fps (frames per second) for the next-generation ultra-high definition television (UHDTV) broadcast system (Super Hi-Vision). This sensor includes FPN (fixed pattern noise) in its signal output as usual CMOS image sensors do. The major source of this noise is the part after column analog to digital conversion in the sensor with column parallel ADCs. We have reduced the FPN by subtracting the FPN data from the image data. In this case we have to measure the dark level, average the noise level, then figure out the FPN data prior to shooting, i.e., off-line operation. We also need an external frame memory and signal processor.

There are several other methods to reduce the FPN by using digital CDS (correlated double sampling) [1-2]. Those methods sample the signals twice, before and after the reading of the signal from the pixel, and subtract both data in a digital circuit, which enables it to reduce device variation and circuit offset that cause vertical FPN (VFPN). Advantages of the digital CDS are: (1) on-line operation and no need for an external frame memory and signal processor, (2) possible to reduce the reset noise caused by the analog CDS circuit, (3) possible to reduce the random noise which may deviate from a row (horizontal) scanning frequency. Disadvantages of the digital CDS are the circuit complexity if it is in the image sensor and the requirement of double speed of the ADC (analog-to-digital convertor). In particular, it is difficult for usual UHDTV sensors to operate at doubled speed to apply the digital CDS methods as they already operates at a high speed due to a large pixel count.

On the other hand, our image sensor has a very fast ADC using “two-stage cyclic ADC” architecture [3] and is capable of A/D conversion less than only 1.92us. This image sensor is basically designed for 120 fps high frame rate driving, however, we can also use this ADC as a double-speed ADC when we use the image sensor for 60 fps frame rate. Then we test the digital CDS driving using this image sensor.

In this experiment, we read the same row twice at 120 fps to sample the two signals, i.e., the signal after resetting the floating diffusion (FD), and the signal after the readout of the pixel data. This time we externally calculated the subtraction. We stored each output data in a PC and calculated the noise. The result showed that the VFPN was effectively reduced from 24.25 e-rms to 0.43 e-rms. The random noise was also reduced from 4.65 e-rms to 3.10 e-rms.

[1] Woodward Yang, et al., “An Integrated 800x600 CMOS Imaging System,” in Proc. ISSCC Dig. Tech. Papers, WA 17.3, pp. 304-305 (1999)

[2] Yoshikazu Nitta, et al., “High-Speed Digital Double Sampling with Analog CDS on Column Parallel ADC Architecture for Low-Noise Active Pixel Sensor,” in Proc. ISSCC Dig. Tech. Papers, 27.5 pp. 2031-2032 (2006)

[3] Kazuya Kitamura, et al., “A 33-Megapixel 120-Frames-Per-Second 2.5-Watt CMOS Image Sensor with Column-Parallel Two-Stage Cyclic Analog-to-Digital Converters,” IEEE Trans. Electron Devices, vol. 59, no. 12, pp. 3426-3433 (2012)

9022-11, Session 3

Pixel structure for asymmetry removal in ToF 3D camera

Byong Min Kang, Jungsoon Shin, Jaehyuk Choi, Dokyoon Kim, Samsung Advanced Institute of Technology (Korea, Republic of)

Recently, ToF cameras have been adopted for real time depth acquisition. ToF cameras capture the 3D geometry of objects by measuring the time delay between emitted and reflected NIR and calculating distance by multiplying speed of light. Most ToF cameras modulate the NIR to 10~50 MHz modulation frequency (MF). In each pixel, transfer gates are modulated by the clock signal that is synchronous with the NIR, and the pixel detects reflected NIR from objects. A photodiode in pixels converts reflected NIR to electrons. By turning on the transfer gate, the integrated charges are transferred to the floating diffusion (FD) node for the readout. The FD node that has capacitance temporarily stores charges before the readout. Many depth sensors employ two-tap pixel architecture, which has two transfer gates (TX0, TX1) and two FD nodes in one pixel. In this architecture, the depth can be calculated with two consecutive frames. In the first frame, two transfer gates are modulated by two clock signals with 180 phase difference (0°, 180°) and two FD nodes stores transferred charges (Q0 in FD0, Q180 in FD1). In the second frame, same operation is repeated but the phase of clock signals are shifted by 90° (90°, 270°). A depth can be calculated from the pixel output that maps integrated charges from different phases: Q0°, Q180°in the first frame, Q90°, and Q270° in the second frame. The depth is proportional to tan-1[(Q90°-Q270°)/(Q0°-Q180°)]. However, the depth accuracy will be degraded from the tap asymmetry between FD0 and FD1, which comes from the capacitance mismatch in FD nodes, offsets in readout circuits and so on. Since this asymmetry is inherent due to the variation of the fabrication process, we need an additional scheme for eliminating the effects from the tap asymmetry. In order to address tap asymmetry, conventional methods re-measure charges stored in FD nodes by applying the clock signal with 180° phase shift to transfer gates. After adding two outputs from two measurements, signals are added up and noises are cancelled out due to the phase inversion in the additional measurement. However, conventional methods need additional frame memories and have low acquisition speed due to additional readout time. Therefore, we propose the pixel architecture for the suppression of tap asymmetry without additional memories and timing budget.

Conventional 2x2 shared pixel of two-tap structure shares four FD nodes and a control signal is applied to four transfer gates (TX) of identical side at each sub pixel. However, conventional shared pixels need additional frame memories and low acquisition speed as mentioned before. Therefore, we propose a novel shared pixel structure for suppressing the asymmetry between two taps. The two-tap pixel is shared with neighbor pixels and transfer gates in the pixel are cross-connected between upper and lower pixels. For example, right TX gate in upper sub pixel are connected to left TX gate in lower sub pixel. Signals are added up and noises are cancelled out. So, we designed a modified shared pixel structure for tap asymmetry removal.

9022-12, Session 3

A stimulated Raman scattering imager using high-speed lateral electric field modulator and lock-in pixels amplifiers

Kamel Mars, Beak Guseul, Sangman Han, Taishi Takasawa, Keita Yasutomi, Keiichiro Kagawa, Shizuoka Univ. (Japan); Mamoru Hashimoto, Osaka Univ. (Japan); Shoji Kawahito, Shizuoka Univ. (Japan)

Stimulated Raman Scattering (SRS) is one of the most emergent biomedical imaging which works by detecting the vibrations in chemical bonds between atoms and could provide a new form of real-time bio-imaging with improved image contrast and free of fluorescent labels that can hinder many biological processes [1]. Increased demand on bio-imagers and the expansion of their area of use require the achievement of a very high dynamic range.

Since the generated SRS signal is very small compared with the offset, a technique for eliminating the offset signal, extracting and amplifying the small SRS signal is necessary. A CMOS image sensor suitable for SRS imaging using the proposed design is implemented. It consists of an array of lateral electric field modulator (LEFM) for charge modulation with lock-in pixels fully differential amplifiers and sampling circuits, a vertical and horizontal scanner, two buffer amplifiers , biasing current circuits and a clock generator to provide the required signals for the sampling circuits.

The modulated light signal which contains small SRS signal and large offset due to the excitation laser pulse need to be demodulated to extract the small SRS signal component The LEFM is used to extract the small SRS signal component from the modulated light signal. The LEFM pixel use pinned photodiode and two sets of gates for applying a lateral electric field. The gates are not used for draining the photo charge but for controlling the electric field of X-X' direction where a relatively small positive voltage (M=1.3V) and negative voltage (L=-2V) are used for this operation. Unlike the conventional charge transfer technique using a transfer gate, the LEFM structure does not have a problem of the creation of potential barrier at the edge of the transfer gate, and charge trapping under the gate (Si-SiO2interface). Therefore, very high-speed electron transfer of less than 1ns is possible [2].

Since the detected signal from the SRS process is very small compared to the offset signal, differential charge accumulation using the LEFM and lock-in pixels fully differential amplifiers is very effective. The phase delay between the modulated pixel output signal and the reading out clock needs to be carefully adjusted to set the differential output of the amplifier to zero if there is no modulated signal detected from the SRS process.

The proposed design is fabricated by 0.11?m 1P4M CIS process. In order to measure the fabricated SRS imager, a near infra-red laser pulse light is used. By adjusting the phase delay, the lock-in pixel amplifier differential output is unchanged to time when there is no incident light. Small SRS signal level is detected and the lock-in pixel amplifier differential output is linearly increased to time.

9022-13, Session 4

Estimating an image sensor's temperature for darksignal-correction

Julian Achatzi, Gregor Fischer, Fachhochschule Köln (Germany); Volker Zimmer, Leica Camera AG (Germany); Dietrich W. Paulus, Univ. Koblenz-Landau (Germany)

An image taken with an image sensor, like e.g. a CMOS-Sensor, is corrupted by noise. One fraction of this noise is the so-called fixed pattern noise, i.e. noise that has a determined mean per pixel. Fixed pattern noise mainly consists of the dark signal, that is the signal which is produced by a pixel in the absence of light falling on this pixel. It is one of the first image processing steps in a camera to eliminate this dark signal from the image.

In order to correct an image for the dark signal, one could a) subtract a dark frame, b) subtract values derived from the signal of optically black pixels (OB-pixels) or c) subtract a dark frame that is calculated using a model. While the first will amplify temporal noise, the second will leave parts or all of the dark signal non uniformity (DSNU) in the image. For the third approach one needs to know the image sensors temperature in order to estimate an accurate dark-frame. Therefore Widenhorn et al. calculate a temperature indicator which is derived from the values of hot pixels, i.e. pixels with high dark current (DOI 10.1117/12.714784). In US7787033B2 Rossi et al. calculate an estimate of the sensor's temperature using some OB-pixels' values in order to control the power supply on the chip.

The goal of our research is to get accurate temperature estimations out of the dark signal of some OB-pixels or the pixels from a dark frame that was made with shorter integration time than the image to be corrected. Accurate in our case means that the temperature estimation should improve the calculations of the dark frame in comparison to the calculation using the temperature measurements of two temperature-sensors that are incorporated into a given imagie sensor.

Therefore we analyse the theoretically achievable precision and derive a measure of confidence for the temperature-measurement of each pixel. We investigate the performance of different methods to calculate the sensor's temperature using this confidence-measure, like e.g. a weighted average. Further effort is made to answer the question, if one could measure and correct for the local variation of the temperature over the sensor's surface.

Let m(t,T,s) be a model for a single pixel's dark signal at integration time t, temperature T and ISO speed s, and let m be solvable for T, so that one can calculate the pixel's temperature for a measured dark signal. Of course, the measured dark signal is corrupted by noise and the model's parameters are erroneous due to noise and errors, generated and made during calibration. Therefore the question arises how accurate and precise the calculated temperature can be and whether there are pixels, that are better temperature-sensors than others. Since the noise depends on the same parameters as the dark signal, the confidence-measure and therefore the answer to the given question also depends on said parameters and the statistical distribution of the pixel's characteristics.

We test our methods using a CMOS-Sensor with 4T-Architecture which has two temperature-sensors integrated onto the die. We show that we can improve the darksignal-correction by using the local temperature estimates computed from some pixels' dark signals upon using the temperature-sensors' values.

9022-14, Session 4

A statistical evaluation of effective time constants of random telegraph noise with various operation timings of in-pixel source follower transistors

Akihiro Yonezawa, Rihito Kuroda, Toshiki Obara, Akinobu Teramoto, Shigetoshi Sugawa, Tohoku Univ. (Japan)

RTN (Random Telegraph Noise) causes the image degradation in CMOS image sensor (CIS) . RTN is caused by the carrier trapping and detrapping phenomenon into oxide traps, and the signal is randomly and discretely changes. RTN has three typical parameters which are RTN amplitude and time constants; time to capture (?c) and time to emission (?e). Statistically analyzing these parameters is critically important to estimate noise distribution of CIS pixels. At readout operation in CIS, the row selected pixel-SFs (source follower) turn on and not selected pixel-SFs operate at different bias conditions depending on the select switch position. When the select switch locates top of the SF driver, SF driver keeps turn on; gate voltage is Vdd, and when select switch locate bottom of the SF driver, SF driver nearly turn off; the source voltage becomes

nearly equal to (Vdd or floating diffusion reset voltage – Vth). The duty ratio and cyclic period of selected time of SF driver frame rate depends on the operation timing determined by column read out sequence. In general, RTN analysis takes place under constant operation condition. Then it is important to understand the dependency of the RTN time constants and amplitude on duty ratio and cyclic period. In this work, using arrayed test circuit, we extract these RTN parameters with various duty ratio and cyclic period statistically.

We employed arrayed test circuit which can detect and analyze in-pixel SF MOSFET in short time [5-6]. In this circuit, selected switch locate bottom of the SF driver then when row are not selected, pixel-SF nearly turn off. Root mean square of output-voltage (Vrms), RTN amplitude and average time constants (<?c> and <?e>) which were extracted in measurement time were extracted from 131072 MOSFETs with gate width / length = 0.28/0.22 um under two measurement conditions; duty ratio = 1, cyclic period = 14ms and duty ratio = 7.1 times 10^-3, cyclic period = 700ms, sampling time was 210 sec for both case.

In our measurement, when duty ratio = 1, 13401 MOSFETs have RTN, while when duty ratio =7.1 times 10^-3, 6659 MOSFETs have RTN. 4204 MOSFETs have RTN in both conditions. For MOSFETs which have RTN amplitude is greater than 1mV in both duty ratios, the MOSFETs with RTN have almost the same RTN amplitudes. The coefficient determination of two amplitudes is 0.93. The scatter plot of average time constant ratio (<?c> / <?e>) of both duty ratio widely distributed. Especially, when duty ratio = 1, the distribution of <?c> / <?e> at duty ratio = 1 tend to be smaller than that of <?c> / <?e> at duty ratio =7.1 times 10-3. It means <?c> becomes smaller or <?e> becomes longer at duty ratio = 1. This indicates that the probability of capture and emission of the carrier to the trap changes depending on the pixel-SF operation duty ratio.

These results are important for the detection and analysis of in pixel-SF with RTN. We will evaluate more duty ratios and bias conditions.

9022-15, Session 4

Correcting high density hot pixel defects in digital imagers

Glenn H. Chapman, Rohit Thomas, Simon Fraser Univ. (Canada); Israel Koren, Zahava Koren, Univ. of Massachusetts Amherst (United States)

Recent investigations of digital imagers' in-field defect development have identified failed sites as hot pixels which accumulate steadily with time, caused by cosmic ray damage, which cannot be protected against by shielding. More importantly our recent studies created an empirical curve fit for defect rates/year/sq mm CCD imagers of a power law with of sensor area proportional to the pixel size to the power of -2.24 times the ISO (gain) raised to the 0.68 times a constant; and for CMOS sensors the pixel size raised to -3.6 times ISO raised to the 0.56. This projects that the current trend in cameras with pixel sizes dropping below two microns, and sensitivities increasing towards those for low light night pictures, will result in defect rates becoming hundreds to thousands per year in typical cameras. At these rates the simple correction methods of replacing the defect with interpolation from the nearest neighbors is poor in many locations and will degrade the picture. This paper then investigates methods to correct images based on the knowledge of hot pixel defect characteristics. Hot pixels have an offset value plus growth with exposure time, both of which increase with the sensitivity. Characteristics of hot pixels are constant with time after formation. We attempt to correct the damage to the image caused by a hot pixel, and claim that the correction method should depend on the severity of the hot pixel, on the exposure time, on the ISO. and on the variability of the pixel's neighbors in the image being corrected. The concept is tested on several cameras with 19 to 90 known hot pixels (at maximum low noise ISO) in two ways. The first set of tests image a nearly uniform background where the light intensity or "gray level" can be set from very dark to bright levels below saturation. In this test interpolation of neighbors always gives a good value of the true (not defective) level at the hot pixel. This is tested at a range of exposure times while adjusting camera F# so the background intensity is constant. The change in exposure times mean the hot pixel characteristics will change in a known manner. The second tests follow the same exposure time/lighting conditions but use a typically complex background where interpolation is known to fail at rapidly changing areas. Moving the camera a known number of pixels allows extracting the true value of the image without the hot pixel. Then we explore correcting the pixels with the calculated hot pixel error removed and via neighbor interpolation. We experimentally compare the improvement in pixel defect correction with this method, relative to the current classic method of simply removing the pixel and doing an interpolation of its value from surrounding image information. In busy images 68% of the pixels are best corrected with the hot pixel estimates, while interpolation methods are better in some slowly changing areas. Current algorithm selects hot pixel correction generally but interpolation when slowly changing background or the hot pixel saturates.

9022-16, Session 4

Comparison of two optimized readout chains for low light CIS

Assim Boukhayma, CEA-LETI (France) and Ecole Polytechnique Fédérale de Lausanne (Switzerland); Arnaud Peizerat, CEA-LETI (France); Antoine Dupret, Commissariat à l'Énergie Atomique (France); Christian C. Enz, Ecole Polytechnique Fédérale de Lausanne (Switzerland)

In this paper, we compare the noise performances of two readout chains based on 4T pixels. The comparison is based on the same pixel transistors count and the same bandwidth of 265KHz, i.e. enough to read 1Megapixel with 50frame/s. Both chains contain a 4T pixel, a column amplifier and a single slope analog-to-digital converter operating a CDS. In one case, the pixel operates in source follower configuration, and in common source configuration in the other case.

Firstly, state of the art noise reduction techniques in CMOS low light image sensors are reviewed. We discuss, analytically the impact of correlated multiple sampling on 1/f and thermal noise. A table of reported works in last few years is presented comparing noise performance of image sensors based on different techniques : Buried channel source follower, multiple sampling with SSADC, in-pixel amplification using pMOS common source and switch biasing.

We start with the analytic noise calculation of a classical source follower based readout chain, confirmed by simulation using a 130nm process. An optimised version is also presented, using a cascode column amplifier to achieve higher gain and enhance noise performance. Finally, an analytic noise calculation for a common source based readout chain is presented. Simulations lead to the following conclusions :

• For submicron in-pixel amplifying transistor, optimized column amplifier and CDS with the given bandwidth, the 1/f noise originating from the in-pixel amplifier dominates the readout chain noise for both configurations and both transistor types.

• If low 1/f noise in-pixel amplifying transistor is used, e.g. pMOS with W/L of 1.5/0.5µm, common source based readout chain shows lower power consumption and 50% total RMS noise reduction compared to the source follower based one. A total input referred noise of 27µV RMS is reached.

We compare the impact of both configurations on the sense node capacitance and discuss the limitations of the common source configuration. We provide the reader with a comparative table between the two readout chains. The table contains several variants (column amplifier gain, in-pixel transistor sizes and type) that help to choose the right scheme for the right application.

9022-17, Session 4

Review of ADCs for imaging

Juan A. Leñero-Bardallo, Jorge Fernández-Berni, Ángel B. Rodríguez-Vázquez, IMSE-CNM (Spain) and Univ. de Sevilla (Spain)

The analog to digital conversion of pixels outputs is a very important process that affects the image quality, the frame rate, or can even impose restrictions to the imager layout. Vision sensor designers are not usually experts in the design of analog to digital converters. On the other hand, ADCs designers are sometimes not aware of the special requirements than ADCs have to satisfy for imaging. General purpose ADCs are not suitable for commercial imaging anymore. Some of their specifications can be oversized leading to unnecessary power or area consumption. On the contrary, if they do not achieve some of the imager requirements for signal digitalization,they will degrade the image quality.

Traditionally, a global ADC was used to convert all the pixel outputs. ADC designers were not quite concern about the power and area requirements. However, modern image sensors demand high frame rates and pixels with fine pitch. More than one ADCs are shared by different group of pixels to increase the read out speed. In that sense, it is crucial to decide how many ADCs are used to convert the pixel outputs, how they are distributed, how the pixel outputs are multiplexed, and the ADC requirements. The specifications of these ADCs can be very different to the converters dedicated to communications, for instance.

In the literature, we find very different ADCs topologies. Based on the requirements for analog to digital conversion in imaging and the most recent relevant publications on the field, we will discuss which ones are more suitable for imaging. Their advantages and disadvantages will be presented.

Emerging 3D integration technologies also open new enticing possibilities for ADC design. The conversion speed could be reduced by placing more ADCs in one tier. ADCs could be shared by reduced groups of pixels with a fill factor close to 100%. The decision of how many ADCs are going to be placed and how the pixel outputs are going to be multiplexed is not a trivial matter.

In our presentation, we will discuss and analise in detail of the ADC requirements for modern imaging. The advantages and disadvantages of the main ADCs topologies recently reported for imaging will be analyzed. We will also give guidelines to optimize the ADC design. Some relevant ADCs aimed for vision sensors will be presented.

Finally, we will discuss the new challenges for digitalization and ADC layout in the emerging 3D technologies.

9022-18, Session 5

Color image sensor using stacked organic photoconductive films with transparent readout circuits separated by thin interlayer insulator

Toshikatsu Sakai, Hokuto Seo, Satoshi Aihara, Hiroshi Ohtake, Misao Kubota, NHK Science & Technical Research Labs. (Japan); Mamoru Furuta, Kochi Univ. of Technology (Japan)

We have been researching a novel image sensor with three stacked organic photoconductive films (OPFs) sensitive to only one of the primary color components (red (R), green (G), or blue (B)), each of which has a transparent signal readout circuit. Color separation optical systems, such as a dichroic prism or a color filter array, can be eliminated by using this stack-type organic image sensor because color separation can be achieved with only the OPFs in its depth direction. This color separation system in an organic image sensor is a great advantage to achieve both compact and high picture quality cameras.

We confirmed that color imaging is feasible using a prototype sensor consisting of three OPFs (for R, G, and B) with three transparent readout circuits that is made with a ZnO thin-film transistor (TFT) array (128x96 pixels, pixel pitch: 100 um). In the prototype sensor, three separate elements that were composed of R-, G-, and B-sensitive OPFs with a readout circuit on three glass substrates were fabricated, and then these elements were stacked to produce a color image sensor. In the prototype sensor, however, not all color component images could come into focus simultaneously because each OPF was separated by a 0.7-mm-thick glass substrate. This was one reason the resolution in the output image degraded. The acceptable focal depth is roughly estimated to be shorter than 20 um when the pixel pitch of the sensor is 5 um.

Therefore, we started developing an organic image sensor consisting of three OPFs with readout circuits that are close to each other (within 10 um in total), separated by thin interlayer insulators, on a single glass substrate. In this case, the interlayer insulator and the TFT fabrication process temperatures are limited up to 150°C because organic materials, which are located under the TFT and interlayer insulator layers, are easily affected by heat. We previously developed basic technologies such as low-temperature fabrication of TFT and interlayer insulator and heat resistant OPF.

In this study, we fabricated a test color image sensor with R- and G-sensitive OPFs by applying In-Ga-Zn-O TFT readout circuits (128x96 pixels, pixel pitch: 100 um) utilizing those basic technologies. The R- and G-sensitive layers were separated by a 10-um-thick interlayer insulator. The whole fabrication process was able to be implemented at 150°C or less without the underlying layer peeling by controlling stress for each thin film, constituting the TFT and interlayer insulator. Even the TFT fabricated at 150°C or less on the interlayer insulator showed operation characteristics excellent enough to read out signal charges accumulated in OPF. Spectral responses of R and G layers were measured, and photoresponses from each OPF were obtained through the TFT array. We also did a shooting experiment using this sensor and successfully took multi-color images with it. This result opens up a way for achieving high-resolution organic image sensor in the future.

9022-19, Session 5

Real-time compact multispectral imaging solutions using dichroic filter arrays

Steve M. Smith, Dave Fish, Pixelteq, Inc. (United States)

The next generation of multispectral sensors and cameras will need to deliver significant improvements in size, weight, portability, and spectral band customization to support widespread commercial deployment for handheld instrumentation and portable field-based cameras. The benefits of multispectral imaging are well established for a growing number of applications including machine vision, biomedical, authentication, and aerial remote sensing environments. However, many OEM solutions demand more compact, robust, and cost-effective production cameras to realize these benefits.

Dichroic Filter Arrays

A novel multispectral imaging implementation uses micro-patterning of dichroic interference filters into Bayer and custom mosaics, enabling true real-time multispectral imaging with simultaneous multi-band image acquisition. Consistent with color camera image processing, individual spectral channels are de-mosaiced with each channel providing an image of the field of view. We demonstrate recent results of 4-9 band dichroic filter arrays in multispectral cameras using a variety of sensors including linear, area, silicon, and InGaAs for applications across UV, visible, near infrared (NIR), and short wave infrared (SWIR) spectral bands. Unlike conventional organic color gels or dyes, dichroic filters can be customized to image discrete narrow spectral bands with bandwidths from ~ 20nm to greater than 100nm. By adjusting the filters' central wavelengths, bandwidths, and other design criteria, the sensor can be optimized to fit a specific application.

Multispectral Camera Examples

Specific multispectral camera implementations range from hybrid RGB + NIR sensors to custom sensors with application-specific VIS, NIR, and SWIR spectral bands. Benefits and tradeoffs of multispectral sensors using dichroic filter arrays are compared with alternative approaches – including their passivity, spectral range, customization options, and development path. Hyperspectral cameras and imaging spectrometers, for example, have their place in the initial research – but are generally to bulky for portable devices and generate too much data to analyze and process in real-time. The dichroic filter array approach can also be combined with other imaging technologies to extend capabilities . Using dichroic filter arrays in multi-sensor cameras can broaden the spectral range and/or number of spectral bands available. Intelligent cameras can provide on-board processing for real-time multispectral results delivered wirelessly from portable cameras and aerial platforms. In addition to showing examples of these multispectral cameras, we will share: (a) images and video demonstrating real-world examples of how multispectral imaging can reveal features beyond our human vision, (b) design considerations unique to multispectral, and (c) current design tradeoffs and production constraints to dichroic filter arrays.

Scalable, Cost-Effective Production

Finally, we report on the wafer-level fabrication of dichroic filter arrays – a key advantage to this method. Patterning these repeating arrays at the pixel level on photodetectors and image sensors enables the scalable production of multispectral sensors, making application-specific cameras much more cost-effective in volume. Bridging the gap from R&D to production, we share a step-wise approach to develop and industrialize a multispectral application along with practical lessons learned along the way.

9022-20, Session 5

A 1024x1 linear photodiode array sensor with fast readout speed flexible pixel-level integration time and high stability to UV light exposure

Takahiro Akutsu, Shun Kawada, Yasumasa Koda, Taiki Nakazawa, Rihito Kuroda, Shigetoshi Sugawa, Tohoku Univ. (Japan)

Linear photodiode array sensors (PDA) are widely used in various typed spectrophotometry, such as absorption based, emission based and combination of them. Following performances are desired to PDA; a fast sampling speed and a high light sensitivity to increase the measurement efficiency, integration time controllability to equalize the photo signal levels of incident light, and a high stability of sensitivity to ultraviolet (UV) light exposure to guarantee the measurement reliability.

We demonstrate a PDA with high stability of sensitivity to UV-light exposure, fast readout speed and integration time controllability using a novel photodiode (PD) structure and circuitry.

Regarding the stability of sensitivity to strong UV-light exposure, it is known that fixed charges and interface states are generated by UV-light exposure, which tends to cause a modulation of Si surface electric field and increase generation/recombination rate. This causes a fluctuation of sensitivity especially to UV-light with short penetration length of several nanometers in Si, and dark current increase. [1] We introduced a surface high concentration layer with a steep dopant profile to a flattened Si surface. This PD structure forms a several nanometer thick surface high concentration neutral layer to suppress the modulation of electric field and to passivate the interface state, and electric field to drift photo carriers. Thus, a high sensitivity and stability to UV-light are simultaneously achieved. [1, 2]

For the readout speed, the shape of PD is a long rectangle. The PD length and the readout speed have a tradeoff. When there is one photo carriers need to move a long distance in the PD one readout path at a short side of the PD, the RC delay as a function of the PD length limits the readout speed. To solve the tradeoff, we designed multiple readout paths and metal wire along the long side of the PD for high readout speed.

Regarding the integration time controllability, in spectrophotometry, PDA is illuminated by dispersed light with a large difference of intensity along wavelength. The developed readout circuitry has a function to flexibly increase or decrease the integration time turned by one clock period per pixel to equalize the photo signal levels of incident light across the whole waveband.

The designed PDA was fabricated through 0.18um 1P3M CMOS technology with buried PD process which has a steep surface dopant profile. The number of effective pixels is 1024 and pixel size is 25umH ? 2500umV . For comparison, a PDA with one readout path at a short side of the PD was also fabricated with the same process. The developed PDAs have achieved 99pC full-well capacity. The time to read the 1024 pixels data is 0.62msec in the developed PDA and 6.8msec in the reference one, respectively. A high sensitivity from 200-1000nm has been confirmed with a discrete PD which has the same dopant profile. Furthermore, the sensitivity degradation did not occur for 10000min strong UV-light exposure.

We will evaluate the stability of the sensitivity of the developed PDA to UV-light up to 10000min exposure.

9022-21, Session 5

A high fill-factor low dark leakage CMOS image sensor with shared-pixel design

Min-Woong Seo, Keita Yasutomi, Keiichiro Kagawa, Shoji Kawahito, Shizuoka Univ. (Japan)

The noise performance of CMOS image sensor (CIS) has been greatly improved since the introduction of the pinned photodiode technology. In addition, the dark leakage in the CISs has been dramatically reduced. But the increasingly advanced functions of scientific cameras require CISs with higher signal to noise ratio (SNR). Therefore the dark leakages component results from the interface defects around field oxide border, e.g., shallow trench isolation (STI) corners, thus become the most critical issue.

A CMOS image sensor using the proposed pixel design is implemented with 0.18-?m standard CMOS technology with pinned photodiodes. It consists of a test pixel pattern array (including the proposed pixel array), a column-parallel folding-integration/cyclic ADC for low-noise read out, a vertical and horizontal shift registers, a reference voltage and current blocks.

A high responsivity and low dark leakage current 1.75 transistors/pixel CMOS image sensor without any process modification is presented. A ring-gate shared-pixel design with a high fill-factor makes it possible to achieve the low-light imaging. As eliminating the STI in the proposed pixel, the dark leakage current is significantly decreased because one of major dark leakage sources is removed.

The each pixel is isolated by a P-well and other highly-doped P-type layers. This means the STI is not used for pixel isolation. This STI-less structure does not only avoid large surface dark leakage contribution but also allows fill-factor improvement by the optimized pixel layout. As a result, the pixel has a 43 % fill-factor and effective number of transistors per pixel is 1.75.

For accurate measurements, a light source dedicated for image sensor evaluation was used with a collimeter. The proposed image sensor has achieved the high sensitivity of 144.6 ke/lx•sec by a high fill-factor. Obviously, the dark leakage current of the proposed structure which does not use the STI structure for pixel isolation is dramatically reduced compared with a conventional type active pixel sensor (APS) which has the STI structure for pixel isolation. The dark leakage current of the proposed pixel is measured to be 104.5 e-/s (median), corresponding to a dark current density J [proposed] of about 30 pA/cm^2. In contrast, the conventional type test pixel has a large dark leakage current of 2450 e-/s (median), corresponding to J [conventional] of about 700 pA/cm^2. Both

pixels have a same pixel size of 7.5?7.5 ?m^2 and are fabricated by a same process.

9022-22, Session 5

Co-integration of a smart CMOS image sensor and a spatial light modulator for real-time optical phase modulation

Timothe Laforest, Antoine Dupret, Arnaud Verdant, CEA-LETI-Minatec (France); François Ramaz, Ecole Supérieure de Physique et de Chimie Industrielles (France); Sylvain Gigan, Institut Langevin (France); Gilles Tessier, Ecole Supérieure de Physique et de Chimie Industrielles (France)

Optical imaging through biological media is strongly limited because of light scattering. This is an issue, especially in medical imaging, when the goal is to detect a millimeter-sized object within a several centimetres thick scattering medium, e.g. for early breast cancer detection. The use of an acousto-optical holographic scheme allows detecting emerging tumors with a resolution of 1 millimeter and foreseeing great progress in breast medical imaging. However, its clinical application is still out of reach because of the complexity of the setup and the limitations of the detection scheme. One of the major problems is that in thick biological tissues, the correlation time of the transmitted intensity through the sample is typically a few milliseconds. Another challenge of this technique is the detection of the relevant signal (i.e. useful to extract optical information), which corresponds to 10% of the total incident light power on the sensor. Moreover, coupling acousto-optic holographic scheme with a spatial phase modulation setup allows phase conjugation and light focusing through the sample on a region of interest. Nevertheless, the phase conjugation setup presents several limitations. Firstly, the modulator and the camera pixels arrays must be perfectly spatially matched which is a complicated, bulky and expensive opto-mechanical process. Secondly, image acquisition, transfer and processing lead to a feedback delay that prohibits controlling the reflection of the wavefront (before changes of the speckle pattern). We present a new light detector-actuator device with in-pixel detection and spatial light modulator (SLM) controlling. The pixel integrates a photodiode with some analog processing circuits and a SLM made of liquid crystals covering the entire pixel. This stacking allows a perfect matching between photodiode and SLM pixels. The image sensor-modulator has been designed using a 130 nm CMOS process. It is able to work in a global shutter mode at 4000 Hz, synchronised with the signal frequency. Such a pixel can thus acts as lock-in system for synchronous detection, which is the goal for an acousto-optic detection. The proposed architecture is also more compact compared to state of the art high speed image sensor. Pixel to SLM connectivity circumvents the latency due to rolling mode of the image readout and the writing of the phase mask. As a consequence this stack allows a 1 ms delay feedback between the end of acquisition signal and the stabilization of the corresponding SLM state. Even if we implement analog operations for optical phase conjugation, more complex wavefront control operations can be integrated.

9022-23, Session 6

(JEI Invited) Compressive sensing underwater active serial imaging systems

Bing Ouyang, Fraser R. Dalgleish, Frank M. Caimi, Anni K. Vuorenkoski, Walter B. Britton, Harbor Branch Oceanographic Institute (United States)

In recent years, the Compressive Sensing (CS) theory has been adopted in many image acquisition applications. In underwater electro-optical imaging systems, through exploiting the significant data sparsity exists due to the lowpass filtering effect associated with the propagation of light through the scattering and absorbing medium (i.e. ocean water), a CS based system enables a more compact, cost-effective and reliable image acquisition hardware to be adopted. Some additional benefits include more adaptive to the changing environment and/or task requirements; less capital intensive system innovation/enhancement due to the algorithmic/software centric paradigm.

In this presentation, the recent development of different types of CS based underwater active serial electro-optic imaging systems currently under investigation at HBOI/FAU will be discussed.

In a frame based CS based serial underwater laser imager [Ouyang et al. 2013b], a series of measurement patterns generated from a Digital Micromirror Device (DMD) based laser illuminator modulates the target plane and a Photomultiplier Tube (PMT) acquires the reflected total photon flux as the corresponding measurements. Multi-scaled measurements matrix design and radiative transfer model based image reconstruction were adopted to mitigate the measurement degradation due to forward scattering and volume backscattering. One constraint for such implementation is that the illuminator and PMT to be on stationary platforms such as the hover-capable autonomous underwater vehicles.

For this reason, a Compressive Line Sensing Underwater Imaging System is also proposed [Ouyang et al 2013a]. It is more compatible with the traditional underwater electro-optical survey platforms, where the imager reconstructs one line of the target at a time, and relies on the forward motion of the platform to complete the image of the entire scene in a whisk-broom fashion. One concept incorporated into this design is the Distributed Compressive Sensing (DCS) [Baron et al.] joint sparsity model to exploit the significant correlation/redundancy among adjacent lines. In this implementation, on the acquisition (encoding) side, each line will be measured independently. The image acquisition hardware setup and degradation mitigation techniques in the aforementioned frame based system are still valid. One difference is that such patterns will be focused in one direction to make it cover one line on the target plane at greater radiant intensity. During reconstruction (decoding), a group of lines will be reconstructed jointly. The number of lines to be clustered together can be a combination of hard reset (i.e., predefined maximum number of lines) and an equivalent of scene change detection (i.e., significant change from previous line).

The current image acquisition system hardware design, simulation results as well as initial test tank experimental results of these two system concepts will be presented. The extension of the designs to aerial platforms as well as some new CS active serial imager concepts will also be discussed.

References

* B. Ouyang et al., “Compressive line sensing underwater imaging system”, SPIE Proceedings Vol. 8717, 2013a.

* B. Ouyang et al., “Compressive Sensing Underwater Laser Serial Imaging System”, JEI, Vol. 22(02), 2013b.

* D. Baron et al., “Distributed compressed sensing”, Rice University TREE-0612, 2006.

9022-24, Session 6

A CMOS time-of-flight range image sensor using draining only modulation structure

Sangman Han, Keita Yasutomi, Keiichiro Kagawa, Shoji Kawahito, Shizuoka Univ. (Japan)

This paper presents new structure and method of charge modulation for CMOS TOF range image sensor using pinned photodiodes. Proposed pixel structure allows us to achieve high-speed charge transfer by generating lateral electric field from the pinned photo-diode (PPD) to the pinned storage-diode (PSD). And generated electrons by PPD are transferred to the PSD or drained off through the charge draining gate (TXD). This structure allows us to realize a trapping less when the charges transfer between PPD and PSD. Therefore, it can reduce the noise that is caused by a transfer gate (TX).

In TOF range image sensors, high speed charge transfer from PPD to PSD is essential. To accelerate the speed of charge transfer, the generation of lateral electric field is necessary. To generate electric field, the width of the PPD is changed along the direction of the charge transfer.

PPD is formed by the p+, n layer on the p-substrate. And PSD is doped by another n type layer for higher concentration than the n layer in PPD. By this pixel structure, the potential difference is created between PPD and PSD. Therefore, photoelectrons generated by PPD can be quickly transferred to PSD. In addition, PSD is doped by additional p layer between n layer and p-substrate for protecting the injection of unwanted charge from the substrate to PSD.

The range is calculated with signals in the three consecutive sub-frames; one for delay sensitive charge by setting the light pulse timing at the edge of TXD pulse, another for delay independent charge by setting the light pulse timing during the charge transfer, and the other for ambient light charge by setting the light pulse timing during the charge draining.

The pixel consists of a PPD, a PSD, a TXD, a TX for readout between the PSD and the floating diffusion (FD) with MOS capacitor for increasing the full well capacity, a reset transistor and a source follower amplifier transistor. To increase the sensitivity of sensor and obtain high speed charge transfer, a pixel is consist of 16 sub pixel, and electrons are merged into one FD connected to MOS capacitor. The pixel array has 313(Row) x 240(Column) pixels and the pixel pitch is 22.4?m. A TOF range imager prototype is designed and implemented with 0.11um CMOS image sensor. The accumulated signal intensity in PSD as a function of the TD gate voltage is measured. The ratio of the signal for the TD off to the signal for the TD on is 33:1. And the response of the pixel output as a function of the light pulse delay has been also measured.

9022-25, Session 6

A high speed 2D time-to-impact algorithm targeted for smart image sensors

Anders Astrom, Combitech AB (Sweden); Robert Forchheimer, Linköping Univ. (Sweden)

The concept of optical flow has been known for more than 30 years [5]. It is also known that this is a memory consuming and computationally demanding task requiring powerful processors.

Recently, it has been described how to implement optical flow for time-to-impact, TTI, detection using the Near-Sensor Image Processing (NSIP) concept [4][6] in a 1D implementation . The resulting performance would be in the order of 10 kHz of time-to-impact calculations. The reason for the high performance is that the TTI algorithm fits very well with the NSIP architecture. TTI is defined as the distance to the object divided by the speed towards the object.

NSIP is a concept described for the first time almost 30 years ago, in which an optical sensor array and a specific low-level processing unit are tightly integrated into a hybrid analog-digital device [1]. Despite its low overall complexity, numerous image processing operations can be performed at high speed competing favorably with state-of-art [2].

In this paper we present a 2D extension of our previously described 1D method [4] for a time-to-impact sensor. As in the earlier paper, the approach is based on measuring time instead of the apparent motion of points in the image plane to obtain data similar to the optical flow. The specific properties of the motion field in the time-to-impact application are used, such as utilizing simple feature points which are tracked from frame to frame. Compared to the 1D case, the features will be proportionally fewer which will affect the quality of the estimation. We will show in the paper how the Local Extreme Points (LEP) are propagating out from the center of the image. There are a number of pixels/positions which will not have a relevant value.

In the paper we give a proposal on how to solve this problem.

In our experiments we have been using three different scenes, ten different speeds, and five different sample lengths. The result shows that we have a good estimate of the true normalized speed. The estimates are in general better than a 10 % error.

The relationship between the estimate normalized speed and the true normalized speed is not linear. In the paper we will show why this is so and how to compensate for this.

Finally, we will give some indication on how a fully custom TTI can be implemented.

REFERENCES

[1] Forchheimer R, Åström A, Near-Sensor Image Processing. A New paradigm. IEEE Trans Image Processing, 3 , 6, 735-746, November, 1994.

[2] Eklund J-E, Svensson C, and Åström A, Implementation of a Focal Plane Processor. A realization of the Near-Sensor Image Processing Concept, IEEE Trans. VLSI Systems, 4, 3, September, 1996.

[3] Åström A, Forchheimer R., Near-Sensor image Processing, Advances in Imaging and Electron Physics, Vol 105, 1999.

[4] Åström A, Forchheimer R., “Low-complexity, high-speed, and high-dynamic range time-to-impact algorithm”, Journal of Electronic Imaging 21(4), 043025 (Oct–Dec 2012)

[5] Horns B. K. P. and Schunk B. G., “Determining optical flow,” Artificial Intelligence, vol. 17, no 1-3, pp 185-204, 1981.

[6] Åström A, Forchheimer R, “Time-to-impact sensors in robot vision applications based on the near sensor image processing concept,” Proc. SPIE 8298, 829808 (2012).

9022-26, Session 6

Real-time 3D millimeter wave imaging using focal plane array of detectors

Daniel Rozban, Avihai Aharon Akram, Amir Abramovich III, Ariel Univ. Ctr. of Samaria (Israel); Natan S. Kopeika, Assaf Levanon, Ben-Gurion Univ. of the Negev (Israel)

Imaging systems in millimeter waves are required for applications in medicine, communications, homeland security, and space technology. This is because there is no known ionization hazard for biological tissue, and atmospheric attenuation in this range of the spectrum is low compared to that of infrared and optical rays. The lack of inexpensive room temperature imaging system makes it difficult to give a suitable implement for the above applications. 3D MMW imaging system based on chirp radar was studied previously using a scanning imaging system of a signal detector. The system presented here proposes to employ chirp radar method with Glow Discharge Detector (GDD) Focal Plane Array (FPA of plasma based detectors). Each point on the object is corresponds to a point in the image and thus it includes the distance information. This will enable 3D MMW imaging system. The radar system requires that the millimeter wave detector (GDD) will be able to operate as heterodyne detector. Since the source of radiation is a frequency modulated continues wave (FMCW), the detected signal as a result of heterodyne detection gives the object depth information in additions to the reflectance of the image. In this work we experimentally demonstrate the feasibility of implementing an imaging system based on radar principles and FPA of detectors. This imaging system is shown to be capable of imaging objects from distances of at least 10 meters.

Conference 9023: Digital Photography X

Monday - Wednesday 3 –5 February 2014

Part of Proceedings of SPIE Vol. 9023 Digital Photography X

9023-1, Session 1

A hardware validated unified model of multi-bit temporally and spatially oversampled image sensors with conditional reset

Thomas Vogelsang, David G. Stork, Rambus Inc. (United States); Michael Guidash, Rambus, Inc. (United States)

We describe a photon statistics based theoretical model of the response to incident light of an image sensor and show that conditional reset and multi-bit temporal oversampling increase the dynamic range significantly. This photon-based modeling approach describes the full image sensor design space of temporal and spatial oversampling either with a binary comparison or with a multi-bit read of each sample. We find excellent quantitative agreement between measurements on custom hardware and our theoretical predictions. We then use this model to show what improvements in dynamic range and low-light response can be achieved by oversampling and what the limits of improvement caused by pixel size and lens parameters are.

We have developed a theoretical model that describes light capture of a photo sensor based on photon statistics, thereby incorporating photon shot noise directly. This model describes the sampling of photons as a series of binary comparisons with a threshold. We showed that multi-bit sampling with an ADC is mathematically equivalent to spatially oversampling the pixel with virtual jots that are sampled with thresholds at the steps of the ADC. Our sensor model can therefore be used to predict and optimize the light response of any binary oversampling sensor, conventional single-sample multi-bit sensors and multi-bit oversampling sensors. The sensor response can be linearized either by a lookup table or by a weighted sum of the results of the individual samplings. We verified this model on hardware using a small test chip. Using the model we demonstrated that sampling policies that use only temporal oversampling (binary or multi-bit) and reset the pixel only conditionally when a threshold has been reached have better low-light response than sampling policies with unconditional reset or spatial oversampling. By calculating the number of photons on the sensor based on target illumination and camera parameters, we were able to compare exposure settings for low-light and bright-light settings of conventional sensors with oversampled sensors both for sensors typical for mobile devices as for DSLR sensors. A significant increase of dynamic range of about 24dB in our example can be seen in all cases. In a typical camera application the dynamic range would be extended to the high end in a low-light situation and to the low end in a bright light situation. The dynamic range of an oversampled mobile camera can be as large as the range of a conventional DSLR in medium or bright light situations. Such a matchup is not possible either at very low-light situations where the pixel size is important to collect as many photons as possible or at very bright light situations where the aperture needs to be changed to let less light on the sensor. While the high end can be further extended in all cases by more oversampling, at the low end an improvement is only possible when more photons can be collected by having a larger pixel area, higher pixel sensitivity or a combination of these approaches. We expect that

the pixel can be designed to achieve higher sensitivity when using our approach as there is no need to have a large full well capacity to handle brightly lit parts of the scene. More generally, low-light response can be improved in camera systems employing sensors having multi-bit oversampling with conditional reset by exposing for the low-light regions of the scene while retaining all of the bright-light information and detail.

9023-2, Session 1

All-glass wafer-level lens technology for array cameras

Palle Dinesen, AAC Technologies (Denmark)

Mobile imaging has undergone a tremendous evolution since the first mobile phone to include a digital camera was introduced more than a decade ago. First generations of mobile phone cameras had very low resolution with fairly large pixels on the image sensor, and the performance of the camera compared to digital still cameras was marginal.

The evolution of CMOS image sensors has enabled a significant downscaling of mobile phone cameras to a point, where pixels are today at or below the diffraction limit of lenses with typical F-numbers. However, at the same time modern day smartphones have also become thinner and thinner, and the camera has in many phones become the factor limiting the thinness of a phone.

Recently, a number of companies have proposed to overcome the thickness challenge by replacing a single-aperture cameras with multi-aperture cameras. There are numerous approaches to array cameras but the basic idea is, that by splitting the image into several channels, each channel can be downscaled, leading to a reduced thickness of the camera compared to a single-aperture camera. In some implementations, the footprint of the array camera will be larger than that of a single-aperture camera, but this is typically not an issue, since smartphone screens keep growing in size, which leaves more room for a camera with a larger lateral footprint inside the smartphone.

Needless to say, the images from the channels of the multi-aperture camera must be “fused” to form a single digital image (or video stream). This can be a computationally very demanding task, and this is where array cameras are helped along by the Moore’s law evolution in smartphone processing power. Fusing images from a multi-aperture camera not only requires smart algorithms, it also needs a lot of computations, which can be done in the application processor, the graphics processor, dedicated image signal processors or a combination of the three. Most recent generations of mobile platforms have very powerful APs, GPUs and even embedded ISPs to support this task

In addition to addressing the thickness challenge of mobile phone cameras, array cameras offer a wealth of additional advantages and features, such as improved low light performance, depth information acquisition, gesture control capabilities, etc.

On top of dedicated sensor architectures and substantial image processing capabilities, a third key ingredient in a successful array camera implementation is the lens system. For conventional single-aperture cameras, the prevailing lens technology is currently injection-molded lenses assembled into a lens barrel along with spacers and apertures stops. The lens barrel is then thread-mounted into a voice-coil motor autofocus structure. This method has been developed and refined to cover every mobile lens from a low-performance 2MP lens to a high-end 13MP lens in a state-of-the-art mobile phone camera.

Array cameras and thread-mounted lenses in barrel is not a good match. An array of plastic lenses in a barrel will require a lot of space in between channels in the array, and image fusion algorithms do in general require a well-defined and well-maintained distance between the channels, which can be difficult with barrel-mounted plastic lenses. Array camera for those and other reasons lend themselves towards wafer-level lens manufacturing technology, where and array of lens stacks is manufactured in one piece.

Wafer-level lens manufacturing has been around for quite some time and 5-6 years was expected to entirely replace plastic lenses in mobile phone cameras, but that has not happened. Wafer-level lenses are today mainly used for low-resolution front-facing cameras.

While “conventional” wafer-level lens technology consist of imprinting polymer lens surfaces onto a glass substrate and then stacking 2 or more layers of such lens wafers on top of each other to form multiple identical lens systems in one process, we have proposed a novel approached of realizing wafer-level lenses monolithically in glass. In this paper, we describe this technology in more detail and demonstrate how it is extremely well suited for array cameras.

Optical design

One of the things that sets glass apart from polymers, when it comes to lenses, is the availability of many different glass types offering a wide span of refractive index and dispersion properties. With polymer lenses, the lens designer is limited to selecting materials from a confined space in the Abbe diagram, whereas glass lens design offers the designer access to a wide variety of optical properties from a much larger area in the Abbe diagram. This allow for more compact lens designs by selecting glass types with a high refractive index for the power lens, and it offers achromatization through selecting glass types with quite different Abbe numbers. This further has the advantage, than lens designs can be made much less sensitive to decentering errors in the manufacturing process that what can be achieved with polymer lenses.

A third advantage worth mentioning is the ability to manufacture thin meniscus lenses, which is a very common lens type in triplet designs. With conventional wafer-level lens technology, because material must be added on both sides of the glass carrier substrate, it is not possible to realize thin meniscus lenses. In contrast, because our wafer-level lenses are molded, we are able to change the base shape of the glass substrate and realize thin meniscus lenses, again something that contributes to the manufacturability of the lens design.

We will present a 3-element lens design suitable for a variety of array camera implementations as well as for HD (720p) single-channel front facing cameras.

Tooling

While an all-glass based approach to wafer-level lenses has numerous advantages as outlined about, it also comes with challenges. Since we replicate the lens surfaces from the tool onto the glass substrate by molding, the mold tools needs to be made of a ceramic hard metal. Machining lens surfaces into a hard metal tool can only be achieved by grinding and not diamond-turning which would be a preferred and more easily controlled process used for tooling for both injection-molding of plastic lenses as well as tools for conventional wafer-level optics.

Grinding of optical surfaces is challenging for a number of reasons, the primary being that the grinding tools wears in the process and this something that needs to be compensated. We are able to achieve form accuracies better than 200 nm (PV) for wafer-level tools manufactured as monoliths. Another critical tolerance is the location of each lens surface on the tool, which will determine the decentering errors in the lens stack. We are able to maintain +/-1 µm absolute position error of lens surfaces on the wafer tools for molding.

Molding

Molding of glass singlet lens is a mature technology, which is routinely used for manufacturing of lenses to be used in combination with plastic lenses for very high-end camera modules. Critical elements such as molding equipment, tool coating, molding process control, and lens centration, are all well developed, and while some of them applies also to our wafer-based technology, there are distinct differences, and so we have developed new solutions in a number of areas.

We have developed proprietary equipment which enables us to exercise much tighter control over the molding process and adapt to minor variations in the process. This is in contrast to conventional glass process, which runs based on a fixed set of parameters without any adaptive control.

In a conventional molding process, lens centration is a matter of relatively simple alignment of the vertices of the two lens surfaces. Now, with a wafer-level technology, all lenses on the wafer must be aligned simultaneously, so that the alignment operation becomes two-dimensional. Tilt also needs to be tighly controlled, and we have developed a 3-axis alignment scheme. A new tool set is run in by doing a first trial run, then thoroughly characterizing the alignment and tilt errors, then adjusting the tool set up after which no further alignment operations are needed.

With our molding technology, we are able to maintain a ± 1µm decentering error between front and back surface and a ±5µm thickness control of the molded wafers.

Stacking

Once wafers are successfully molded, two or three wafers must be stacked in order to form the full wafer stack. Again, we have developed proprietary technology for wafer stacking, as we find that conventional wafer bonding equipment, which is commonly used for wafer alignment, does offer sufficient precision. We have successfully developed an interferometric alignment technology with 6 degrees of freedom of alignment of each of the wafers to be stacked. Using this technology, we are able to achieve ±2µm alignment in the stacking process

Applications for array cameras

As highlighted above, wafer-level lenses are extremely well suited for array cameras. In fact, it can be argued that array cameras will be difficult to realize without the existence of wafer-level lenses. There are, however, several challenges that needs to be addressed for a successful implementation of wafer-level lenses in array cameras. When an array of lenses are attached as one monolith, it is important that all of the lenses in the array share the same back-focal length, such that all channels in the array is in focus at the same time. Furtermore, for an array camera it is of particular importance that the MTF performance throughout the entire image field. In particular, it is important to be able to maintain good MTF performance in the corners of the image, in order for the image fusion to be successful.

We will demonstrate how our all-glass wafer-level technology addresses those challenges paving the way for successful deployment of array cameras in smartphones.

9023-3, Session 1

Real time algorithm invariant to natural lighting with LBP techniques through an adaptive thresholding implemented in GPU processors

Sergio A. Orjuela Vargas, Jennifer C. Triana, Andres Rodriguez, Univ. Antonio Nariño (Colombia); Wilfried Philips, Univ. Gent (Belgium); Juan Pablo Yañez, Univ. Antonio Nariño (Colombia)

Real time applications in video processing require low computational cost algorithms that allow processing a considerable number, commonly 25, of frames per second. A challenge in outdoor visual scenes is to deal with environmental conditions such as natural lighting. We propose in this approach an adaptive threshold for extracting robust features with invariance to natural lighting when applying LBP techniques. LBP techniques characterize local intensity variations in the neighborhood of a pixel by performing Boolean comparisons of the intensity values on the neighborhood of a pixel with the intensity value of the pixel. We use the residuals of the Boolean comparisons to compute the adaptive threshold. For this, the probability distribution of the residuals is first modelled with a generalized Gaussian distribution. Then, the parameters of the theoretical distribution model are used to compute the adaptive threshold. Real time implementation of the LBP algorithms can be performed by computing the Boolean comparisons in parallel for blocks of pixels using a Graphic Processing Unit (GPU). GPUs have been used to accelerate the execution of image analysis algorithms thanks to the use of a vast number of simple, data-parallel, deeply multithreaded cores and high memory bandwidths. These main features make the use of the Graphic Process Units (GPU) suitable for processing large-scale data-parallel load of high-density computing, turning it into an attractive area for research and development. The device’s programming is possible thanks to the use of the Compute Unified Device Architecture (CUDA) programing, only supported by NVidia Graphics Cards. In

this approach we implemented a set of classic image algorithms sequentially in a typical c++ implementation first, making use of the OpenCv Libraries and then we compare the speed performance of the same algorithms for evaluation of texture by using CUDA programing in a Windows 7x64 system with 2.90 GHz core processor, 4GB RAM and a GeForce GT 525M. We estimate the parameters of the generalized Gaussian distribution for different number of lines on an image. We conduct experiments on images where the lines and the background have different contrasts. This experiment suggests that the parameters change with the number of features in the image independently while are independent with the contrast. We use the LBP technique based on Symmetries. This is an alternative approach to group LBP-codes based on symmetry and group theory, and topology. It includes mirror, rotational and complement invariants. This is particularly useful for feature extraction in videos since the main features of the visual scene can be codified with different colors based on the symmetries using a look up table. We test the methods on four videos captures during day and night. We obtained an average of 0.005 in terms of the error between the experimental and the theoretical distributions. Experiments suggest that the parameters of the generalized Gaussian distribution change with the number of features in the image independently while are independent with the contrast. The results of this approach are of interest to determine patterns, to identify objects or to detect background in a further step. However, an extra step for blur correction must be still included considering that the images of the frames at night are commonly blurred.

9023-4, Session 1

Embedded FIR filter design for real-time refocusing using a standard plenoptic video camera

Christopher Hahne, Univ. of Bedfordshire (United Kingdom)

A novel and low-cost embedded hardware architecture for real-time refocusing based on a standard plenoptic camera is presented in this study. We introduce a simple solution to synthesize refocused image planes directly from micro images by omitting the process for the commonly used sub-aperture extraction. Therefore, intellectual property cores, containing systolic FIR filters, are developed and applied to the field programmable gate array XC6SLX45 from Xilinx. Enabling the hardware design to work economically, the FIR filters are composed of stored product as well as upsampling and interpolation techniques in order to achieve an ideal relation between image resolution, delay time, power consumption and the demand of logic gates. This research aims to provide real-time refocused video content for a standalone video camera device. Nevertheless, the proposed hardware design can certainly be utilized for refocusing still plenoptic pictures in real-time as well. The video output is transmitted via High-Definition Multimedia interface having a resolution of 720p at a frame rate of 60 fps according to the HD ready standard. Examples of our synthesized refocusing planes are exposed.

9023-5, Session 2

Mobile multi-flash photography

Xinqing Guo, Univ. of Delaware (United States); Jin Sun, Univ. of Maryland, College Park (United States); Zhan Yu, Univ. of Delaware (United States); Haibin Ling, Temple Univ. (United States); Jingyi Yu, Univ. of Delaware (United States)

Multi-flash (MF) photography takes successive photos of a scene, each with a different flashlight located close to the camera's center of projection (CoP). Due to the small baseline between the camera CoP and the flash, a narrow sliver of shadow would appear attached to each depth edge. By analyzing shadow variations across different flashes, we can robustly distinguish depth edges from material edges. MF photography hence can be used to remove the effects of illumination, color and texture in images as well as highlight occluding contours. Previous MF cameras, however, tend to be bulky and unwieldy in order to accommodate the flash array and the control unit. In this paper, we present a mobile MF photography technique suitable for personal devices such as smart phones or tablets.

Implementing mobile MF photography is challenging due to restricted form factor, limited synchronization capabilities, low computational power and limited interface connectivity of mobile devices. We resolve these issues by developing an effective and inexpensive pseudo flash-camera synchronization unit as well as a class of tailored image processing algorithms.

Our prototype mobile MF device consists of four LED flashes, a photocell, and a micro-controller, which is used to trigger the LED flashes. To control the micro-controller, the simplest approach would be to directly use the mobile device's external interface. However, it requires additional wiring on top of the already complex setup and will occupy the USB interface and limit the use of other application. Our strategy is to implement a cross-platform solution: we use the original flash on the mobile device to trigger the LED flashes. Specifically, the mobile flash first triggers the flash ring via an auxiliary photocell. It then activates a simple micro-controller to consecutively trigger the LED flashes in sync with the mobile camera's frame rate, to guarantee that each image is captured with only one LED flash on.

To avoid the device's flash to interfere with the LED flashes, we initiate the image acquisition process only after the device's flash goes off. The frame rates of the camera and the LED flash ring are set to be identical by software (e.g., the AVFoundation SDK for iOS) and by micro-controller respectively. After acquiring four images, we turn on the device's flash to stop the acquisition module. We also provide a quick preview mode to allow users to easy navigate the captured four images. If the user is unsatisfied with the results, with a single click, he/she can reacquire the image and discard the previous results.

Conceptually, it is ideal to capture images at the highest possible frame rate of the device (e.g., 30 fps on iPhone 4S). In practice, we discover that a frame rate higher than 10 will cause the camera out-of-sync with the flash. This is because the iPhone and the Arduino micro-controller use different system clocks and are only perfectly sync'ed at the acquisition initiation stage. In our implementation, we generally capture four flash images at a resolution of 640x480 images in 0.4s. The low frame rate can lead to image misalignment since the device is commonly held by a hand. We compensate for hand motion by applying image registration directly on mobile devices.

A unique feature of our system is its extensibility, i.e., we can potentially use many more flashes if needed. The Arduino pro mini microcontroller in our system has 14 digital I/O pins: one serves as an input for the triggering signal and the others as output for the LED flashes. Therefore, in theory, we can control 13 flashes with minimum modification.

To process the acquired MF images, we further develop a class of fast mobile image processing techniques for image registration, depth edge extraction, and edge-preserving smoothing. Traditional MF photography assume that the images are captured from a fixed viewpoint. In contrast, our mobile MF photography uses a hand-held device and the images are usually shifted across different flashes as we capture with a low frame rate. Extracting depth edges without image alignment will lead to errors. In particular, the texture edges are likely to be detected as depth edges. We therefore implement a simple image registration algorithm by first detecting SIFT features and then use them to estimate the homography between images. This scheme works well for scenes that contain textured foreground or background. It fails in the rare scenario that the scene contains very few textures and the shadow edges become the dominating SIFT features in homography estimations.

Once we align the images, we adopt the shadow traversing algorithm in [Raskar et al. '04] to extract the depth edge. We first convert the color images to grey scale and apply Gaussian smoothing. Then we construct a maximum composite image and take the ratio of a shadow image with the maximum composite image to detect shadow regions. The ratio is close to 1 for non-shadow pixels and is close to 0 for shadow pixels. A pixel on the depth edge must transition from the non-shadow region to

the shadow region and we apply Sobel filter on each of the ratio images to detect such transitions. In the final step, we apply a median filter to the depth edge image to further suppress the noise.

From the depth edge image, we can further perform post-processing techniques for synthesizing various non-photorealistic effects.

LINE-ART RENDERING Line-art image is a simple yet powerful way to display an object. Raskar et al. convert the edge image to a linked list of pixels via skeletonization and then re-render each edge stroke. However, it is computationally expensive. We adopt a much simpler approach using simple filtering. We first downsample the image by bicubic interpolation, then apply the gaussian filter, and finally upsample the image. Both bicubic interpolation and gaussian filter serve as low pass filters, which will blur the binary depth edge image. Also users are capable of adjusting the kernel size to control the smoothness. Our processing pipeline is simple, making it suitable for implementation on the mobile platform.

IMAGE ABSTRACTION The most straightforward approach is to use edge-preserving filters such as bilateral filters or anisotropic diffusion to suppress texture edges while preserving depth edges. A downside of this approach is that the result may exhibit color blending across the occlusion boundaries. This is because bilateral filters do not explicitly encode the boundary constraint in the blurring process. To avoid this issue, we apply anisotropic diffusion instead. Specifically, we diffuse the value of each pixel to its neighboring pixels iteratively and use the depth edge as constraints.

IMAGE THUMBNAILING Image thumbnailing reduces the size of the normal image for better organizing and storing. By using bicubic interpolation, we can downsample the texture de-emphasized image to create a stylized thumbnail image. The depth edges are preserved while the texture regions are blurred, making it suitable for creating icons.

we can also apply depth edge to enhance the performance of object category classification. Specifically, depth edges can serve as feature filter which help high-level vision tasks to get “purified” shape related features. Here we use similar bag-of-visual-word classification framework for evaluation on a dataset collected by the proposed mobile multi-flash camera. The main idea of bag-of-visual-word (BOW) approach is to represent image as histogram of visual words. 128-dimensional SIFT descriptor is used as independent feature. The dictionary of visual words is learned from training data using clustering method such as k-means. Each training and testing image is represented by histogram of visual words in the dictionary. A classifier is then learned in the space of these visual words for classification task. In this experiment we use Support Vector Machine (SVM) due to its simplicity and discriminative power. As for implementation detail, we chose the LibSVM package and Gaussian kernel.

We have implemented our mobile MF system on an iPhone 4S. iPhone 4S features a 1 GHz dual core, a 512 MB RAM and an 8 megapixel camera with a fixed aperture of f/2.4. All examples in this paper are captured and rendered at an image resolution of 640x480. The images are captured under indoor conditions to avoid outdoor ambient light overshadowing the LED flash light which would make it difficult to identify shadow regions.

We showcases the potential of exporting computational photography techniques onto mobile platforms. We test our system on several different objects, including a cowboy model in front of a white background, a headstand mannequin in front of a highly textured background and a complex plant that are covered by leaves and branches. In all senarios our mobile MF camera produces comparable results as conventional MF camera.

For object category classification, we created a dataset containing 5 categories similar to the “Category-5” dataset used in [Sun et al. ‘10]. Each of the 5 categories contains 25 images (accompanied with depth edge images) taken from 5 objects. For each object, images are taken from 5 poses with 5 different background. Each image is generated along with depth edges using the proposed mobile multi-flash camera. Standard bag-of-visual-word (BOW) and BOW with depth edge filtering (BOW+DE) are compared to evaluate the effectiveness of proposed camera. Training and testing sets are randomly divided into half for each run and the experimental result is summarized over 100 such random splits. The result has shown that using depth edge images has significant improvement (about 10%) in recognition rate. This result is consistent with that found in [Sun et al. ‘10]. It suggests that the proposed mobile multi-flash camera shares the similar performance with traditional multi-flash camera system but it is much compact and light-weighted.

9023-6, Session 2

Stereo vision based depth of field rendering on a mobile device

Qiaosong Wang, Zhan Yu, Univ. of Delaware (United States); Christopher Rasmussen, University of Delaware (United States); Jingyi Yu, Univ. of Delaware (United States)

The Depth of Field effect is a useful tool in photography and cinematography because of its aesthetic values. However, capturing and displaying dynamic DoF effect was until recently a quality unique to expensive and bulky movie cameras. In this paper, we propose a computational approach to generate realistic Depth of Field effects for mobile devices such as tablets. We first calibrate the rear stereo cameras and rectify stereo image pairs through FCam API, then generate a low-res disparity map using graph cuts stereo matching and subsequently upsample it via Joint Bilateral Upsampling. Next we generate a synthetic light field by warping the raw color image to nearby viewpoints according to corresponding values in the upsampled high resolution disparity map. Finally, we render dynamic DoF effect on the tablet screen with light field rendering. The user can easily capture and generate desired DoF effects with arbitrary aperture sizes and focal depths with the tablet only and no additional hardware or software required. The system has been tested in a variety of environments with satisfactory results.

9023-7, Session 2

Comparison of approaches for mobile document image analysis using server supported smartphones

Suleyman Ozarslan, P. E. Eren, Middle East Technical Univ. (Turkey)

Recent developments in mobile device technology have increased capabilities of these devices. Majority of current mobile devices have powerful CPUs, large RAMs, fast cellular and Wi-Fi networks, high resolution displays and megapixel cameras. This trend enables emergence of various new approaches in solving problems, such as analyzing document images captured by mobile device cameras, which requires powerful processing and high resolution image capturing capabilities. On the other hand, mobile devices also have some resource constraints, such as power and bandwidth consumptions. Researchers propose different methods to overcome these limitations. One method is reducing resource consumption using in-device solutions, such as dynamic voltage scaling, dynamic power management of disks, middleware based adaptations, and application level adaptations. Another approach is offloading resource-intensive parts of the process to external sources, such as remote servers.

Majority of the document image analysis applications use Optical Character Recognition (OCR) for recognition of text in documents. Some applications also apply post-OCR processes after OCR text recognition. For instance, text correction is applied after OCR in order to correct erroneously recognized characters caused by image imperfections, such as uneven lighting, perspective distortion, text skew, text misalignment, and focus loss. In this study, we use our developed receipt scanning application as a sample document image analysis application. This application automatically extracts information from store receipt images captured by mobile phones. Extracted information includes the store name, address, phone number, the purchase date and time, the receipt

number, the list of products and the corresponding prices, the total price and the paid taxes. This application can be used for different purposes, such as expense tracking and price comparison. The receipt scanner application includes two phases. The first phase includes recognition of information in store receipt images by using OCR. The OCR process includes several resource-incentive sub-processes, such as preprocessing, segmentation, feature extracting, and character recognition. In this application, we use the Tesseract OCR engine for text detection and recognition. The Tesseract OCR engine has implementations running on both mobile and desktop operating systems. The Tesseract engine is trained with a training set of receipts in order to obtain better OCR results. In the second phase, a text correction algorithm is applied on the extracted text in order to correct erroneously recognized text pieces. The second phase includes segmentation, row determination and correction sub-processes. We use the Levenshtein distance in order to replace wrong information with the right one in the correction algorithm. The correction algorithm uses a database which includes information about products and stores.

In this study, we explore tradeoffs of three approaches for mobile document image analysis in terms of power consumption, consumed CPU resource, bandwidth, time, and OCR accuracy. The first approach is performing the document image analysis process completely on the mobile phone, and the second approach is performing this process on the remote server. The third approach includes splitting the document image analysis process between the mobile device and the remote server. All of the three approaches analyze images captured by the mobile device and show the results on the display of the mobile device.

First, we evaluate the in-phone approach. As mentioned before, in this approach, all processes are performed in the mobile phone. At first, the user captures an image of the store receipt. Then, information on this receipt image is extracted using the OCR engine running on the mobile phone. Next, the text correction process running on the mobile phone corrects erroneous characters on the output of the OCR process. Finally, corrected text is displayed on the mobile phone. In this evaluation, we measure the following metrics: power consumption, consumed CPU resource, consumed time, and character accuracy rate.

Second, the remote server approach is evaluated. In this approach, all resource consuming processes are offloaded to the remote server. Initially, the user captures an image of the store receipt. Next, this image is sent to the remote server. Then, the OCR engine running on the remote server extracts the text on the received image. After that, the text correction algorithm running on the remote server corrects the output text of the OCR process. Finally, the corrected text is sent to the mobile phone and displayed to the user. In the evaluation of this approach, we measure metrics in the first channel: power consumption of the mobile phone, consumed CPU resource of both the mobile phone and the remote server, character accuracy rate, and consumed time on the mobile phone, the remote server and communication channel. Moreover, the consumed bandwidth during sending images from the mobile phone to the remote server, and sending the text from the remote server to the mobile phone are measured.

Third, the hybrid approach is evaluated. In this approach, some of the tasks are performed on the mobile device, while other tasks are performed on the remote server. Different scenarios are identified for these experiments. In the first scenario, the OCR process is performed on the mobile phone and the text correction process is performed on the remote server. Only texts are transferred in this scenario. In the second scenario, the OCR process is performed on the remote server, and the text correction process is performed on the mobile device. The image is transferred to the remote server and the text output is transferred to mobile device in this scenario. The third scenario incorporates downscaling of the image resolution and sending the downscaled image to the remote server. Then, the server performs the OCR and text correction processes and sends the corrected text to the mobile phone. In this scenario, the consumed bandwidth and communication delay between the server and the mobile device are reduced. However, downscaling of the image is an extra task for the mobile device, and the character accuracy rate is expected to reduce. In this approach, same metrics in the second approach are measured. Finally, the results of all the experiments regarding the in-phone, remote server and hybrid approaches are compared.

In addition to the use of still images for information extraction, the same comparisons are carried out for mobile phone captured videos as well, where the resource demands are more significant. The results obtained are compared against the image-based results in order to highlight the trade-offs.

9023-8, Session 2

UV curing adhesives optimized for UV replication processes used in micro optical applications

Andreas Kraft, Markus Brehm, Kilian Kreul, DELO Industrial Adhesives (Germany)

Nowadays, consumer electronics cannot be imagined without miniaturization anymore. The devices' variety of functions is ever-increasing, while production costs are being cut down. Optical sensors, LED Flash light and array cameras use special microoptical elements with lenses made of an innovative, UV-curing optical material.

In the last decade, wafer level technologies for micro optics manufacturing have reached a high level of confidence and are supposed to be one of the manufacturing technologies to address future requirements for even smaller and thinner devices.

The challenges now for material suppliers is in the optimization of the formulations for high optical transmission with good thermal stability in the processes like lead free soldering (260°C). On the other hand, to reach the volumes and cost requirements in mobile phone applications, automatic processing with a high output must be targeted.

This presentation will give an overview on developments and achievements of UV curing, reflowable optical materials, used to manufacture micro optical components.

In this presentation we will go through adhesive properties relevant in micro optics manufacturing, and then evaluate the following development challenges:

• Material selection: Acrylic versus epoxy materials: Benefits against disadvantages

• Curing speed optimization for epoxy based materials and the influence to transmission

• Optimization of interaction for different stamp type materials (e.g. PDMS)

In conclusion we found that there are many correlations between parameters and a close relation to the customer is necessary to clearly define the optical and processing requirements and then fit the best material to his application.

9023-9, Session 2

Mobile microscopy on the move

Woei-Ming S. Lee, Australian National Univ. (Australia)

Traditional optical materials are ill-suited for mobile microscope because of rigid material properties and high fabrication cost. The mechanical properties, material cost and optical efficiency of elastomer (polydimethylsiloxane, PDMS) satisfy the requirements of a field portable microscope lens. However, existing soft lithography step requires a pre-fabricated master mold for imprinting which adds considerable cost and time towards lens fabrication. To overcome this limitation, I recently established a revolutionary approach, where high quality PDMS lenses are fabricated without any imprinting step: mold-free. In contrast to existing mobile microscope device, my existing microscope device (elastomer lens ~ USD$0.002 and light emitting diode ~USD $2) costs a mere USD$ 2.002 and weighs ~0.01 grams. The entire microscope, occupied only ~15 mm in diameter, ~5 mm thickness, which is around two times smaller than any existing designs In this presentation, I

shall illustrate the fabrication steps, the imaging performance and the integration of the lens to a regular smartphone device.

9023-10, Session 3

No training blind image quality assessment

Ying Chu, Xi'an Jiaotong Univ. (China) and Shenzhen Univ. (China); Xuanqin Mou, Xi'an Jiaotong Univ. (China); Zhen Ji, Shenzhen Univ. (China)

Description of purpose:

Current general purpose blind image quality assessment (BIQA) methods usually learn to evaluate the image quality by regression from human subjective scores of the training samples [1]. The most popularly adopted learning tool is support vector machine (SVM). However, the training and learning process in SVM is like a black box, we have no explicit explanation of how the image quality is affected by the characteristic of image features.

In our recent work, a novel BIQA metric based on statistical independence was proposed and called STAIND [2]. We observed the joint histogram of neighboring divisive normalization transform (DNT) coefficients in natural and degraded images, and represented the statistical dependency between neighbor DNT coefficients by mutual information of the joint histogram, and defined the mutual information as perceptual feature. The features and the corresponding subjective scores of the training images were sent into the SVM to learn the regression modal, which was used to predict the objective scores of the testing images.

Although the extracted features reflect the image quality well, just like the other state of the art SVM based BIQA metrics, STAIND has some inherent drawbacks. First, it needs training and learning, which are complicated and time-consuming. Secondly, the IQA metric has no explicit expression, which blocks us to further understand the mechanism of how the perceptual features influence the image quality. The purpose of this paper is to find the way to avoid training and learning, and to figure out a BIQA metric with explicit expression.

Methods:

Take the horizontal spatial DNT neighbors for example, Fig. 1 implies the statistical independence in natural image. Despite the variation of values for one of the neighbors, the conditional histogram curves look almost the same. We recently found that this curve conformity will be changed in distortion images. For instance, Fig. 2 exhibits the variation trend for JPEG distortion. From level 1 to level 5, the more the degradation is, the more severe the variations are. A counter example comes from the white noise contamination, where the curves become more and more consistent when the noise level increases. However, in either case the degree of the curve conformity could serve as an intuitive IQA perceptual feature.

To quantitatively present the curve conformity, the mean conditional histogram (red thick line in Fig. 3) is defined:

where represents the joint conditional histograms (blue dashed lines in Fig. 3). To reduce the influence from the edge area, only the central half conditional histograms are selected from the joint histogram.

The Kullback-Leibler (KL) distance between the mean conditional histogram and each of the joint conditional histogram is calculated, and the whole curve conformity is defined:

where represents the th perceptual feature representing the curve conformity of the th joint conditional histogram group, is the dimension of the feature space, is the KL distance between the conditional histogram and the mean conditional histogram .

According to the feature selection strategy in Ref. [2], there are three neighborhood relationships: scale, orientation and space. In our former work we concluded that the spatial neighbor relationship is crucial to the performance of the BIQA model design. In this paper, we extract perceptual features from the four first order spatial neighborships, as shown in Fig. 4. Since the wavelet decomposition includes three scales and four orientations, the dimension of the feature space equals to 4?3?4 = 48. For each image, we can extract a 48 dimensional vector containing its perceptual quality information.

We observed the perceptual vector of the natural reference images in the three popular public databases: LIVE, CSIQ and TID2008. Although the vectors look quite different in the feature space, they do have some latent characteristics in common. For example, the norm of the vectors , i.e., , are very similar to each other. The mean values and standard deviations of in each database are compared in Table 1. Thereinto, , and represents the norm of the subcomponents , and , which represent the perceptual vectors in scale1, 2 and 3 respectively:

From Table 1 we can see that not only the of the natural images in one special database tends to be constant, the aggregation occurs in all the three databases, and the mean values in LIVE, CSIQ and TID2008 are quite close. In single scale, similar phenomenon is also observed. Therefore, we assume that for all images capturing natural scenes, there exists a "natural" clustering center in the characteristic space we explored.

Anyway, the change rule for distortion images in this space is what we concern more, so does the relevance with BIQA model design. To better observe the clustering phenomenon and to find the change law, the spatial point for all the natural and distorted images in the LIVE database are illustrated in Fig. 5.

From Fig. 5, an intuitive idea to design the BIQA metric is to define the Euclidean distance between and the clustering center of the natural images in the LIVE database, i.e., :

We call the proposed metric as CC-DIST, since it calculates the distance in Eq. 6 according to the curve conformity in degraded and natural images. Meanwhile, there is a better way to define the metric by constructing a feature-subjective score dictionary:

where represents the th subjective score of the natural or distorted image in the LIVE database, which has 29 natural images and 779 distortion images in total, hence equals to 808.

For each test image, we compare its calculated with the in the (see Eq. 8), and sorted in ascending order to select the nearest neighbors, together with their subjective scores , to construct the overall quality label:

where is a parameter to control the decay rate, which is set to 32; and equals to 80, i.e., ten percent of.

We call the improved metric in Eq. 10 as CC-DICT, and compared the CC and SROCC values in the four common distortion types (JPEG2000, JPEG, white noise and Gaussian blur) among the three non-SVM BIQA metrics in the LIVE, CSIQ and TID2008 databases respectively (see Table 2, 3 and 4).

From Table. 2 and Table. 3, we can see that the performance of CC-DICT in LIVE and CSIQ databases are fairly well, the mean CC and SROCC values are above 0.83, remaindering the fact that our method needs no reference image, as well as training and learning phase. Although CC-DIST and CC-DICT are not comparable to QAC [1], they perform better than pLSA [3] in WN, and CC-DICT defeats pLSA in the overall situation. The proposed metrics behave a little weak in the TID2008 database, especially for the white noise and Gaussian blur distortion.

New or breakthrough work to be presented:

1. The degree of the curve conformity of the joint conditional histograms between neighboring DNT coefficients is a useful perceptual feature for BIQA model design;

2. The proposed perceptual vectors for all natural images cluster to one natural centroid in the feature space;

3. The Euclidean distance between the perceptual vector of the distortion image and the common natural centroid could be used to design the BIQA metric;

4. The perceptual vectors and the corresponding subjective scores for the natural and distortion images in the LIVE database could be utilized to construct a dictionary, to further interpolate the IQA label for the test image.

Conclusions:

By comparing the curve conformity of the joint conditional histograms between neighbor DNT coefficients, a novel explicitly expressed BIQA metric is presented. The proposed method needs no training and learning, and reflects the relationship of the image quality and the perceptual features well. Simulation results in LIVE, CSIQ and TID2008 databases all prove its availability and effectiveness.

References:

[1] W. Xue, L. Zhang and X. Mou, "Learning Without Human Scores for Blind Image Quality Assessment," CVPR 2013, Jun. 2013.

[2] Y. Chu, X.Q. Mou, W. Hong and Z. Ji, "A Novel Blind Image Quality Assessment Metric and Its Feature Selection Strategy," IS&T/SPIE 2013, Feb. 2013.

[3] A. Moorthy, G. Muralidhar, J. Ghosh and A. Bovik, "Blind Image Quality Assessment without Human Training using latent quality factors," IEEE Signal Processing Letters, 19(2):75-78, 2012.

9023-11, Session 3

Description of texture loss using the dead leaves target: current issues and a new intrinsic approach

Uwe Artmann, Leonie Kirk, Image Engineering GmbH & Co. KG (Germany)

The computing power in modern digital imaging devices allows complex denoising algorithms. The negative influence of denoising on the reproduction of low contrast, fine details is also known as texture loss. Over the last years, several approaches have been presented to describe this important image quality aspect.

Using the dead leaves structure is a common technique in this field and is currently discussed as a standard method in workgroups of ISO and CPIQ (Cell phone image quality group, part of IEEE).

The so called SFR_DeadLeaves approach compares the power spectrum of the target (which can be modeled) with the measured power spectrum in the image. An additional measurement of the noise power spectrum corrects for the added noise by the imaging system.

We present our experience using this method. Based on real camera data of several devices and additional simulation data, we can point out where weak points in the SFR_DeadLeaves method are and why results should be interpreted carefully. The image processing as a combination of contrast enhancement, sharpening and denoising can be tuned that way, that the device gets good numerical results without actually improving the texture loss or the image quality in general.

The SFR_DeadLeaves approach follows the concept of a semi-reference-method, so statistical characteristics of the target are compared to statistical characteristics in the image. In the case of SFR_DeadLeaves, the compared characteristic is the power spectrum. The biggest disadvantage of using the power spectrum is, that the phase information is ignored, as only the complex modulus is used.

We present a new approach, our experience with it and compare it to the SFR_DeadLeaves method.

The new method follows the concept of a full-reference-method, which is an intrinsic comparison of image data to reference data. As we maintain the full image information, we can describe the influence of noise reduction directly. Additionally we gain information about other aspects of image quality like sharpening and image noise. The noise level is measured on structures. This is especially interesting as the standard methods for noise measurement are easily fooled by noise reduction.

9023-12, Session 3

Electronic trigger for capacitive touchscreen and extension of ISO 15781 standard time lags measurements to smartphones

François-Xavier Bucher, Frédéric Cao, Clément Viard, Frédéric Guichard, DxO Labs (France)

This article presents a new electronic device that simulates a user touching the capacitive touchscreen of a smartphone and synchronizes trigger action with the LED timer to measure shooting time lag and shutter lag according to ISO 15781:2013. The device and protocol also extends the time lag measurement beyond this standard with negative shutter lag, a feature that is more and more common on smartphones.

Shooting time lag is the delay between the time the shutter is pressed and the time the image is taken. It is one of very important aspect of photographic user experiences as less than 1 second delay makes a big difference in capturing the image intended by the photographer whether it is a child playing, a beautiful smile or any scene with action. This delay is coming from many time consuming operations within the camera such as camera controls (auto-exposure, autofocus) or stabilization start-up.

The measurement principle is to synchronize the trigger event with a timing device that is being photographed and to measure the elapsed time between the trigger and the moment the scene is actually captured by the camera. This synchronization can be done either by inserting a switch between the trigger button and the finger of the operator, or using a device to trigger the camera synchronously with the timing device.

However, most recent mobile devices camera have two features requiring changes in the measurement protocol described in ISO 15781:2013 to achieve relevant results. First, the negative shutter lag have been described in a companion paper also published in this conference and secondly, the wide use of camera trigger directly from the touchscreen requires a new device to control the camera as well as some changes in the protocols. It finally allows large automated series of captures with a device compatible with most smartphones available today, and thus allows a more accurate estimation of time lags.

This electronic only solution takes advantage of the working principle of capacitive touchscreens and very accurately simulates a user touching the screen of a smartphone to take a picture; since it is purely electronic, the accuracy is better than 10µs when 1ms is typically achieved with mechanical solutions.

As visible in the above photograph this capacitive trigger is placed on the virtual shutter button (the whole preview on the above smartphone). This trigger is also sending a synchronization signal to the LED timer, presented in a companion paper in this conference and visible in the above photograph. The timer records its state defined by the positions of the LEDs when receiving this signal. This state is read on a PC via a USB port. Time lag is calculated by comparing the state of the timer at trigger with its state in the captured image. Recording the state of the timer also allows to measure zero shutter lag or negative time lags. It is worth noticing that since the current ISO standard recommends the timer to start with the trigger, measuring negative shutter lag is impossible, while our system is designed to provide measurements even with negative shutter-lag. The electronic trigger also allows more versatile user interface: the duration of the simulated push can be user defined, and the signal can be synchronized either on pushing or releasing the trigger button. All the components of this hardware are controlled by computer. Therefore, it is possible to control any device with a capacitive touchscreen with this application that is independent from the device operating system. For instance, it is possible to automatize a series of captures. With an automatic image analysis algorithm, it is possible to make a large number of unit measurements, and derive statistical analysis, since the operator does not have to interact with the camera.

As an experimental illustration, we thereafter comment a series of measurements performed on a device with zero shutter lag. These measurements show that shutter lag is not repeatable. Its average value and its variation are correlated with exposure time and thus with scene illumination. This is illustrated in the graph below that shows the

probability density for measured exposure time of 60, 50 and 10ms on series of more than 200 images each.

This behavior is explained by the capture is synced with the sensor frame rate, so shutter lag repeatability cannot be better than +/- one frame time. So, a large amount of measurements is necessary to reach an accuracy better than the frame-time by statistical analysis. For example, to achieve 5ms accuracy on the shutter lag average value in usual conditions, more than 100 shots are necessary, even though ISO 15781:2013 Standard only requires 10 shots. Furthermore, even if the camera application of this smartphone freezes a preview that seems post-trigger, experimental measurements show that the shutter lag is negative, which means that the recorded image is actually captured before the capture command.

Shutter lag and shooting time lag measurements were also performed on other smartphones, DLSR and DSC. Time lag variability was found on many devices, in particular on all devices with Live View (and therefore a continuously working sensor). Shooting time lag measurements that also include autofocus show even wider dispersions. The comparison of various imaging devices eventually highlighted the relevance of using several statistical indicators to evaluate the performance of a system.

9023-13, Session 4

Space-varying blur kernel estimation and image deblurring

Qinchun Qian, Bahadir K. Gunturk, Louisiana State Univ. (United States)

No Abstract Available

9023-14, Session 4

Super-resolution restoration of motion blurred images

Qinchun Qian, Bahadir K. Gunturk, Louisiana State Univ. (United States)

No Abstract Available

9023-15, Session 4

To denoise or deblur: parameter optimization for imaging systems

Kaushik Mitra, Rice Univ. (United States); Oliver Cossairt, Northwestern Univ. (United States); Ashok Veeraraghavan, Rice Univ. (United States)

In recent years smartphone cameras have improved a lot but they still produce very noisy images in low light conditions. This is mainly because of their small sensor size. Image quality can be improved by increasing the aperture size and/or exposure time however this make them susceptible to defocus and/or motion blurs. In this paper, we analyze the trade-off between denoising and deblurring as a function of the illumination level. For this purpose we utilize a recently introduced framework for analysis of computational imaging systems that takes into account the effect of (1) optical multiplexing, (2) noise characteristics of the sensor, and (3) the reconstruction algorithm, which typically uses image priors. Following this framework, we model the image prior using Gaussian Mixture Model (GMM), which allows us to analytically compute the Minimum Mean Squared Error (MMSE). We analyze the specific problem of motion and defocus deblurring, showing how to find the optimal exposure time and aperture setting as a function of illumination level. This framework gives us the machinery to answer an open question in computational imaging: “To deblur or denoise?”.

9023-16, Session 4

Depth from defocus using the mean spectral ratio

David P. Morgan-Mar, Matthew R. Arnison, Canon Information Systems Research Australia Pty. Ltd. (Australia)

1. Introduction

Depth from defocus (DFD) is a technique used to estimate scene depth from two or more photos captured with differing camera parameters, first proposed by Pentland[1]. Typically the photos are captured at different lens aperture or focus settings. Either of these changes between two photos results in the images having different image blur characteristics. The magnitude and character of this difference in image blur varies across the image, dependent on the distance from the camera to the objects being imaged. DFD aims to extract distance information from this blur difference.

One DFD approach is to quantify the amount of blur in regions of each image. The difference in the amounts of blur between corresponding regions in the two images depends on object depth. An example of this method is by McCloskey[2]. A serious problem with this approach is the difficulty of objectively quantifying the amount of blur in a single image. Another approach is to analyze corresponding regions of both images jointly. For example, a cross-correlation of the image regions gives a quantity related to the blur difference between the regions[3]. For similar reasons to the above approach, this example method can be shown to depend strongly on the texture of the object.

2. Method

In simple terms, a photographic image f1 is formed by the two-dimensional convolution of the scene light distribution s by the camera point spread function PSF1:

Taking the Fourier transform, the spatial frequency spectrum F1 of the image f1 is given by the product of the Fourier transform S of the scene with the optical transfer function (OTF):

where the OTF1 is the Fourier transform of the PSF1 and (u, v) are spatial frequency coordinates. If we consider small patches of two images f1 and f2 taken of the same scene s, the ratio of the spatial frequency spectra F1 and F2 is given by

where we have assumed that OTF1 and S are non-zero at all points (u, v), allowing S to be cancelled. Equation (3) implies that the ratio F2/F1 is independent of the scene, depending only on the OTFs of the camera optics under the two capture conditions. We call this ratio the spectral ratio (SR) (Fig. 1). The assumption that OTF1 and S are non-zero is in general false. Defocus OTFs can have zeros, and natural scene images often have near-zero Fourier components. These produce substantial noise in the SR of real world images.

Where PSFr can be considered to be a relative PSF which when convolved with f1 produces f2. PSFr varies with focus, aperture, the depth of the object imaged in patch s, and with field angle across the image plane. The scene independence of the SR has been noted before[4,5]. However, previous attempts to derive a depth measure from the SR have used only the modulus, as well as applying additional strong constraints, such as assuming a Gaussian or pillbox relative PSF[6], or using only a one-dimensional slice of the SR[4]. The true behavior of the SR and PSFr with object distance (in the absence of noise or aberrations other than defocus) is shown in Fig 2.

For objects near the focal plane, PSFr can be characterized by its width. However, the height of PSFr at the origin is a more stable measure over a greater defocus range (Fig. 2). We calculate the mean value of the SR pixels over the area of an image patch; the imaginary components cancel due to Hermitian symmetry. By properties of the Fourier transform, the mean spectral ratio (MSR) is equal to the value of the relative PSF at the origin. The MSR is a more robust measure of the shape of PSFr than attempting to fit a height or width to noisy data in the spatial domain.

It is important to take the mean of the complex SR values rather than their moduli. By the triangle inequality, the latter will generally be greater than the MSR. A modulus calculation implies a higher peak value in the PSFr than reality, thus underestimating the true relati

ive defocus blur. This can lead to errors in depth estimation. Similarly, calculations based on a ratio of the moduli of the OTFs (the MTFs) will also produce unrealistic physical results and hence depth estimation errors. In practice, performance of the MSR can be improved by filtering or weighting the SR. Pixels beyond the OTF support cutoff of the aperture may be filtered out. Observation of the SR of many natural image patches shows that the shape of the SR is generally less noisy at lower frequencies and more noisy at higher frequencies. This can be accounted for using a weighting based on the radial spatial frequency, or by using a low-pass smoothing filter on the image patches before Fourier transforming.

3. Results

We tested our MSR method with a static portrait scene. Images were captured with a Canon EOS 40D digital SLR using an EF 50mm f/1.4 lens. Shots were taken focused on the mannequin's face, with the aperture set to f/14 and f/22. The raw depth scores of various methods were calculated from the green Bayer channel. Scores from overlapping tiles were assembled into defocus maps, in which the scores ideally have a monotonic relationship with physical depth.

Results for our MSR method are shown in Fig. 3, with results from Subbarao's S-transform method[8], McCloskey's sample correlation coefficient[2], and Aydin's normalized cross correlation[3]. The MSR shows several advantages. Firstly, the MSR score is largely independent of object texture, as expected from the theoretical basis, except for regions with very low texture such as the scarf, the mannequin's face, and the guitar body. The other methods also have trouble with these areas, but additionally show strongly biased scores in regions with strong edges (the plant shadow on the back wall, the blinds in the upper right corner, and the black and white cloth near the guitar). The best performing prior method is McCloskey's sample correlation coefficient, but this shows significant errors on the blue striped pattern of the shirt, where the MSR finds the same depth as the brown jacket. The MSR result also shows a noticeably higher spatial resolution than the other methods, best seen in the structures of the flowers and plants. There is a speckly noise in the MSR results, but this can be reduced by filtering such as with a bilateral filter.

Figure. 3. Defocus maps produced by various methods. (a) Photo of the test scene. (b) Subbarao's S-transform method. (c) McCloskey's sample correlation coefficient method. (d) Aydin's normalized cross correlation method. (d) Our mean spectral ratio method. The false color scale indicates relative depths in the defocus maps; the marked distances are approximate as the calibrations vary between methods.

In conclusion, we have developed a DFD depth estimation method using two photos of a scene. Our MSR method is based on characterizing the ratio of Fourier transforms of image patches, without applying any simplified optical models. By considering the full complex SR, we obtain depth maps with significantly improved accuracy over previous methods.

References

[1] Pentland, Alex Paul., "A New Sense for Depth of Field," Transactions on Pattern Analysis and Machine Intelligence, PAMI-9, 523–531 (1987).

[2] McCloskey, Scott; Langer, Michael; Siddiqi, Kaleem. "The Reverse Projection Correlation Principle for Depth from Defocus," in Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06), IEEE Computer Society, 607–614, (2006).

[3] Aydin, Tarkan; Akgul, Yusuf Sinan. "An occlusion insensitive adaptive focus measurement method," Optics Express, 18, 14212–14224, (2010).

[4] Subbarao, M.; Wei, T. C. "Depth from defocus and rapid autofocusing: a practical approach," in Proceedings of the 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '92), IEEE Computer Society, 773–776, (1992).

[5] Pentland, A., Scherock, S., Darrell, T., & Girod, B. "Simple range cameras based on focal error," Journal of the Optical Society of America A, 11(11), 2925-2934, (1994).

[6] Subbarao, M. "Parallel Depth Recovery by changing Camera Parameters," in Proceedings of the Second International Conference on Computer Vision, 149–155, (1988).

[7] Stokseth, P. A., "Properties of a Defocused Optical System," Journal of the Optical Society of America, 59(10), 1314–1321, (1969).

[8] Subbarao, M.; Surya, G. "Depth from defocus: A spatial domain approach". International Journal of Computer Vision, 13, 271–294, (1994).

9023-17, Session 4

An extensive empirical evaluation of focus measures for digital photography

Hashim Mir, Peter Xu, Peter van Beek, Univ. of Waterloo (Canada)

Automatic focusing of a digital camera in liveview mode, where the camera's display screen is used as a viewfinder, is done through contrast detection. In focusing using contrast detection, a focus measure is used to map an image to a value that represents the degree of focus of the image. Many focus measures have been proposed and evaluated in the literature. However, previous studies on focus measures have either used a small number of benchmarks images in their evaluation, been directed at microscopy and not digital cameras, or have been based on ad hoc evaluation criteria. In this paper, we perform an extensive empirical evaluation of focus measures for digital photography and advocate using two standard statistical measures of performance, precision and recall, as evaluation criteria.

9023-18, Session 4

Out-of-focus point spread functions

Henry G. Dietz, Univ. of Kentucky (United States)

No Abstract Available

9023-19, Session 5

Automating the design of image processing pipelines for novel color filter arrays: local, linear, learned (L3) method

Qiyuan Tian, Stanford Univ. (United States); Steven P. Lansel, Olympus America Inc. (United States); Joyce E. Farrell, Brian A. Wandell, Stanford Univ. (United States)

The high density of pixels in modern color sensors provides an opportunity to experiment with new color filter array (CFA) designs. A significant bottleneck in evaluating new designs is the need to create demosaicking, denoising and color transform algorithms tuned for the CFA. To address this issue, we have developed a local, linear, learned (L3) algorithm that automates the process of creating an imaging system's processing pipeline.

We describe the L3 algorithm and illustrate how we created a pipeline for a CFA with an RGB/W design, in which the W pixel is clear and thus far more sensitive than the RGB pixels. Under low light conditions, the images produced by the L3 pipeline are superior to those produced by a matched Bayer RGB sensor. We will also apply the algorithm to other CFAs. The L3 algorithm simplifies and shortens the tasks of image pipeline development for novel CFA designs.

9023-20, Session 5

Minimized-Laplacian residual interpolation for color image demosaicking

Kiku Daisuke, Yusuke Monno, Masayuki Tanaka, Masatoshi Okutomi, Tokyo Institute of Technology (Japan)

No Abstract Available

9023-21, Session 5

Image sensor noise profiling by voting based curve fitting

Sebastiano Battiato, Giovanni Puglisi, Rosetta Rizzo, Univ. degli Studi di Catania (Italy); Angelo Bosco, Arcangelo R. Bruna, STMicroelectronics (Italy)

The extended abstract containing the details related to the proposed approach together with the experimental results can be found in the attached document ("extended_abstract.pdf").

9023-22, Session 5

(JEI Invited) Color correction pipeline optimization for digital cameras

Simone Bianco, Univ. degli Studi di Milano-Bicocca (Italy); Arcangelo R. Bruna, Filippo Naccari, STMicroelectronics (Italy); Raimondo Schettini, Univ. degli Studi di Milano-Bicocca (Italy)

The processing pipeline of a digital camera converts the RAW image acquired by the sensor to a representation of the original scene that should be as faithful as possible. There are mainly two modules responsible for the color-rendering accuracy of a digital camera: the former is the illuminant estimation and correction module, and the latter is the color matrix transformation aimed to adapt the color response of the sensor to a standard color space. These two modules together form what may be called the color correction pipeline. We design and test new color correction pipelines that exploit different illuminant estimation and correction algorithms that are tuned and automatically selected on the basis of the image content. Since the illuminant estimation is an ill-posed problem, illuminant correction is not error-free. An adaptive color matrix transformation module is optimized, taking into account the behavior of the first module in order to alleviate the amplification of color errors. The proposed pipelines are tested on a publicly available dataset of RAW images. Experimental results show that exploiting the cross-talks between the modules of the pipeline can lead to a higher color-rendition accuracy.

9023-28, Session 5

Analysis of a 64x64 matrix of direct color sensors based on spectrally tunable pixels

Giacomo Langfelder, Alessandro Caspani, Antonio F. Longoni, Edoardo Linari, Politecnico di Milano (Italy)

In the past years the concept of spectrally tunable direct color sensors, based on the principle of the Transverse Field Detector (TFD) [1] was conceived and developed. Such a tunable pixel enables a lot of features that allow the adaptation of the sensor spectral sensitivities to the scene characteristics, or multi-band image capture [2, 3]. In this work we analyze the performance of a 64x64 (x3 colors) matrix of such a sensor built in a 150-nm CMOS technology for demonstrative purposes.

The matrix features (10 um)^2 equivalent pixels arranged with a pixel-level electronics based on a single transistor charge amplifier [4]. A picture of the matrix can be seen in Fig. 1.

The matrix is mounted on a PCB board that provides the biasing (maximum voltages in the order of 9 V) and interfaces with an off-chip analog-to-digital acquisition board.

For the sensor characterization and calibration, the board is inserted into a suitably modified film slot (magazine) of a Hasselblad 500C with an 80mm lens. During the tests the camera F# number is kept to 2.8 with the B exposure setting. The choice of such a camera is motivated by the easy and quick possibility of mounting different slots with different TFD implementations (e.g. large-area single passive pixel, 8x8x3 matrices or 64x64x3 matrices) for comparison purposes.

The camera is aligned in front of a 24-patch transparency Color Checker illuminated by an integrating sphere (Image Engineering LE6-100). A suitable IR-cut filter is added in the optical path as direct color sensors are widely sensitive to IR radiation (see e.g. [5]). Pictures of the camera and the whole setup are reported in Fig. 2.

After reviewing in the next Section the TFD working principle, its pixel-level arrangement and the matrix-level implementation, electrical, colorimetric and noise performance, which are the main object of this analysis, will be discussed in the final Section.

The color reconstruction results will be compared to those obtainable through large-area (150 um)^2 devices based on the same concept (and built in the same technology), which, void of side effects and almost free of electronic noise, represent the reference for performance comparisons.

The Transverse Field Detector basic working principle relies on the physical properties of visible light absorption in Silicon, in particular on the dependence of its penetration length from the wavelength. With respect to other sensors that rely on this property (see e.g. [6-8]), the TFD capability of discriminating radiations absorbed at different depths relies on a suitable electric field configuration developed in a depleted layer beneath the Silicon surface. The device can be thus built through surface doping implants only (see Fig. 3a and [1] for details). The electric field drives photo-charges generated at different depths toward different surface contacts: by changing the field configuration one changes the collection trajectories and therefore the spectral sensitivity of the collecting contacts.

The typical spectral sensitivities (or equivalent filters as defined in [9]) are shown in Fig. 3b, characterized – as in the case of other direct color sensors [8] – by a relatively large overlap with respect e.g. to color filter arrays transmittance [10-11]. As shown in the close up of Fig. 1, each TFD is surrounded by the three electronics for the color channels, so to form the active pixel. The transistor-level schematic, based on a single-transistor charge amplifier, is reported in Fig. 3c (the reader can refer to [4] for more details).

The 64x64 sensor is formed by the sensor array, a row-selecting decoder and a column-selecting multiplexer (right and bottom side of the 64x64 matrix in Fig. 1). As no column-wise amplifier is implemented on the chip, during data readout the load capacitance for the pixel is represented not only by the column capacitance, but also by the out-of-chip parasitic; for this reason, though out-of-chip buffers are provided immediately at each color channel output on the PCB, the matrix scanning time is slow in order to wait for the time needed by the pixel-level circuit to charge a capacitance in the order of a few pF. The designed PCB includes the said analog amplifiers, a ceramic carrier that hosts the TFD, and suitable digital electronics for the sequential addressing of decoders and multiplexers.

First, the overall electrical operation of the matrix was tested. For this purpose, only the gray-scale patches of the Color Checker were used. The camera with the TFD sensor inside was put in front of each patch to have the entire sensor evenly illuminated. Relatively low-light levels were used due to the slow scanning time required as described above. The Correlated Color Temperature of the source inside the integrating sphere was 3240K, but a 6500K Color Changing Filter was used to obtain a daylight illumination.

Fig. 4 reports the obtained result of the grayscale acquisitions. The residual spatial noise in the responses of the different pixels can be appreciated to be proportional across the different illumination levels (e.g. this is clearly evident in the region enclosed in the ellipse for the blue pixels). This behavior suggests this noise to be due to photoresponse nonuniformity (PRNU). A suitable calibration and compensation algorithm will be developed to reduce these effects.

In Fig. 5a the circle, square and diamond markers report the average of the output voltages of Fig. 4, to be compared with the cross markers which plot the calibrated intensity of the different patches (normalized to the brightest patch). The obtained data are in quite good agreement with predictions.

Fig. 5b shows the experimental output voltages for the three color

channels, this time for all the 24 patches and in the situation of the reference TFD of Fig. 1. The inset shows the reconstructed color checker after this acquisition. A similar kind of measurements is ongoing also for the 64x64 matrix and will be used, after a PRNU calibration and compensation, to quantify the color reconstruction performance of the TFD matrix, both in terms of absolute deviation from the mean and in terms of noise propagation through the operations of color conversion [12].

9023-35, Session PWed

A method of mobile display (OLED/ LCD) sharpness assessment through the perceptual brightness and edge characteristic of display and image

Min Woo Lee, Seul Ki Jang, Samsung Electronics Co., Ltd. (Korea, Republic of)

After displaying specific image on th the device, display's resolution can briefly displayed using Optical Testing Device such as CA-2000. Preintroduced AETS(Average Edge Transition Slope) method has provided means to measures sharpness of greyscale digital image with 0~255 bit-depth. However, it is uncapable to compare displays which has variation in lightness scale. this paper suggests methods to measure sharpness of display with different chharacterists by modifing and correcting AETS methodolgy. Brightness of each display and uses angle computed from height and width of display's edge. We use the PCL(Perceptual Contrast Length) instead of Brightness on the ICDM(International Committee for Display Metrology)'s IDMS(International Display Measurement Standard).

This Sharpness measurement method called ECS(Edge Contrast Slope). It shows that sharpness characteritic of edge considered perceptual brightness. If low PPI display has brighter display, It could be more than sharper the high PPI display.

9023-36, Session PWed

Spatial adaptive upsampling filter for HDR image based on multiple luminance range

Qian Chen, Guan-Ming Su, Peng Yin, Dolby Labs., Inc. (United States)

To capture greater dynamic range is the future of image and video technology. The emerging video compression standard HEVC (H.265) has already supported 10-bits per color with 4:2:0 chroma subsampling in its Main 10 profile, and a 12-bits extension will be also available. It can be predicted scalable high dynamic range (HDR) image/video processing is also possible once the HEVC extended SVC (scalable video coding) standard is established. On the other hand, large HDR pictures may have to be downsized before compression to save bits in transmission, while post-processing needs to upsample the pictures for display. Therefore, proper strategies to spatially downsample and upsample HDR images are necessary.

In this paper, we propose an adaptive upsampling filter to spatially upscale HDR image size based on luminance range of the HDR picture in each color channel. It first searches for the optimal luminance range values to partition a HDR image to three different parts: dark, mid-tone and highlight. Then we derive the optimal set of filter coefficients both vertically and horizontally for each part. The sense of optimality is general. It can be defined to any picture quality metric depending on the application, from the traditional mean square error (MSE) based peak signal-to-noise ratio (PSNR) to structure similarity index measure (SSIM), or other perceptual based quality metrics. When the HDR pixel is within the dark area, we apply one set of filter coefficients to vertically upsample the pixel. If the HDR pixel falls in mid-tone area, we apply another set of filter for vertical upsampling. Otherwise the HDR pixel is in highlight area, another set of filter will be applied for vertical upsampling. Similarly, horizontal upsampling will be carried out in the same way for each pixel based on its luminance.

The inherent idea to partition HDR image to different luminance areas is based on the fact that most HDR images are created from multiple exposures. Due to the limitation in digital image sensors, it is generally impossible to capture the full dynamic range of a scene in a single exposure. However, a standard camera can take a sequence of pictures of the same scene of different exposure time. Each pixel will have a proper, under- and overexposed version in the sequence. Assume the scene is completely static (no moving object is observed), and the camera is perfectly aligned, the multiple low dynamic range (LDR) exposures can be combined to capture the full dynamic range. Various methods can be used to combine multiple exposures to HDR image. Corresponding pixels can be weighted averaged across exposures [1]. Typical weighting functions include Mitsunaga-Nayar weighting[2], and its extension of multiplying a broad hat function, which works better as it avoids dubious pixels near the extremes, where gamut limitation and clamping may affect the output values unpredictably[1]. [3] divides one picture into rectangular regions, and always takes the exposure that contains the most details from multiple exposures for each region. In general, this method is a subset of weighting average, as it chooses weighting factor 1 for the selected exposure and 0 for the rest of the exposures in each particular region. In HDR images created from multiple exposures, one exposure contributes differently in different regions, and usually every pixel will have one dominating exposure that weighs significantly more than others. Under most circumstances, we observed that dark regions in HDR come from over-exposed LDR, highlight regions come from under-exposed LDR, and properly exposed LDR contributes most in mid-tone regions. Even with the same camera, different exposures may demonstrate slight variation in captured signal statistics, such as noise level, subtle misalignment etc. And this variation increases with the exposure time difference, and can become noticeable between two extremes of under- and over-exposed LDR. Hence, to group different regions to three luminance partitions actually helps to eliminate the variation between signals, and to derive optimal filter for each group with signals of lesser variation is certainly more efficient than for the entire HDR image.

To verify our idea, we conduct experiments on a set of HDR images from [4]. We first reduce HDR images to quarter size of the original, and apply different filters to upscale the quarter HDR to its original size. Then we calculate PSNR between the upsampled HDR and original HDR as performance metric. And MSE is hence used as the sense of optimality to be consistent with PSNR. Note that we consider 1-D separable filter in our simulation to obtain the spatial upsampling filter coefficients for both optimal and proposed filters: two phases vertical filters are designed first, and then one phase horizontal filter is derived based on the vertical upscaling results. However, the concept can be easily extended to 2-D non-separable kernel based filter. Other upsampling filters for comparison in the simulation include fixed coefficients filters (Lanczos3, JVT-I019[5]) and explicit filter (MSE sense based optimal filter over entire HDR image). Experimental results show that the proposed adaptive upsampling filter based on luminance ranges outperforms the optimal upsampling filter around 0.57dB for R channel, 0.44dB for G channel and 0.31dB for B channel. Compared with other upsampling filters with fix coefficients, the performance of the proposed filter is far more superior. Its average gain over Lanczos3 is 5.51dB for R channel, 5.68dB for G channel, and 5.85dB for B channel.

9023-38, Session PWed

A classification-and-reconstruction approach for a single image super-resolution by a sparse representation

YingYing Fan, Masayuki Tanaka, Masatoshi Okutomi, Tokyo Institute of Technology (Japan)

In this paper, we propose a classification-and-reconstruction approach with multiple dictionaries. In a learning phase of the proposed method, first, a classifier is learned with training patches. Then, multiple dictionaries are learned with classified patches. In a reconstruction phase of the propose method, input patches are classified by the learned classifier. Then, input patches are reconstructed with the dictionary associated to the classification results. Multiple dictionaries can provide more appropriate choice to sparse representation. Experimental comparisons show that the proposed method outperforms existing single image super-resolution.

9023-39, Session PWed

LoG acts as a good feature in the task of image quality assessment

Xuanqin Mou, Wufeng Xue, Congmin Chen, Xi'an Jiaotong Univ. (China); Lei Zhang, The Hong Kong Polytechnic Univ. (Hong Kong, China)

In the previous work, the LoG (Laplacian of Gaussian) signal that is the earliest stage output of human visual neural system was suggested to be useful in image quality assessment (IQA) model design. This work considered that LoG signal carried crucial structural information in the position of its zero-crossing and proposed a non-shift edge (NSE) based IQA model. In this study, we focus on another aspect of the properties of the LoG signal, i.e., LoG whitens the power spectrum of natural images. .Here our interesting is that: when exposed to an unnatural image, or a distorted image, how does the HVS whitening this new signal? In this paper, we first investigate the whitening filter for natural image and distorted image respectively, and then suggest that the LoG is also a whitening filter for distorted images. Based on this fact, we deploy the LOG signal in the task of image quality assessment (IQA)model design by applying two very simple distance metrics, i.e., the MSE (mean square error) and the correlation. The proposed models are analyzed according to the evaluation performance on three subjective databases. The experimental results validate the usability of the LoG signal in IQA model design and that the proposed models stay in the state-of-the-art IQA models.

9023-40, Session PWed

Sharpness enhancement for mobile devices

Seul Ki Jang, Hyun Hee Park, Jee Young Yeom, Jong Ho Kim, Min Woo Lee, Samsung Electronics Co., Ltd. (Korea, Republic of)

As mobile devices usage increase, consumer demands for image quality specification of mobile devices are gradually increasing. Various techniques are applied to meet the needs for improved image quality of mobile displays. This paper is focused on sharpness enhancement technique. Human visual experiments are performed to analyze viewers'preference for sharpness. With increase in PPI, the perceived sharpness determining the good image quality is increased. However, the perceived sharpness decreases as the luminance increases. Relationship between various factors such as display and human visual characteristics and the values of sharpness control parameter are modeled by functions based on the results of human visual experiments. Experimental results indicate that proposed sharpness control parameter yields better image quality.

9023-41, Session PWed

White constancy method for mobile displays

Ji Young Yum, Hyun Hee Park, Seul Ki Jang, Jae Hyang Lee, Jong Man Kim, Ji Young Yi, Min Woo Lee, Samsung Electronics Co., Ltd. (Korea, Republic of)

In these days, smart phone usage is increased, consumer requirements for image quality of mobile devices are increasing. For example, colors may perceived differently when displayed contents under different illuminants. Displayed white in incandescent lamp perceived bluish, while same content in LED light perceived yellowish. When changed in perceived white under illuminant environment, image quality would be degraded. Objective of the proposed white constancy method is restricted to maintain consistent output colors regardless of the illuminants utilized. Human visual experiments are performed to analyze viewers'perceptual constancy. Participantes are asked to choose the display white in a variety of illuminants. Relationship between the illuminants and the selected colors with white are modeled by mapping function based on the results of human visual experiments. White constancy values for image control are determined on the predesigned functions. Experimental results indicate that propsed method yields better image quality by keeping the display white.

9023-23, Session 6

(JEI Invited) Improved method of finding the illuminant direction of a sphere

Richard Dosselmann, Xue Dong Yang, Univ. of Regina (Canada)

An improved means of finding the direction of a light source illuminating a sphere is presented. In particular, the vertical slant angle of a light source is computed using a new approach. The planar tilt angle of a source is known to be related to the average of the local gradient vectors of an image of a lit sphere. This research uncovers a straightforward link between the length of the average of the local normalized gradient vectors of an image and the slant angle of the corresponding light source. The new technique is far more effective when compared with existing approaches over a variety of test images.

9023-24, Session 6

Light transport matrix recovery for nearly planar objects

Niranjan Thanikachalam, Loïc A. Baboulaz, Paolo Prandoni, Martin Vetterli, Ecole Polytechnique Fédérale de Lausanne (Switzerland)

We present a new method for the recovery of light transport matrix (LTM) of nearly planar scenes. We exploit the structure of the light transport matrix to selectively sample it in a non-adaptive fashion using compressive sensing techniques and thus obtain more than 90% reduction in the number of required illumination vectors. Though our method is restricted to objects with nearly planar geometry, our specific interest lies in relighting artworks like stained glass and oil paintings. Such artworks exhibit complex microstructures resulting in intricate reflection or scattering patterns. This justifies the need for such specialized algorithms that allow very fast acquisition while maintaining a good SNR.

The LTM is a 4D slice of the 8D bidirectional scattering distribution function, obtained by fixing the incident and exitant directions. Let l ? R^M be the vectorized illumination image projected on the scene by a programmable light source. Let y ? R^N be the vectorized image captured by a camera.The LTM, T ? R^{M ?N}, then defines the transport of light between the l plane and the y plane such that, y = l* T

Current state of art methods for LTM recovery either use compressive sensing on the whole matrix or use low rank matrix recovery methods by imposing symmetry constraints on the LTM using a co-axial configuration for the camera and the projector. Reddy et al were the first to study in detail the structure of the light transport matrix for a general scene, and they proposed a frequency domain reconstruction, by assuming a low frequency light transport. By restricting ourselves to objects with almost planar geometry, we employ compressive sensing to further reduce the acquisition time, while still preserving the SNR.

For our analysis of the structure of the light transport matrix, we note that the light transport matrix is the dual of the light field, i.e., it is obtained by the translations of a light source along a plane, while the camera remains fixed. This lets us derive two complementary views of the light transport matrix : the array of point scattering functions view, obtained by reordering each row of the LTM as an image, and the array of light transport functions (LTF) view, obtained by reordering each column of the LTM as an illumination image. The light transport matrix can be split into a direct component, a subsurface scattering component and an inter- reflections or caustics component, based on the phenomenon causing image formation. The direct component is a diagonal matrix for reflection while the subsurface scattering component is a band matrix. The inter-reflections component on the other hand occurs anywhere along the off diagonal elements and can be divided into a banded 'near range' and an unstructured 'far range' component. However since we are only interested in almost planarscenes, only near range inter reflections are present. Thus, the LTM for nearly planar objects is always a band matrix, where the bandwidth s is related to the surface regularity in case of a reflective object or the material scattering properties in case of a refractive or translucent scene.

When represented as an array of LTFs , the banded nature of the LTM manifests itself as a light transport function with finite spatial support, s, defined about pixel n ? [1...N]. Let Y = [y1 ...yp] and L = [l1 ...lp] be the ensemble of observed image and illumination vectors respectively, while m ? [1...M] be the index of light elements in the illumination vector. Let Ws^n be a square window of area s defined about the pixel n. We then define an indicator function ?n [m] that describes which light elements from the illumination plane contribute to the pixel n, ?n [m] = 1 if m ? Ws^n 0 otherwise

The illumination vector, Ln , that contributes to any given pixel in the image plane can then be calculated as Ln = L.*?n where .* represents element-wise multiplication. The light transport function at pixel n, tn , is then the solution of Yn = Ln'* tn

Case : (s < p) The system of equations is overdetermined, when the number of observations is larger than the spatial support, s, of tn , we minimize the l2 -norm, while imposing a non-negativity constraint on tn .

This simplified problem by itself provides considerable savings in both computational power and acquisition time, compared to out-of-the-box compressive sensing.

Case : (p < s < m) In the under-determined case, we pose light transport recovery as a sparse recovery problem. Since each LTF, tn is essentially a point spread function, it can be compressed in some basis ? ? Rs?s .

tn = arg min |?'*tn |1 s.t|Yn ? Ln'* tn |2 ≤ and tn ≥ 0 ?n ? [1...N]

where Yn ? R^p and Ln ? R^{s?p} . Our acquisition setup consists of a programmable area light source projecting gaussian random patterns on the scene of interest and a static camera capturing the resulting scene.

At this stage, we have provided relighting results for a stained glass artifact and a rendered scene with nonconvex surface geometry . The SNR variation as a function of undersampling factor p/M is provided. An example of the two array of array representations discussed here is also provided for the sake of clarity.

With the advent of digital museums and virtual tours, truthful digitization of artworks gains importance.

Given the large variety of paintings, engravings and stained glass artifacts, it is interesting to have a common class of algorithms that exploit the particular geometric structure of the object, by taking advantage of the structure of the light transport matrix. We have presented an approach that aids in high fidelity image based relighting, while minimizing both the computation time and the acquisition time.

9023-25, Session 6

The color of water: using underwater photography to assess water quality

John W. Breneman IV, Henryk Blasinski, Joyce E. Farrell, Stanford Univ. (United States)

The same filtering of sunlight that tints the sky a pale blue is present and magnified underwater reducing most marine scenes to a greenish or bluish tinge. Many divers must bring a flashlight to re-illuminate a scene. Despite the limitations, there is much information to be gained via color underwater. Extracting this color information requires detailed models describing an illuminating light spectrum's evolution as it passes through seawater. Although the basic principles of the absorption of light are described by the Beer-Lambert law [2], the specific absorption constants may vary by location or season. To provide a means to easily characterize an underwater environment's illumination, a simple color rig was designed using a commercial point-and-shoot camera in an underwater housing and a calibrated color target. The rig was used to estimate the variation of illumination with depth near Stanford University's Hopkins Marine Station in Monterey CA.

9023-26, Session 6

Surveillance system of power transmission line via object recognition and 3D vision computation

YuanXin Zhang, Xuanqin Mou, Xi'an Jiaotong Univ. (China)

Security precaution of power transmission system is crucial to maintain power system operating safely and stably. There are a variety of risk factors causing security problems, among which the construction activity in the vicinity of power transmission line has always been a major one. Video surveillance is usually used for security precaution in those cases. Unfortunately, the automatic object detection of the currently used surveillance systems surfers from high error rate because they are generally adopt the technique of invasion detection that is based on difference analysis between two adjacent images in video sequence. This type technique has at least two limitations: 1) it doesn't recognize what the object exactly is while in practice only a few of objects, e.g., crane, is realistically dangerous to power transmission line; and 2) it doesn't determine what location the object is in the 3D space by which we can identify the dangerous strength.

In this paper, we propose a video surveillance method for the security precaution of power transmission line via the techniques of object recognition and 3D spatial location detection. First, we use the SSIM metric to detect the motion object between two adjacent pictures in video sequence because the SSIM metric can remove the influence caused by varying light conditions so that the motion detection is more reliable. Secondly, we use the Adaboost classifier with Haar features to detect the motion object. In this paper, we have implemented a classifier to identify crane vehicles. Thirdly, we propose a single camera based 3D spatial measurement method to determine object location and height so that the risk the object is dangerous to the power line can be estimated. Experimental results show that our proposed method can exactly discriminate the dangerous strength of an invasive object to the power transmission line and the developed system is feasible and practical.

9023-27, Session 7

Metamer density estimation using an identical ellipsoidal Gaussian mixture prior

Yusuke Murayama, Pengchang Zhang, Ari Ide-Ektessabi, Kyoto Univ. (Japan)

No Abstract Available

9023-29, Session 7

Absolute colorimetric characterization of a DSLR camera

Giuseppe Claudio Guarnera, Simone Bianco, Raimondo Schettini, Univ. degli Studi di Milano-Bicocca (Italy)

In digital photography the incident photons on the sensor pixel area cause charge to accumulate at each pixel location, thus forming a picture. The relation between the sensor irradiance and the RGB triplet is called Opto-Electronic Conversion Function (OECF). In analogy with the human visual system, a sensor should have a set of color filters in order to mimic the trichromatic color matching functions, actually mosaicing the captured image. However, consumer Digital Single Lens Reflex (DSLR) cameras are typically designed to produce pleasing images, in which the importance of obtaining good contrast and vivid colors is more important than an accurate colorimetric reproduction of the scene. A specific design of the spectral sensitivities of the color filter arrays is usually employed with this purpose. It affects the way the sensor collects the charges which will form the RAW picture, on which a set of algorithms for demosaicing, white balancing, gamut mapping, color enhancement and so on is subsequently applied.

Color characterization of imaging devices establishes the relationship between the camera responses to a set of colors and the corresponding colorimetric values. Various techniques have been proposed to find such a relationship, which typically either requires the acquisition of a reference color target (e.g. a GretagMacbeth ColorChecker) with known spectral reflectance [1] or the use of specific equipment such as monochromators, as recommended by the standard ISO:17321-1:2006 [2] . Empirical DSLR characterization directly relates measured colorimetric data from a target and the corresponding camera RGB data, obtained from a picture of the target itself. Many of these techniques are limited to linear regression on training samples, whereas more advanced methods use a white point preserving maximum ignorance assumption. In [3] several empirical techniques to compute the optimal 3 x 3 transformation matrix are compared. It is an useful approach when the camera spectral sensitivities are unknown, assuming that the characterize camera will be used in illumination conditions very similar to those encountered during the characterization [4]. However, the limited set of color patches on the target can lead to inaccurate results and they are generally limited to low dynamic range imaging, since the measurement devices and the model used discard the intensity scale of the illuminant, preserving only the relative spectral power distribution. In [5] a transparent target is proposed, which allows the use of the characterization method also for High Dynamic Range (HDR) imaging.

Characterization approaches based on a monochromator generally prove to be more accurate and of general application. Since a white integrating sphere is illuminated with a monochromatic light, with a single wavelength at a time, the main drawback of monochromator-based techniques is the time required to collect the pictures of the sphere for each single wavelength. Moreover, the absence of an absolute scale for the luminance represents a limitation. To overcome this limitation in [6] an adaptation algorithm is described, to obtain absolute measurements of the tristimulus values by exploiting the variation of the lens aperture. The inverse OECFs are obtained from a simulated spectrally neutral grayscale pattern, assuming the equal energy illuminant E. The estimated data are further fitted with true measurements to derive a linear model of color correction. In this paper we propose a simple but effective technique, which offers a large dynamic range requiring just a single, off-the-shelf target and a commonly available controllable light source for the characterization. We separate the characterization task into two modules, an absolute luminance estimation module and a colorimetric characterization matrix estimation. The characterized camera can be effectively used as a tele-colorimeter, giving an absolute estimation of the XYZ data in cd/m2, just varying the f-number of the camera lens or the shutter time t. The estimated absolute tristimulus values closely match the values measured by a professional spectro-radiometer, as confirmed by preliminary results.

The usefulness of camera characterization is often limited by the low dynamic range of the sensor, further reduced by the noise floor. A better solution would allow the possibility to change the integration time or to vary the lens aperture, extending the measurement range of the system. We separate the characterization task into two modules, an absolute luminance estimation module and a colorimetric characterization matrix estimation.

To obtain an absolute estimation of the luminance we must link the measured RGB values with a set of known radiance values, including the intensity scale and not only the relative spectral power distribution. Most reflectance-based techniques for camera characterization make use of a gray-scale target, either real or simulated, illuminated by a known light source. These techniques are inherently low dynamic range. A better strategy would make use of an equal tristimulus values (CIEXYZ) light source, with increasing intensity. Such an illuminant would allow an accurate modeling of the OECFs, without introducing a bias due to the combined effect of the light chromaticity and the channel-dependent sensor spectral sensitivities.

We propose a simple but effective technique, which offers a large dynamic range employing just a single, of-the-shelf target and a commonly available controllable light source. A white diffuse patch is placed in front of a DLP projector, roughly in the center of the light beam. A DSLR camera is aligned with the light beam, with the white patch placed in the middle of the image formation plane, to reduce the influence of the vignetting effect. The lens should allow the selection of the f -number to use, with standard full-stop scale. The projector should have been previously calibrated with a spectro-radiometer, to correct for the gamma and to obtain an equal tristimulus scale. The projected light beams hence have the same relative spectral distribution but a different absolute intensity. A sequence of pictures of the white patch is acquired using the widest entrance pupil available, with increasing luminance. All images are saved as RAWs files, with no debayering or white balancing. The ISO setting should be fixed beforehand to the native ISO value of the sensor. Once saturation is detected on a pixel, the entrance pupil is shrunk by a full-stop, and the gray scale is projected back from the lowest value.

The process is repeated until a picture of the patch illuminated by the brightest gray value is captured.

The noise floor is estimated with a sequence of pictures acquired with the cap on the lens, at different times and averaging them. We need to account for the different areas of the spectral sensitivities curves, which we need to estimate. We used the technique described in [7], which just requires a pictures of a reflective color target with known relative spectral reflectances (e.g. a Color Checker), acquired under daylight illumination. RAW pixel values in the acquired images are normalized and fitted with the measured luminance, to derive the inverse OECFs .

To derive the color characterization matrix M we propose the use of the same setup used for luminance estimation, in which we can modulate the relative and absolute spectral distribution of the light beam thus enabling a virtually infinite number of illuminants useful to characterize the colorimetric profile of the camera. A set of illuminants are then projected on the white reflective patch, and the corresponding absolute tristimulus values are annotated. For each illuminant a picture is captured, and used to estimate the luminance values. The matrix M is found by solving the optimization problem proposed in [8]. Since the objective function is non-linear and non-differentiable the optimization problem is solved using the Pattern Search Method (PSM). PSMs are a class of direct search methods for nonlinear optimization [9]. PSMs are simple to implement and do not require any explicit estimate of derivatives. Furthermore, global convergence can be established under certain regularity assumptions of the function to minimize.

A Canon 40D camera, with a standard 28-135mm f=3:5-5:6 zoom lens, was used in this work; the focal length was set to 30mm and kept fixed during the experiments. The lens f-number ranges from f=4 to f=22, with a standard full-stop scale; a slip-on lens hood is used on the lens, in order to reduce stray light, thus increasing the measurements accuracy. A previously calibrated Dell S300 DLP projector, with a maximum brightness of 2200 ANSI lumens was aligned with the camera and employed as controllable light source. The reflective targets are fixed to a blackboard placed in front of the projector, about 0.6 meters away from

it. A spectro-radiometer with a luminance measurement range of 0.2-1200cd/m2 and spectral range 380-730nm is employed for the system characterization and to obtain the ground truth in our experiments. In order to reduce the number of acquisitions and at the same time to allow a fair comparison with previous work [1, 6], we decided to mimic the color target approach by projecting the 24 illuminants which would produce on the white reflective patch a set of tristimulus values similar to a physical ColorChecker target. A schematic representation of our setup is reported.

While a single reflective white target can suffice for the characterization, we evaluated our method using a set of 8 different color targets, illuminated by a set of 30 illuminants which differ in terms of spectral power distribution. The camera sensor is equipped with a standard color filter array, hence we need to account for the reduced spatial resolution camera of each color channel, which have a sampling factor of 1:2 for the Green and 1:4 for both the Red and Blue values. Since the targets we used can be considered roughly lambertian surfaces and they are aligned with the center of the projector beam and orthogonal, the surface reflectance varies smoothly and a simple interpolation, performed individually for each channel, suffice to recover the missing data.

To evaluate the goodness of the proposed characterization we computed the Delta E94 and the CIEDE2000 color differences between our estimated absolute tristimulus values and the ground-truth, for all the 8x30 samples in our test set. In order to compute the color differences formula all values are converted into the CIELab color space. In particular, the mean Delta E94 is below 1.77 and the mean CIEDE2000 is less than 1.64; on the luminance Y we obtained a median relative error of 2.37% and a median absolute error of 2.7 cd/m2.

9023-30, Session 7

Simultaneous capturing of RGB and additional band images using hybrid color filter array

Daisuke Kiku, Yusuke Monno, Masayuki Tanaka, Masatoshi Okutomi, Tokyo Institute of Technology (Japan)

No Abstract Available

9023-31, Session 8

Badly exposed object recovery using images captured under disparate conditions

Florian M. Savoy, Univ. of Illinois at Urbana-Champaign (Singapore) and École Polytechnique Fédérale de Lausanne (Switzerland); Vassilios Vonikakis, Stefan Winkler, Advanced Digital Sciences Ctr. (Singapore); Sabine Süsstrunk, Ecole Polytechnique Fédérale de Lausanne (Switzerland)

In this paper we consider the problem of clipped-pixel recovery over an entire badly exposed image region. In order to retrieve the missing information, we use two correctly exposed reference images of the scene captured under different imaging conditions. The first one is used to recover texture and is taken from a similar viewpoint. Feature point locations are extracted along the boundaries of both objects in the source and reference images and a warping function is then computed to deform the reference object to fit inside the source. The second is taken under similar illumination conditions and is used to recover color by replacing the mean and the variance of the texture reference image by those of the color reference image object. The typical application scenario is thus the following : a tourist visits a monument and takes a photograph of it. However, due to inappropriate illumination conditions or camera settings, clipped pixel values appear in the object. When coming back home, he searches the web for two images of the monument satisfying the above constraints and can thus improve his photograph. Our preliminary results outperform the previous limitation to simple patches from state-of-the-art clipped-pixels recovery methods.

9023-32, Session 8

Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays

Jan Froehlich, Andreas Schilling, Eberhard Karls Univ. Tübingen (Germany); Stefan Grandinetti, Simon Walter, Bernd Eberhardt, Hochschule der Medien (Germany); Harald Brendel, Arnold & Richter Cine Technik GmbH & Co. Betriebs KG (Germany)

No Abstract Available

9023-33, Session 8

Cost-effective multi-camera array for high quality video with very high dynamic range

Joachim Keinert, Marcus Wetzel, Fraunhofer-Institut für Integrierte Schaltungen (IIS) (Germany); Michael Schöberl, Friedrich-Alexander-Univ. Erlangen-Nürnberg (Germany); Peter Schäfer, Frederik Zilly, Fraunhofer-Institut für Integrierte Schaltungen (IIS) (Germany); Michel Bätz, Friedrich-Alexander-Univ. Erlangen-Nürnberg (Germany); Siegfried Fößel, Fraunhofer-Institut für Integrierte Schaltungen (IIS) (Germany); André Kaup, Friedrich-Alexander-Univ. Erlangen-Nürnberg (Germany)

Temporal bracketing can create images with higher dynamic range than the underlying sensor. Unfortunately, moving objects cause disturbing artifacts. Moreover, the combination with high frame rates is almost unachievable since one video frame requires multiple sensor readouts.

The combination of multiple cameras equipped with different attenuation filters promises a remedy, since all exposures can be performed at the same time with the same duration using the playout video frame rate. However, a disparity correction is needed to compensate the spatial displacement of the cameras.

Unfortunately, the requirements for a high quality disparity correction contradict the goal to increase dynamic range. When using two cameras, disparity correction needs an object to be properly exposed in both cameras. In contrast, a dynamic range increase needs the cameras to capture different luminance ranges.

As this contradiction has not been addressed in literature so far, this paper proposes a novel solution based on a three camera setup. It enables accurate determination of the disparities and an increase of the dynamic range by nearly a factor of two while still limiting costs. Compared to a two camera solution, the mean opinion score (MOS) is improved by 13.47 units in average for the Middleburry images.

9023-34, Session 8

The effect of split pixel HDR image sensor technology on MTF measurements

Brian M. Deegan, Valeo Vision Systems (Ireland)

Split-pixel HDR sensor technology is particularly advantageous in automotive applications, because the images are captured simultaneously rather than sequentially, thereby reducing motion blur. However, split pixel technology introduces artefacts in MTF measurement.

To achieve a HDR image, raw images are captured from both long

and short exposure pixels, and combined to make the HDR output. In some cases, a large subpixel is used for long exposure captures, and a small subpixel for short exposures, to extend the dynamic range. The relative size of the photosensitive area of the pixel (fill factor) plays a very significant role in the output MTF measurement. Given an identical scene, the MTF will be significantly different, depending on whether you use the long or short pixels i.e. a smaller fill factor (e.g. in the short exposure subpixel) will result in higher MTF scores, but significantly greater aliasing.

To demonstrate the effect of having two subpixels on MTF50 measurement, the following simulation was performed. A squared chirp signal was generated in Python. The starting frequency was 0.15Hz, and the end frequency was 15Hz, over a period of 10 seconds. The sample rate of the input signal is 500Hz

The signal was downsampled to 10Hz. For each 50 point downsampling bin, the mean of the first 40 samples was taken as the long channel output, and the mean of the last 10 samples was taken as the input for the short channel output. This simulates the fact that the fill factor (the relative size of the photosensitive part of the pixel) is 4 times larger for the large subpixel than for the small subpixel, and that the large and small subpixels are physically separated. The contributions of the large and small subpixels are combined using alpha blending. The formula used is:

Output = alpha*large subpixel + (1-alpha)*small subpixel

Where alpha is the absolute luminance level.

As the spatial frequency increases, the large subpixel MTF decreases, because the fill factor is too large to accurately capture the black to white transition. However, for the small subpixel, the fill factor is smaller, resulting in significant aliasing effects. The combined HDR output will suffer from a combination of reduced MTF50 and aliasing effects.

In a separate simulation, a split-pixel model was cross-correlated with a rectangle function. The shape of the output function was significantly different for the leading edge transition, compared with the falling edge. For the falling edge, the output has a smooth transition between the high and low points. In contrast, the rising edge has an initial peak, followed by a local trough at ~80% of the maximum peak value, and then finally stabilizing.

The simulated results matched the measured output from a split pixel HDR image sensor. Experimental results showed a difference of 41% in measured MTF50 between the falling and rising edges of a slanted edge test chart.

To summarize, as the spatial frequency increases (e.g. at edges in an image), the way in which long and short pixels reproduce the image is significantly different. Large subpixels smooth black to white transitions, returning a grey value. Small subpixels in contrast will have significant aliasing artifacts at high spatial frequencies.

In split pixel HDR sensors, the long and short channel information from each pixel is combined, to produce the final image. How the pixels are combined is dependent on the scene content. Depending on the luminance level, the output pixel may contain entirely large subpixel data, entirely small subpixel data, or a combination of both. MTF50 values can therefore be significantly different, depending on the nature of the scene presented to the image sensor i.e. light levels, scene dynamic range, and direction of luminance transitions. This should be taken into account for all split pixel HDR sensor applications.

Conference 9024: Image Processing: Machine Vision Applications VII

Monday - Tuesday 3 –4 February 2014

Part of Proceedings of SPIE Vol. 9024 Image Processing: Machine Vision Applications VII

9024-2, Session 1

Machine vision based quality inspection of flat glass products

Gerald Zauner, Martin Schagerl, FH OÖ Forschungs & Entwicklungs GmbH (Austria)

This paper presents a machine vision solution for the quality inspection of flat glass products. The flat glass or insulation glass parts vary in size but can be as large as 6 x 3 meters. The quality inspection at the end of the production line aims to classify 5 different error types (which are categorized in eliminable and non-eliminable errors with respect to production costs). These error classes are: water drops, bubbles (inclusions or material defects), coating defects, scratches and fingerprints. A contact image sensor (CIS) is used to generate digital images of the glass surfaces. A first simple pre-processing step then extracts defective image regions from the high-resolution raw images resulting in 'sub-images' representing defective glass regions of different size (see Fig.1).

These defect images are usually characterized by very little 'image structure', i.e. homogeneous regions without distinct texture. Additionally, these defect images often consist of only a few pixels. At the same time the appearance of certain defect classes can be very diverse (e.g. water drops). As the image quality usually suffers from various scanning artefacts and image intensity variations due to changing optical properties of the glass coating, a number of necessary pre-processing steps are applied (e.g. removal of periodic scanning artefacts or contrast optimization) followed by a segmentation step (Otsu's method).

The main contribution of this work now lies in the systematic evaluation of different state-of-the-art image feature extraction methods followed by an extensive parameter study of various machine learning (classification) algorithms. The image features used in this study were: statistical histogram based features (std. deviation, curtosis, skewness), geometric features (form factor/elongation, excentricity, Hu-moments) and texture features (gray level run length matrix, co-occurrence matrix). The following machine learning algorithms were compared: decision tree (J48), random forest, JRip rules, naive Bayes, Support Vector Machine (multi class), neural network (multilayer perceptron) and k-Nearest Neighbour. We used a representative image database of 1500 defect images and applied cross validation for evaluation purposes. Special focus was also put on the data preparation (i.e. normalization of feature attributes).

The cost evaluation (supported by ROC analysis and a run-time analysis) clearly shows certain types of classification algorithms to be more appropriate for the proposed task (see Fig.2), namely: random forest, SVM and multilayer perceptron. These classification algorithms usually reach classification rates (cross validated) of approx. 95% in the present case.

9024-3, Session 1

Stain defect detection for mobile phone camera modules

Sehee Hong, Chulhee Lee, Yonsei Univ. (Korea, Republic of)

In this paper, we propose a stain defect detection algorithm based on difference of mean brightness using windows. It is important to detect stain defects in camera modules which are largely used in mobile phones. Stain defect detection is very difficult during camera module inspection process since stain has extremely low contrast and various sizes. They also include black points, white points, hot pixel defects and texture issues [1]. Stain defects occur due to foreign substances which may exist on IR filters or inside of lens in camera modules. Recent stain defect rates are from 2% to 5%. However, there are few quantitative inspection methods.

A key idea of the proposed method is using the maximum square value of the difference brightness in divided windows (MAX WDMS: maximum window difference mean square). The proposed method consists of three steps: window design, stain location detection using MAX WDMS and WDMS threshold determination. The proposed method has been successfully applied to stain defect detection using real-world image data obtained from camera module manufacturing.

To evaluate the proposed stain defect detection algorithm, we used 30 images of LED 5100K, 750 lux which may include stain defects. There were 11 images of 1-3 gray level differences between the background and stain defects, 7 images of 4-6 gray level difference and 12 images of 7-15 gray level difference. The proposed method achieved about 15% improvement compared to existing methods. To evaluate stain detection accuracy, we used 170 images and manually classified them as good, fair and poor grades. Then we used quantitative methods to classify them. The proposed method achieved about 10~15% improvement compared to existing methods. Experimental results also show that the proposed method produced improved performance in terms of perceptual sensibility estimation.

The proposed detection algorithm is fast and produced noticeable improved performance compared to existing methods.

[1] Sobral, Joao Luis, "Optimized filters for texture defect detection," IEEE International Conference on Image Processing, Vol. 3, 2005.

9024-4, Session 1

A novel automatic full-scale inspecting system for banknote printing plates

Jian Zhang Sr., Li Feng, Jibing Lu, Qingwang Qin, Security Printing Institute of People's Bank of China (China)

Printing plates quality assurance is an important issue for workers who produce banknote printing plates. Every plate must be checked carefully and fully before it's sent to banknote printing factories. Previously the checking work is done by specific workers, usually with the help of powder and a magnifier. The checking work usually lasts 3 to 4 hours for a 5?7 plate with the size of about 650?500mm. Similar systems have low resolving imaging results and can't be used to inspect the plates. Yula has a laser checking system which only checks parts of intaglio plates. Now we have developed an automatic inspecting system to replace human work. The system includes a flat platform, an electrical system which holds a microscope. The microscope can move to anywhere in the X-Y plane over the platform. A digital camera fastened to the microscope captures the gray image from the microscope output. Each digital image has the size of 2672?4008, and each pixel corresponds to about 2.9?2.9um area of the plate. 4200 images are captured for a 5?7 plate. The inspecting model is formed from images of qualified plates, and then used to inspect indeterminate plates. Every image of the model and the corresponding one of the indeterminate plate are captured at the same (x,y) position. The system takes about 64 minutes to inspect a plate, and identifies obvious defects. Moreover, digital images are saved and can be reviewed.

9024-5, Session 1

Possible future trends and developments in industrial machine vision

Kurt S. Niel, Upper Austria Univ. of Applied Sciences (Austria)

If some follows the actual developments regarding implementations of machine vision there seems to be no borders for future developments: Calculating power constantly increases, every week there is a new idea, e.g. face recognition got into the huge consumer market, 3D capturing became cheap, due to the huge community SW-coding gets easier using sophisticated development platforms. But there will be a remaining gap between applications and industrial applications. The first ones have to be entertaining, the second reliable. There are some studies (e.g. VDI, Germany) which show moderate increasing budgetary market for machine vision in the industry. What are the industrial needs: Revenue, thus simple usage and reliable for the process, quick support, fully automation, self/easy adjustment at changing process parameters, “forget it in the line”. Another approach goes for supporting quality control: The machine vision system shall increase the quality, “which yet we are not quite sure to exactly name parameters for”.

We state three main actual and future topics for industrial machine vision: Metrology supporting automation, quality control (inline/offline) and visualization of huge datasets.

Measuring geometric parameters has got near to the optical/physical light limits. There are many available systems which have built in standard functionalities – outer/inner diameter of circles, edge angles, relative object positioning etc. There are smart cameras, cameras with micro controllers, at least PC-systems. And there are different levels for maintaining the systems. In this case the trend shows a stabile level; just some decreasing costs and going smarter.

Quality control is an issue since years and still will be one for the developing community. There is a huge amount of image data entering the system. Feature segmentation is necessary prior to the capturing unit by sufficient illumination. Feature registration has to combine different lines for evaluating continuing strips. Good/proper quality has to be qualified fast with a better knowledge than the experienced operators. Actually the operators are under pressure of lot of possible bad features recognized by machine vision systems in the patterns. So the systems have to learn like humans for distinguishing between false/proper alarms.

Due to increasing automation levels there is need for simple visualization of huge datasets. The operators have to get the information about the status of a whole plant within a moment and without reading lots of tables. Additionally there are a lot of single measurement data available representing a single work peace. For quickly understanding the status of the process there is need for simple but efficient complete data visualization.

The general trend for industrial machine vision (except the first task of metrology) goes from the pixel orientated view to the object orientated evaluation. At least by increasing the system stability, the reliability, and by decreasing the maintaining stuff as well as the general costs.

9024-6, Session 2

Symbolic feature detection for image understanding

Sinem Aslan, Ege Üniv. (Turkey); Ceyhun B. Akgul, Bülent Sankur, Bogaziçi Üniv. (Turkey)

Automatic image understanding is a challenging problem in computer vision. Current partial solutions have been already finding applications ranging from industrial inspection, to medical diagnosis, from remote sensed imaging to media content analysis [1-2]. The dominant paradigm in the literature is an approach of pyramidal hierarchy. At the bottom layer, lies pixels or interest points derived from them. At the intermediate layer of representation, image structures such as lines, regions, segments takes place. As we move up the pyramid, higher level representations enable interpretation of the scene content. Typical of tasks that can be accomplished at this level one can cite scene recognition, defect detection, object identification, etc...

From a historical perspective, Biederman was the first to propose in 1981 scene emergent features which were introduced as the very basic geometrical forms of the scenes [3]. Accordingly, scene emergent features do not define objects by themselves, but once they are put together in a particular relationship with each other, they emerge as being objects. In this study, we tackle the very first step of analyzing images for automatic understanding by generating a primitive label map consisting of semantic interest points, in a way reminiscent of Biederman’s scene emergent features. Each point in a such a map is associated with an attachment score and can be used in a Bag-of-Words (BoW)-based image description for subsequent stages. This will make the scheme to move up from pixels or pixel-based primitives to intermediate level of representation exploiting both local geometrical and semantic structure of images. As such, the approach can be named as semantic interest point detection.

Existing local interest point detectors are developed from purely geometrical arguments, not directly taking into account the local semantics. They can be basically formulated as thresholding operators running on some geometrical “interestingness” measure. In our case, on the other hand, interest points are detected by scanning over the image a statistical classifier measuring the attachment of the local image patch to predefined primitive appearance classes. Judicious selection of primitive classes takes into account the local semantics more directly than conventional interest point detectors. The primitive classes are characterized by a handful of prototypical appearance seeds and populated by parametrically manipulating a set of variations and this approach makes the method dataset-independent.

In this preliminary study, we have used five basic appearance forms delineating flat regions, ramps, ridges/valleys and circular/elliptical peaks pits. We have parametrically generated photometric and geometric variations of these main forms in order to produce rich and varied primitive appearance patch datasets. Smoothed colored noise is added on the generated image patches in order to render them more realistic. Training, Validation and Testing sets are generated in order to measure the classification performance of a simple k-nearest neighbor classifier. Each set contains five shape classes each of which contains 16384 image patches of size pixels. Parameters such as intensity, azimuth angle, slope (a concept can be thought similar to the elevation angle), eccentricity of ellipse, variance of the Gaussian noise are sampled from the same statistical distribution for both of the three sets for training, validation and testing.

The image representation has been taken as a B-bin feature histogram accumulated from N feature measurements per patch. The feature has been kept as simple as in Eq. (1), computed as the slope of the line joining two randomly sampled intensity points within the patch.:

$$F(p_i,p_j) = \arctan\{(I(x_i,y_i)-I(x_j,y_j))/\sqrt{(x_j-x_i)^2+(y_j-y_i)^2}\};\ I = 1,..,N;\ j = 1,...,N \qquad \text{Eq. (1)}$$

where I(x_i, y_i) denotes the intensity value on the coordinate (x_i, y_i) for the point p_i. In order to classify a patch, we used two histogram distance measures: euclidean and chi-squared. Finally the classifier was k-NN where k was determined on the validation set rather than using cross-validation, because we have the opportunity of producing as many image patches as we needed. Classification accuracy on the test set is examined for different combinations of (N,B) with optimal k which was computed for each (N,B) combination on the validation set. The best classification accuracy of 84.36% is obtained when (N,B)=(2000,40), k=16 and chi-squared metric was used. Observing the confusion matrix obtained with (N,B)=(2000,40), it is determined that the confusion mostly occurs in point and ellipse classes.

References

[1] A. Ramanan and M. Niranjan, “A review of codebook models in patch-based visual object recognition,” Journal of Signal Processing Systems, vol. 68, no. 3, pp.333-352, 2012.

[2] Y.G. Jiang, J. Yang, C.W. Ngo, and A.G. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," Multimedia, IEEE Transactions on, vol. 12, no. 1, pp. 42-53, 2010.

[3] I. Biederman. "On the semantics of a glance at a scene," 1981.

9024-7, Session 2

Depth and all-in-focus images obtained by multi-line-scan light-field approach

Svorad ?tolc, AIT Austrian Institute of Technology GmbH (Austria) and Institute of Measurement Science (Slovakia); Reinhold Huber-Mörk, Branislav Holländer, AIT Austrian Institute of Technology GmbH (Austria); Daniel Soukup, Austrian Research Ctrs. GmbH (Austria)

Acquisition of depth information from images is typically obtained using specific setups such as time-of-flight sensors, configurations based on laser triangulation or multi-camera systems. At the same time, in order to observe a larger or varying depth of field, approaches to extend or adapt the focal range become necessary. In this paper, we address both issues related to industrial applications. We present a light-field multi-line-scan image acquisition system intended for the 3D inspection of fine surface structures, such as small parts, printed circuit boards, intaglio print, etc. The system consists of an area-scan camera, which allows for a small number of sensor lines to be extracted at very high frame rates (in our experiments, we used the AVT BONITO CL-400C camera), and a mechanism for transporting the inspected object (e.g., a conveyor belt). During the inspection, the object is moved at a constant speed orthogonally to the camera's optical axis as well as the orientation of the sensor lines. In each time step, a region of interest consisting of several sensor lines is read out and stored. Afterwards, by collecting all corresponding sensor lines acquired over time (i.e., all 1st lines, all 2nd lines, etc.), one can produce a 3D light-field structure comprising multiple views of the object as if observed from different angles (with respect to the optical axis of the system) at the same time. As the field of view of the camera is limited in the transport direction by the number of the extracted sensor lines (typically less than 5 degrees), the proposed device can be considered as an extremely narrow-baseline multi-view stereo system, which is nevertheless providing information-rich light-field data instead of a single stereo pair. Note that the nature of the relative motion between the object and the acquisition device as well as the approach to the construction of the light field guarantees the epipolar constraint in stereo vision automatically. This property allows for easy construction of so called epipolar plane images (EPIs) for each sensor column and subsequent EPI-based analysis. Beside the detailed hardware description of the proposed system, we also address the two objectives regarding the efficient processing of the obtained light-field data: (i) the construction of an all-in-focus color intensity image with an improved signal-to-noise ratio and (ii) the reliable estimation of a dense depth model. To achieve both goals, the light-field based processing is performed in the EPI domain. In both cases, the algorithmic solutions are proposed and analyzed from both the quality and performance points of view. Regarding the depth estimation algorithm, we also provide a comparison to existing stereo-matching techniques. The overall performance of the proposed system is demonstrated on several real world examples, such as the analysis of small deformations of printed matter, inspection of small parts and highly reflective metallic surfaces.

9024-8, Session 2

An image projection system that avoids image projection on all-except-of-the-screen objects that are in the area covered by output projector light

Viacheslav Sabirov, Skolkovo Institute of Science and Technology (Russian Federation)

The Problem.

The most common application of projector is to help make a lecture or any other presentation more attractive and interesting. However a projector introduces a significant disadvantage into the presentation process. The disadvantage is the intense light coming from extremely bright light bulb that is a parcel part of each device.

When a person looks at a usual light bulb, about 3 microwatts of power enters his eyes while a projector causes 300 microwatts (!) to enter the eyes. For people who make presentations often it results in:

- headaches,
- increased nervousness,
- increased chances of developing eye lens and retina diseases.

The human eye can only protect itself when a person looks at the projector directly. Therefore, the eye is extremely vulnerable during presentations whenever the projector's light source is in peripheral vision area during extended periods of time. Retina cannot send hurt signal to brain and a person does not feel damage that is done to it.

The Method.

Product determines the position of the presenter's face from the video captured by either web-camera or smartphone camera that has to be set up in the presentation room so that the screen or the presenter's face is visible at all times.

Authors of the invention suggest three options of what image the system can capture and process to achieve the goal.

Option 1: A camera captures the image of all objects (including the screen) that are in the area covered by output projector direct light rays. The image captured by said camera is processed by image processing unit that defines the approximate object position and shape. Finally, the decision of how to modify the projector output image is done by image changing unit based on the information received from image processing unit. Thus the image that is output by a projector is modified by turning to black the areas of original image that correspond to position and shape of said objects. The option is the best if the camera is placed in front of the screen center and under 0 degrees angle corresponding to the screen normal.

Option 2: If the camera is installed under a large angle (above 45 degrees) corresponding to the screen normal the proposed system can process image of shadow cast by the object on the screen. Thus the system can make an input image change decision without estimation of distance between the obstacle and projector.

Option 3: Option three is a combination of both previous options. Such system will have the best performance if camera faces the direction that is under the 45 degrees angle corresponding to screen normal.

Results & Summary.

So far our prototype works satisfactory: the light intensity entering eyes drops by a factor that equal to the projector device contrast (>2000). The System is very flexible regarding both the camera position and architecture flexibility as it will be able to use a smartphone for video capture. It means the system may work with any existing projector.

9024-9, Session 2

Line detection in a noisy environment with weighted Radon transform

Pavel V. Babayan, Nikita Y. Shubin, Ryazan State Radio Engineering Univ. (Russian Federation)

Line detection algorithms are widely used in many machine vision techniques especially in artificial object detection, recognition, tracking and parameters estimation. One of the most popular approaches to detect lines is based on the Radon transform. The typical Radon-based line detection algorithm consists of the calculation of the magnitude of the gradient of the image, Radon transform and local maximums

localization. Each local maximum belongs to the line. But in real-world applications Radon-based approach suffers from the noise and clutter, because they decrease the sharpness of the local maximums. There are different ways to improve the robustness of the approach in a noisy environment. One of the most effective ones is to take into consideration the direction of the image gradient, not only its magnitude. The common technique is to project an image gradient on different directions. The set of images is calculated, where each image belongs to the specified projection direction. Then Radon transform is applied to each image. The disadvantage of this algorithm is a high computational complexity because the projections and Radon transform need to be calculated for each direction.

In this paper we suggest a new approach to computational effective line detection using Radon transform. It is nearly 3 times more complex than a Radon-based approach using a gradient magnitude. It consists of four steps. The first step is a calculation of the image gradient in horizontal and vertical directions. The second step is performing of three Radon transforms for the horizontal, vertical gradients and their multiplication. The third step is a weighted summation of three transforms. On the last step the local maximums are localized. We call the calculation process from the first to the third step as a Weighted Radon Transform (WRT). The suggested WRT-based approach uses gradient direction information, so only the differences that are perpendicular to the line direction are integrated to make a local maximum corresponding to the line. It sharpens local maximums in comparison to a Radon-based algorithm.

From the point of view of the object tracking problem the local maximums become more temporary stable. This increases the object tracking quality.

The theoretical and experimental studies were performed to show the effectiveness of the WRT-based line detection. The theoretical study shows that the increase of the local maximum sharpness is determined by the cosine squared multiplier between the WRT and Radon transforms.

The experimental study was performed on the sets of the synthesized and real-world images and image sequences. The analysis of the line detection quality shows the improvement of the quality of the line detection in noisy environments. The line-based object tracking algorithm was implemented to study the effectiveness of the suggested approach in an object tracking task. The algorithm is based on the temporal filtering of the local maximums localized after the Radon transform or WRT. The improvement of the object tracking quality was approximately 40% in comparison to a simple Radon-based technique.

The suggested WRT-based algorithm can be effectively implemented in real-time systems using parallelization and FFT-based techniques.

9024-10, Session 3

Efficient adaptive thresholding with image masks

Young-Taek Oh, Young-Kyoo Hwang, Jung-Bae Kim, Won-Chul Bang, Samsung Advanced Institute of Technology (Korea, Republic of)

Problem Statement and Motivation

Adaptive thresholding is an useful technique for document analysis. In medical image processing, it is also useful for segmenting structures such as the diaphragm or blood vessels. This technique sets a threshold value using local information around a pixel, then binarize the pixel according to the value. Although it is robust to changes in illumination, it takes much time in computing threshold values, which requires to compute sum of all neighbor pixels. The use of the integral images can help to reduce this overhead. Using only 4 pixels, the sum of the given rectangular area can be computed very fast. However, medical images such as ultrasound often come with image masks, where ordinary algorithms do not yield expected results. For example, the threshold value at the boundary of the mask is different from the true value because pixels outside the mask image is also counted. Main problem is the shape of summing area is not rectangular near boundaries of the image mask. In this paper, we propose a novel adaptive thresholding technique resolving this problem.

Methods

Computation of integral images are intrinsically parallelizable. For 2D images, prefix sums (or scan) are computed for each row of an image, then computed for each column of the image. Because the prefix sum is a well-known algorithm, there are several implementations available. We use 'thrust' library which is included in CUDA SDK. In this process, pixel values outside the image mask are considered as 0.

Our solution to cope with an image mask is computing the integral image for the image mask. Because pixel values outside the image mask is assumed to be 0, the summed value contains only pixel values inside the image mask. Therefore, only computation of the valid pixel count is necessary. Let a pixel value be 0 or 1 whether the pixel is outside the image mask or not. Then area sum obtained from the integral image gives the number of valid pixels.

Preliminary Experimental Results

First, we measured the performance of our integral image generating algorithm. Time is measured with a machine having Intel i7 X990 CPU and GTX 580 graphics card. GPU based algorithm is 4 times faster than CPU algorithm in average, and it takes only 44.05 ms for a 10M pixel image.

We then compared the performance of adaptive thresholding algorithm with an image mask between CPU based naive algorithm and our proposed algorithm.

Image Size	0.16M	1M	4M	10M	40M
CPU (ms)	256.20	2190.75	6006.6	15337.10	63298.10
GPU (ms)	4.00	12.30	29.6	82.75	334.65

Conclusion

We propose an efficient adaptive thresholding algorithm with image masks. Our algorithm uses only one additional integral image of an image mask, which can be rapidly computed using GPU. In further research, we will show the performance of our algorithm on segmenting diaphragms and blood vessels from 3D US image volumes. More performance comparison including CPU multi-threaded algorithm will be also conducted.

9024-11, Session 3

A shape-preserving image binarization method

Jingu Heo, Samsung Advanced Institute of Technology (Korea, Republic of)

Extraction of essential information from images is a fundamental task in computer vision and machine learning. Since color and gray-scale images often contain a large amount of redundant information, retrieving crucial information about the identity of these images has been a longstanding goal for image processing, computer vision, and pattern recognition researchers. An extreme way of depicting an image can be realized by using only two values (0 or 1), known as image binarization, where the desired processed image only contains edges or sharp shapes whose information can be directly useful for determining the identity or category of the image. Although there are many edge detection methods and image thresholding approaches, it is still unclear which methods should be utilized for representing images to accomplish certain high-level tasks in an automated way.

In order to overcome common problems in image binarization methods, we propose a shape-preserving image binarization method. The overall procedure can be explained as follows. Firstly, an input image is processed by using bilateral filtering, followed by Gaussian smoothing. Then, pixel-wise division is applied for extracting a shape preserving and noise removed representation. Depending on a pixel distribution of an

image, non-important pixels are removed based on histogram. Finally, a clustering method based on k-means is applied for retrieving a threshold. Based on the automatically calculated threshold value which bisects two regions based on the intensity of the resulting image after pixel-wise division, shape-preserving edge detection results can be obtained. Our proposed work is mainly motivated by Self Quotient Image (SQI), an illumination normalization method. However, unlike SQI, extreme data compression while maintaining consistent information under different settings can be achieved by using the proposed work.

In order to show the effectiveness of the proposed algorithm, we provide quantitative experimental results by using human faces. Our goal is not to develop or compare state-of-the-art algorithms in face recognition; rather, we try to show the effectiveness of our proposed shape representation for recognition. For this reason, we focus on providing high-discrimination power in face recognition despite the use of binary images and only use one-to-one distance matching, without requiring any training or subspace learning steps. The proposed binarization scheme outperforms original illumination normalization (SQI) (despite the use of only binary information) with a margin of 3% VR and 1.3% EER (98% VR, 2% ERR), respectively. Here, we compute a total of 4980 by 4980 similarity matrix. In order to show the effectiveness of the proposed work qualitatively, we compare our proposed work against other edge detection methods by showing the proposed approach are more intuitive to humans. We also observed that we can obtain both very consistent results although the input images acquired under severe changes in image conditions.

In conclusion, the proposed method retrieves shape information consistently under a wide of image conditions and can directly used for recognition in a fast and automated way, without the need of subspace modeling or any training steps for dimensionality reduction.

9024-12, Session 3

Illumination-invariant pattern recognition using fringe-adjusted joint transform correlator and monogenic signal

Paheding Sidike, Vijayan K. Asari, Univ. of Dayton (United States); Mohammad S. Alam, Univ. of South Alabama (United States)

The joint transform correlator (JTC) technique has shown attractive performance for real-time pattern recognition applications. The reference image used in a JTC is usually prestored in computer while the unknown input scene from a live world which may or may not contain the object of interest. Accordingly, the input scene illumination could be entirely different from that of the reference. This degrades the performance of JTC in terms of discrimination sensitivity. A number of JTC architectures have been proposed to accommodate variance in illumination and noise. Among the various JTC techniques proposed in the literature, the fringe-adjusted JTC (FJTC) yields remarkable promise for object detection, and it has been shown that the FJTC has a better performance than do the classical and the binary JTCs under varying illumination conditions of the input scene; however, it has been found that FJTC is not illumination invariant. Therefore, to alleviate this drawback of FJTC, an illumination-invariant joint transform correlator, based on the fringe-adjusted joint transform correlator and the monogenic signal concepts, is presented. The performance of the classical FJTC, phase-only FJTC and the proposed local phase based FJTC technique in unknown input-image with varied light illumination is investigated and compared. The proposed detection algorithm makes use of the monogenic signal from a two dimensional object region to extract the local phase information as a preprocessing. Computing the monogenic signal enables us to split the local phase information from the local amplitude thereby achieving illumination invariance. In the preprocessing step, both reference image and unknown input scene are transformed into the local phase representation, which helps to significantly reduces illumination sensitivities. Next, the FJTC technique is implemented to detect and identify targets in the input scene under varying illumination condition. Fourier plane image subtraction technique is also applied before using fringe-adjusted filter to the joint power spectrum for eliminating undesired zero-order term and cross-correlation terms. In the testing stage, a real-life New Era Technology dataset, which contains varied background illumination, is used for object detection. Computer simulation results show that the proposed technique can significantly improve the performance of FJTC with varying illumination of the input scene and effectively detect targets from input scenes with higher pattern discriminability, sharper peaks, and more-robust detection, whereas alternative FJTCs produce false detection when the input scene is under severe low and high illumination conditions. In addition, it is also found that using the monogenic signal representation as a preprocessing enables FJTC to maintain a sharp correlation peak, a narrow correlation width and a high peak-to-sidelobe ratio both in poor illumination background and in normal light condition. Optical implementation for the proposed scheme is also suggested. The proposed technique may be used as a real-time region-of-interest identifier in wide-area surveillance for automatic object detection under dark or bright condition that beyond human vision.

9024-13, Session 3

Illumination invariant 3D change detection

Yakov Diskin, Vijayan K. Asari, Univ. of Dayton (United States)

We present a 3D change detection technique designed to support various applications in changing environmental conditions. The novelty of work lies in our approach of creating an illumination invariant system tasked with detecting changes in a changing environment. Previous efforts have focused on image enhancement techniques that manipulate the intensity values of the image to create a more controlled and unnatural illumination. Since most applications require detecting changes in a scene irrespective of the time of day, lighting or weather conditions that may be present at the time of the frame capture, image enhancement algorithms fail to suppress the illumination differences enough for Background Model (BM) subtraction to be effective. A more effective change detection technique utilizes the 3D scene reconstruction capabilities of structure from motion to create a 3D background model of the environment. By rotating and computing the projectile of the 3D model, previous work has been shown to effectively eliminate the background by subtracting the newly capture dataset from the BM projectile leaving only the changes within the scene. Although previous techniques have proven to work in some cases, these techniques fail when the illumination significantly changes between the capture of the datasets. Our approach completely eliminates the illumination challenges from the change detection problem. The algorithm is based on our previous work in which we have shown a capability to reconstruct a surrounding environment in near real-time speeds. The algorithm, namely Dense Point-Cloud Representation (DPR), allows for a 3D reconstruction of a scene using only a single moving camera. Utilizing video frames captured at different points in time allows us to determine the relative depths in a scene. The reconstruction process resulting in a point-cloud is computed based on SURF feature matching and depth triangulation analysis. We utilized optical flow features and a single image super resolution technique to create an extremely dense model. The accuracy of DPR is independent of the environmental changes that may be present between the datasets, since DPR only operates on images within one dataset to create the 3D model for each dataset. Our change detection technique utilizes an Iterative Closest Point (ICP) scheme to register the two 3D models. The ICP technique uses the mean squared cost function to compute the optimal homography needed to transform a 3D point-cloud model from one dataset to align with the 3D model produced by another. Next, in order to eliminate any effects of the illumination change we convert each point-cloud model into a 3D binary voxel grid. A one is assigned to voxels containing points from the model while a zero is assigned to voxels with no points. In our final step, we detect the changes between the two environments by geometrically subtracting the registered 3D binary voxel models. This process is extremely computationally efficient due to logic-based operation, XOR, available when handling binary models. We measure the success of our technique by evaluating the detection outputs, false alarm rate and computational

expense when compared with two state-of-the-art change detection techniques.

9024-14, Session 4

High throughput imaging and analysis for biological interpretation of agricultural plants and environmental interaction

Hyundae Hong, Jasenka Benac, Daniel Riggsbee, Keith A. Koutsky, Monsanto Co. (United States)

The agriculture of today's world is facing challenge of producing increased crop yields with limited and weakening resources like water and soil nutrition as the world population grows and food demand is escalated while the natural environment for agriculture becomes deteriorated by climate change. In developing higher yielding crops, it is essential to understand the relationship between genomics and phenomics. The greenhouse offers merits for crop screening and development: (1) the environmental variability is controlled so the genotype to phenotype links are better discerned, (2) high end analytical technologies like imaging can be efficiently used with higher precision than field cases, (3) testing various conditions of resource scarcity like water and nutrient deficient can be easily implemented, (4) turn around of one study is shorter than field study, (5) it is feasible to handle large number of plants by automation technology - high throughput phenotyping.

We have state of the art automated greenhouse (AGH) facility; automation was made to key operations – automatic handling of pots, water and nutrient supply, and imaging of plant phenotypes. Each imaging station has both broadband color digital cameras and hyperspectral cameras. The AGH operation system senses and tightly monitors light intensity of imaging lamps for quality control. The acquired broadband images are preprocessed for white-balance and exposure compensation; the hyperspectral images are normalized for reflectance calibration at each wavelength. Our customized software analyzes those preprocessed images for the measurement of plant geometry and biochemical properties. Morphometric information is gathered such as plant area, height, width, canopy area and color hue scale. With these core metrics, we also measure the plant biomass which is modeled by the correlation between image dimension and physical plant mass at various growth stages and crop types. The hyperspectral image analysis is done to characterize biochemical compositions of plant material; we measure chlorophyll, anthocyanin, and foliar water content; these metrics are calculated by ratios of reflectance profiles of plant detected at various wavelengths of light and corresponding broadband images.

Imaging and analysis on crops like corn, soybean, and cotton that were conducted for last 4 years at our AGH facility have demonstrated successful applications of imaging and analysis technologies on high throughput plant phenotyping. Plant biomass and stress related biochemical signatures were found to be biologically relevant by distinctive trends complying with given abiotic stress. Those distinctive patterns of responses were confirmed to be not only aligned in scientifically meaningful ways with given environmental conditions like water and nutrient deficient, but genetic backgrounds of crops as well.

9024-15, Session 4

Interactive quantification of complex objects in microscopy images

Reid B. Porter, Christy Ruggiero, Neal Harvey, Los Alamos National Lab. (United States)

In material science and bio-medical domains there is great value in automatically detecting, delineating and quantifying particles, grains, cells and neurons within digital microscopy images. These are challenging problems because the objects of interest are complex, compound objects with multiple parts, and there is large variability in the parts as well in how parts group into the larger object. In bio-medical imaging, structures such as Glomeruli, are defined by particular configurations of tissue and cell types that are common throughout the Kidney. In material science, particles of interest are often agglomerations of smaller sub-particles that form in particular ways. Standard detection and segmentation routines are often insufficient for these types of problems. Another common characteristic of these problems is that inexperienced users often find them difficult as well, and domain expertise is required to resolve ambiguous cases. Experts often bring greater knowledge of the imagery and context to the problem. They also have a more detailed understanding of the application level objectives which helps them make better decisions with limited data. The problem of course, is that the expert's time is expensive.

We suggest interactive image segmentation and quantification tools are a promising way to solve these difficult problems because they can provide two complementary benefits: 1) a way for experts to include their domain knowledge while solving analysis problems they care about, and 2) a way to build upon the expert input to improve the tools over time (minimizing the amount of expert time required). Historically, image processing has developed interactive tools to provide benefit 1 (e.g. Interactive graph cuts, marker based segmentation and active contour methods). We suggest recent advances in machine learning, and specifically, structured output prediction, enable a new generation of tools that can also provide benefit 2.

In this paper we develop an interactive image quantification solution that has both benefits. We present expert's with an initial segmentation and quantification, and then provide a number of intuitive editing tools (merge/split/labeling) that let users interact and refine the segmentation until it satisfies their objectives (benefit 1). Given multiple examples of this interaction, we then show how tools can improve over time (benefit 2). Tools are improved by combining a graph-based image representation (with sparse coding and Hu moment attributes) with structured output prediction methods. We exploit a property of connected component analysis (the fact that connected components commutes with thresholding) to derive structured output prediction methods for gradient descent and structured support vector machines. We also develop a way to introduce hierarchy into the method to better characterize objects with long-range dependencies. Our experiments compare the various design choices in the solution method and also evaluate the application level performance in terms of efficiency gains for the expert.

9024-16, Session 4

On the use of MKL for cooking action recognition

Simone Bianco, Gianluigi Ciocca, Paolo Napoletano, Univ. degli Studi di Milano-Bicocca (Italy)

Automatic action recognition in videos is a challenging computer vision task that has become an active research area in recent years. Most diffused approaches to this problem consider (i) the use of local as well as global features by computing them over 2D frames or over a 3D video volume [1] (ii) the use of factorization techniques over video volume tensors and defining similarity measures over the resulting lower dimensional factors [2]. In addition, existing strategies usually use kernel-based learning algorithms that considers a simple combination of different features completely disregarding how such features should be integrated to fit the given problem. Since a given feature is most suitable to describe a given image/video property, the adaptive weighting of such features can improve the performance of the learning algorithm.

In this paper, we investigated the use of the Multiple Kernel Learning (MKL) algorithm to adaptive search for the best linear relation among the considered features. MKL has been introduced by Lanckriet et al. in 2004 as extension of the support vector machines (SVMs), which are intrinsically single kernel, to work with a weighted linear combination of several single kernels [3]. This learning approach allows to simultaneously estimate the weights for the multiple kernels combination as well as the

underlying SVM parameters.

In order to prove the validity of the MKL approach, we considered the descriptor proposed in a recent work by Wang et al [1]. The authors use multiple features aligned with dense trajectories. The trajectories are extracted by tracing initial key points, set on a regular/dense grid of pixels, over a given time interval, while the aligned multiple features are extracted from a window surrounding each point of the trajectory. More in details the descriptor includes features that characterize the appearance (histograms of oriented gradients, HOG), motion (histograms of optical flow, HOF) and motion boundary (histograms of differential optical flow, MBH). MKL will automatically provides the best weights to fuse the multiple features of the descriptor.

We experimented our approach on cooking videos acquired in the context of the Feed for Good project [4]. The main goal of this project is the creation of a cooking assistant application to guide the users in the preparation of the dishes relevant to their profile diets and food preferences, and illustrating the actions of the cook. The videos have been recorded in a professional kitchen with stainless steel worktop using multiple cameras. With respect to other domains, our cooking domain presents particular challenges such as frequent occlusions and food appearance changes. Results confirm that the use of MKL improves the classification performance.

[1] Wang, H., Klser, A., Schmid, C., and Liu, C.-L., “Dense trajectories and motion boundary descriptors for action recognition,” International Journal of Computer Vision 103(1), 60–79 (2013).

[2] Lui, Y. M., “Tangent bundles on special manifolds for action recognition,” Circuits and Systems for Video Technology, IEEE Transactions on 22(6), 930–942 (2012).

[3] Lanckriet, G. R. G., Cristianini, N., Bartlett, P., Ghaoui, L. E., and Jordan, M. I., “Learning the kernel matrix with semidefinite programming,” J. Mach. Learn. Res. 5, 27–72 (Dec. 2004).

[4] Bianco, S., Ciocca, G., Napoletano, P., Schettini, R., Margherita, R., Marini, G., Gianforme, G., and Pantaleo, G., “A semi-automatic annotation tool for cooking video,” in [Image Processing: Machine Vision Applications VI], 8661, 866112, SPIE (2013).

9024-17, Session 4

Video-based automatic vehicle classification based on axle detection

Xinhua Xiao, Zhuo Yao, Chia Yung Han, Xuefu Zhou, William G. Wee, Wei Heng, Univ. of Cincinnati (United States)

Problem Statement:

Vehicle classification data is critical to almost all fields of transportation engineering and management applications, including pavement and roadway design, road management and maintenance, environmental impact analysis, etc. Although there are many other axle base vehicle classification methods and tools available, they are all limited by their intrusive axle detection nature. Pneumatic tube traffic counters usually provide the total number of axles at the end of a count, but an adjustment factor is usually required to convert the total number of axles into total number of vehicles. Noting the difficulties lied in the above discussed intrusive length-based and axle-based vehicle classification methods and models, video traffic data outperformed with its capability of generating more accurate vehicle classification and speed data. That is, to promote the application of video-based detection systems, an efficient and fast method for processing video to produce accurate traffic information under varied traffic conditions is needed to greatly improve the quality of traffic data acquisition.

The goal of this research is to explore an axle-based vehicle classification method using video processing techniques for vehicle classification. To achieve the goal, two objectives are designated to fulfill: (1) to design a vehicle classification system, Rapid Video-based Vehicle Identification System (RVIS), based on vehicle axles parameters; and (2) to calibrate and validate the proposed axle parameter based vehicle classification system using case study.

Methodology:

The RVIS has four modules, namely, video acquisition, vehicle axle parameters extraction, feature-based vehicle classification and the calibration and validation module. The video acquisition module enhances, splits, and resizes raw video data into a common standard size and length so the program can read. The vehicle axle parameter extraction module is the core module in this system. It detects, extracts the vehicle axles and their configurations. The vehicle classification module, which contains predefined FHWA vehicle classification axle configurations, matches and classifies the outputs from the axle parameter extraction module. The classification is based on number of axles and their configurations on axle groups and spacing. Finally, a calibration and validation module is designed to guarantee the performance of the proposed RVIS system. This module uses the ground truth video to correct any possible misclassifications and errors.

Experimental results:

Preliminary experimental results show that the proposed method can automatically detect the vehicle axles in real-time and calculate the required axle parameters for vehicle classification. More experiments are still needed to further validate the proposed method. Nonetheless, the research proposed a prototype of video-based vehicle classification system using axle parameters extracted from image processing and computer vision techniques, and provides valuable insights to the video based vehicle classification using axle based classification scheme.

Summary and Conclusion:

This study explores a prototype of vehicle classification using image processing techniques. The results show that it is a robust method to detect vehicular axles and extract the axle configurations. The proposed RVIS system is capable of handle video inputs and produce reliable results once the parameters are set. However, the automatic parameter determination will be explored in future study.

9024-18, Session 4

Hyperspectral estimation using RGB color for foodborne pathogen detection on agar plates

Seung-Chul Yoon, Tae-Sung Shin, William R. Windham, Bosoon Park, Kurt C. Lawrence, Young-Wook Seo, Agricultural Research Service (United States)

Detection and identification of foodborne pathogens are increasingly important for development of intervention and verification strategies for the food industry and regulatory agencies. Traditional culture-based direct plating methods are still the “gold standard” for presumptive-positive pathogen screening in many microbiology laboratories, where agar media are routinely used for isolation, enumeration, and detection of pathogenic bacteria. In practice, highly skilled technicians visually screen and manually select presumptive-positive colonies by trial and error for microscopic, biochemical, serological and molecular confirmation tests whose results may or may not be obtained rapidly. Therefore, the culture methods are labor intensive and prone to human subjective errors. Another challenge with direct plating is that competitive microflora often grow together with target microorganisms on agar media and can appear morphologically similar.

Researchers at the Agricultural Research Service of the U.S. Department of Agriculture, have been developing machine vision techniques based on hyperspectral imaging for detection and identification of pathogenic colonies including Campylobacter, Salmonella and Shiga toxin-producing Escherichia coli. In an attempt to reduce human error and increase the screening throughput, visible and near-infrared (VNIR) hyperspectral imaging in the spectral range from 400 to 1000 nm has been studied for detecting the foodborne pathogens.

Recently, the VNIR hyperspectral imaging studies have been expanded to develop multispectral imaging techniques in order to reduce the cost and complexity involved with the deployment of hyperspectral imaging system. This paper reports the latest development of color vision

techniques for detecting pathogens on agar plates with hyperspectral image classification models that were developed using full hyperspectral data. Thus, the objective of the study was to test the feasibility of a RGB color-based prediction of pathogens on agar plates with hyperspectral classification models based on the spectral range from 400 and 700 nm (240 narrow spectral bands).

A multivariate linear regression method was used to estimate hyperspectral curves with 240 band points from RGB bands with an assumption of a linear relationship between RGB bands and visible spectrum between 400 and 700nm. The regression method does not require capturing images of color references with known spectral responses in order to estimate regression coefficients. The performance of the RGB-to-hyperspectral estimation was evaluated with the hyperspectral data in terms of its estimation accuracy and pathogen detection accuracy. To simulate RGB image data, 450nm, 550nm and 650nm are extracted from the hyperspectral data cubes.

For Salmonella detection, a training set was obtained from pure cultures with known background organisms including two Salmonella serotypes (Enteritidis and Typhimurium), whereas a test set for validation was obtained from mixed cultures with unknown background flora in chicken carcass rinses where Salmonella was added into chicken carcass rinses. The mean R-squared value for hyperspectral estimation was 0.90 and the detection accuracy of the estimated hyperspectral prediction model with quadratic discriminant analysis for the test set was 92% (96.5% with the original hyperspectral model). The results of the study suggested that color-based detection may be viable as a multispectral imaging solution without much loss of prediction accuracy compared to hyperspectral imaging.

9024-19, Session 4

Improved wheal detection from skin prick test images

Orhan Bulan, Xerox Corp. (United States)

The skin prick test (SPT) is one of the most commonly used methods for diagnosing allergies to food, pollen, dust, pet dander and fungus, among others. SPT's are typically applied to forearms or back, where multiple allergens are introduced to patient's skin simultaneously. The test region on the skin is marked with a pen corresponding to number of allergens to be applied. After dropping small amount of allergens onto the marked regions, each drop is pricked with a metallic sterile pin/needle and skin reaction to the allergens within 15-20 minutes are evaluated. A wheal is provoked in the test region surrounded by erythema depending on the skin reaction to a specific allergen. The physical size of the wheal is the key parameter for allergy diagnosis.

The wheal size is typically measured by a nurse or medical personnel using a ruler. This manual method for measuring wheal size is a cumbersome process and is prone to errors and inter-observer variations. The SPT results of a patient can be interpreted differently due to measurement errors and inter-observer variations of different medical personnel, which can easily cause inconsistencies in the diagnosis of the same patient. The manual measurement process can be automated by capturing images from the test region using a digital camera and estimating the wheal size from the captured images. As a prerequisite of this automated method, an accurate wheal detection algorithm from SPT images is required.

In this paper, we propose a method for improved wheal detection from SPT images. Our method operates by first localizing the test region from captured images by detecting the calibration marks drawn on the skin. The calibration marks, later on, can also be used for mapping image coordinates to real-world coordinates to estimate the physical size of the detected wheal using a camera calibration technique. After localizing the test region on the image plane, we perform a color transformation from RGB to YCbCr and discard the luminance channel (Y). This process eliminates the illumination variation across the test region, which otherwise adversely impacts our wheal detection algorithm. The contrast between the wheal and the surrounding erythema is maximized by performing principal component analysis (PCA) on the Cb and Cr color channels. We then perform morphological operations on the contrast enhanced image to detect the wheal on the image plane.

The performance of the proposed method was evaluated across 36 allergy patients. The calibration marks were drawn on the skin of the patients as the corners of a square, where the smallest distance between the corners was set to 2 cm. The results of the SPT were captured in a typical room lighting conditions (indoor, fluorescent lighting) with a mobile phone, an iPhone 5 with an 8 megapixel camera. The images were captured in an uncontrolled environment (varying illumination, geometry, orientation) with default settings (auto focus, auto flash) of the camera. From the acquired images, we first localized the test region by detecting the calibration marks drawn on the skin. The localized test regions were cropped out from the rest of the images. We performed median and low pass filtering on the localized images to reduce the impact of noises due to hair/freckles in the test region. The filtering operations were performed in all the color channels and the size of the localized images were reduced by 2 after the filtering operations. After localizing the test region and pre-processing the localized image, we evaluated the performance of the proposed wheal detection algorithm by comparing its results with the manually segmented images (groundtruth). Our experimental results show that the proposed method provides an efficient solution for wheal detection from SPT images and achieves 94% accuracy. Our proposed method can be integrated as an application to a smartphone, which can be used in allergy clinics to efficiently interpret SPT results.

9024-20, Session 5

Face recognition by detection of matching cliques of points

Fred W Stentiford, Univ. College London (United Kingdom)

Many approaches to face recognition are reported in the literature [1,2]. Graph matching approaches provide attractive alternatives to the feature space solutions in computer vision. Identifying correspondences between patterns can potentially cope with non-rigid distortions such as expression changes, pose angle and occlusions. However, graph matching is an NP-complete problem and much of current research is aimed at solving the associated computational difficulties [3-12]. The approach taken in this paper detects fully connected graphical structure that is common between pairs of images and uses the extent of such structure to measure similarity.

A pictorial structure is represented as a collection of parts and by a graph where the vertices correspond to the parts and there is an edge for each pair of connected parts. In this paper parts correspond to individual pixels. Given a set of pixels in image 1 that correspond to a set of pixels in image 2 the following values are matched by all pixels to form a clique : (1) grey level gradient (2) grey level value and (3) angles between all pairs of pixels.

Clique generation begins with the selection of a random pair of pixels from reference image 1 and a pair from candidate image 2 that satisfy (1,2,3). A new pair of points is added that satisfy (1,2) and the angle condition with the nearest pixel in image 1. It is noted that distant points in image 1 are very likely to satisfy the angle condition if the local condition is met. Further candidate points are selected randomly and added to the clique if similar conditions are satisfied. Up to N attempts are made to find a new point after which the current clique is completed and the construction of a new clique started. After the generation of P cliques the largest is retained. The candidate is classified according to the class of the reference that obtains the largest clique in the candidate.

The process allows more than one point in the first image to be mapped into the same point in the second image, but not the reverse. This gives the search more freedom to navigate around occlusions. The relationship between points is not dependent upon their separation or absolute position and therefore the similarity measure is translation and scale invariant.

The approach is evaluated on the Yale Face Database A [16] where the category Normal is used as a reference set when measuring the similarity of the 15 faces in the expression, illumination and occlusion categories.

A 100% correct result was obtained without a training stage that was an improvement on other published results. 100% correct results were also obtained on occluded versions. The extraction of nodes in each clique is an independent operation and may therefore be conducted in parallel. This means that the overall potential for a speedup of several orders of magnitude is possible in an appropriate implementation.

References

[1] Patel, R, Rathod, N., Shah, A.: Comparative analysis of face recognition approaches a survey. Int. J. of Computer Applications, vol. 57, no. 17, pp. 50-61, (2012)

[2] Naruniec, J.: A survey on facial features detection. Int. J. of Electronics and Telecommunications, vol. 56, no. 3, pp. 267-272, (2010)

[3] Leordeanu, M., Hebert, M., : A spectral technique for correspondence problems using pairwise constraints. ICCV, (2005)

[4] Felzenszwalb, P. F., Huttenlocher, D. P. : Efficient matching of pictorial structures. CVPR, (2000)

[5] Fergus, R., Perona, P., Zisserman, A. : A sparse object category model for efficient learning and exhaustive recognition. CVPR, (2005)

[6] Kim, J., Grauman, K. : Asymmetric region-to-image matching for comparing images with generic object categories. CVPR, (2010)

[7] Duchenne, O., Joulin, A., Ponce, J. : A graph-matching kernel for object categorization. ICCV, (2011)

[8] Duchenne, O., Bach, F., Kweon, I., Ponce, J. : A tensor-based algorithm for high-order graph matching. IEEE Trans PAMI, vol. 33, no. 12, pp 2383-2395, (2011)

[9] Celiktutan, O., Wolf, C., Sankur, B. : Fast exact matching and correspondence with hyper-graphs on spatio-temporal data. Technical Report LIRIS-RR-2012-002, INSA-Lyon, (2012)

[10] Kolmogorov, V., Zabih, R. : Computing visual correspondence with occlusions using graph cuts. ICCV, (2001)

[11] Berg, A. C., Berg, T. L., Malik, J. : Shape matching and object recognition using low distortion correspondences. CVPR, (2005)

[12] Cho, M., Lee, K. M. : Progressive graph matching: making a move of graphs via probabilistic voting. CVPR, (2012)

[13] Wiskott, L., Fellous, J-M., Kruger, N. von der Malsburg, C., Face recognition by elastic bunch graph matching, IEEE Trans. Pattern Anal. Machine Intell., 16, 775-779, (1997).

[14] Cootes, T. F., Edwards, G. J., Taylor, C. J.: Active Appearance Models, IEEE Trans. Pattern Anal. Machine Intell., 23, 681-685 (2001)

[15] Matthews, I., Baker, S.: Active appearance models revisted. Int. J. Computer Vision, vol.. 60, no. 2, pp 135-164, (2004)

[16] Yale Face Database, http://cvc.yale.edu/projects/yalefaces/yalefaces.html.

9024-21, Session 5

Scoring recognizability of faces for security applications

Simone Bianco, Gianluigi Ciocca, Giuseppe Claudio Guarnera, Univ. degli Studi di Milano-Bicocca (Italy); Andrea Scaggiante, Bettini S.r.l. (Italy); Raimondo Schettini, Univ. degli Studi di Milano-Bicocca (Italy)

In security applications the human face plays a fundamental role, since it conveys a person identity. It offers several strong points, such as being a non-contact process which does not require a direct interaction with the subject. We have to assume non-collaborative subjects, hence a face can be partially visible (or possibly occluded) because common-use accessories such as sunglasses, hats, scarves are worn but also because of the use of items with the aim to cover the whole face or to distort facial outlines (e.g. stockings masks). Another source of occlusion could be the pose of the head relative to the camera, if it differs significantly from the frontal pose. Given a video sequence in input, the proposed real-time system is able to establish if a face is depicted in a frame and to determine the degree of recognizability, in terms of clearly visible facial features that allow subject recognition (i.e. eyes, nose, mouth) and to assign a score. Our system allows the selection of the facial features required to classify a face as visible and also to provide an assessment which is not limited to just a single frame, but also takes into account the trend over a customizable time window, to increase the reliability and to allow a high level analysis of the subject behavior. For each frame the first step is face detection, which is performed using a standard implementation of the Viola and Jones algorithm; a frame without a face could be simply classified as “Occluded” or with a different label, depending on the application. On the sub-image defined by the face bounding box we search for the mouth, nose and eyes. To take into a account multiple detections and false positives, a filtering is performed based on the size and location of the bounding boxes of the facial features previously found. Multiple detections of the same feature are addressed pairwise, checking for each couple of bounding boxes the ratio of the intersection area w.r.t. the area of the smallest bounding box within the pair. If the ratio is greater than an empirical threshold the bounding boxes are merged and the pair of features is considered as a single one, repeating the process until all detections of the same type are considered. To address false positives we rely on the geometric properties of the face, which allow us to derive some simple rules, such as “the nose should be lie within the eyes, just below them”. A small set of rules proves to be useful enough to filter almost all erroneous detections. A possible issue of the face detections algorithm, which is based on the Haar features, is that it can be misled by a cartoon-like image. We hence apply on the face bounding box a skin-based face detector with lighting compensation, which is able to correctly discriminate cartoon and human faces even in presence of a complex background. The detected facial features are then compared with the customizable face visibility criteria. Even the behavior of a collaborative-subject, which performs a sudden change of pose, could be classified as attempting to disguise the system, using a simple frame-based analysis. We propose to track the recognizability score of the face over a sliding time window, where each frame gives a vote in the domain {no detection, occluded, partially visible, visible}. We evaluated our system both in qualitative and quantitative terms, using a challenging data set of manually annotated videos. Preliminary results confirm the effectiveness of the proposed system.

References:

[1] Viola, P.; Jones, M., “Rapid object detection using a boosted cascade of simple features”, Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol.1, no., pp.I-511,I-518 vol.1, 2001

[2]Yu-Ting Pai; Shanq-Jang Ruan; Mon-Chau Shie; Yi-Chi Liu, “A Simple and Accurate Color Face Detection Algorithm in Complex Background”, Multimedia and Expo, 2006 IEEE International Conference on, vol., no., pp.1545,1548, 9-12 July 2006

[3] Peter, K.J.; Nagarajan, G.; Glory, G.G.S.; Devi, V.V.S.; Arguman, S.; Kannan, K.S., “Improving ATM security via face recognition”, Electronics Computer Technology (ICECT), 2011 3rd International Conference on, vol.6, no., pp.373,376, 8-10 April 2011

[4] Sungmin Eum; Jae Kyu Suhr; Jaihie Kim, “Face recognizability evaluation for ATM applications with exceptional occlusion handling”, Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on, vol., no., pp.82,89, 20-25 June 2011

[5] Jae Kyu Suhr, Sungmin Eum, Ho Gi Jung, Gen Li, Gahyun Kim, Jaihie Kim, “Recognizability assessment of facial images for automated teller machine applications”, Pattern Recognition, Volume 45, Issue 5, May 2012, Pages 1899-1914

9024-22, Session 5

Structure-from-motion reconstruction based on weighted hamming descriptors

Guoyu Lu, Vincent Ly, Chandra Kambhamettu, Univ. of Delaware (United States)

Structure-from-Motion (SfM) reconstruction is largely used as a technique to recover the 3D shape of an object based on 2D images. Based on 2D images captured by the same or different cameras, the 3D coordinate of the feature points in an object can then be calculated. Together with the 3D position information, the camera pose of each image can also be computed. Compared with Stereo Reconstruction which uses at least 2 calibrated cameras towards the same object, SfM does not require the cameras to be calibrated. For this less strict restriction, images used for SfM is easier to obtain, which allows the SfM technique to be applied to large-scale reconstruction tasks. With the development of mobile phone, the number of images uploaded to photo collection websites has been largely increased. Usually, images obtained from photo collection websites(e.g. Flicker) are used for large-scale reconstruction. In reconstructing a large building or even a city, several hundred to thousand of pictures are used in building a dense reconstruction. Structure-from-Motion reconstruction analyzes the distributed images and combines the images' information as a 3D model.

Local features are used in finding correspondences of the images. Because of the good properties such as invariance to rotation, translation and scaling, SIFT is usually used as the local feature for finding correspondences. In each image, usually several thousand features will be detected. However, as SIFT feature contains 128 dimension real-value numbers, a large memory is required to store all the image descriptors. The purpose of this paper is to find a solution which can reduce the memory requirement for large-scale Structure-from-Motion reconstruction.

In this paper, we learn a projection matrix that projects the high dimensional feature into a low dimensional space in reducing the curse of dimensionality. After projection, the distance of descriptors belonging to the same point is decreased and the distance of the descriptors belonging to different points is increased. With the lower dimensional descriptor, the computation speed is also accelerated. To further reduce the memory requirement, we map the lower dimensional real-value descriptors into Hamming space which also simplifies the distance computation. After projecting to Hamming space, several descriptors usually has the same distance to the query descriptor. To deal with this problem, we give weight to each dimension of the Hamming descriptor. In original Euclidean space, some dimensions of the real-valued descriptor contribute more than other dimensions to the final correct result. In Hamming space, we will give higher weight to these dimensions. This will help to avoid the situation that several learned Hamming descriptors have the same distance to the query descriptor.

In experiments, we present the result of both learned lower dimensional real-valued descriptors and the Hamming descriptors. For both descriptors, we show the performance of different dimensional descriptors. Experiments show that we can achieve comparable reconstruction result with much less memory consumption. The dataset we use is Notre Dame dataset, which collects 715 images from Flickr. The photos capture the view of Notre Dame Cathedral. As we use a single computer instead of the computer cloud, we randomly select 50 images to do the reconstruction for verifying our method. Using the original SIFT descriptor, there are 16895 points get reconstructed. For low-dimension real-value descriptors, we use 32 dimensional projected descriptors: it is just 1/4th of the original dimensionality. As after projection, the original descriptor will be projected to the range of [-1721, 1687] in floating value, which original unsigned char data format cannot store. So we scale the descriptor to the range of [0,255] and round the value into integer. There are 14611 points in the reconstruction model, which is similar to the one using SIFT. Comparing the original reconstruction result with the new reconstruction result, the visual effect is not readily distinguishable, indicating that our learned descriptor performs well on finding the correct correspondences. However, the memory cost is just 25% of the original system. By changing the distance computation method from Euclidean distance to Hellinger distance, further improvement can be achieved. After changing the distance computation method, there are another 253 points reconstructed in 3D space, which is an improvement compared with original distance computation method. For Hamming descriptor, 96 dimensional Hamming descriptor usually gives the best trade-off between the memory consumption and accuracy. The reconstruction model contains 2656 reconstructed points. Though the reconstruction model is much sparser than the one using the original SIFT descriptor and the learned 32 dimensional descriptor, the memory consumption is just (1/8)*(96/128)=9.38%, which is less than 10 percent of the original memory consumption.

In this study, we discussed the use of Structure-from-Motion reconstruction and its heavy memory consumption problem. For solving this problem, we derive the projection matrix which could best increase the distance of negative descriptor pairs and decrease the distance of positive descriptor pairs. Further, we compute the Hamming descriptor by learning a threshold which maps the real-value descriptor to Hamming space. For each kind of descriptor, we change the distance computation method for correspondence searching. Though the reconstruction model for Hamming descriptor is sparser, the memory consumption is greatly decreased. And the learned low dimensional real-value descriptor can achieve reconstruction model as dense as the original SIFT high dimensional descriptor. Based on different memory conditions and the various reconstruction tasks, both kinds of learned descriptors can have excellent applications.

9024-23, Session 5

(JEI Invited) New online signature acquisition system

Messaoud Mostefai, Adel Oulefki, Abbadi Belkacem, Samira Djebrani, Abderraouf Bouziane, Univ. of Bordj Bou Arreridj (Algeria); Youssef Chahir, Univ. de Caen Basse-Normandie (France)

Signature-based authentication systems are an attractive method for biometric authentication because they are cheaper and less constraining than fingerprint or iris based methods. In addition, signatures are a socially and legally acceptable means for personal authentication.

Several systems have been developed to perform an online signature acquisition. Most of them are based on video camera systems, or on digital tablets (tablet PCs, PDA…). Although they guarantee satisfactory results, the systems which use one or two cameras for the dynamic capture of the signature are sensitive to variations of lighting (shade produced by the movements of the hand), and cannot be easily adapted to real world systems (such as control access systems). Moreover, the systems using digital tablets seem to be more practical than those based on cameras. However, they are relatively more expensive and require more dedicated material for the exploitation of acquired data. Other types of acquisition systems based on a Data glove (conceived initially for virtual reality) have also been developed for the same purpose. The latter are powerful, but very constraining and not suitable for a general public use.

This article will show that it is possible to have a system in which the user signs with his hand, leaving no trace of the signature in order to prevent possible forgers from knowing it. Indeed, by regarding the signature as a specific dynamic hand activity and not as the result of this activity on paper or digital support, we have been able to propose a new online signature acquisition device allowing the construction of low cost and non-constraining signature authentication systems.

A primary laboratory prototype has been developed for this purpose. It is initially composed on a high resolution camera placed in front of a transparent signing glass. Signers perform signatures on the glass by moving their index. Acquired movements are used to generate the correspondent signature features (x(t), y(t), (x,y)). Developed method captures and includes the curves where the index is not on the glass.

This additional information represents a specific and non-visible dynamic behavior of the signatory.

Important modifications were made to the initial prototype in order to make the online signing process similar to the offline process.

Instead of being robust to signature imitation attacks, improved system opens the field to various non-constraining handwriting applications dedicated principally to people with motor or emotional state problems. Moreover, developed signature reconstruction techniques have low computational complexity and are therefore well suited for a hardware implementation within a dedicated smart system.

The improvements made to our initial prototype as well as its advantages will be presented in this conference paper.

9024-1, Session PTues

An attentive multi-camera system

Paolo Napoletano, Francesco Tisato, Univ. degli Studi di Milano-Bicocca (Italy)

Intelligent multi-camera systems have become very popular in the field of video surveillance [1] and ambient intelligence [2]. An intelligent multi-camera system is a kind of context aware system: it extracts salient information from multiple video stream by automatically detecting, tracking and recognizing objects of interest (for instance people), and understanding and analyzing their activities (for instance people desires, needs, emotions etc) [2]. However, multi-camera systems that integrate computer vision algorithms are not error free, and thus both false positive and negative detections need to be revised by a specialized human operator. Human supervision is not simple, especially when the number of cameras increases. In fact, traditional multi-camera systems usually include a control center with a wall of monitors displaying videos from each camera of the network. As the number of cameras increases, switching from a camera to another becomes hard for a human operator.

In this work we propose a new method that dynamically selects and displays the content of a video camera from all the available contents in the multi-camera system. The proposed method is based on a computational model of human visual attention. These models simulate the human visual system by determining the temporal sequence of salient points (focus of attention) that are focused by a human being that observes a given scene [3]. The salience of a point can be defined in several ways: color contrast of the surrounding region, motion, object presence etc. The proposed model is inspired by the work of Boccignone et al [4]. It describes how the attentive process integrates top-down and bottom-up cues: at the highest level, the attention is driven by prior information and by relevant objects, for instance human faces, within the scene; at the same time, local saliency together with novel and abrupt visual events contribute by triggering lower level attention. The resulting attentive multi-camera system works as follows: at time t the system shows the camera n because salient points (objects or activities) have been detected in the video stream of the camera n. At time t + 1 the system passes from the camera n to the camera m if other points have been detected in the video stream of the camera m that are more salient of previous points.

Several other approaches to the dynamic selection of the camera view have been proposed in literature, for instance game-theoretic, geometric, information-theoretic, probabilistic based approaches etc [1]. We believe that this is the first work that tries to use a model of human visual attention for the dynamic selection of the camera view of a multi-camera system. The proposed method has been experimented in several scenarios and has demonstrated its effectiveness with respect to the baseline methods and manually generated ground-truth. The effectiveness has been evaluated in terms of selection of the best-view while minimizing the number of view switches.

REFERENCES

[1] Wang, X., “Intelligent multi-camera video surveillance: A review,” Pattern Recognition Letters 34(1), 3 – 19 (2013).

[2] Alves Lino, J., Salem, B., and Rauterberg, M., “Responsive environments: User experiences for ambient intelligence,” J. Ambient Intell. Smart Environ. 2, 347–367 (Dec. 2010).

[3] Borji, A. and Itti, L., “State-of-the-art in visual attention modeling,” Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(1), 185–207 (2013).

[4] Boccignone, G., Marcelli, A., Napoletano, P., Di Fiore, G., Iacovoni, G., and Morsa, S., “Bayesian integration of face and low-level cues for foveated video coding,” IEEE Transactions on Circuits and Systems for Video Technology 18, 1727–1740 (ISSN: 1051-8215, 2008).

9024-24, Session PTues

Object detection in MOUT: evaluation of a hybrid approach for confirmation and rejection of object detection hypotheses

Daniel Manger, Juergen Metzler, Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung (Germany)

Military Operations in Urban Terrain (MOUT) require the capability to perceive and to analyze the situation around a patrol in order to recognize potential threats. A permanent monitoring of the surrounding area is essential in order to react appropriately to the given situation, where one relevant task is the detection of objects that can pose a threat. Especially the robust detection of persons is important, as in MOUT scenarios threats usually arise from persons. This task can be supported by image processing systems. However, depending on the scenario, person detection in MOUT can be challenging, e.g. persons are often occluded in complex outdoor scenes and the person detection also suffers from low image resolution. Furthermore, there are several requirements on person detection systems for MOUT such as the detection of non-moving persons, as they can be a part of an ambush. Existing detectors therefore have to operate on single images with low thresholds for detection in order to not miss any person. This, in turn, leads to a comparatively high number of false positives detections which renders an automatic vision-based threat detection system ineffective. In this paper, hybrid detection approaches are presented addressing these challenges. Combinations of discriminative and generative models are examined with respect to improve existing person detectors. The objective is to increase the accuracy of existing detectors by integrating a separate hypotheses confirmation and rejection step which is built by discriminative and generative models. These models enable the overall detection system to make use of both the discriminative power and the capability to detect partly hidden objects with the models. Furthermore, we examine the applicability of the different model combinations to general object detection tasks. The approaches are evaluated on benchmark data sets generated from real-world image sequences captured during MOUT exercises and our extensions show a significant improvement of the false positive detection rate.

9024-25, Session PTues

Improving the classification of texture by multifractal analysis with discrepancy, distance, and Lacunarity spectra

Mohamed Khider, Univ. des Sciences et de la Technologie Houari Boumediene (Algeria)

Although the interpretation of the results in the form of a spectrum is much easier than an image. In texture classification by multifractal analysis, the problem is losing information about the pixels position of the same singularity set. In fact, the multifractal analysis by the spectrum of large deviation or multifractal formalism or other procedures that involve multiresolution analysis by box counting or wavelets transform to estimate the local singularity exponent and calculate the fractal dimension of the spatial distribution for the sets that have the same

degree of singularity. Indeed, the multifractal spectrum gives the fractal dimension of the sets with the same Hölder exponents, but according to Mandelbrot, this dimension is not sufficient, he then proposed to associate the fractal Lacunarity.

In this work, we propose in addition to the two previous parameters : dimension and Lacunarity, the using of average Euclidean distance of separation and discrepancy and represent that in form of spectra, it was found that it can be crucial in some applications such as segmentation, pattern recognition and classification. Indeed, the classification procedures by multifractal analysis in the scientific literature exploit the multifractal spectrum mode or index of multifractality. So, we propose the use of the average distance between the positions of singularities of the same set, and also the fractal Lacunarity of these sets and the discrepancy, preliminary results taken from the KTH-TIPS2 database indicates that it is possible to improve the classification rate.

The procedure we propose is based on the calculation of singularity exponents in each position. To do this, the continuous wavelet transform is used; the Gaussian derivative is chosen for its remarkable properties, in order to estimate the singularity exponent, a linear regression adjustment is introduced, the slope of the point cloud in the bi-log plan that gives the modules of the CWT depending on scales of analysis represents the Hölder exponent. We then construct the sets of same holder exponent hx given by Sh(x) with hx ? [hmin, hmax], in the conventional multifractal analysis, we must calculate the fractal dimension for each set Shx to obtain the spectrum D(Shx), but this makes us lose the spatial information, to avoid this problem, we propose to include three other spectra in addition to the multifractal spectrum, namely the average Euclidean distance of separation between the pixels of the same set represented in a standard plane [0, 1] x [0, 1], we use also the discrepancy that characterizes the uniformity of the set, the Lacunarity spectrum is obtained from the interpolation by an exponential function of the fractal Lacunarity curve, the association of these three spectra with the multifractal spectrum has allowed us to improve the classification rate for the KTH-TIPS2 database texture.

9024-26, Session PTues

Image thresholding using standard deviation

Jung-Min Sung, Dae-Chul Kim, Bong-Yeol Choi, Yeong-Ho Ha, Kyungpook National Univ. (Korea, Republic of)

Threshold selection using the within-class variance in Otsu's method is generally moderate, yet inappropriate for expressing class statistical distributions. Otsu uses a variance to represent the dispersion of each class based on the distance square from the mean to any data. However, since the optimal threshold is biased toward the larger variance among two class variances, variances cannot be used to denote the real class statistical distributions. The dispersion of classes is represented as the distance from the mean level of a class to any intensity level and this distance is proportional to the standard deviation rather than variance. Therefore, the standard deviation is used as a more accurate expression of the statistical distribution of each class. As a result, the optimal threshold is expected to be less biased toward the large dispersion of classes than Otsu's method. Thus, to express more accurate class statistical distributions, this paper proposes the within-class standard deviation as a criterion for threshold selection, and the optimal threshold is then determined by minimizing the within-class standard deviation. Experiments were conducted using two normal distributions and two shadowed images, and the performance of the proposed method evaluated based on the misclassification error(ME). Experimental results confirm that the proposed method produced a better performance than existing algorithms.

9024-27, Session PTues

Soft brush touch image sensing based on planar optical waveguide sensor system

Jeong Dae Suh, Joo-Hang Lee, Ahyun Lee, Electronics and Telecommunications Research Institute (Korea, Republic of)

There is a need for soft object detection such as painting brush for digital canvas development. We have explored soft brush touch image sensing system that employ planar optical wave guide sensors. Recently, optical touching imaging based on the planar wave guide devices based on frustrated total internal reflection (FTIR) have been considered as one of the promising techniques for touch interactive device applications; multi touch screens for electronics, digital arts and human computer interactions in various fields. Up to date hard touch image sensing based on finger of hand is usually done, but they don't convey soft object image detections. In this study, we present the results for soft brush touch image detections by using planar waveguide sensor structures. Soft brush touch image sensor system consisted with a transparent planar waveguide plate, infrared light emitting diodes, high resolution optical camera and computers for image acquisition and processing. When the waveguide surface touched by smooth brush, we clearly obtained the bright optical blob patterns on the screen that indicate the soft brush touch imaging. For the soft brush touch imaging surface, the parameters of planar waveguide sensor structures were analyzed to optimize the waveguide sensor reactivity.

9024-28, Session PTues

Hyperspectral imaging applied to process and quality control in waste recycling plants

Silvia Serranti, Giuseppe Bonifazi, Univ. degli Studi di Roma La Sapienza (Italy)

In secondary raw materials and recycling sectors, the products quality represents, more and more, the key issue to pursuit in order to be competitive in a more and more demanding market, where quality standards and products certification play a preeminent role. These goals assume particular importance when recycling actions are applied. Recovered products, resulting from waste materials, and/or dismissed products processing, are, in fact, always seen with a certain suspect.

The possibility to utilize HyperSpectral Imaging (HSI) techniques for waste materials inspection in industrial recycling plants opens new interesting scenarios for the development of on-line process control and/or product quality control strategies. Despite the advantages of HSI, the technique is still difficult to be systematically applied, especially in real-time industrial applications, because of the huge amount of data constituting a spectral image. In this paper, different applications of HSI techniques, with reference to the solid waste recycling are presented, and critically analyzed.

The problems arising when suitable HSI based procedures have to be developed and implemented in solid waste products characterization, in order to define time efficient compression and interpretation techniques, are analyzed and discussed in this paper. Particular attention was also addressed to define an integrated (HW&SW) platform able to perform a non-intrusive, non-contact and real-time analysis and embedding a core of analytical logics and procedures to utilize both at laboratory and industrial scale.

Investigations have been carried out utilizing different HSI devices from SPECIM Ltd (Finland) working in different wavelength ranges: i) ImSpector™ V10E acting in the range 400-1000 nm, ii) NIR Spectral Camera™, embedding an ImSpector™ N17E working in the range 1000-1700 nm, iii) SisuCHEMA XL? Chemical Imaging workstation, a complete and high speed HSI system operating in the SWIR region (1000-2500 nm).

Spectral data analysis was carried out utilizing the PLS_Toolbox (Version 6.5.1, Eigenvector Research, Inc.) running inside Matlab® (Version 7.11.1, The Mathworks, Inc.), applying different chemometric techniques, depending on the materials under investigation.

An adequate response of the industry to the market can only be given through the utilization of equipment and procedures ensuring pure, high-quality production, and efficient work and cost. All these goals can be reached adopting not only more efficient equipment and layouts, but also introducing new processing logics able to realize a full control of the handled material flow streams fulfilling, at the same time, i) an easy management of the procedures, ii) an efficient use of the energy, iii) the definition and set up of reliable and robust procedures, iv) the possibility to implement network connectivity capabilities finalized to a remote monitoring and control of the processes and v) a full data storage, analysis and retrieving.

9024-29, Session PTues

Eye Gaze Tracking using Correlation Filters

Mahmut Karakaya, David S. Bolme, Christopher B. Boehnen, Oak Ridge National Lab. (United States)

In this paper, we study a gaze estimation method based on the distances between top point of the eyelid and eye corner detected by the correlation filters. There is extensive literature related to the video-based gaze tracking applications, ranging from human-computer interaction to automotive industry. In these applications, intrusive and non-intrusive techniques are used depending upon the required level of accuracy.

A gaze estimate with a deviation of ±0.5 degrees or lower accuracy is achieved by using additional intrusive equipment (e.g., reflective dots, electrodes and head-mounted devices) which are physically attached to the user. Compared to intrusive methods, non-intrusive techniques are more comfortable, more practical, less risky, and less accurate. These techniques are classified into four main categories: corneal reflection based methods, mapping functions based methods, model based methods, and appearance based methods. To get an accurate gaze estimation with these non-intrusive methods, a set of parameters (such as camera intrinsic parameters, locations and orientations of cameras, lights and monitors, subject's cornea curvature, angular offset between optical and visual axis, etc.) needs to be determined at the initial setup by conducting well-controlled calibration and training phases.

In some eye tracking scenarios, it is not provided additional information or equipment such as well-controlled light sources, mounted intrusive devices, or multiple cameras to determine the set of parameters required for state of the art gaze estimation methods. The only available information to estimate the gaze is the iris image. Therefore, we focus on the distance between the eye corner and the top point of the eyelid to estimate the eye gaze.

In order to estimate the gaze, the distance metrics are computed between eye corners and eyelid points by subtracting their positions. When users change their gaze from one point to another on the screen, the distance metrics change accordingly. For example, the distance metrics for the x-axis changes as user changes their gaze horizontally and the distance metrics for the y-axis changes based on the vertical gaze change. To find the correspondence between distance metrics and gaze, we compute a transfer function between distance metrics and screen coordinates. In this purpose, we use second order polynomial function to generate the mapping.

Since the accuracy of the gaze estimation is based on the accuracy of the correlation filter, we need to capture good quality images. By attaching a 5X telephoto lens in front of the camera on the tablet computer, we get a higher image resolution and fewer background artifacts such as the white balance problem, since the field of view of the camera becomes smaller, less amount of background region appears on the captured image.

In order to test our proposed method, we captured video frames of three subjects while they change their gaze by looking at different images shown on the screen after the calibration procedure. Correlation filters detects the eyelid and eye corner positions for each frame and a mapping function generated in the calibration is used to transfer the distance metrics to screen coordinates. In order to visualize the gaze points on the screen, a heatmap is generated for each image slides shown on the screen for testing. We observed that gaze points are densely collected around the user gaze positions on the image slides and are consistent where the user mentioned.

9024-30, Session PTues

An efficient automatic object recognition method based on region similarity: application to roof detection from orthophotoplans

Abdellatif Elidrissi, Univ. de Technologie de Belfort-Montbéliard (France) and Univ. Abdelmalek Essadi (Morocco); Youssef Elmerabet, Univ. de Technologie de Belfort-Montbéliard (France) and Univ. Ibn Tofail (Morocco); Yassine Ruichek, Univ. de Technologie de Belfort-Montbéliard (France); Ahmed Moussa, Univ. Abdelmalek Essadi (Morocco); Cyril Meurie, Univ. de Technologie de Belfort-Montbéliard (France)

We present an efficient method for automatic and accurate multiple detection of objects of interest from images using a region similarity measure. The method we propose requires the construction of two knowledge databases.

The first one contains several significant textures of the objects to be extracted. The second database is composed with textures representing the background. These two databases are provided by some examples of images. The proposed method involves then two main steps. The first one consists in segmenting the image into homogeneous regions. In order to separate the objects of interest and the image background, an evaluation of the similarity between the regions of the segmented image and those of the constructed knowledge databases is then performed.

The proposed approach presents several advantages in terms of applicability, suitability and simplicity. To show its performance, the method is applied to extract building roofs from orthophotoplans.

9024-31, Session PTues

An uniformity algorithm for high-speed fixed-array printers

Kartheek Chandu, Mikel J. Stanich, Larry M. Ernst, Ricoh Production Print Solutions, LLC (United States); Sreenath Rao Vantaram, Intel Corp. (United States)

Current high speed inkjet printers commonly employ fixed-array printheads. The fixed-array system is constructed from multiple printhead modules stitched together edge-to-edge, each comprising several hundred nozzles from which ink is ejected onto the media/paper. The final array spans the entire width of the web of paper. The array remains stationary while the media moves across the heads, as they selectively jet ink, creating a process capable of printing at web speeds of thousands of feet per minute. After a thorough literature search, it is our view that most of the automatic print inspection and calibration systems have been developed in-house by manufactures or third party vendors for specific applications and not described in published scientific literature. Many automatic print defect detection systems using in-line scanners have been proposed, however none of these methods compensate/calibrate for the identified defects, rather they only report the identified faults back to the system/operator. An automatic system, such as the one we propose, which utilizes an in-line scanner to identify print artifacts and perform calibration for a high-speed fixed printhead array inkjet printer was not found during the search. The system we are

proposing compensates for the identified defects and also calibrates the tonal response. Calibration is employed to produce distinct gray levels, matched to a defined target, while compensation addresses print irregularities. Calibration accounts for dot gain to avoid saturated printing in the shadow tone range. In this massively parallel, partially redundant ink jet array system, a variety of artifact producing mechanisms are present. One mechanism is created by redundant nozzles located at the ends of each printhead module. Printing of the same data by the redundant nozzles may produce artifacts due to the physical overlaps. The combined jetting from the redundant nozzles must be accounted for in the compensation process to avoid artifacts. Further artifact producing mechanisms are created because each printhead module and its nozzles are unique. Macroscopic variations between printheads and localized variations of jetting from every nozzle may create artifacts. Fundamentally the ink coverage over an area drives the Optical Density globally and locally. It is our objective to make the average ejected ink coverage constant, so as to provide uniform density across the entire width of the printed web. In this paper we describe the development of an automated calibration procedure for compensating non-uniformities to a degree where the printed output is perceived as uniform. The proposed calibration process uses a “Golden” functional form target for the final tonal characteristics. Ultimately a calibrated halftone threshold array is generated for each color that includes compensation at the nozzle level. An initial multi-bit stochastic halftone mask is modified to account for the non-uniformities and gray scale calibration. The process considers the Point Spread Function (PSF) of the printer/inline scanner “system”. The proposed algorithm initially pre-processes scanned image of printed output to isolate the desired regions of interest (ROI). The ROI information is utilized to extract non-uniformities across the entire printed area for each tint level of the calibration chart. The algorithm concludes in a calibration step that enables compensation of the identified non-uniformities and provides the desired tonal response that matches the Golden target.

Conference 2025: Intelligent Robots and Computer Vision XXXI: Algorithms and Techniques

Wednesday - Thursday 5 –6 February 2014

Part of Proceedings of SPIE Vol. 9025 Intelligent Robots and Computer Vision XXXI: Algorithms and Techniques

9025-31, Session PWed

OpenCLIPP: OpenCL integrated performance primitives library for computer vision applications

Moulay A. Akhloufi, Antoine Campagna, Ctr of Robotics and Vision (Canada)

In recent years, we see an increase of interest for GPGPU computing (General-Purpose computation on Graphics Processing Units). This domain aim to using the processing power of the GPU (Graphics Processing Units) in order to accelerate general processing like mathematics, 3D visualization, image processing, etc.

In the past years, CUDA (Compute Unified Device Architecture) a parallel computing platform and programming model invented by NVIDIA was the main driver of this interest and the most used architecture for GPGPU computing. With the recent advent of Open Computing Language (OpenCL), we see more and more work conducted using this new platform. OpenCL is an open standard maintained by the non-profit technology consortium Khronos Group. It has been adopted by multiple companies including NVIDIA (the inventor of CUDA).

With this increase of interest, the availability of a set of performance primitives for general purpose applications can help accelerate the work of the research and industrial communities. Intel, for example, develops Intel Integrated Performance Primitives (Intel IPP), a multi-threaded software library of functions for multimedia and data processing applications. In the other hand, NVIDIA offers the NVIDIA Performance Primitives library (NPP), a collection of GPU-accelerated image, video, and signal processing functions that deliver faster performance than comparable CPU-only implementations.

In this work, we present the architecture and development of an open source OpenCL integrated performance primitives library called OpenCLIPP. This library aim to provide a free and open source set of OpenCL functions with a simple interface similar to Intel IPP and NVIDIA NPP. The first release includes mainly image processing and computer vision algorithms: Convolution filters, Thresholding, Blobs, etc. The developed functions are introduced and benchmarks with equivalent Intel IPP and NVIDIA NPP functions are presented. This library will be made available to the open source community.

9025-32, Session PWed

An intelligent hybrid behavior coordination system for an autonomous mobile robot

Chaomin Luo, Mohan Krishnan, Mark Paulik, Samer Fallouh, Univ. of Detroit Mercy (United States)

Wall-following, obstacle avoidance and navigation are required for autonomous mobile robots and many other robotic applications. The capability to perform wall-following, obstacle avoidance and navigation is a critical competence for successful search and exploration in real-world environments of autonomous mobile robots in cleaning, transportation, medical, and rescue robotics applications.

By employing PID controller associated behavior coordination system, the design of complex and nonlinear control systems without having a model of the system is able to be simplified. In this paper, the development of a low-cost PID controller with an intelligent behavior coordination system for an autonomous mobile robot is described that is equipped with IR sensors, ultrasonic sensors, regulator, and RC filters on the robot platform based on HCS12 microcontroller and embedded systems.

An efficient hybrid PID controller and behavior coordination system is developed for wall-following, obstacle avoidance and navigation of an autonomous mobile robot. Adaptive control used in this robot is a novel hybrid PID algorithm associated with template and behavior coordination models. Software development contains motor control, behavior coordination intelligent system and sensor fusion in this paper. In addition, the module-based programming technique is adopted to improve the efficiency of integrating the hybrid PID and template as well as behavior coordination model algorithms.

The hybrid model is developed to synthesize PID control algorithms, fault tolerance schemes, template and behavior coordination technique for wall-following, obstacle avoidance, and navigation systems. Dynamic behavior performance of the obtained hybrid model is validated by experiments on the self-developed actual robot under real world environments. A bubble technique with virtual obstacle or virtual robot is developed to ensure safer and more reasonable wall-following, obstacle avoidance and navigation performance.

The hardware design is composed of motor control with encoders, wheel and chassis installation, a sensor interface, and sensor configuration. An H-Bridge with pulse width modulation (PWM) technique is utilized to control the speed of the motors. The software development is focused on wall-following, obstacle avoidance, navigation, motor control, and sensor fusion. These applications of the embedded system usually include the reading of sensor data which can be used to interpret real world conditions. Processing this data requires an embedded microprocessor able to drive other circuits to respond or communicate this data to the real world in this paper. The system utilizes this data to respond to the environment through the use of various actuators.

The motor control, obstacle avoidance, wall-following and navigation algorithms are developed to propel and steer the autonomous mobile robot. Environmental information of the real-world collected by IR and ultrasonic sensors is transmitted to the PID controller and behavior coordination system.

Experimental results validate how this novel PID controller and behavior coordination system directs an autonomous mobile robot to perform wall-following and point-to-point search with obstacle avoidance. Hardware configuration and module-based technique are described in this paper. Experimental results demonstrate that the robot is successfully capable of being guided by the hybrid PID controller and behavior coordination system for wall-following and navigation with obstacle avoidance.

9025-33, Session PWed

Increasing signal-to-noise ratio of registered images by using light spatial noise portrait of camera's photosensor

Nikolay N. Evtikhiev, Pavel A. Cheremkhin, Vitaly V. Krasnov, Vladislav G. Rodin, Sergey N. Starikov, National Research Nuclear Univ. MEPhI (Russian Federation)

Increase of signal-to-noise ratio (SNR) of registered images is important in such fields as image encryption, digital holography, pattern recognition and etc. The method of image SNR increasing by using light spatial noise portrait (LSNP) of camera's photosensor is presented.

LSNP is array of photosensor pixels photo response non-uniformities. It is used for camera identification, determination of images origin and any post-processing done. We propose method of image SNR increasing by spatial noise suppression using compensation of photosensor LSNP. Usually spatial noise is about 0.5 % of camera signal value that is 2-4 times less than temporal noise. So use of the proposed method

is effective after application of others methods of SNR increasing that suppress temporal noise. We investigated application of the LSNP compensation method in conjunction with the most widely utilized averaging over frames method.

Proposed method requires knowledge of used photosensor LSNP. The procedure of LSNP measurement is described. Using the described procedure LSNP of photosensor of camera Canon EOS 400D was obtained. Relative error of LSNP measurement is estimated as 15-20 %.

Analytical expressions for estimation of achievable increase of image SNR using the LSNP compensation method were derived for two cases: application of LSNP compensation method only and LSNP compensation method in conjunction with the averaging over frames method. Using characteristics of obtained LSNP, numerical experiments on estimation of SNR increase for camera Canon EOS 400D were performed. Solo use of the averaging over frames method allows to increase SNR up to 2 times. Application of the LSNP compensation method in conjunction with the averaging over frames method allows 10 times SNR increase. These numerical experiments results confirm derived analytical expressions. It is shown that in the case of using more accurate LSNP compared to the obtained, SNR can be increased up to 50 times.

Using cameras Canon EOS 400D and MegaPlus II ES11000, test experiments were performed. For the LSNP compensation method in conjunction with the averaging over frames method experimental results on both temporal and spatial noise suppression are in good agreement with the numerical ones. The mean difference between experimental and calculated values of SNR increasing is only 3.4 %. Experimentally obtained SNR increase is equal to 5 times.

9025-34, Session PWed

Color back projection for fruit maturity evaluation

Dong Zhang, Sun Yat-Sen Univ. (China); Dah Jye Lee, Alok Desai, Brigham Young Univ. (United States)

This paper presents an efficient color back projection and image processing technique that is designed specifically for real-time maturity evaluation of Medjool dates. This color processing method requires very simple training procedure to obtain the frequencies of colors that appear in each of four different maturity stages. This color statistics is used to back project colors to predefined color indexes. Fruit maturity is then evaluated by analyzing the projected color indexes. This method has been implemented and used for commercial production and proven to be efficient and accurate.

9025-35, Session PWed

Unmanned ground vehicle: controls and dynamics

Ebrahim F. Attarwala, Pranav Maheshwari, Kumar Keshav, Pranjal Jain, Arpit Gupta, Kriti Gupta, Ravi Yadav, Indian Institute of Technology Bombay (India)

The field of autonomous unmanned vehicles has caught everyone's attention recently not only as an important research topic but also because of its immense applications. Vehicles that can move autonomously avoiding hurdles in a disciplined way becoming a part of everyday traffic on all types of terrains will soon become a reality. Along with the advantages of automation, integrating the vehicle with various intelligent features such as GPS navigation, environment friendly power source, etc will be the solution to many global concerns such as fuel consumption, pollution, vehicle tracking and road safety. Not only public transport, but autonomous vehicles have military applications like unmanned rescue operations and autonomous inspection vehicle

We have developed a ground vehicle capable of maneuvering in an open environment negotiating outdoor obstacle course autonomously, carrying a payload by finding colored lanes, obstructions and navigating through GPS waypoints.

The machine's perception of surroundings and its pose is mainly based on the use of Laser range finder (LIDAR), inertial sensors (IMU), Camera, GPS and Encoders.

MACHINE VISION:

In real time Image Processing, effect of sunlight is very dominant. So we have used Discrete Cosine Transform (DCT) for removing the noise in the image. By dividing input frame into many blocks, we have taken the DCT values corresponding to each block. Now if for any block its DCT value is greater than the average value of the image, then that block of image is considered as noise and we scale down its corresponding DCT value to eliminate the effect of sunlight. We then used color thresholding followed by Hough Transform to explore the possibility of connecting line.

MAPPING:

We did an extensive literature research and came with an algorithm which makes use of LIDAR and camera for localization and obstacle detection. LIDAR detects obstacles and Camera detects lanes. The algorithm builds upon by segmenting space around the vehicle into angular sectors of pre-defined resolution. It then uses distance data from LIDAR and camera, and creates clusters of consecutive sectors which are free of obstacles and white lanes and computes mean angle for each cluster.

NAVIGATION:

The angle nearest to the current GPS heading is chosen. The algorithm also alternates between two different navigation modes. When the area in front of the robot is unobstructed, the robot navigates straight towards the goal and when the path is obstructed, the robot follows the contours of the obstacles until the way is clear.

MECHANICAL DESIGN:

The purpose of our mechanical design is to improve the robot's efficiency and dynamic ability to navigate. It has a front castor and 2 driving rear wheels connected to motors via universal joints. This significantly reduces play and direct load on motor shaft which improves motor performance and shaft life. Machine has an indigenously developed suspension system designed to travel smoothly and distribute equal load to wheels. We analyzed distortion capability of the suspensions which made our machine adaptive to any terrain and eliminated spikes in the sensor data caused by vibrations in the robot.

9025-36, Session PWed

A super-fast algorithm for self grouping in a 2D binary picture

Chialun John Hu, SunnyFuture (United States)

Novel real-time image processing, super fast pattern extraction in 2D picture space, local polar edge detection (as reported by this author in the last 4 years.)

9025-1, Session 1

Adaptation of human routines to support a robot's tasks planning and scheduling

A. Tikanmäki, S. Troyano Feliu, J. Röning, Univ. of Oulu (Finland)

Home robots usually share their workspace with people. Therefore, there exists the need to take into account the presence of humans when planning their actions and it is indispensable to have knowledge of robots' environments. It means knowing when (time and events duration) and where (workspace) robot's tasks can be performed. This research paper deals with the obtaining of the spatial information required to execute a software to plan tasks to be performed by a robot. With this aim, a program capable to define meaningful areas or zones in the

robot workspace by the use of a clustering is created. This software is tested using real data obtained from different cameras located along the corridors of Department of Computer Science and Engineering at the University of Oulu.

9025-2, Session 1

A novel lidar-driven two-level approach for real-time unmanned ground vehicle navigation and map building

Chaomin Luo, Mohan Krishnan, Mark Paulik, Bo Cui, Xingzhong Zhang, Univ. of Detroit Mercy (United States)

The basic navigation problem for unmanned ground vehicle is concerned with finding a safe and good-quality collision-free path from an initial point to a destination while a local map is constructed. In this paper, a LIDAR-driven two-level hybrid real-time navigation approach for unmanned ground vehicles is proposed. Top level is newly designed enhanced Voronoi Diagram (EVD) method to generate a global trajectory for an unmanned vehicle. Bottom level employs Vector Field Histogram (VFH+) algorithm based on the LIDAR sensor information to guide the vehicle locally to autonomously traverse from one node to another within the planned enhanced Voronoi Diagram with obstacle avoidance. In order to generate safer, more reasonable collision-free trajectories, novel heuristic algorithms are developed to optimize the path among the potential trajectories planned within the Voronoi Diagram in search of angles and distance. The enhanced Voronoi diagram is utilized for modelling the environment populated with various obstacles. The path planning is based on EDV, where obstacles in the environment are considered to generate points of the diagram, and a heuristic algorithm is developed to search a path without collisions from the robot initial to target position.

In this paper, the two level hybrid LIDAR-driven enhanced Voronoi Diagrams and VFH+ algorithm is effectively integrated to enable a vehicle to plan and navigate a path successfully under complicated circumstances while a local map is dynamically built up. The unmanned ground vehicle is equipped with a camera, digital compass, LADAR, and GPS. The enhanced Voronoi Diagram is built up by utilizing processed data from these sensors. A local map composed of square cells is created with the VFH+ algorithm with necessary sensor data during the movement of the unmanned ground vehicle with limited sensory information. From the measured sensory information, a map of the robot's immediate limited surroundings is dynamically built up for the vehicle navigation.

Additionally, graph-based merging techniques are also proposed to categorize and integrate various obstacles so as to construct effective enhanced Voronoi Diagrams for path planning purpose. Under the best of circumstances, this method works reasonably well. The algorithm, software design, and hardware configuration of the unmanned ground vehicle are described this paper.

The path planning model has been successfully demonstrated in a Player/Stage simulation environment. With the enhancement of the Voronoi Diagrams by using the "goal-search", a safe, short, and reasonable trajectory is successfully planned in a majority of situations. Its simplicity, versatility, effectiveness, and efficiency of real-time navigation and map building for unmanned ground vehicles have been successfully validated by simulation, comparison studies and experiments. The proposed approach is successfully experimented on an actual unmanned ground vehicle to demonstrate the real-time navigation and map building performance of the proposed method, both in static and moderately dynamic environments. The vehicle appears to follow a very stable path while navigating through various obstacles. Comparison studies of the proposed approach with the other path planning approaches demonstrate that the proposed method is capable of planning more reasonable and shorter collision-free trajectories autonomously.

9025-3, Session 1

The 21st annual intelligent ground vehicle competition: robotists for the future

Bernard L. Theisen, U.S. Army Tank Automotive Research, Development and Engineering Ctr. (United States)

The Intelligent Ground Vehicle Competition (IGVC) is one of four, unmanned systems, student competitions that were founded by the Association for Unmanned Vehicle Systems International (AUVSI). The IGVC is a multidisciplinary exercise in product realization that challenges college engineering student teams to integrate advanced control theory, machine vision, vehicular electronics and mobile platform fundamentals to design and build an unmanned system. Teams from around the world focus on developing a suite of dual-use technologies to equip ground vehicles of the future with intelligent driving capabilities. Over the past 21 years, the competition has challenged undergraduate, graduate and Ph.D. students with real world applications in intelligent transportation systems, the military and manufacturing automation. To date, teams from over 80 universities and colleges have participated. This paper describes some of the applications of the technologies required by this competition and discusses the educational benefits. The primary goal of the IGVC is to advance engineering education in intelligent vehicles and related technologies. The employment and professional networking opportunities created for students and industrial sponsors through a series of technical events over the four-day competition are highlighted. Finally, an assessment of the competition based on participation is presented.

9025-4, Session 1

Surveillance and detection technology research based on panoramic depth vision system for public emergencies on-site inspection

Weijia Feng, Tianjin Univ. (China); Juha Röning, Univ. of Oulu (Finland); Yi Tian, Huazhi Sun, Xirong Ma, Tianjin Normal Univ. (China)

Figure 1 is a diagram of the Panoramic Depth Vision (PDV) system which is composed of a novel combined fish-eye lenses module and a specially designed embedded image processor. The combined fish-eye lenses module is organized by four 185??185? FOV fish-eye lenses which will be mounted on a same horizontal plane with 90 degrees interval between each other. In this way, two adjacent fish-eye lenses can overlap all of the 90??185? FOV information. So, using four fish-eye lenses with horizontal interval of 90? can realize 360??360? FOV stereo information. Embedded technology is applying to build the special image processor, like a camera in the central position of PDV system shown in Figure1. This design can insure the body of camera will be not appeared in the captured images. Through this design, we can get no blind area 360??360? panoramic image.

In this paper, PDV system is mounted on mobile platform to complete the mission of real-time recording and on-line analysis of Public Emergencies On-site Inspection. Here, PDV has been used to replace the traditional vision module of Mobile Mapping System (MMS). To achieve the function of Public Emergencies On-site Inspection, the theoretical model of PDV is developed firstly, The methods of panoramic image generation and mapping with the depth image are achieved based on a specialized feature points matching algorithm; An unique image detection method explores a new kind of texture feature descriptor to extract the features from the panoramic image and the depth image, and judge the attribute of the Public Emergencies by texture segmentation and classification; A feedback mechanism is established with the geographic information with GIS and GPS to transfer the on-site inspection results to the control and decision center. The processes of Public Emergencies On-site Inspection based on PDV system is shown in Figure2:

This paper is focus on discussing the surveillance and detection methods for Public Emergencies On-site Inspection. The remainder of this paper is organized as follows: First, in Section 2, we introduce a special Mobile Mapping System organized by PDV. Thereafter, an improvement Local Binary Pattern algorithm for extracting features form the panoramic image and the depth image which created by PDV system is explored. Following that, the procedure for segmenting and classifying the information of public emergencies is described. Finally, the experimental results and conclusions are presented in Section 5and Section6, respectively.

9025-5, Session 1

Self-localization for an autonomous mobile robot based on an omni-directional vision system

Shu-Yin Chiang, Ming Chuan Univ. (Taiwan); Kuang-Yu Lin, Tsorng-Lin Chia, Ming Chuan University (Taiwan)

In this study, we designed an autonomous mobile robot based on the rules of the Federation of International Robot-soccer Association (FIRA) RoboSot category, integrating the techniques of computer vision, real-time image processing, dynamic target tracking, wireless communication, self-localization, motion control, path planning, and control strategy to achieve the contest goal. The self-localization scheme of the mobile robot is based on the algorithms featured in the images from its omni-directional vision system. In previous works, we used the image colors of the field goals as reference points, combining either dual-circle or trilateration positioning of the reference points to achieve self-localization of the autonomous mobile robot. However, because the image of the game field is easily affected by ambient light, positioning systems exclusively based on color model algorithms cause errors. To reduce environmental effects and achieve the self-localization of the robot, the proposed algorithm is applied in assessing the corners of field lines by using an omni-directional vision system. Particularly in the mid-size league of the RobotCup soccer competition, self-localization algorithms based on extracting white lines from the soccer field have become increasingly popular. Moreover, white lines are less influenced by light than are the color model of the goals.

Therefore, we propose an algorithm that transforms the omni-directional image into an unwrapped transformed image, enhancing the extraction features. The process is described as follows:

First, radical scan-lines were used to process omni-directional images, reducing the computational load and improving system efficiency. The lines were radically arranged around the center of the omni-directional camera image, resulting in a shorter computational time compared with the traditional Cartesian coordinate system.

However, the omni-directional image is a distorted image, which makes it difficult to recognize the position of the robot. Therefore, image transformation was required to implement self-localization.

Second, we used an approach to transform the omni-directional images into panoramic images. Hence, the distortion of the white line can be fixed through the transformation.

The interest points that form the corners of the landmark were then located using the features from accelerated segment test (FAST) algorithm. In this algorithm, a circle of sixteen pixels surrounding the corner candidate is considered and is a high-speed feature detector in real-time frame rate applications.

Finally, the dual-circle, trilateration, and cross-ratio projection algorithms were implemented in choosing the corners obtained from the FAST algorithm and localizing the position of the robot. The results demonstrate that the proposed algorithm is accurate, exhibiting a 2-cm position error in a soccer field measuring 600 cm2 x 400 cm2.

9025-6, Session 2

High-speed object matching and localization using gradient orientation features

Xinyu Xu, Peter van Beek, Xiao-Fan Feng, Sharp Labs. of America, Inc. (United States)

Motivation and problem statement

Robots are nowadays used more and more widely in manufacturing inspection applications for improved inspection throughput and quality, and reduced labor cost. In these applications, robots are often needed to perform object detection and pose estimation: given a target object in a model image, find the position and orientation of the target object in input images automatically. This is a challenging task because (1) the target object in input images often undergo geometric transforms (rotation, translation, etc) and the input image may have photometric changes (brightness, blur, noise, etc) with respect to the given model image, (2) robot has to finish the task very quickly, and (3) the estimated pose should be highly accurate.

Existing methods include template matching and feature point-based matching. Standard template matching methods employ exhaustive position and orientation search and are usually slow. Feature point-based methods fail for objects that lack sufficient feature points, which occur often for industrial objects common in inspection applications.

Methods

In this paper, we proposed a new template matching method that improves efficiency over existing approaches by decomposing orientation and position estimation into two cascade steps. In the coarse search step, an initial position and orientation is found by matching with Histogram of Oriented Gradients (HOG), reducing orientation search from with 2D templates to 1D correlation matching. In the next middle search step, a more precise orientation and position is computed by matching based on Dominant Orientation Template (DOT), using robust edge orientation features. The cascade combination of the HOG and DOT feature for high-speed and robust object matching is the key novelty of the proposed method. Depending on the precision requirement, the orientation and position obtained by the middle search stage can be further refined by a fine search stage. In addition to the principle method, we proposed a novel strategy to resolve the rotation angle ambiguity estimated by the HOG matching. What's more, we proposed an optimized solution for computing the HOG with O(1) time complexity which significantly improves the speed of coarse search stage.

Experiment Results

We evaluated the method with real-world single-object and multi-object product inspection datasets using software implemented on an Atom CPU platform. We compared the performance of this new method with an already accelerated template matching method on the same platform. Our results show that the proposed method achieves significant speed improvement at comparable accuracy performance. And the proposed method achieves high robustness to photometric changes (brightness, blur, noise) and geometric changes (rotation, translation and small scale change). As can be seen from Table 1, the proposed method is 2.37 times, 4.92 times and 6.27 times faster than the baseline template matching method for spring, transistor and screw objects in the coarse search stage.

Conclusion

In this paper, we proposed a new template matching method that improves efficiency over existing approaches by decomposing orientation and position estimation into two cascade steps. Our results show that the proposed method achieves significant speed improvement at comparable accuracy performance. And the proposed method achieves high robustness to photometric changes (contrast, background illumination, blur and noise) and geometric changes (rotation, translation and small scale change).

9025-7, Session 2

Automatic lip reading by using multimodal visual features

Shohei Takahashi, Jun Ohya, Waseda Univ. (Japan)

This paper deals with a vision based lip-reading method that can be used in noisy environments and is useful for people with hearing disabilities. Previous lip-reading methods used only single feature such as the shape of lips, optical flaw or another feature. Therefore, the recognition accuracy was not high enough.

Active Shape models (ASM) are statistical model of the shape of objects to fit and track the object in a temporal image sequence. Sets of points aligned in the objects are used to train the distribution of the objects shape.

Suppose we have s sets of points x_i aligned in the shape of the objects. Principal Component Analysis (PCA) is applied to the s sets of points x_i?to reduce the dimension. Then, ASM can approximate the statistical shape model of the objects using Eigenvectors and parameters obtained by PCA: that is, ASM can represent a new shape in a low dimension. ASM can apply rotations, scales and translations to fit and track the object in a new image.

To track and detect the object, statistical model of gray-level values in training set is used. ASM sample k pixels along the profile normal to the boundary of model from either side of the model in the every training image. The derivatives are also used to normalize the samples to reduce the effect of global intensity changes. Then, the mean and covariance values of the grey-level pixels are calculated. As the cost function to fit the objects, the Mahalanobis distance between the new object in the new image and the object in the training image set is used. During searching a new object, ASM changes the parameter values of shape, rotation, scale and translation.

After tracking and detecting the face and lip in video frames, we extract the features for machine learning. We use the height and width of the lips, optical flow of the points of ASM model around the lip and spatial frequencies of the lip. Those features vary, depending on individuals and geometrical relationships between the camera and lip. To solve this problem, we normalize the height, width and optical flow of the lips by using the initial state. The lip' spatial frequencies are also normalized by scaling to defined image size.

As a machine learning, support vector machine is applied to the multimodal features obtained from video sequences that capture lip motions. Late fusion is used for the first SVM. Each multimodal feature chronologically ordered is classified by the SVM trained by only each feature. The results of the first SVMs are features for the final SVM for the lip motion on uttering words. To train the final SVM, the fact that positive samples are far less than the negative samples is also considered.

Experiments for recognizing lip motions on uttering numerals and words are performed. As a result, most of the uttered numerals and words are classified by multimodal features more accurately than by single feature. The experimental results show the effectiveness of the proposed method.

9025-8, Session 2

A Viola-Jones based hybrid face detection framework

Thomas M. Murphy, Randy Broussard, Robert C. Schultz, Ryan Rakvic, Hau Ngo, U.S. Naval Academy (United States)

Face detection -- the localization and size determination of (principally human) faces in digital images, often as a component of a broader facial recognition system -- is a mature technology, yet its operational performance, even in less difficult frontal face tests, is generally sub-optimal. An improvement would be beneficial in many applications. The OpenCV library is widely used in the image processing community. It implements a standard face detection solution, commonly known as the Viola-Jones detector (Viola and Jones, 2004). Here, the AdaBoost learning algorithm (Freund and Schapire, 1995) is employed to construct a statistically boosted rejection cascade of binary classifiers (face vs. no face) using Haar-like input features with high detection rates and low rejection rates at each stage. Despite its overall superior performance and desirable properties, empirical evidence has shown that the Viola-Jones methodology underdetects in some instances. The goal of maximizing its already high detection rates while preserving or decreasing its already low false alarm rates and exceptional execution times motivated investigations of manipulating, truncating, and merging techniques. This research shows that a number of true faces survive many stages of the rejection cascade, but not the final stages, resulting in misses. A method to retain these missed faces was explored. To this end, a hybrid framework was constructed, with a truncated Viola-Jones cascade followed by a complex classifier -- an artificial neural network -- used to augment and fine tune the face decision in a particular rectangular region-of-interest. Optimally, a truncation stage that captured all faces and still allowed the neural network to remove the false alarms was selected. A feedforward backpropagation network with one hidden layer was utilized -- standard, yet the solution of choice for many machine learning applications. It was trained to discriminate faces based upon a few simple features, namely the thresholding (detection) values of a subset of intermediate stages of the full rejection cascade. A clustering algorithm was used as a precursor to the neural network, to group possible detections that overlapped significantly. For these grouped detections, information was combined into region-of-interest rectangles by statistical methods such as averaging, minimizing, maximizing, or by selecting the best (largest) detection values. Experiments revealed improved performance, with the neural network able to discover feature dependencies and handle the high-dimensional input data with low storage requirements. The framework was evaluated on the CMU/VASC Image Database (vasc.ri.cmu.edu/idb), a widely used benchmark that dates back to some of the earliest high performing face detection systems (Rowley et al., 1998; Klette et al., 2001). Using this data, a comparison of the hybrid methodology implementation -- Viola-Jones truncated to an intermediate stage plus neural network -- with an unmodified OpenCV approach showed: (1) a 37% increase in detection rates if constrained by the requirement of no increase in false alarms, (2) a 48% increase in detection rates if some additional false alarms were tolerated, and (3) an 82% reduction in false alarms with no reduction in detection rates. These results demonstrate improved face detection and could address the need for such improvement in various applications.

9025-9, Session 2

Towards automatic identification of mismatched image pairs through loop constraints

Armagan Elibol, KAIST (Korea, Republic of) and Y?ld?z Teknik Üniv. (Turkey); Jinwhan Kim, KAIST (Korea, Republic of); Nuno Gracias, Rafael Garcia, Univ. de Girona (Spain)

2D image registration defined as the process of overlaying two or more views of the same scene taken from different viewpoints is one of the crucial steps and plays very important role in vast amount of computer vision and robotics processes (e.g., mapping, 3D reconstruction, visual servoing, visual navigation, SLAM, among others). Image registration is mainly accomplished by using feature or featureless methods. Over the last decade, impressive progress (such as SIFT and SURF) has been made on detecting and extracting distinctive salient points (called as features) in the image, which leads to foster and promote the usage of feature-based methods more than the featureless intensity based methods. Compared to previously proposed schemes, SIFT and its successor methods such as SURF show substantially greater invariance to image scaling, and rotation, and robustness under change in illumination and 3D camera viewpoint. Detected features are usually matched using descriptor vectors, which are obtained utilizing gradient information at a particular orientation and spatial frequencies. This

matching frequently produces some incorrect correspondences, which are called outliers. Outliers are identified and removed typically with a robust estimation algorithm (e.g., RANSAC or LMedS) to estimate the 2D planar transformation (homography) between image coordinate frames. These probabilistic methods can fail when there are highly repetitive textures and structural similarities. This failure can cause a mismatched image pairs although resulting homography accuracy can be within the error bounds used in RANSAC and/or LMedS. While processing a sequence of images, the probability of occurrence of such cases becomes high and these mismatched image pairs provides misleading information about the camera trajectory and prevents having accurate final outcome. In this paper, we present preliminary results of identifying such image pairs by using loop constraints. Our method relies on the fact that images forming a loop should have identity mapping when all the homographies multiplied Let assume that 5 images are formed a closed-loop in an order of 1->2->3->4->5->1. Let 1H2 denotes the 3x3 transformation matrix between images 1 and 2. The equality I= 1H2 . 2H3 . 3H4 . 4H5 . 5H1 should hold where I represents 3x3 identity matrix. Our algorithm starts with finding cycles (closed-loops) in the sequence. Next errors (deviation from identity) for each matched pair were computed. To have meaningful descriptive statistics, this process repeated several times with generating different cycles. After obtaining a series of errors for each image pair, we compute error histograms and compared them using a match distance metric. Mismatched pairs are identified as the ones that have a higher distance. To validate our proposal, we prepare a test scenario using underwater image sequences obtained by robotic platforms. We have added some mismatched image pairs where the estimated homography between them lies within the error bound defined in RANSAC. Our proposal was able to identify up to %80 of mismatched pairs.

9025-10, Session 2

Using short-wave infrared imaging for fruit quality evaluation

Dong Zhang, Sun Yat-Sen Univ. (China); Dah Jye Lee, Alok Desai, Brigham Young Univ. (United States)

This paper presents an efficient histogram analysis and image processing technique that is designed specifically for real-time surface quality evaluation of Medjool dates. This approach, based on short-wave infrared imaging, provides excellent image contrast between the fruit surface and delaminated skin, which allows significant simplification of image processing algorithm and reduction of computational power requirements. The proposed quality grading method requires very simple training procedure to obtain a gray scale image histogram for each quality level. Using histogram comparison, each date is assigned to one of the four quality levels and an optimal threshold is calculated for segmenting skin delamination areas from the fruit surface. The percentage of the fruit surface that has skin delamination can then be calculated for quality evaluation. This method has been implemented and used for commercial production and proven to be efficient and accurate.

9025-12, Session 3

Disaster scene partial reconstruction driven by attentive search of an operator wearing the gaze machine *(Invited Paper)*

Fiora Pirri, Bruno Cafaro, Valsamis Ntouskos, Manuel Ruiz, Univ. degli Studi di Roma La Sapienza (Italy)

Scene reconstruction is a critical problem in rescue robotics applications for several reasons. First of all the robot has to navigate and explore a totally unstructured scenario to take measurements, look for specific item, possibly people, reporting to the operators the state of affair, therefore scene reconstruction is a central issue for accomplishing these tasks.

However, the reconstruction of the scene can be a very complex issue mainly because the robot has to transmit the images to a remote base, not having on board all the computing facilities to perform the required elaboration. Furthermore, the unstructured scenario is often a hostile environment where it is quite difficult to navigate, because there are piles of rubble, collapsed walls and roofs, beams that bar the way, pits and several other obstacles difficult to interpret.

On the other hand, under these conditions an unmanned ground vehicle (UGV) has hard time understanding what to look at so as to determine what can be crossed and what should be avoided. Furthermore it is extremely difficult to look for some items and for people when there is no clear spatial reference. These difficulties can be alleviated if the robot can learn where to look at and where to focus at, discerning what is most important on its path. Learning where to look at implies that the robot has to learn what saliency amount to in a disaster scenario.

In this presentation we discuss how an UGV can learn how to partially reconstruct the disaster scene by learning from an operator where to focus its attention, in order to accomplish the required tasks. We therefore address 3 problems in one.

1. How the Gaze Machine, a wearable device that tracks the point of regard in the scene, while the subject is freely moving in it, is used reconstruct what the operator looks at while exploring a disaster environment passing from light to dark and having to get into tunnels, and other complex places.

2. How the method provided ensures to obtain the camera pose so as to reconstruct the whole scene, projecting on it the operator point of regard.

3. How the method proposed is able to extract the set of features that identify the important aspects necessary to navigate the scene, using the operator fixations. How from these features the UGV can learn what to look at and how to correctly separate the space, for example by discerning the terrain from the collapsed not navigable areas.

We complete our presentation by discussing the experiments we have done in the Fire Fighters training area of Prato. Here the fire fighters, who are the first responders in case of disasters, train both human rescue operators and rescue dogs. We illustrate the point cloud inferred by structure and motion from the video stream gathered by a fire fighter, the surface reconstruction and the projected PORS. Finally we show how the robot infer its own interesting points, on the terrain, given the learned gaze scan path.

9025-13, Session 3

Planning perception and action for cognitive mobile manipulators *(Invited Paper)*

Andre Gaschler, Svetlana Nogina, Technische Univ. München (Germany); Ronald P. A. Petrick, Univ of Edinburgh (United Kingdom); Alois Knoll, Technische Univ. München (Germany)

We present a general approach to plan perception and manipulation for cognitive mobile manipulation. Rather than hard-coding single purpose robot applications, a robot should be able to reason about its basic skills and solve complex problems autonomously.

Our “knowledge of volumes” approach to robot planning, in short KVP, is guided by two main principles, making it particularly useful for planning robot perception and manipulation with uncertain or incomplete knowledge, real-world geometry, and multiple robots and sensors: (i) As an underlying symbolic planner, we use PKS (Planning with Knowledge and Sensing), a general-purpose planner that operates at the knowledge level. In contrast to other off-the-shelf planning engines, it can represent both known and unknown information. Therefore we can model sensing actions with clear and concise domain descriptions, and reason in partially known environments that are typical for mobile manipulation. (ii) Rather than discretizing the search space, we represent many geometric predicates---preconditions for perception and manipulation---by continuous volumes, specifically sets of convex polyhedra. This notion of volumes serves as a powerful intermediate representation for modeling

perception and action both on the geometric and on the symbolic level. Our approach is among the first to treat sets of convex polyhedra as such an intermediary representation between continuously-valued robot motions and viewing cones, and discrete symbolic actions, building a combined geometric and symbolic cognitive architecture.

Early work on cognitive mobile manipulators date back to systems like Shakey in 1984. Since that time, the field has made significant developments, and cognitive planning architectures have been proposed from different research communities, including probabilistic techniques from artificial intelligence, closed-world symbolic planning, formal synthesis, and sampling-based manipulation planning.

A relevant contribution most closely related to our approach is the belief space planner by Kaelbling and Lozano-Perez, which models a belief space of probability distributions over states, making it robust against uncertainty and change. In contrast to belief states, our work instead relies on discrete knowledge and is designed for structured environments with incomplete information and sensing. Furthermore, while Kaelbling and Lozano-Perez use octrees to represent swept volumes of robot motion, we use sets of convex polyhedra, allowing very efficient collision detection in the deterministic case.

Perception is to be formulated as a necessary precondition for manipulation, and is not to be hard-coded as a task itself: Our planning approach allows us to define geometric preconditions for sensing and manipulation actions, involving gain and loss of knowledge, and yields an automatized solution of interleaved sensing and manipulation actions.

As a result, this work makes a number of strong contributions to the problem of planning combined perception and action for mobile manipulation. Our approach is novel in using 3D geometric volumes as the underlying representation for symbolic planning and motion planning, and in combining this idea with an off-the-shelf, general-purpose AI planner supporting deterministic planning with incomplete information and sensing.

In contrast to more specialized approaches, which inherently limit the scope of application, our use of an automated planning engine is more general and expressive enough for handling arbitrary numbers of mobile manipulators. Besides this, we can profit from future improvements of new planning engines that become available from the artificial intelligence community.

Finally, our work demonstrated with a mobile manipulator, and both described scenarios were executed on a real Robotino mobile platform with a Katana manipulator, rather than solely in a simulated environment.

9025-14, Session 3

Continuous 3D recovery of human gaze using multisensory tracking

Lucas Paletta, JOANNEUM RESEARCH Forschungsgesellschaft mbH (Austria)

The estimation of human attention has recently been addressed in the context of human robot interaction (Newcombe & Davison, 2010). Today, joint work spaces for the interaction between robots and human agents already exist (Wiese et al., 2012) and challenge cooperating systems to jointly focus on objects and scenes, such as in industrial or in social robotics. Joint attention is a basis for communication (Sumioka et al., 2007) and defines the frame of reference for contextual computing (Carpenter, M., & Liebal, 2011). Another application objective is to develop computational models of human attention for humanoid robots (Wyart et al., 2005) or for the evaluation of perception by robots (Klank et al., 2012).

(Munn et al. 2008) introduced 3D gaze estimation from monocular eye-tracking and triangulation of 2D gaze positions of subsequent key frames. They reconstructed observer positions as well as single 3D gaze points in the environment without the reference to a complete 3D model, achieving angular errors of ≈3.8°. (Voßkühler et al. 2009), similarly to (Pirri et al. 2011), who proposed indoor 3D gaze reconstruction with special head tracking units, achieving accuracies ≈3.6 cm at 2 m distance. (Paletta et al., 2013) proposed a SLAM methodology using RGB-D information, reconstructing the 3D gaze within an automatically reconstructed 3D model, with precision of ≈1 cm at 2-3 m distances. This work described for the first time the estimation of human fixations in 3D environments without the requirement of artificial landmarks in the field of view and enabled – partially continuous - automated attention mapping in 3D models, opening new opportunities for joint attention studies as well as for bringing new potential into automated processing for human factors technologies.

In our work we significantly improve the previous approach in terms of continuous coverage of service during in studies using eye tracking glasses. The main challenge is to bridge the gaps in the largely continuous but vision based 6 DOF pose reconstruction, by bridging effects of motion blur and insufficient coverage of texture with high spatial frequency.

We achieve a continuous coverage of 3D gaze reconstruction by making full exploitation of multisensory information, complementing the head worn eye tracking glasses (ETG) sensor with a high accuracy but still wearable accelerometer sensor which is being attached to the glasses frame.

Using a particle filter approach (Durrant-Whyte & Bailey 2006; Gustaffson 2010) for multisensor based information fusion in the 6 DOF localization – using the camera sensor, an accelerometer, the floor plan provided by the 3D model and WLAN based positioning, we achieve an average position accuracy of about +/- 5cm versus ground truth data, over a course of about 50 m attention service task, in contrast to +/- 20 cm using the purely vision based approach.

We visualize how the new methodology enables to bridge outages of the vision based approach, for example in case of image blur, viewpoint changes that include previously not modeled areas of the environment, and focusing at regions with not sufficient texture information for purely vision based approaches.

9025-15, Session 4

Motion lecture annotation system to learn Naginata performances

Daisuke Kobayashi, Ryota Sakamoyo, Yoshihiko Nomura, Mie Univ. (Japan)

Sports, especially way of practice budo starts acquirement of basic motion called “kata”.

Generally, player run over simulate trainer’s kata and be collected his kata by trainer. But, there is a problem. Trainer is able to correct slowly motion of one part such as hand or leg at a time at most. If trainer can take a look at player’s whole body, he is difficult to teach and correct motion coinstantaneous and interlock; we called “accommodativeness”. And it is difficult to even only one part in real-time and dictation. So, we develop “motion annotation system” what player can carefully looks budo or sports trainer’s motion and teach as letter form. Player understands easily by use that system. We expect our system help understand trainer’s intend to see our image content (include “read” content).

The way of trainer teach the motion to player, instead of teaching on times, having pointed out to leaders to recorded video.

The advantage of this way player can look clear to player’s own appearance. But, this is difficult to advice face to face on time. So, we use annotation that add comment to video, trainer can available to use video’s advantage; watch careful and over and over again.

Since comment remains in the video, player should not keep listen by trainer. We describe using motion capture data to teach naginata performance. There are some video annotation tools such as YouTube. However these video based tools have only single angle of view. Our approach that uses motion captured data allows us to view any angle. The trainer can write annotations related to parts of body. The player can correct the motion by reviewing motion data and annotations. We have made a comparison of effectiveness between the annotation tool

of YouTube and the proposed system. The experimental result showed that our system triggered more annotations than the annotation tool of YouTube. That reason is because the proposed system can change view angles. The data in the results will be used to create new tools for practicing real-time motion which player can own practices.

9025-16, Session 4

Illumination-robust people tracking using a smart camera network

Nyan Bo Bo, Peter Van Hese, Junzhi Guan, Sebastian Gruenwedel, Jorge Oswaldo Niño-Castaneda, Dimitri Van Cauwelaert, Dirk Van Haerenborgh, Peter Veelaert, Wilfried Philips, Univ. Gent (Belgium)

Visual tracking of multiple persons in difficult lighting conditions is an essential component of smart meeting rooms, and intelligent surveillance systems. Robust tracking of multiple people is very challenging because the appearance of a person changes with the body movements, changes in pose and orientation, and lighting changes. Practical systems must be robust against sudden light changes and difficult lighting conditions. For instance, in smart meeting rooms the light(s) may be dimmed or switched off when a presentation is projected. In a living room, the light is heavily influenced by outside light from the sun. Moreover, a person may sometimes be occluded by other person(s) or by objects in the scene making the problem even more difficult. Our system uses a decentralized tracking architecture, i.e., the computational load is distributed over smart cameras and a fusion center. Each smart camera tracks all persons in its view locally in image coordinates and all estimates are sent to the fusion center to jointly estimate the position of each person in world coordinates. These joint estimates are fed back to all cameras, which are then used as a prior for local tracking in the next frame. Each camera extracts foreground objects by detecting texture changes between the current picture and a static background picture, rather than changes in gray value as most existing foreground/background segmentation methods do. The texture based approach makes our method very robust to illumination changes. The result of foreground detection is a binary image in which white pixels represent foreground objects (persons). From the foreground image and the known positions of all persons in the previous frame, as communicated by the fusion center, the local tracker constructs a person and a background model for each person. The position of a person in the current frame is estimated by finding the position in the current foreground image where both the person and background models match best. After receiving estimates from the cameras, the fusion center defines the hypothesis points around the know position of a person in previous frame. Each hypothesis point is projected into the image coordinates of each smart camera. The position of a person is at the hypothesis point where the sum of the Euclidean distances between the projected point and camera's local estimate of all cameras is the minimal. We use a simple occlusion handling strategy: we simply discard those views in which a person is occluded. As we use six camera the number of remaining views almost always suffices to correctly locate the person. We evaluate our method in six video sequences of up to 4 people walking in a room of 8.8?9.2 m2. The sequences have total duration of 25 minutes and are captured using six calibrated cameras with overlapping views. Persons are often occluded by other persons or furniture. During video capture, lights are switched on/off repeatedly. The average tracking error, defined as the average Euclidean distance between the estimates of the tracker and manually annotated ground truth positions over all sequences, is 11cm. We also compared performance of our system to a state-of-the-art tracking system with respect to robustness to sudden illumination changes and find that our method outperforms.

9025-17, Session 4

Image-based indoor localization system based on 3D SfM model

Guoyu Lu, Chandra Kambhamettu, Univ. of Delaware (United States)

Indoor localization has attracted intensive research for both mobile and Robot communities. Most indoor localization systems use cellular base stations or WiFi ratio for the localization task. However, the localization accuracy is largely dependent on the beacons' distribution. Also, for capturing these signals, users may need to carry extra equipments. With the development of camera techniques, especially on the use of mobile phones, image-based localization system is largely used in outdoor environment for its high localization accuracy in the large building area where the GPS information is weak. In recent years, image-based localization is also employed in indoor environment for the easy availability of the necessary equipments. By capturing an image and sending it to an image database, the best matching image can be returned with the navigation information. With the development of Structure-from-Motion (SfM) techniques on the use of reconstructing 3D scene cloud points, 3D SfM model is used in image-based localization techniques. By allowing further camera pose estimation, the image-based localization system with the use of SfM model can achieve higher accuracy than the methods of searching through a 2D image database. However, this emerging technique is still only on the use of outdoor environment. In this paper, we introduce the 3D SfM model based image-based localization system into the indoor localization task. We capture images of our department and reconstruct the 3D model of the department. On the localization task, we simply use the images captured by a mobile to match the 3D reconstructed model to localize the image. In this process, local features are on the use of finding correspondences between 2D query images and the 3D model. As the 3D model usually contains several million descriptors, finding the correspondences for the query descriptors is a time-consuming process. In dealing with this problem, two techniques are used in fast searching correspondences. One is to use K-D tree search to find the approximate nearest neighbours. Another one is to cluster all the descriptors into a large amount of visual words and search the query descriptor's correspondence in each assigned small visual word. By this way, the searching scope is reduced. We will analyze these two methods separately on the aspects of both speed and accuracy. Usually several thousand local features are extracted from an image. For reducing the localization time, when a certain number of correspondences are found, we stop correspondences searching. Further, we use RANSAC to find the inliers between the query image and the 3D model. Only if 12 inliers are found within the query image, we consider this image is registered to the 3D model, which means the image is correctly matched to the 3D model. 6-point-direct-linear-transformation (6P DLT) algorithm is applied for estimating the camera pose. We localize the image if the camera pose estimation is successful. As the 3D model allows camera pose estimation, this indoor localization method is more accurate than the methods searching though the whole image database. Meanwhile, from the 3D SfM model, the building indoor structure is more obvious to the users, facilitating users to better plan their visiting route. The whole system shows multiple advantages than other existing indoor localization systems.

9025-18, Session 4

Using probabilistic model as feature descriptor on a smartphone device for autonomous navigation of unmanned ground vehicles

Alok Desai, Dah Jye Lee, Brigham Young Univ. (United States)

In past few decades there has been significant research on the

development of feature descriptors. The research mainly emphasized real-time applications. This paper presents the development of an affine invariant feature descriptor for low resource applications such as UAV and UGV embedded systems, a small microprocessor, a field programmable gate array (FPGA), and a Smartphone device. UAV and UGV have proven suitable for many promising applications such as unknown environment exploration and search and rescue operation. Those applications required on board image processing for obstacle detection, avoidance and navigation. All real-time vision applications required a camera to grab images and match features using feature descriptor. A good feature descriptor will uniquely describe a feature point thus allowing it to be correctly identified and matched with similar images.

We have seen a growing number of systems from micro unmanned vehicles to smart phones and other devices that when compared to commonly used image processing hardware, have very limited resources. For these devices, few feature description algorithms are readily available, but they require too much of the device's resources and require too many simplifications and result is reduction in accuracy. This research is aimed at meeting the needs of these systems without the sacrifice of accuracy. We can bring high quality computer vision algorithms into the real-time low-resource systems. This paper introduces novel feature descriptor PRObabilistic model (PRO) for a navigation purpose of UGVs. It is a compact and efficient binary descriptor. We divided our process of implementation into two parts.

1. Off-line dictionary creation

The first step is to find feature points by using a feature detector. Small pixel regions around detected feature points called Feature Region Images (FRIs), are saved as individual images. From thousands of FRIs datasets, K-SVD is used to create a set of basis functions. The final step is to create Basis Dictionary Image sets (BDIs) and use them in real time.

2. Online feature matching.

The first features were found using a Harris detector in image 1. PRO descriptors were then computed for each feature. The same process was carried out for the second image 2. Next, the similarity indicator was calculated between each feature in 1 and each feature in 2 and the results were sorted according to ascending value of the similarity indicators.

For the validation of our algorithm, we use an Idaho dataset which is publicly available at www.roboticvision.groups.et.byu.net/. We compare our algorithm with SIFT, SURF, and BASIS for the accuracy. We implemented our algorithm on a Smartphone device to navigate unmanned ground vehicle.

9025-19, Session 4

Relative localization in urban environment using exclusively RGB-D cameras

Marie-Anne Mittet, Pierre Grussenmeyer, Tania Landes, Institut National des Sciences Appliquées de Strasbourg (France)

Typically, GNSS technology gives an indication of direct absolute positioning when enough satellites are visible. In urban environment, urban canyon conditions due to the tall nearby structures obstruct the visibility of satellites. The current solution to overcome this problem is to use additional sensors in order to replace the GNSS positioning and to provide not only position but also orientation parameters of the vehicle. The results currently offered by commercially available mobile mapping solutions are not up to the application requirements and the high cost significantly limits the diffusion of mobile data acquisition systems. The goal of our work is to provide a fast and accurate method to estimate the position and orientation of a mobile platform in urban environment, as an alternative to conventional systems. RGB-D cameras seem to offer interesting characteristics that haven't been studied with the purpose of localization in urban areas. Today they are used in many applications, including 3D modeling or pattern recognition, in the fields of geomatics as well as robotics. We intend to develop a new type of sensor integration solution based on multiple RGB-D cameras with the capability of relative localization.

The developed algorithm takes into account the urban furniture to evaluate the displacement of the platform. The RGB-D cameras placed on top of the moving mobile acquire RGB images, as well as depth images. These acquisitions are used to generate the so-called "ortho-projection" images, which might facilitate the recognition of the displacement occurring between successive images. The main goal of our algorithm is to determine the set of transformations between series of ortho-projection images and then to combine these transformations to determine the set of successive positions occupied by the mobile. The first stage of the developed approach consists of ortho-projecting the point clouds obtained by the camera on the bottom plane, which represents the ground. For the second stage, points of interest are detected in the ortho-projections images. Then the link between these points of interest with those of the preceding image is computed. Finally, an estimation of the mobile displacement is computed based on these two series of points. In this way, the trajectory of the mobile can be estimated in real-time during the displacement.

In this paper, we present the key features of our approach and evaluate its performance. In particular, we evaluate the accuracy, robustness, and processing time for different feature descriptors. The installation of three cameras on board of the vehicle allows covering a larger field of view than using only one, so the coupling of the data obtained by each camera is also detailed. The experiments are performed on two kinds of datasets: a simulated dataset and a real dataset. The study area for the experiments made on simulated data is a large public open space, located in Paris and previously modeled in 3D. The real dataset has been collected in a street of Paris, at dusk, using a draft version of our prototype. For both simulated and real datasets, a ground truth is available for assessing the accuracy of the calculated trajectory.

9025-20, Session 4

Classification and segmentation of orbital space based objects against terrestrial distractors for the purpose of finding holes in shape from motion 3D reconstruction

Terrell Nathan Mundhenk, Arturo Flores, Heiko Hoffman, HRL Labs., LLC (United States)

3D reconstruction of objects via Shape from Motion (SFM) has made great strides recently. Utilizing images from a variety of poses, objects can be reconstructed in 3D without knowing a priori the camera pose. These features can then be bundled together to create large scale scene reconstructions automatically. A short coming of current methods of SFM reconstruction is in dealing with specular or flat low feature surfaces. The inability of SFM to handle these places creates holes in a 3D reconstruction. This can cause problems when the 3D reconstruction is used for proximity detection and collision avoidance by a space vehicle working around another space vehicle. As such, we would like the automatic ability to recognize when a hole in a 3D reconstruction is in fact not a hole, but is a place where reconstruction has failed. Once we know about such a location, methods can be used to try to either more vigorously fill in that region or to instruct a space vehicle to proceed with more caution around that area. Detecting such areas in earth orbiting objects is non-trivial since we need to parse out complex vehicle features from complex earth features, particularly when the observing vehicle is overhead the target vehicle. To do this, we have created a Space Object Classifier and Segmenter (SOCS) hole finder. The general principle we use is to classify image features into three categories (earth, man-made, space). Classified regions are then clustered into probabilistic regions which can then be segmented out. Our categorization method uses an augmentation of a state of the art bag of features method for object categorization (Vedaldi & Fulkerson, 2008). This method works by first extracting PHOW (dense SIFT like) features which are computed over an image and then quantized via KD Tree. The quantization results are then binned into histograms and results classified by the PEGASOS support

vector machine solver (Singer & Srebro, 2007). This gives a probability that a patch in the image corresponds to one of three categories: Earth, Man-Made or Space. Here man-made refers to artificial objects in space. To categorize a whole image, a common sliding window protocol is used. Here we obtained 90 high resolution images from space shuttle servicing missions of the international space station. We extracted 9000 128x128 patches from the images, then we hand sorted them into one of the three categories. We then trained our categorizer on a subset of 7500 patches. Testing on 1500 testing patches yielded 96.8% accuracy. This is basically good enough because detection returns a probabilistic score (e.g. p of man-made). Detections can then be spatially pooled to smooth out statistical blips. Spatial pooling can be done by creating a three channel (dimension) image where each channel is the probability of each of the three classes at that location in the image. The probability image can then be segmented or co-segmented with the visible image using a classical segmentation method such as Mean Shift. This yields contiguous regions of classified image. Holes can be detected when SFM does not fill in a region segmented as man-made. Results are shown of the SOCS implementation finding and segmenting man-made objects in pictures containing space vehicles very different from the training set such as Skylab or the Hubble space telescope.

9025-21, Session 5

Discrete and continuous curvature computation for real data

Dirk J. Colbry, Michigan State Univ. (United States); Neelima Shrikhande, Central Michigan Univ. (United States)

This paper describes two methods for estimating the minimum and maximum curvatures for a 3D surface and compares the computational efficiency of these approaches on 3D sensor data. The classical method of Least Square Fitting (LSF) finds an approximation of a cubic polynomial fit for the local surface around the point of interest P and uses the coefficients to compute curvatures. The Discrete Differential Geometry (DDG) algorithm approximates a triangulation of the surface around P and calculates the angle deficit at P as an estimate of the curvatures. The accuracy and speed of both algorithms are compared by applying them to synthetic and real data sets with sampling neighborhoods of varying sizes. Our results indicate that the LSF and DDG methods produce comparable results for curvature estimations but the DDG method performs two orders of magnitude faster, on average. However, the DDG algorithm is more susceptible to noise because it does not smooth the data as well as the LSF method. In applications where it is not necessary for the curvatures to be precise (such as estimating anchor point locations for face recognition) the DDG method yields similar results to the LSF method while performing much more efficiently. We cite algorithms and describe effect of key parameters such as sampling radius.

9025-22, Session 5

An evaluation of attention models for use in SLAM

Samuel Dodge, Lina Karam, Arizona State Univ. (United States)

Simultaneous Localization and Mapping (SLAM) is the problem of simultaneously determining a robot's position and constructing a map of the environment. Although it is possible to use a multitude of sensors for this problem, in this work we only consider visual SLAM where only video cameras are used. This can be more desirable than using laser scanners or other sensors because of lower power and lower cost requirements. However, visual SLAM relies on the detection of accurate repeatable keypoints. These keypoints must be easy to detect and track over several frames. It has been shown that a model of human attention can be used to generate such reliable keypoints [1], however the previous work only considers one formulation of visual saliency.

There are many different formulations of saliency. For example, the classical Itti model formulates saliency as a combination of center-surround features of orientation, color, and intensity [2]. Koostra et al. computes a saliency map using a symmetry-based feature [3]. Bruce et al. pose saliency as a pursuit of information maximization [4]. Judd et al. incorporate several top down features such as a horizon detector and a face detector [5]. To the best of our knowledge it has not been tested which such approaches perform the best in the SLAM framework.

To this end, we propose to use a machine learning algorithm, namely Adaboost, to determine which saliency features perform the best. We consider the following features: orientation, color and intensity center-surround differences; symmetry features; attention maximization; and the SIFT keypoint detector. Adaboost will assign a weight to each feature based on its relative importance. In this process we learn a suitable saliency model that uses an optimal combination of features for the SLAM problem.

SLAM by nature needs to be real time. Furthermore, mobile robots are often subject to limited computational resources. Thus although a combination of many features could yield more stable keypoints, the computational complexity could make such an approach infeasible. To this end, we analyze the computational burden as more features are added and find a suitable tradeoff.

9025-23, Session 5

3D vision system for intelligent milking robot automation

Moulay A. Akhloufi, Ctr of Robotics and Vision (Canada)

In a milking robot, the correct localization and positioning of milking teat cups is of very high importance. The milking robots technology has not changed since a decade and is based primarily on laser profiles for teats approximate positions estimation. This technology has reached its limit and does not allow optimal positioning of the milking cups. Also, in the presence of occlusions, the milking robot fails to milk the cow. This problems, have economic consequences for producers and animal health (e.g. development of mastitis).

To overcome the limitations of current robots, we have developed a new system based on 3D vision, capable of efficiently positioning the milking cups. A prototype of an intelligent robot system based on 3D vision for real-time positioning of a milking robot has been built and tested under various conditions on a synthetic udder model (in static and moving scenarios).

Experimental tests, were performed using 3D TOF and RGBD cameras. The proposed algorithms permit the online segmentation of teats by combing 2D and 3D visual information. The obtained results permit the teat 3D position computation. This information is then sent to the milking robot for teat cups positioning. The vision system has a real-time performance and monitors the optimal positioning of the cups even in the presence of motion. The obtained results, with both TOF and RGBD cameras, show the good performance of the proposed system. The best performance was obtained with RGBD cameras. This latter technology will be used in future real life experimental tests.

9025-24, Session 5

SDTP: a robust method for interest point detection on 3D range images

Shandong Wang, Lujin Gong, Hui Zhang, Yongjie Zhang, Haibing Ren, Samsung Advanced Institute of Technology (China); Seon-Min Rhee, Hyong-Euk Lee, Samsung Advanced Institute of Technology (Korea, Republic of)

In applications of intelligent robots and computer vision, interest point detection and feature description are highly focused and investigated.

This paper presents a novel method for interest point detection on 3D range images called SDTP (Signed Distance to Tangent Plane). The method is based on the construction of the repeatable local reference frames (LRFs) and tries to select interest points in the areas that are locally stable but have large surface variation in the vicinity. The overall algorithm contains several steps: 1) Pre-filter the range image obtained from depth cameras, here mainly referring to noise smoothing and outliers removal. 2) Identify border points indicating depth discontinuity in the range image, which are used to accurately estimate the geometrical characteristics by the neighboring points on the continuous surface. 3) Compute the repeatable LRF for each image point, of which the z axis is robustly estimated by the plane fitting with a small neighborhood surrounding the point, and the determination of the x axis considers only a subset of points lying at the periphery of the support and is based on the largest signed distance to the tangent plane. For detail algorithm readers can refer to the paper by Alioscia P. et al. (A repeatable and efficient canonical reference for surface matching, 3DimPvt2012). 4) Calculate an interest value for each considered image point. Firstly accumulate the approximate curvatures of the small local neighborhood to achieve a score representing how locally stable the surface is, and the curvature mentioned here are obtained while estimating the Z axis of the LRF. Secondly compute a score to measure the surface change in the vicinity. Look at each point in the periphery of the neighboring region and construct a D curve, of which the horizontal coordinate indicates the angle between the LRF's X axis and the point's projection on the tangent plane, and the vertical coordinate indicates its signed distance to the tangent plane. After this, perform a statistic analysis over the D curve with the goal of that the high amplitude change and the small angle distance lead to the high surface change. Thirdly define the interest value as the product of the two scores. 5) Perform smoothing on the interest values and do non-maximal suppression to select the final interest points.

The proposed detector of SDTP theoretically benefits from the construction and analysis of the D curve, and thus it is robust to point density variations and missing regions and also can better deal with the nuisances of occlusions and clutter accompanying with the range image. Compared to the methods based on the statistic of the normal or other higher derivatives, the SDTP directly uses 3D coordinates of neighboring points to measure the surface variation, and thus tends to be more insensitive to noise and much faster in computation. Moreover, the method is easily extended for the successive development of feature descriptor. The algorithm is tested on publicly available datasets containing synthetic objects and scanning scenes acquired with different depth cameras. Experimental results validate the proposed method with high repeatability comparable to the state of the art. The average run-time for the overall computation is 150 ms on the range image with 320x240 resolutions without any optimization and acceleration on an Intel i5 CPU.

9025-25, Session 5

Real time moving object detection using motor signal and depth map for robot car

Hao Wu, Wan-Chi Siu, The Hong Kong Polytechnic Univ. (Hong Kong, China)

Moving object detection from a moving camera is a fundamental task in many applications, especially for robot application like obstacle tracking and avoidance [1]. Many vision based schemes have been developed for moving object detection under moving background, such as by differentiating the foreground moving object motion pattern with background motion pattern [2-4], or modeling the background movement in 2D affine transform by feature points matching [5-6]. However, the existing moving object detection algorithms are either too complex to implement into a real time system or the 2D background modeling assumption does not fit well with a robot car application.

In the vision of a robot car, the background movement brought by camera moving is actually a 3D motion structure [7]. Utilizing the 3D transformation matrix, the coordinate relationship after camera movement between two frames can be identified by a 3 by 3 perspective transform matrix [8], whose variable includes camera rotation angle, translational moving distance and depth value of the corresponding pixel. Then the moving object can be detected for the areas that do not align with the perspective transform matrix.

For the robot car situation, we found that the above three variables can be obtained through motor control signal and stereo camera image matching. The perspective transform matrix which represents the camera motions can be predicted by the available signal actively instead of conventional passive methods.

In our approach, a relationship between raw motor control signals, normalized depth map and perspective transform matrix is obtained through modeled offline training. Then a pixel on current frame can map to a new position on reference frame with the perspective transform matrix predicted from the effective depth intensity and current motor control signals. If the pixel belongs to static background, a similar pixel can be found on the obtained position in the reference frame, or else this pixel should be classified as moving object pixel. After the pixel classification, some further processing steps are needed including de-noising, connection analysis and segmentation to identify the full foreground object.

To increase the robustness of pixel classification, a tolerance should be given on the rotation angle, translational moving distance and pixel depth in the aspect transform matrix to moderate the system noise, including the camera vibration and signal quantization distortion. At the same time, multi reference frames are used to further eliminate the noise effect brought by error in the temporal domain.

The proposed scheme can detect the moving object during the normal movement, including forward movement and horizontal rotation, of our robot car in real time. Moreover, the proposed scheme can be further applied to cars with a 3D freely moving camera. Compare with other vision based algorithms, the proposed algorithm find a practical way on the moving background modeling with 3D structure, especially for the zoom-in and zoom-out moving background, which has not sufficiently studied in the literature. Besides, the perspective transform matrix is obtained through the offline training and prediction, instead of online modeling, therefore, the computation complexity and memory requirement is low, which makes it possible to implement this scheme in real time.

References:

[1] Lee, Seungwon, et al. "Moving object detection using unstable camera for consumer surveillance systems." Consumer Electronics (ICCE), 2013 IEEE International Conference on. IEEE, 2013.

[2] Sheikh, Yaser, Omar Javed, and Takeo Kanade. "Background subtraction for freely moving cameras." Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009.

[3] Yue, Qin, et al. "Real-time moving target Detection in the dynamic background." Computing Technology and Information Management (ICCM), 2012 8th International Conference on. Vol. 1. IEEE, 2012.

[4] Li, Haojie, et al. "Automatic detection and analysis of player action in moving background sports video sequences." Circuits and Systems for Video Technology, IEEE Transactions on 20.3 (2010): 351-364.

[5] Xiaowei Zhou; Can Yang; Weichuan Yu, "Moving Object Detection by Detecting Contiguous Outliers in the Low-Rank Representation," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.35, no.3, pp.597,610, March 2013

[6] Xiaochun Zou; Xinbo Zhao; Zheru Chi, "A robust background subtraction approach with a moving camera," Computing and Convergence Technology (ICCCT), 2012 7th International Conference on , vol., no., pp.1026,1029, 3-5 Dec. 2012

[7] Szeliski, Richard. Computer vision: algorithms and applications. Springer, 2011.

[8] Hoi-Kok Cheung; Wan-Chi Siu,; Lee, S.; Poon, L.; Chiu-Shing Ng, "Accurate distance estimation using camera orientation compensation technique for vehicle driver assistance system," Consumer Electronics (ICCE), 2012 IEEE International Conference on , vol., no., pp.227,228, 13-16 Jan. 2012

9025-26, Session 6

Research and development of Ro-boat: an autonomous river cleaning robot

Aakash Sinha, Omnipresent Robot Technologies Pvt. Ltd. (India) and Lockheed Martin Corp. (United States) and Carnegie Mellon Univ. (United States); Prashant Bhardwaj, Omnipresent Robot Technologies Pvt. Ltd. (India) and Manav Rachna College Of Engineering (India); Bipul Vaibhav, Omnipresent Robot Technologies Pvt. Ltd. (India) and Indian Institute Of Technology Guwahati (India); Noor Mohommad, Omnipresent Robot Technologies Pvt. Ltd. (India) and National Institute Of Technology, Warangal (India)

In this paper we explore Ro-Boat, an Autonomous river cleaning intelligent robot in terms of environmental effect, mechanical design and computer vision algorithm to achieve autonomous river cleaning. Ro-boat is designed to clean great Indian rivers like the Ganges, which is in bad shape due to constant contamination by various pollutants which are floating, submerged and dissolved in nature. In order to efficiently clean such pollutants, we have designed Ro-Boat in a modular fashion incorporating design details such as mechanical structural design and kinematic analysis in CATIA V5. Force analysis on rudder design is done to measure the impact of high speed jets of water as Ro-Boat traverses through a stream of water; the data is extrapolated for different density of water and different hydrodynamic calculations for buoyancy and motion are done with the estimation of metacentric height to assess the stability of the two hull boat system in different streamlines and stability in case of oscillation. Vibration analysis for the structure is also done to measure the effect of motor vibration and damping by fluid. After we achieve a stable mechanical system with propulsion, robotic arms and energy source, we proceed to make the system autonomous. Computer Vision is used for detecting and recognizing pollutants from a gimbaled camera with two axis rotation. Basically, two pollutants tracking algorithms are proposed to use as measurements for a Kalman Filter. Filter is trained for dynamically identifying when each algorithm is weak and penalize it accordingly with a high measurement variance. First algorithm is “HSV color space” for color based image segmentation where multiple pollutants are detected on the color basis and “moments” method is used for calculating center of the object. This is fast, robust, occlusion-free and applicable for multiple pollutants tracking. But this algorithm loses out on detecting the complex pollutants which are multicolored or having same color as the background. “SURF (Speeded Up Robust Feature)” is used for sorting out this problem by matching key points of template of the desired pollutant with the real time pollutant. Given images of a pollutant and a scene, SURF is very good at finding the location of the target pollutant even under suboptimal lightning. This method is robust and capable of tracking pollutants having complex structures and colors.The use of feature vectors allows SURF to be scale and rotation invariant. Finally, Kalman Filter is employed for tracking the pollutants and generates command for the controller to maneuver the Robot towards the waste and collect it. Kalman filter ensures that we get the best of both the algorithms resulting in an extremely robust pollutants tracking. We have tested the system with successful results in the Yamuna River in New Delhi. We foresee that a system of Ro-boats working autonomously 24x7 can clean a major river in a city on about 6 months time , which is unmatched by alternative methods of river cleaning.

9025-27, Session 6

Real-time, resource-constrained object classification on a micro-air vehicle

Louis Buck, Laura E. Ray, Dartmouth College (United States)

The ability to classify objects in an environment is essential to many high-level autonomous robotic tasks or coordinated maneuvers in multi-agent systems. This task is made more difficult on a resource-limited platform like a quad-rotor micro-air vehicle (MAV), the computational limitations of which prevent the use of many of the state-of-the-art techniques like spatial-pyramid matching with SVMs [1]. Recently, Chandrashekar and Granger [2] developed a novel supervised classifier, called the Cortical-Striatal Loop (CSL) algorithm that is particularly promising for operation in real-time on a resource-constrained platform, as it boasts classification times 100 times faster than a standard classifier while using less memory. The algorithm is based on a learning mechanism used by 80% of the human brain. In the paper, a Bag of Visual Words (BoVW) model with densely-extracted SIFT features was used to achieve 100x lower classification times on a 39 class object dataset from the Caltech-256 dataset with accuracy comparable to classification with the SVM. CSL’s 100x improvement in classification time over the commonly-used SVM makes it a promising candidate for use in real-time on an embedded platform, but other parts of the classification pipeline still are too computationally intensive. In particular the extraction and quantization of SIFT feature descriptors take 70 and 20 percent of the classification time, respectively, yielding embedded classification times of over 8 seconds and making the system impractical to use in real-time.

This paper investigates the possibility of using new and faster binary descriptors to accelerate object classification using CSL with the goal of real-time computation on the Gumstix Overo FireSTORM COM. To provide a direct comparison with the results in [2], the same densely-sampled bag of visual words classification technique is used. The accuracy and embedded classification times of the algorithm are compared using SIFT, BRIEF, ORB and FREAK feature descriptors on four different data sets. Additionally, the effect of chi-squared feature mapping and opponent-color descriptors is investigated.

The paper provides an overview of object classification techniques. In particular the bag-of-words model and the range of choices for its various system components are discussed in detail. An overview of the CSL learning algorithm and opponent-color descriptors is provided. The small body of work specifically related to the use of binary descriptors for object classification is described, and the implementation details of the CSL and binary descriptor BoVW classifier are provided. The details and results of its testing on a variety of classification datasets are given along with extensive analysis and include performance, classification time, and memory requirements. The BRIEF descriptors provide classification performance of 65% as well as SIFT on a 39-Class Caltech dataset, 80% as well on a 5-Class Caltech set, 95% as well on a 5-Class Vehicle set, and 95% as well on a 2-Class set of civilians vs. soldiers. When SIFT features are used for classification the process takes 10 s, while BRIEF descriptors perform with times of 0.15-0.2 s. For a 300 x 300 pixel image with step size of 4 pixels, SIFT requires 128 B per feature descriptor (720 kB total) and 65.6 kB for a codebook of size 512, while BRIEF requires 32B per descriptor (180 kB total) and a 16.4 kB codebook. These results show that the BRIEF feature descriptor provides the best tradeoff between resources required for classification and performance.

References

[1] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan, “Contextualizing object detection and classification,” in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1585–1592.

[2] A. Chandrashekar and R. Granger, “Derivation of a novel efficient supervised learning algorithm from cortical-subcortical loops,” Frontiers in computational neuroscience, vol. 5, 2011.

9025-28, Session 6

ARM-based system integration and testing for ROBO: application for precision agriculture

Aditya Goyal, Shubham Mehrotra, Birla Institute of Technology and Science, Pilani (India); Shashikant Sadistap, Sai K. Vaddadi, Central Electronics Engineering Research Institute (India)

Over 7000 Indians die of hunger every day – the major reason not being inadequacy of food but rather its wastage. Wastage of large amounts of grains and vegetables in govt. storages have been reported recently. This is a very serious issue given that India has 212 million undernourished people and wastage of even a grain of food cannot be afforded. This has been the major motivation behind our work towards developing a system to properly monitor the storage areas and avoid wastage.

The project involves the use of FriendlyARM mini2440 board, based on ARM9 microprocessor, to develop an autonomous robot for agro-based applications. The robot developed is capable of moving independently, sensing obstacles and deciding upon future course of movement along with the ability of image acquisition when required or as pre-determined.

For image acquisition, a USB camera has been mounted on the robot capable of rotation in both horizontal and vertical plane. This enables the camera to take images of almost every part of the region to be monitored, thus, providing effective monitoring of the storage area.

The images may be of perishable food items such as grains, fruits and vegetables etc., stored in a storage chamber or of any article that needs monitoring. Wireless Connectivity enables quick transmission of gathered images to another location where they can be processed and the status of the food item can be determined –

• Whether the food item is fresh and consumable?

• Whether in some parts of the storage area the food item is getting spoilt?

• Whether the growth is proper?

Thus, according to the information processed , the monitoring person can take some actions required to keep the stored item fresh and avoid wastage.

An ADC channel expansion peripheral board has also been interfaced with the system using SPI to increase the number of ADC channels that can be interfaced with the FriendlyARM MINI2440 board. Availability of higher number of ADC channels enables interfacing additional sensors for light intensity, pH, temperature, pressure and carbon dioxide content to the system thereby making the robot more functional and useful in precision agriculture and greenhouse industry.

The system developed can also be coupled with Electronic-Nose, systems that detect and identify odours and vapours, and hence can be used for quality control, to perform the task of monitoring perishable food items more efficiently. Incorporation of other functionalities such as gesture recognition etc. with the system can make it more powerful and suited for many other applications.

9025-29, Session 6

New vision system and navigation algorithm for an autonomous ground vehicle

Hokchhay Tann, Bicky Shakya, Alex Merchen, Ben Williams, Abhishek Khanal, Jiajia Zhao, David J Ahlgren, Trinity College (United States)

This paper presents improvements made to the intelligence algorithms and hardware employed on Q, an autonomous ground vehicle, for the 2013 Intelligent Ground Vehicle Competition (IGVC). In 2012, the IGVC committee combined the formerly separate autonomous and navigation challenges, into a single AUT-NAV challenge. In this new challenge, the vehicle is required to navigate through a grassy obstacle course and stay within the course boundaries (two white painted lines) that guide it toward a given starting GPS waypoint. Once the vehicle reaches this waypoint, it enters an open course where it is required to navigate to 8 other GPS waypoints while avoiding obstacles. After reaching the final waypoint, the vehicle is required to traverse through another obstacle course before completing the run. To accommodate for these rule changes, the software on Q had to be thoroughly revised. The modular parallel software architecture on Q which features the image processing, navigation and sensor control algorithms running concurrently, was kept the same as previous years. However, changes were made to ensure smooth switching between autonomous and GPS navigation modes at the starting and final GPS waypoints. In addition, the modified Vector Field Histogram (VFH) algorithm, employed on Q as the main navigational algorithm, suffered from poor jerking motions in previous years, due to crude motor command thresholds. The 2013 revision of Q featured a tuned VFH algorithm that allowed Q to smoothly decelerate upon encountering obstacle fields and traverse them with relative ease. A new vision system was also implemented this year. In previous years, two webcams with a small field of view and inflexible gain were used for the vision system. As a result, Q's performance for course-boundary detection varied significantly with different lighting conditions. For the 2013 competition, a Basler Scout camera with wide angle lens was used in place of the two webcams. The new camera's automatic gain and shutter speed controls enabled a consistently high level of performance. In addition, new features such as the Hough transform and color-plane multiplication were added to the image processing algorithm for better detection of the course boundaries. Also, with the new camera, Q is able to see further down the course, allowing for better path planning. With these changes, Q was able to successfully complete the basic AUT-NAV course and finish among the top ten teams to reach the advanced AUT-NAV challenge.

9025-30, Session 6

An effective trace-guided wavefront navigation and map-building approach for autonomous mobile robots

Chaomin Luo, Mohan Krishnan, Mark Paulik, Univ. of Detroit Mercy (United States); Gene Eu Jan, bInstitute of Electrical Engineering, National Taipei University (Taiwan)

Robotic path planning and map building is one of the problems in the field of robotics that attempts to find and optimize the path from the initial position to the final position while the local map is dynamically built up as the robot moves. In this paper, a pheromone-guided real-time navigation and map building approach for an autonomous mobile robot is proposed. Wave-front-based global path planner is developed to generate a global trajectory in workspace. Pheromone traces are remained along the planned global trajectory. As a result, the pheromone traces provide the autonomous mobile robot with a sequence of starting points and goals by a developed local navigator. An autonomous mobile robot like an ant is capable of traversing along the pheromone traces to gradually reach the final destination. A commonly used local navigation algorithm, the Vector Field Histogram (VFH), is relatively fast and thus suitable when computational capabilities on a robot are limited. In this paper, a modified Vector Field Histogram (M-VFH) is developed as a local navigator to guide the robot to perform point-to-point navigation locally with obstacle avoidance by means of LIDAD sensor information to follow pheromone markers, which is not sensitive to sensor noise and other disturbances.

The wave-front propagates from the goal position, marking all free space grids with an incrementing distance value. Once all non-obstacle grids have been marked, the resulting path is created by gradient descent. The M-VFH algorithm outputs a preferred target sector for the robot to move towards. The recommended direction is derived from an analysis of a polar obstacle density histogram constructed from sensor scans of the obstacle field in front of the robot.

While the robot transverses in the workspace, the global environmental information is unavailable for the robot due to the inability of the sensor range. Square grid map representations are suggested for real-time map building and navigation. The proposed method does not need any templates, even in unknown environments. A local map composed of square grids is created through the local navigator while the robot traverses with limited LIDAR sensory information. From the measured sensory information, a map of the robot's immediate limited surroundings is dynamically built for the robot navigation.

In order to generate safer, more reasonable collision-free trajectories, novel heuristic algorithms are developed to optimize the path by effectively integrating the global path planner and local navigator. Due the

ability of the LIDAR with the local M-VFH navigator, the robot is able to avoid moving and unforeseen obstacles while the robot traverses in the workspace.

The real-time wave-front-based navigation and map building methodology has been successfully demonstrated in a Player/Stage simulation environment. With the wave-front-based global path planner and M-VFH local navigator, a safe, short, and reasonable trajectory is successfully planned in a majority of situations without any templates, without explicitly optimizing any global cost functions, and without any learning procedures. Its effectiveness, feasibility, efficiency and simplicity of the proposed real-time navigation and map building for an autonomous mobile robot have been successfully validated by simulation, comparison studies and experiments. Comparison studies of the proposed approach with the other path planning approaches demonstrate that the proposed method is capable of planning more reasonable and shorter collision-free trajectories autonomously.

Conference 9026: Video Surveillance and Transportation Imaging Applications 2014

Monday - Wednesday 3 –5 February 2014

Part of Proceedings of SPIE Vol. 9026 Video Surveillance and Transportation Imaging Applications 2014

9026-1, Session 1

PHACT: parallel HOG and correlation tracking

Waqas Hassan, Philip M. Birch, Rupert C. Young, Christopher R. Chatwin, Univ. of Sussex (United Kingdom)

Histogram of Oriented Gradients (HOG) based methods for the detection of humans have become one of the most reliable methods of detecting pedestrians with a single passive imaging camera. However, they are not 100 percent reliable. This paper presents an improved tracker for the monitoring of pedestrians within images. The Parallel HOG and Correlation Tracking (PHACT) algorithm utilises self learning to overcome the drifting problem. A detection algorithm that utilises HOG features runs in parallel to an adaptive and stateful correlator. The combination of both acting in a cascade provides a much more robust tracker than the two components separately could produce.

9026-2, Session 1

Improved edge directed Super resolution with hardware realization for surveillance, transportation, and multimedia applications

Yue Wang, Osborn F. de Lima, Eli Saber, Rochester Institute of Technology (United States); Kurt R. Bengtson, Hewlett-Packard Co. (United States)

This paper presents an improved edge directed super resolution approach to achieve enhanced edge definition and image quality in the high resolution outcome. With the basic premise to interpolate along the direction of the edge, multiple images shifted by a sub pixel amount are taken, with one of those images chosen as reference. A cross-correlation based Discrete Fourier Transform approach is utilized to register the images to the reference image. The gradient magnitude and direction of each input image at the target high resolution, which is two times the input image resolution, is obtained by using a Sobel mask. The edge direction map is obtained from the gradient direction by quantizing each direction over a 5 x 5 block into 8 possible directions instead of 4 possible directions as was done previously, thereby reducing noise caused due to quantization error. The multiple pass iterative interpolation utilizes spatial information on a high resolution grid as well as gradient information obtained from the gradient magnitude and the quantized edge direction map. The first pass consists of evaluating values for those locations that have known data values from the reference image in the direction of the edge while the second iterative pass exhaustively calculates pixel values that do not have known data values in the direction of the edge. A stopping criterion which is based on the number of pixel locations filled with an interpolated value controls the number of second pass iterations. In order to remove the effects such as high frequency noise resulting from the super resolution process, an Edge Preserving Smoothing Function (EPSF) is employed that maintains edge fidelity but removes the aforementioned noisiness. Additionally, in an attempt to determine the optimal super resolution parameters for the case of still image capture, a hardware setup was utilized to investigate and evaluate those factors. In particular, the number of images captured as well as the amount of sub pixel displacement that yield a high quality result was studied. A XY stage capable of sub-pixel movements in combination with a camera that captures uncompressed images was used to study the effect of varying the number of images and the relative shift between them. For color images, the L channel in the L*a*b* color space was super resolved while the chrominance channels were resized to the target resolution using a conventional bicubic interpolation method. The proposed algorithm showed favorable results both qualitatively and quantitatively using various low resolution datasets in comparison to the previous work. Also, the number of images and the amount of displacement for the technique to yield the best result was determined through experimentation. The possible benefits of this algorithm range from surveillance and transportation applications, to improving image quality in scanners and camera capture in mobile devices. All this is done through software implementation, thereby maintaining or even reducing costs involved in hardware improvements.

9026-3, Session 1

Rotation-invariant histogram features for threat object detection on pipeline right-of-way

Alex Mathew, Vijayan K. Asari, Univ. of Dayton (United States)

We present a novel algorithm that automatically detects anomalies in the pipeline right of way (ROW) regions from aerial surveillance videos. Detection of anomalies on pipeline ROW is gaining more importance as such anomalies can be signs of greater impending failures. Monitoring such anomalies also helps managing potential environmental hazards. Pipelines span over large areas across urban regions and less populous country sides. The far and wide lines are constantly under the threat of construction works. The pipeline companies are required to be self-regulating these days to handle these anomalies. This requirement demands that the pipeline ROWs are under constant video surveillance. Monitoring is typically done by flying aircrafts over the region of interest, along with human assistance to identify/detect possible threats. This is a very expensive and impractical method of dealing with this problem, as pipelines span far and wide. A better way is to use computer vision algorithms to analyze videos captured by surveillance systems mounted on aircrafts. The most serious threats that can be seen along pipeline right of ways are the presence of construction vehicles and heavy duty vehicles, which can be signs of construction in the pipeline corridor. Identifying such equipments or construction vehicles is not a trivial task. Vehicles vary in size, shape and color. The same vehicle may look dissimilar in different flying heights and viewing angles. Other factors that make detection a challenging problem are atmospheric aberrations, the presence of noise in captures from such high altitudes, variation in scale, clutter and motion blur. The detection algorithm should be fast to enable detection in real-time. The method is tested on the data provided by the Pipeline Research Council International (PRCI). Experiments are carried out on three different data bases, provided by three different vendors.

In this method, we propose a rotation-invariant gradient histogram based descriptor in CIELab color space. CIELab is a perceptually uniform color space. We use a linear SVM as classifier. The combination of a robust feature and a linear classifier is a fast and accurate learning system. We have observed that the choice of color space plays an important role in detection accuracy. We have investigated the performance in the RGB, YcbCr, CIElab and HSV color spaces. The best performance is attained when Lab color space is used. We use only the ‘a’ and ‘b’ components since they represent the color values.

The ‘a’ and ‘b’ channels of the Lab color space representation of target image (in this case, the equipment/construction vehicle image) is generated. The region of interest is divided into concentric square regions. The number of such concentric regions is based on the size of the target. The inner square regions capture the local characteristics and finer details of the image. Larger square regions capture the global characteristics. A noise reducing differentiation kernel is used to compute the gradient of the region to cope with noise introduced by atmospheric aberrations and motion blur. The gradient orientation histogram is constructed in each region by voting the squared magnitude of gradients.

Bilinear interpolation is used to vote the squared magnitudes into angle bins. Four quadrant arc tangent function is used to computed gradient angles, since the sign of the gradient is important in this case. It is found that normalizing the histogram decreases recognition rate. The number of angle bins determines the length of the final descriptor. The final descriptor is built by concatenating the magnitude of DFT of orientation histograms collected from 'a' component and 'b' component. The magnitude of Discrete Fourier Transform (DFT) of the histogram is invariant to rotations. DFT can be efficiently computed as Fast Fourier Transform (FFT). The number of points taken for FFT calculation is the same as the number of angle bins. The orientation of the target can be determined from the Fourier phase although it is not required in this particular application. The algorithm is more tolerant to decreasing the number of points in FFT computation than decreasing the number of angle bins. The fundamental difference between the popular Histogram of Oriented Gradients (HOG) algorithm and ours is that the spatial constraint is introduced in HOG by using overlapping regions in a linear fashion while in our algorithm it is done by taking overlapping concentric regions and computing the magnitude of DFT. The advantage of our method is that the final descriptor is fully rotation invariant. This obviates the need for search in multiple orientations. The method is also moderately robust to scale variations. A linear SVM is trained to detect targets in unseen images. These can be construction equipments such as backhoes and excavators. Since the number of available images is limited, we generate new training images by blurring, introducing noise and rotating the available images. Sketches of equipments are also added to the training set. The system is initially trained with a few negative examples. Those negative examples that are misclassified are added to the negative training set. With this iterative process the most representative negative examples can be found. We show that, with a few training examples, we can build a near-perfect detection system. A sliding window detector is used to search for targets in each frame in captured videos. The descriptor in each window is computed and classified as positive or negative, based on the output of the SVM. The target regions may generate multiple detections because of the robustness of the method. Such multiple detections are then combined using non-maxima suppression to fuse several overlapping detections into a single detection. This process also eliminates false detections. The identified construction vehicles and equipments are marked in the output frame as anomalies. The proposed method is tested on the PRCI data set and is shown to offer a recognition rate of 99.8%. Since the algorithm uses a sliding window detector, it can easily be parallelized. Future work includes extending the algorithm as a general object detection method.

9026-4, Session 1

Development of a multispectral active stereo vision system for video surveillance applications

Sanjeev Kumar Malik, Balasubramanian Raman, Indian Institute of Technology Roorkee (India)

Modern video surveillance is an application of computer vision that has attracted much research interest in recent years. These systems have played an increasing role as tools for preventing and investigating crime, traffic control, protecting public safety, and safeguarding national security. Video surveillance and tracking systems typically use a single modality of sequence in the visible spectrum for their input. These systems work well in controlled conditions but often fail with low lighting, shadowing, smoke, dust, unstable backgrounds or when the foreground object is of similar coloring to the background.

In this paper, a multispectral active stereo vision system is developed for tracking the motion of moving objects in a complex environment. The development of such a multispectral surveillance system composed by combining visible (color) and thermal PTZ sensors. The aim is to behind proposing such a network of visible and thermal sensors is to give an optimal performance in various weather conditions (fog, snow, dark, and rain) and increase the detection performance. The other novelty in this system is to localize the moving object on to a 2D map using this multispectral stereo system instead of traditional single camera system. Such a stereo based localization overcomes the problem of accurate localization even in the case of partially occluded targets. A chain of homographies based approach is introduce to establish correspondence between the visible and thermal stereo images. Here, very few matches are required only in a pair of stereo images and the correspondence between rest pairs of the images is obtained by deriving successive homographies with the help of cameras' motion parameters (pan and tilt). The compensation between the unequal zoom setting in the two non-homogeneous spectrums is achieved based on a look-up-table based approach. This compensation is required to perform error-free tracking of moving objects in terms of the optical.

Finally, a number of experimental results (optical flow from stereo images) are given to prove the applicability of such a system in various applications and situations where a robust surveillance and tracking system is needed. The main aim is to conducted these experiments using modalities from both the visible and thermal infrared spectra, allows us to obtain more information from a scene and overcome the problems associated with using visible light only for surveillance and tracking. The experimental results show that the usability of a stereo camera system in video surveillance can enhance the performance in partial occluded cases.

9026-5, Session 1

Extrinsic self-calibration of multiple cameras with non-overlapping views in vehicles

Frank Pagel, Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung (Germany)

Due to decreasing sensor prices and increasing processing performance, the use of multiple cameras in vehicles becomes an attractive possibility for environment perception. Here, we focus on non-overlapping camera configurations. Especially for long term drives camera positions and orientations (the extrinsic parameters) can change due to temperature changes, vibrations or impacts. Ideally, we are able to estimate and correct extrinsic changes in an online manner. So, we aim for a continuous estimation of the extrinsic calibration parameters of mobile multi-camera platforms to be able to detect and correct calibration errors as quick as possible. Here, we assume that the cameras' positions and orientations are fixed relative to each other, that the intrinsic parameters (focal length, principal point, lens distortion) are known and that the cameras' fields of view (fov) do not overlap (otherwise other approaches would be more appropriate). Without overlapping fields of view no direct matching of feature correspondences between different cameras is possible. One would have to match features between cameras at different times (after a given feature in one camera appears in another camera), but this is a practically difficult and error-prone approach, because appearances of features can be perspectively affected, lighting conditions could have been changed and you do not know when the feature will be visible again. Instead, we can use the mobility of the platform. Based on the Hand Eye Calibration (HEC) we can estimate the extrinsic position and orientation – under certain circumstances - given the motion of each single camera. So, a robust motion estimation is the base for this calibration approach. Hence, the solution of the calibration problem is basically devided into the problem of visual odometry and – based on that – the determination of the extrinsic parameters. Visual odometry is calculated via bundle adjustment for each camera based on corresponding image features in consecutive frames. To estimate the parameters continuously we embedded the HEC model into an Iterated Extended Kalman Filter (IEKF) that merges and filters parameter vectors. Furthermore, the IEKF calculates the uncertainties of the transformation parameters. These covariance matrices can be used for merging redundant information on a higher level. By estimating the extrinsic parameters continuously the system is able to perform a self calibration.

We assume, that the mobile platform basically moves on a planar ground, which is the most frequently performed motion by road vehicles. This

kind of motion affects the degree to which the translational parameters can be estimated with the HEC model. So, for this special case we need to adopt the well-known HEC model. To be able to estimate all three translational dof, we make the additional assumption, that all cameras see a common ground plane. This assumption can be used to estimate the longitudinal difference between the camera positions.

In practice it turned out that we could yield very good results in estimating the rotational motion, but the estimated translational parameters suffered from drifts in the translational magnitude over time. But these magnitudes are crucial for the HEC calculations and hence for the whole calibration process. This circumstance in the context of HEC has only been rarely discussed in literature. However, the motion direction (which is the scaled translation vector) could be estimated quite reliably. So, we implemented the so called Sparse Motion Adjustment (SMA). SMA is motivated by the Sparse Bundle Adjustment (SBA). It estimates the relative translational magnitudes as well as the extrinsic parameters simultaneously. With SMA we are able to perform an extrinsic calibration up to scale even with no knowledge about the translational scale.

Especially for an online algorithm, the whole system should remain scalable in the number of cameras to keep the computational effort low, even for an increasing number of cameras. Estimating all unknown parameters in a single parameter vector, containing all motion and calibration parameters of all cameras, leads to a high dimensional state vector that is increasing linearly in the number of cameras. So, the central usage of all sensor data in a single estimation step will result in computational expensive calculations. For that reason, we consider each camera as a separate processing unit that can communicate with and exchange information between other modules. The motion and calibration parameters are estimated consecutively and separately for each camera.

Most calibration routines in literature cover camera systems consisting of 2-3 non-overlapping cameras. But none of them consider the fact that 3 cameras contain more information than a 2-camera system. By explicitely considering the redundancy of a multi-camera system we can yield more robust results. Here, we make the important assumption that the transformations that describe the extrinsic camera configuration are affected by Gaussian noise. Based on the modular structure of the camera system each module produces its own and hence redundant system state. By fusing redundant state vectors, the overall uncertainty of the system can be reduced and the system becomes more robust against bad estimations of single modules (e.g. caused by a bad lighting conditions or close independently moving objects). In the presented solution, both the motion parameters (which are fundamental for the HEC) and extrinsic camera parameters can be merged. A estimation of a single camera module considers only its own (local) sensor data for the state estimation, independent of the other cameras, whereas the merged (global) state considers implicitely the information of all modules. For merging the state vectors we use Gaussian propagation and Gaussian fusion. This is why the covariance matrices of the local states play such an important role. Local states from modules with reliable estimations have a higher portions in the resulting global state than uncertain local estimations. To keep the communication bandwidth between camera modules low, only states instead of image data or features are shared between the modules.

To demonstrate the online calibration and the benefit of the fusion framework the approach was evaluated with simulated and real data.

9026-6, Session 2

Video anomaly detection for transportation applications *(Invited Paper)*

Raja Bala, Xerox Corp. (United States)

This is an invited presentation that will begin with a survey on video anomaly detection techniques that have been developed in transportation applications including traffic law enforcement, security and surveillance. A recent class of techniques based on sparse signal representations will then be described in detail. The talk will conclude with a discussion of future challenges in the field.

9026-7, Session 2

Real-time anomaly detection in dense crowded scenes

Habib Ullah, Univ. degli Studi di Trento (Italy); Mohib Ullah, University of Trento (Italy); Nicola Conci, Univ. degli Studi di Trento (Italy)

Large gatherings of people at different events present challenges of paramount importance to public safety, since such gatherings may be erupted with sudden movement arising from high-density crowd. These situations represent the abnormal behavior in terms of riots and chaotic acts of crowds. In surveillance scenarios, the early detection of abnormal behaviors occurring in the crowd can reduce the potential dangerous consequences and can also alert a human operator for monitoring the ongoing situation more effectively. Abnormal behavior detection in crowded scenes can be categorized into two types: (1) local abnormal event, indicating that behavior of crowd in local region is different from that of its surrounding regions; (2) global abnormal event, indicating that crowd is considered as a single entity instead of treating each individual separately. In this work, we address the latter issue, by considering the crowd as a single entity.

Problems such as crowd motion segmentation [1], crowd density estimation [2], and determining the goal of individuals within a crowd [3], have all been subjects of research. In many of these, however, the goal is not to analyze the behavior of crowd. In [4][5][6], for example, anomalies are detected in terms of circulation of non-pedestrian entities in the crowd by considering the variations of objects appearance to infer abnormality information. However, the main objective of crowd analysis involves not only the modeling of crowd dynamics but also to detect abnormal behaviors in the scene to ensure public safety. Limited research effort has been put into this direction. One reason for the lack of effort in this direction is the complexity of the problem.

In [7][8][9], abnormal events are detected in terms of escape panics. For this purpose, a fixed grid of particles is initialized [7] and then Gaussian Mixture Model [10] is adopted to learn the behavior of motion features extracted from the particles instead of modeling the values of all the pixels as a mixture of Gaussians. In [8], the social force model (SFM) is exploited. After the superposition of a fixed grid of particles on each frame, the SFM is used to estimate the interaction forces associated to the crowd behavior. After that, a bag of words method and a Latent Dirichlet Allocation are exploited to discriminate between normal and abnormal frames, localizing the abnormal areas as those representing the highest force magnitude. In [9], spatio-temporal interest points are exploited to detect the behavior of the crowd. For each interest point, an energy potential is calculated based on the positions and velocities of its neighbor points. However, these methods are very demanding in terms of computational cost.

In this paper we propose a novel approach to detect anomalies in crowded scenes. This is achieved by analyzing the crowd behavior by extracting the corner features. For each corner feature we collect a set of motion features. The motion features are used to train a MLP neural network during the training stage, and the behavior of crowd is inferred on the test samples. Considering the difficulty of tracking individuals in dense crowds due to multiple occlusions and clutter, in this work we extract corner features and consider them as an approximate representation of the people motion. Feature points are then advected over a temporal window through optical flow tracking.

Corner features well match the motion of individuals and their consistency and accuracy are higher both in structured and unstructured crowded scenes compared to other detectors. In the current work, corner features are exploited to extract motion information, which is used as input prior to train the neural network. The MLP neural network is subsequently used to highlight the dominant corner features that can reveal an anomaly in the crowded scenes. The experimental evaluation is conducted on a set of benchmark video sequences commonly used for crowd motion analysis. In addition, we show that our approach outperforms similar state of the art technique proposed in [7].

[1] Mehran, Ramin, Brian E. Moore, and Mubarak Shah. "A streakline representation of flow in crowded scenes." Computer Vision–ECCV 2010. Springer Berlin Heidelberg, 2010. 439-452.

[2] Ge, Weina, and Robert T. Collins. "Marked point processes for crowd counting." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[3] Mazzon, Riccardo, Syed Fahad Tahir, and Andrea Cavallaro. "Person re-identification in crowd." Pattern Recognition Letters 33.14 (2012): 1828-1837.

[4] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, in: International conference on computer vision and pattern recognition, IEEE CVPR, (2010), pp. 1975-1981.

[5] L. Seidenari, M. Bertini, Non-parametric anomaly detection exploiting space-time features, in: International conference on Multimedia, ACM, (2010), pp. 1139-1142.

[6] M. Bertini, A. Del Bimbo, L. Seidenari, Multi-scale and real-time non-parametric approach for anomaly detection and localization, Computer vision and image understanding, Elsevier 116(3) (2012) 320-329.

[7] Ullah, Habib, Lorenza Tenuti, and Nicola Conci. "Gaussian mixtures for anomaly detection in crowded scenes." IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2013.

[8] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in: International conference on computer vision and pattern recognition, IEEE CVPR, (2009), pp. 935-942.

[9] Cui, Xinyi, et al. "Abnormal detection using interaction energy potentials." Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011.

[10] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In IEEE CVPR, pages 246–252, Jun. 1999.

9026-8, Session 2

Enhancing event detection in video using robust background and quality modeling

John M. Irvine, Richard J. Wood, David Reed, Brian K. Collins, Draper Lab. (United States)

Automated event recognition in video data has numerous practical applications for security and transportation. The ability to recognize events in practice depends on precisely detecting and tracking objects of interest in the video data. Numerous factors, such as lighting, weather, camera placement, scene complexity, and data compression can degrade the performance of automated algorithms. As a preprocessing step, developing a set of robust background models can substantially improve system performance. Our object detection and tracking algorithms estimate the object position and attributes within the context of this model to provide more reliable event recognition under challenging conditions. We present an approach to robustly modeling the background as a function of the data acquisition conditions. One element of this approach is automated assessment of the image quality which informs the choice of which background model to use for a given video stream. The video quality model rests on a suite of image metrics computed in real-time from the video, whereas the background models are constructed from historical data collected over a range of conditions. We will describe the formulation of both models. Results from a recent experiment will quantify the empirical performance for recognition of events of interest.

9026-9, Session 2

Driver workload monitoring in heavy good vehicles and on powered two-wheelers

Pasi Pyykönen, Matti H. Kutila, Timo Niemeläinen, VTT Technical Research Ctr. of Finland (Finland); Andrea Saccagno, Ficomirrors Italia S.r.l. (Italy); David Daurenjou, Serge Boverie, Continental Automotive France SAS (France)

Driver monitoring systems have been developed previously in the automotive sector even if only few applications exists in the markets in these days. However, heavy vehicles are part of the daily traffic with having equal problems to the passenger car drivers. In some cases, driver's attention is even more critical due to nature of being very vulnerable in cases of traffic accidents. With heavy vehicles consequence of traffic accident to other road users is more serious compared to passenger cars. The aim is to ensure that critical safety information has been registrated by the driver. However, monitoring is more challenging since the driver needs to turn head to see out from the large cabin of a heavy vehicle.

In addition to automotive vehicles, riders of the powered two-wheelers attention is maybe even more critical due to nature of being very vulnerable in cases of traffic accidents. The aim is to ensure that critical safety information has been registrated by the rider. This means for example alerting the driver when driver's attention has been degraded or the driver is not taking into account the actual road environment. In motorcycle environment, monitoring is more challenging since the rider eyes are protected by a visor of the helmet demanding new monitoring technique to be applied.

Driver monitoring is also needed in driving simulators and driving lessons for training purposes. For driving supervisor, it is challenging to detect situations when training candidate is not focusing enough to traffic in a simulator or during driving session. With driving monitoring system, it is possible to gather data to help analysis of the education event. In this paper, we will show example how to gather data online and offline from driving session and show student how to improve driving.

In this paper, we will show and demonstrate multiple-camera based driver monitoring system in cabin of a heavy vehicle. In heavy vehicles, driver monitoring cameras are situated so that, drivers face is able to detect, even if the driver is not turning head to see out from the side windows. This implementation requires new way of adapting the driver monitoring system to detect drivers gaze direction and calculate activity index of driver awareness.

We also show how to scale this implementation for smaller environment in two-wheelers. In two-wheeler implementation, we will monitor driver only with single camera, installed on hand bars of the motorcycle. Camera is pointing to the drivers helmet, so that best possible image of drivers eyes and face is gathered. In some scenarios, rider's eyes are protected with visors of the helmet. This will block any use of camera to detect rider's eyes for calculating gaze direction. The helmet orientation is measured in the cases when facial features (eyes) are not visible. Orientation of the helmet can be calculated from orientation of static helmet components i.e. visor and reflector stamps. The main idea is to add a software module to the system which detect orientation of a helmet in cases when eye recognition is impossible. The system needs automatic adaptation procedure in order to analyze orientation of and its' influence to the rider eye contact in road ahead.

In these all scenarios the target is to calculate activity index of the driver's awareness. Activity index is calculated from the detected gaze direction changes. With multiple-camera driver monitoring system, gaze direction changes are calculated as combination of detection from each camera. In the motorcycle environment with a helmet on, activity index is calculated as combination of gaze direction and helmet orientation changes. The higher the activity index, the more aware driver is about driving situation. A low activity index is used as output for alarms from driving monitoring.

In this paper we will show, how to use and adapt driver monitoring system in all these scenarios. The final objective of developing driving monitoring system is to adapt it to the motorcycle and heavy vehicle environment. In future, the driver monitoring system can be adapted also more demanding environments i.e. race cars or work machines and all big cabins, were multiple-camera driver monitoring system would be beneficial.

9026-10, Session 2

(JEI Invited) Video retargeting based on group of frames

Chee-Sun Won, Hai Thanh Nguyen, Dongguk Univ. (Korea, Republic of)

Temporal coherence of retargeted video is an important requirement. Even a mild temporal inconsistency among subsequent frames can result in visually annoying artifacts such as jittering or flickering. To reduce these artifacts, we will assign a common seam for a group of consecutive frames (GoF). The video sequence is divided into GoF's and a representative frame for each GoF is formed. Then, both the spatial saliency in each representative frame and the temporal coherence among the consecutive representative frames are considered to determine a common seam for each GoF. Since a fixed seam of the representative frame is applied to the whole frames in a GoF, there will be no jittering and flickering artifacts within GoF. Also, since the temporal coherence between the two consecutive representative frames (i.e., two consecutive GoF's) has been taken into account, the transition between two consecutive GoF's will be smooth.

9026-11, Session 3

Person detection: state of the art and applications

Adrien Descamps, Cyril Carincotte, Multitel A.S.B.L. (Belgium); Bernard Gosselin, Faculté Polytechnique de Mons (Belgium)

In recent years, significant academic and industrial work has been about the ability to detect and recognize objects and events using images from surveillance systems. Years of research have shown that bridging the gap between the raw video data and high-level semantic concepts they contain (objects and relations between them, human activities and associated events, etc..) is still a largely open question. In this work, we will focus on the problem of person detection focusing on various surveillance contexts and applications related to the public transportation (train and metro stations).

Person detection is indeed a necessary component for the implementation of any sophisticated algorithms to monitor people, and an important amount of work has already been focussed on this problem. However, this is still considered as a difficult problem, and current methods still lack of robustness in real world conditions. The state of the art algorithms for detecting people in video can be divided into two categories: algorithms based on background-extraction methods, and algorithms that detect directly people in the images, based on their appearance and movement.

The first category of methods is based on the extraction of moving objects by background extraction; the detected moving objects are then analyzed according to various criteria (size, shape, color, movement, ...) to be classified into different categories (human , vehicle, etc.). Most of commercial solutions and analytics products currently used in video surveillance market belong to this category. If the background extraction can easily and effectively extract moving objects, it generally proves very sensitive to the environment, limited to the case of fixed cameras, and requires a calibration step.

Today, the majority of academic work in the area belongs to the second category, and uses the principle of a classifier scanned on the image. The general principle of these methods is to build a classifier able to distinguishing between the image of a person and any other picture, using machine learning algorithms. This class of techniques is generally more robust than the first, at the cost of greater complexity.

In this paper, we evaluate these two categories of approaches in real metro/railway surveillance systems, for various applications and scenarios, and analyse how they can be applied and compete in different use-cases. While the two approaches give similar results in the "easiest scenarios", the pattern recognition methods based on object detection allow in many contexts to lift the limitations of the former ones.

The two approaches are first evaluated for a people counting application (i.e. determining how many persons are visible in the image), on several video streams with limited person density (no crowd). The relatively low person density is needed for both approaches to operate effectively, because of the important inter-person occlusions that occur at high density. In this context, we compare a background-extraction methods against a state of art person detector, and notice that they can both detect quite reliably the position and the number of persons in the scene. As expected, we can note a under-evaluation when density increases, due to inter-person occlusion, and, more importantly, we note that this under-evaluation is comparable between the two approaches.

Then, we extend our experiments to crowded situations. In this case, background methods are completely ineffective, due to the occlusions. In this context, the use of appearance information is required, but usual person detector are also ineffective, because only parts of people are completely visible. Instead, a head detector is much less affected by occlusion, since the head is almost always visible, up to a very high person density (depending of the view point). We show that a counting approach based on head detection, even if less discriminative and less accurate for low person densities, can give good counting results, up to very high person densities.

Last, we focus on another application, the flow monitoring of escalators (i.e. counting how many persons per minute pass though the escalator). This application is especially challenging, because of the very wide occupancy level of level (from a few persons to a fully crowded escalator), the moving background, and the low resolution due to the distance between escalators and cameras (no specific cameras were used). Again, background based and person detector methods could not be used due to high densities, and again head detection revealed to be the best option. By combining a head detector with a simple tracking algorithm, we show that it is possible to measure with a good accuracy the people flow in escalators, even in such difficult conditions.

All these algorithms were tested on real transportation CCTV systems, and proved to be interesting from an operational point of view. First, for the station supervisors, it can help them to get an overall view of the station usage in real time, and several abnormalities in normal passenger flow could be detected with the results of the algorithm, such as an overcrowd, a disturbance on the metro lines, or a counter flow in the escalators. Second, for the station managers, it can be used to provide statistics about station usage, and also give cues to improve efficiency of the traffic.

These results demonstrate that, although the background based and detection based approaches may be comparable in the easy cases, a benefit of the latter is that they allow in difficult contexts to lift the limitations of background approaches, or even to consider news applications.

Of course, the head detector approach we adopted to solve the high density issues is only a first step to a generic person detector able to operate in any conditions, by modelling a person as a set of parts (head, shoulders, torso, legs), and seeking to independently detect each other, before merging to confirm the detection. These approaches are built to be more robust to occlusions and to changes in appearance due to the articulation of the human body, but this task, given its complexity, is still wide open in the scientific community.

9026-12, Session 3

Real-time detection of small faces in HD video

Seungji Yang, Kyung Hoon Bae, SK Telecom (Korea, Republic of)

Human face is an important biometric to identify a person from others. Face detection has been mainly used as a part of face recognition in the areas of intelligent video surveillance and access control. In recent decades, security camera has been widely spread for a diversity of public

and private purposes. We also see that HD camera gets increasing more and more on sites for the sake of lower price. From this trend, facial pictures become more eligible for having a clear quality even when their sizes are small due to capturing at a distance. Meanwhile, bigger storage is required to stack those HD data and many of the data have been rarely monitored and thrown away due to the deluge of copious information. Instead of keeping all the original data, the collection of facial data only enables to save video storage as well as to help ones browse facial events in a fast way. More recently, the usage of face detection is getting diverse to many other applications such as for auto-focusing in consumer cameras, personalized advertisement and demographic statistics in retail business intelligence, and even energy saving in building energy management system. All those systems are supposed to calculate the presence, race, gender, and/or age range of the face.

Many literatures have been successfully adopted for access control [1-7] while a few for public surveillance and retail business intelligence that have usually tackled face detection problems under wild environments [8-13]. In surveillance videos, we see that human faces often appear some different aspects compared with those in generic access control videos. There are several matters to be issued in face detection for surveillance, such as pose variation, various light conditions, and low facial image quality due to capturing at a distance. Many face researchers have been interested in the way to recognize faces in wild environments such as public surveillance. Meanwhile, the way to detect small faces at a distance from HD camera has been much little addressed in view of detection performance (represented by recall and false-alarm rates) and processing time [11-13].

The simplest way to enable real-time face detection in HD video is down-sampling the original picture before searching facial regions on that. For example, in our experiment, the face detection provided by Viola-Jones [4] took about 20ms to process QVGA-sized frame while it did more than 600ms in 720P-sized frame. The reduction of spatial resolution is obviously working well for fast processing while it usually produces loss of textures, more specially edges. By doing so, it could cause some degradation in detection accuracy. In case of surveillance video, a human face usually takes a small portion of the entire image area since they had been captured by a wide-angled, long distant camera to cover wide area surveillance. Moreover, since most of the existing face detection schemes are constrained with a large enough face size (usually 20x20 and more [4, 7]) they could fail to find the faces having a size of smaller than the minimum requirement.

Another observation is that face detection would often make higher false-alarm rate in HD videos. Like Viola-Jones approach, most of the texture-based face detections use a binary texture pattern classifier [4, 5, 7]. So the fidelity of image texture is definitely a key factor to determine the performance of face detection. As an ordinary video would keep more detailed texture of objects in HD quality, the vivid texture information could preserve better facial components in picture. But one primary side effect is that it would also increase the false-alarm rate due to texture improvement in background and non-facial objects. This can be one reason why ones make the original images smaller and smoother before yielding face detection at a close distance from camera, where there is less constraint for face size.

From the reasoning above, in this paper, we propose a novel approach for real-time face detection for wild surveillance environments and for such cases that the system should process multiple HD video streams simultaneously. The major goal is to detect small faces at a distance for HD videos as well as to preserve low false-alarm rate. The proposed face detection scheme is composed of three layers; the first one to fast scan face candidates in the resized chrominance channels by using skin color features, the second one is to precisely filter true facial objects from the face candidates which location and size are subjected to be mapped into the original luminance channel by using texture features such as Harr-like feature or local binary patterns (LBP). And the third one is to track the detected faces by comparing their color features on the skin color blobs. For the face candidate detection, we spatially reduce the entire region with high quality to small one so that we can fast filter out non-facial regions. The spatial reduction is subject to be done in such way that the minimum size of true face detectable by a certain face detector should be preserved as being one pixel at least. For example, if the minimum face size is '20x20', the spatial reduction rate could be 1/400 at most. Given the candidates, the second layer allows detecting true faces in high quality texture with aids of the positions of face candidates. In our experiment, we showed that the proposed face detection method archived about 20ms/frame in HD videos, meaning that the proposed method is more than 30 times faster than the baseline Viola-Jones one in processing time. And the false-alarm rate was extremely reduced as 1/20 of the baseline approach.

REFERENCES

[1] Rowley, H., Baluja, S., and Kanade, T., "Neural Network-based Face detection," IEEE Trans. On Pattern Anal-ysis and Machine Intelligence, 20, 22-38 (1998)

[2] Feraud, R., Bernier, O. J., Viallet, J. E., and Collobert, M., "A Fast and Accurate Face Detector Based on Neural Networks, " IEEE Trans. on Pattern Analysis and Machine Intelligence, 23(1), 42-53 (2001)

[3] Fleuret, F. and Geman, D., "Coarse-to-fine Face Detection," Int'l Journal of Computer Vision, 41, 85-107 (2001)

[4] Viola, P. and Jones, M. J., "Robust Real-time Face Detection," Int'l Journal of Computer Vision, 72(2), 137-154 (2004)

[5] Froba, B., and Ernst, A., "Face Detection with the Modified Census Transform," IEEE Int'l Conf. on Automatic Face and Gesture Recognition, 91-96 (2004)

[6] Heisele, B., Serre, T., Prentice, S., and Poggio, T., "Hierarchical Classification and Feature Reduction for Fast Face Detection with Support Vector Machines," Pattern Recognition, 36, 2007–2017 (2003)

[7] Castrillon, M., Deniz, O., Hernandez, D., and Lorenzo, J., "A Comparison of Face and Facial Feature Detectors Based on the Viola-Jones General Object Detection Framework," Machine Vision Appl., 22(3), 481-494 (2011)

[8] Jian, Z., and Wan-juan. S., "Face Detection for Security Surveillance System," Int'l Conf. on Computer Science and Education, 1735-1738 (2010)

[9] Kim, T., Lee, S. Lee, J., Kee S., and Kim S., "Integrated Approach of Multiple Face Detection for Video Sur-veillance," Int'l Conf. on Pattern Recognition, 2, 394-397 (2002)

[10] Hazar, M., Mohamed, H., and Hanene, B., "Real Time Face Detection Based on Motion and Skin Color Infor-mation," IEEE Int'l Symp. on Parallel and Distributed Processing with Applications (ISPA), 799-806 (2012)

[11] Back, K., Jang, H., Han, Y., and Hahn, H., "Efficient Small Face Detection in Surveillance Images Using Major Color Component and LDA Scheme," Int'l Conf. on Computational Intelligence and Security, 3802, 285-290 (2005)

[12] El-Barkouky, A., Rara, H., Farag, A., and Womble, P., "Face Detection at a Distance Using Saliency Maps," IEEE Int'l Conf. on Computer Vision and Pattern Recognition Workshops, 31-36 (2012)

[13] Yeom, S., and Lee, D., "Face detection at a Distance with AdaBoost Filtering and Color-shape Information," Proc. of SPIE Mobile Multimedia/ Image Processing, Security, and Applications, 8755 (2013)

9026-13, Session 3

Human behavior understanding for assisted living by means of hierarchical context free grammars

Andrea Rosani, Nicola Conci, Francesco G. De Natale, Univ. degli Studi di Trento (Italy)

Introduction – Video analysis is often used among the different approaches to detect activities and behaviors because of its lower obtrusiveness and limited cost of installation and maintenance. Location, posture, motion, as well as interaction with objects, other people and the environment are significant information that can be inferred from video data [1-2].

In many situations the information coming from video is higher with respect to the real needs of the system. In this context, simpler sensors can provide the relevant data without the need of computational expensive processing and in a more privacy-compliant way. Sensors like contact switches, pressure mats, passive infrared are able to provide information about the interaction of a user with the considered environment, thus resulting in an indirect information about his current position and movement over time, as well as on the possible actions that he is performing.

In the considered situation the different problems to be faced are challenging and can be summarized into this four statements [4]: start and end time of activity is unknown; there is ambiguity between observation and behavior; activities can be performed in different ways; data are noisy.

Semantics and context could help solve these problems, thus improving the recognition performance. In particular, long-term analysis allows discovering common patterns able to describe the behavior of the person on the basis of his previous history.

Proposed Solution - In this paper we propose a solution based on a two levels of control for real time analysis of human behavior. First we generate a grammar of actions from positive and negative samples [3]. Then a second grammar is created taking into account the sequence of actions performed in a period of observation and is used as a high level control for behavior recognition. The introduction of a higher level of control provides better accuracy in behavior understanding.

In the following, the main components of our framework are reported as a sequence of processing phases:

Preprocessing phase: processing of the incoming path; conversion into the symbolic domain.

Learning phase: training set definition; short time rule discovery; long time rule discovery.

Classification phase: parsing of the incoming string; behavior understanding.

In order to test the developed framework in a real environment, we consider as reference the “Ubicomp dataset” [4], where actions are collected using sensor network technology. It is composed by a set of data collected by 14 state-change sensors installed in a home environment where a single person lives, acquired over 28 days. The dataset has been manually annotated by the person living in the environment, distinguishing among seven activities, chosen on the basis of the so called Katz ADL index [5], commonly used as a reference for elderly care. Results presented in [4] use standard probabilistic graphical models for action recognition, in particular Hidden Markov Models and Conditional Random Fields.

Preliminary results confirm that the proposed method outperforms state-of-art techniques so far proposed.

References

[1] N. Piotto, N. Conci, and F.G.B. De Natale, “Syntactic matching of trajectories for ambient intelligence applications,” Multimedia, IEEE Transactions on, vol. 11, no. 7, pp. 1266 –1275, 2009.

[2] A. Rosani ; G. Boato ; F. G. B. De Natale, “Weighted symbolic analysis of human behavior for event detection”. Proc. SPIE 8663, Video Surveillance and Transportation Imaging Applications, March 19, 2013.

[3] K. Nakamura, “Incremental learning of context free grammars by bridging rule generation and search for semi-optimum rule sets,” in Grammatical Inference: Algorithms and Applications, Yasubumi Sakakibara, Satoshi Kobayashi, Kengo Sato, Tetsuro Nishino, and Etsuji Tomita, Eds., vol. 4201 of Lecture Notes in Computer Science, pp. 72–83. Springer Berlin / Heidelberg, 2006.

[4] Tim Van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben Kröse, “Accurate activity recognition in a home setting,” in Proceedings of the 10th international conference on Ubiquitous computing. ACM, 2008, pp. 1–9.

[5] Sidney Katz, Thomas D Downs, Helen R Cash, and Robert C Grotz, “Progress in development of the index of ADL,” The Gerontologist, vol. 10, no. 1 Part 1, pp. 20–30, 1970.

9026-14, Session 3

Human interaction recognition through two-phase sparse coding

Bo Zhang, Nicola Conci, Francesco G. De Natale, Univ. degli Studi di Trento (Italy)

1. INTRODUCTION

Recognition of human activities has emerged a hot research topic in computer vision community in recent years. The potential applications include visual surveillance, video retrieval and event detection. At the beginning, people focused on the recognition of human activities in very simple and controlled scenarios, where single person presents a certain kind of action under a static background. Only recently, research has shifted the interests towards more realistic scenarios, where human activities are collected under unconstrained conditions, like videos from surveillance cameras, YouTube, movies. This has raised a number of new issues, due to the large variability in the temporal and spatial scales, the exponential nature of all possible movement combinations, and also the mutual occlusion of different body parts, as well as background clutter.

For activity recognition, bags-of-words (BOWs) approach has been widely adopted in the past. In [1], Laptev proposes a 3D Harris corner detector to capture the spatio-temporal interest points (STIPs) in videos. Then HOG and HOF are computed to describe the motion features around each STIP. In [2], Wang et al. adopt the dense trajectory to describe motions in videos. Then, motion boundary histogram (MBH) is computed as the descriptor for each trajectory. Following the pipeline of bags-of-words approaches, and after extracting the motion features of human activities, the so-called ‘visual codebook’ is built by applying the k-means clustering on all the motion descriptors, and the center of each cluster is used to represent a visual word. Each motion descriptor in a video is then assigned the label of the closest visual word using Euclidean distance. Each video sequence can be modeled through the normalized histogram of visual words in the codebook.

However, k-means clustering is a very coarse representation of data, as each sample can only be represented by a single visual word in the codebook, therefore losing a considerable portion of essential information. In order to overcome the drawbacks of k-means clustering, sparse model [3] is adopted to represent human activities. In a sparse model, signals of a given type can be represented by the sparse linear combinations of several atoms in the overcomplete dictionary, in which the number of atoms is larger than the signal dimensionality. The sparse model consists of two parts: sparse coding and sparse modeling. In recent years, with the advent of effective overcomplete dictionary-learning algorithms [4], the sparse model has become applicable.

In our work, we are interested in human interaction categorization. Particularly, we adopt a two-phase sparse coding approach to recognize two-person interactions in realistic environment. At the first stage, we exploit the non-negative sparse model to construct the interaction-specific dictionary. At the second stage, we incorporate the label-consistent constraints [5] in the dictionary learning procedure to further increase the discrimination power of the video feature vector. The implementation details are presented in the following sections.

2. PROPOSED SOLUTION

In our work, we propose a two-phase sparse coding strategy to recognize two-person interactions in realistic videos.

Firstly, we adopt the spatio-temporal interest point (STIP) to represent the variations in videos. Then, the concatenation of HOG and HOF are used as the motion features for each STIP.

Secondly, we build the interaction-specific dictionaries by applying the non-negative sparse coding strategy. All the STIP descriptors are then encoded on the concatenated interaction-specific dictionaries. After doing sum-pooling and L2 normalization operation, we construct the feature vector for each video, with the dimensionality reduced to the number of interaction categories.

In order to further improve the classification performance, we adopt the label consistent K-SVD to learn a more discriminative dictionary at

the second phase. After encoding the video feature vectors on the new dictionary, we classify the video through a linear predictive classifier.

3. EXPERIMENT

3.1 Comparison of non-negative sparse coding and k-means clustering in video feature construction

In this section, we compare the discrimination power of video feature vectors derived from the bags-of-words approach and sparse coding. This experiment is carried out on the UT dataset (Set 1) [6]. As the number of sample videos is limited, we adopt the 50-fold leave-one-out strategy. STIP descriptors [1] are used to detect the variations in videos. The concatenation of HOG and HOF are computed as the motion descriptors. We then apply k-means clustering and non-negative sparse coding, respectively, on all the STIP descriptors to construct the visual dictionary.

3.2 The results of two-phase sparse coding

As the first-phase feature is not always discriminative enough in some complicated scenarios, we adopt a second-phase sparse coding by adding the label-consistent constraints in the dictionary learning procedure. This experiment is carried out on TVHI dataset (TV human interaction dataset) [7]. This dataset is very challenging due to a high degree of variability between videos, in terms of the number of persons in the scene, camera angle, and viewpoint. We split the dataset into training and testing sets according to the configuration in [7]. Negative samples are not included in the classification task.

REFERENCES

[1] Laptev, I.: On space-time interest points. International Journal of Computer Vision 64 (2005) 107-123.

[2] Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 3169-3176.

[3] Castrodad, A., Sapiro, G.: Sparse modeling of human actions from motion imagery. International Journal of Computer Vision 100 (2012) 1-15.

[4] Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: Proceedings of the 26th Annual International Conference on Machine Learning, ACM (2009) 689-696.

[5] Zhuolin Jiang, Zhe Lin, Larry S. Davis: Label consistent K-SVD: learning a discriminative dictionary for recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15, 2013.

[6] Ryoo, M., Aggarwal, J.: UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). http://cvrc.ece.utexas.edu/SDHA2010/Human Interaction.html

[7] Patron-Perez, A., Marszalek, M., Reid, I., Zisserman, A.: Structured learning of human interactions in TV shows. IEEE Transactions on Pattern Analysis and Machine Intelligence (2012).

9026-15, Session 4

Representing activities with layers of velocity statistics for multiple human action recognition in surveillance applications

Fabio Martínez Carrillo Jr., Univ. Nacional de Colombia (Colombia); Antoine Manzanera, Ecole Nationale Supérieure de Techniques Avancées (France); Eduardo Romero Castro M.D., Univ. Nacional de Colombia (Colombia)

A novel multiple human action recognition strategy in a video-surveillance context is herein presented. The method starts by computing a multiscale dense optical flow, from which coherent movements are clustered as Regions of Interest (RoIs) and then characterized by the corresponding orientation histogram. A multilayer structure stores the orientation histograms associated to any of the found RoIs in the scene and some cumulated temporal statistics serve to label that RoI using a previously trained support vector machine model. The method is widely evaluated using classic human action dataset and public surveillance datasets, with two different tasks: (1) classification of short sequences containing individual actions, and (2) Frame-level recognition of human action in long sequences containing simultaneous actions. The accuracy performance measurements are: 96.7 % (sequence rate) for the classification task, and 95.3 % (frame rate) for the recognition in surveillance scenes

INTRODUCTION

Human action recognition is a challenging task in many surveillance applications [1]. Previous surveillance applications are typically based on motion global descriptors that use optical flow strategies. For instance, Ikizler et al [2] used histograms of orientations of (block-based) optical flow combined with contour orientations. This method can distinguish simple periodic actions but its temporal integration is too limited to address more complex activities. In addition, Chaudhry et al [3] proposed histograms of oriented optical Flow (HOOF) to describe human activities, invariant under vertical symmetry but missing many times relevant local information. Other methods use sparse optical flow reducing the representation to a few salient features.

The idea behind this work is to recognize human actions using motion descriptors. This approach starts by computing a dense optical flow which is then spatially clustered into RoIs and described as orientation histograms. The histograms are collected within a multi-layer structure and labeled by trained SVM classifiers using some temporal cumulated statistics. Evaluation was performed with Weizmann dataset [4] and ViSOR video-surveillance database [5].

PROPOSED METHOD

The computed multiscale dense optical flow estimation [6] consists in projecting every pixel to a feature space composed of spatial derivatives of different orders, at several scales (the local jet). Then, for each frame and every pixel, the apparent velocity vector is estimated by searching the pixel associated to the nearest neighbor in the space of local jet vector at the precedent time. The dense optical flow is used to coarsely segment potential human actions by morphologically closing those pixels whose velocity norm is above a certain threshold and connecting spatio-temporally the resulting regions according to a distance criterion.

An orientation histogram is then calculated for each of the found RoIs. Unlike the HOOF descriptor of Chaudhry et al [3] our descriptor is variant under vertical symmetry. This property makes the HOOF descriptor independent to the main direction of transverse motions missing some crucial motion information. The RoI histograms are stored in a FIFO multi-layer data structure, being the x axis the computed histogram, the y axis the temporal dimension and the z axis the actions (layers) in the video. For each activity layer, if the layer contains n histograms, a set of temporal cumulated statistics (maximum, mean and standard deviation) are calculated for every orientation. Afterwards, temporal means are calculated from three consecutive intervals of histograms.

Finally, the recognition of each potential motion stored in a layer is performed using a Support Vector Machine (SVM) classifier, a one-against-one SVM multiclass classification. The SVM model was trained with a set of motion descriptors, extracted from hand labeled human activity sequences. The Radial Basis Function (RBF) kernel was used. A sensitivity parameter analysis (gamma,C) , was performed under a grid-search using a cross-validation. Additionally, a simple rule was introduced to detect complex activities. If the system had detected two simple human actions and they are grouped as a single region, then a new activity is defined and tagged as complex.

EVALUATION AND RESULTS

Different public dataset were used. The Weizmann dataset [4], composed of 9 subjects and 10 actions recorded in 93 sequences. The ViSOR: Video Surveillance Online Repository dataset, from a real world surveillance system [5]. The ViSOR dataset is composed of 150 videos with 5 human activities. In addition, several long videos of the dataset have different number of actors and activities occurring simultaneously.

In the first experiment, we tested our approach in an action classification task, using a leave-one-out cross validation scheme. Our approach achieves a 95 % of accuracy. For the ViSOR dataset, each video was divided into two parts and a k- fold cross validation scheme was used: for each split, 60 % of the data were used for training and 40 % for

testing, obtaining an averaged accuracy of, 96.7 %. The obtained results demonstrate both good performance using a very compact action descriptor.

In a second experiment, the accuracy in an action recognition task for 5 long videos was evaluated (each of 400 frames long). A first raw prediction made at each frame achieves an average accuracy of 90.81 % with a delay time of three frames. A time smoothing of the prediction was useful to improve the recognition rate, and averaging of the SVM votes for each class in a non causal interval Delta_t = [t - 1, t + 4] . Using this strategy, we achieved an average accuracy of 95.3 %.

CONCLUSIONS

This paper presented a novel approach for multiple human action recognition that uses robust multiple action representation (96 dimension) into a multi-layer data structure, and action classification based on velocity orientation statistics (average accuracy of 95.3 %).

REFERENCES

1. Aggarwal, et al, Human activity analysis: A review. ACM Computing Surveys, 2011, 43, 1-43 (16), 3.

2. Ikizler, et al, Human Action Recognition with Line and Flow Histograms, (ICPR) (2008).

3. Chaudhry, et al, Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions., (CVPR), 2009

4. Schuldt, et al, Recognizing Human Actions: A Local SVM Approach, ICPR 2004.

5. Ballan, L et al., Effective Codebooks for Human Action Categorization, ICCV. 2009.

6. Manzanera A, Local Jet Feature Space Framework for Image Processing and Representation, SITIS, 2011

9026-16, Session 4

Optical flow based Kalman filter for body joint prediction and tracking using local binary pattern matching

Binu M. Nair, Univ. of Dayton (United States); Kimberley D. Kendricks, Central State Univ. (United States); Ronald F. Tuttle, Air Force Institute of Technology (United States); Vijayan K. Asari, Univ. of Dayton (United States)

We propose a real-time novel framework for tracking the specific joints in the body on low resolution stereo vision images using optical flow based Kalman tracker. Body Joint tracking is necessary for a variety of surveillance based applications such as recognizing gait signatures of individuals, identifying the motion patterns associated with a particular action and the corresponding interactions with objects in the scene to classify a certain activity. The proposed framework consists of two stages; the initialization stage and the tracking stage. In the initialization stage, the joints to be tracked are either manually marked or automatically obtained from other joint detection algorithms in each of the stereo images and subsequently appropriate image descriptions of each joint are computed. One of the important criteria to look for in describing a body joint is that the region descriptor representing the joint should be invariant to the rotation of the joint. Another criterion is the issue of brightness of the joint region which must remain constant throughout the tracking scenario. Thus, we employ the use of a well-known image coding scheme known as the Local Binary Patterns (LBP). This image coding removes the variance to lighting conditions as well as enhances the underlying edges and corners necessary to represent that joint. Therefore, each of the stereo images are LBP-Coded before the tracking stage is commenced. The image descriptions of the joint region would then include the LBP-coded region of interest (ROI) surrounding the joint and a rotation invariant image histogram computed from the LBP-code ROI. Next the tracking stage can be divided into three phases: Optical Flow based detection of joints in corresponding lbp-coded stereo images, projection of 2D stereo image joint coordinates onto 3D world coordinates and prediction/correction phases of Kalman tracker with respect to the 3D joint coordinates. Lucas Kanade optical flow is used to locate the individual joints in consecutive lbp-coded frames for each camera in the stereo system based on their location in the previous frame. But more often, mismatches can occur due to the rotation of the joint region and the rotation variance of the optical flow matching technique. The mismatch is determined by comparing the rotation invariant lbp-coded image histogram of the joint ROI using Chi-squared metric for each stereo image and depending on this statistic, either the prediction phase or the correction phase of the corresponding Kalman filter is called. When there is no mismatch of a joint region between instant k and instant k+1, the estimated location of the joint in frame k+1 from each stereo image is transformed to the prior estimate of the location in 3D world coordinates. The Kalman filter then corrects itself using the prior joint coordinates to get the posterior or true coordinates of the joint in each of the stereo images at instant k+1. When there is a mismatch between instant k and k+1, then the Kalman filter enters the prediction phase for every instant after k. The prediction phase gives a prior estimate of the joint location in 3D world coordinates and this depends on the joint motion trajectory modeled by the corresponding Kalman Filter. As soon as the joint region gets matched correctly, the Kalman filter again enters the correction phase. The Kalman filter for each joint is modeled based on its actual trajectory in the 3D world coordinates which is mostly sinusoidal in fashion. So, the design of the tracking mechanism involves the setting of the transition and measurement matrix of the Kalman filter to take into account the non-linearity of the motion. The framework is tested on the two datasets, a private dataset provided by Air Force Institute of Technology and the other a public dataset which is the Microsoft Cambridge Kinect-12 Gesture dataset. The private dataset consists of a total of 38 video sequences captured by a stereo vision camera. Each sequence contains an individual walking across the face of the building and climbing up/down a flight of stairs. The challenges associated in this dataset are the low-resolution imagery along with interlacing effects which makes joint region description a very difficult task. Moreover, there are occlusions in some of the joints when the person climbs up due to the railing present along the staircase. The Kinect gesture dataset originally used for testing gesture recognition algorithms, contains 594 video sequences of human movements collected from 30 different people performing 12 gestures and also contains motion files which has tracks of 20 joints estimated from the Kinect sensor. These joint tracks provided by the dataset will help in validating the proposed framework since these provide the truth data similar to a VICON motion capture system and so, provides a good reference basis for system evaluation.

9026-17, Session 4

Application-driven merging and analysis of person trajectories for distributed smart camera networks

Eduardo Monari, Juergen Metzler, Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung (Germany); Colin Kuntzsch, Leibniz Univ. Hannover (Germany)

Tracking of persons and analysis of their trajectories are important tasks of surveillance systems as they support the monitoring personnel. However, this trend is accompanied by an increasing demand on smarter camera networks carrying out surveillance tasks autonomously. For instance, it is desirable that cameras slew autonomously in order to get an overlap of the camera fields of view so that an individual can be hand-off from camera to camera. In general, using non-static cameras allow a better coverage of surveillance areas without the need of additional hardware. Unfortunately, at the same time the requirements on algorithms are increasing as well, due to the higher complexity of video analysis.

In this paper, we present a system concept and application for gathering, processing and analysis of trajectories in distributed smart camera networks. The system is split into three level of data processing:

On a first level (signal level), an approach for person detection based on background subtraction is applied. Although classical detectors assume a static camera, the presented approach has been enhanced to be used for non-static cameras as well. In order to do this, the images from a video stream are registered on each other first. The result is a joint static panorama image that represents the learned background. Once the background panorama is generated, foreground (motion) segmentation is achieved by a real-time frame-to-panorama registration, followed by a classical background subtraction. After motion segmentation, connected foreground regions are detected and analyzed. They are verified for persons by analyzing their geometric properties.

On a second processing level (feature/object level), a single-camera tracking approach is presented. The tracking is performed locally by individual cameras embedded within a dynamically reconfiguring camera network. It performs a multi-object-tracking together with an explicit occlusion handling. A continuous tracking of persons across merge & split situations is important for trajectories analysis. It is solved through a merge & split detection process prior to the update of the tracks. Whenever two tracks have been registered as merged, the tracking module keeps track of the merged tracks maintaining the number of the persons inside the joint track. If a split is detected, the persons are reassigned to the corresponding single tracks again. Thus, “anonymous” identities can be retained so that there are no gaps during merge situations which is an important advantageous for the trajectory analysis.

Finally, at the third level (mulit-camera / object level), the post-processing of single-camera trajectories to global ones by position- and appearance-based fusion of them is proposed. Individual tracks from single-camera tracking, that are naturally spatially restricted to a single camera field of view, need to be integrated into common trajectory knowledge about tracked objects moving within the camera network. Adding multi-camera tracking capabilities to the system allows active tracking and re-identification of persons through the camera network by collaborative reconfiguration of pairs of cameras in order create a single trajectory spanning a spatially extended area covered by multiple fields of view. A purely geometric approach exploits properties of the spatio-temporal relations between trajectory data observed by different sensors within the camera network. By employing an appearance-based model for observed persons on top of our first solution, we are then capable of reliably tracking persons through the camera network, resulting in reliable global trajectories.

This system approach allows a multitude of analysis techniques such as inspecting individual properties of the observed movement (location, time) in real-time with or without taking into account certain static (e.g. movement through a specific shelf setup in the context of retail store layout planning) or dynamic contexts (e.g. interaction with other persons in order to identify group dynamic in a security application). Each specific domain/application of these analysis techniques requires extensive prior knowledge depending on the complexity of the required type of analysis. Additionally, the anonymous movement data allows long-term storage and big data analyses for statistical purposes. The data may even pose as context to itself, as security-applications often work with anomaly-detection approaches where a certain amount of prior knowledge is required as reference with respect to what types of movement behavior are common respectively uncommon in a stable context.

The system described in this paper has been implemented as prototype system and deployed for proof of concept under real conditions at the entrance hall of the Leibniz University Hannover. As proof of concept the system has been evaluated during a three day evaluation phase. It shows an overall stable performance, particularly with respect to significant illumination changes over hours, as well as regarding the reduction of false positives by post processing and trajectory merging performed on top of the panorama based person detection module.

9026-18, Session 4

Real time human versus animal classification using pyro-electric sensor array and Hidden Markov Model

Jakir Hossen, Eddie L. Jacobs, Univ. of Memphis (United States); Srikant Chari, Consultant (United States)

In this paper, we propose a real–time human versus animal classification technique using a pyro-electric sensor array and Hidden Markov model (HMM). The technique starts with the variational energy functional level set segmentation technique to separate the object from background. After segmentation, we convert the segmented object to a signal by considering column-wise pixel values and then finding the wavelet coefficients of the signal. HMMs are trained to statistically model the wavelet features of individuals through an expectation-maximization (EM) learning process. Human versus animal classifications are made by evaluating a set of new wavelet feature data against the trained HMMs using the maximum-likelihood (ML) criterion. Human and animal data acquired-using a pyro-electric sensor in different terrains are used for performance evaluation of the algorithms. Failures of the computationally effective SURF feature based approach that we develop in our previous research are because of distorted images produced when the object runs very fast or if the temperature difference between target and background is not sufficient to accurately profile the object. We show that wavelet based Hidden Markov Models (HMMs) work well for handling some of the distorted profiles in the data set. Further, HMM achieves improved classification rate over the SURF algorithm with almost the same computational time.

9026-19, Session 5

Vision for intelligent vehicles: holistic perception of dynamic vehicle surround and driver behavior *(Invited Paper)*

Mohan M. Trivedi, Univ. of California, San Diego (United States)

Machine vision technology is an essential enabler in the development and growth of intelligent vehicles. These vehicles can be either fully autonomous or capable of advanced driver assistance features for safe, stress?free operation of the vehicle. Designing fully autonomous robotic vehicles, which can drive on regular roads, requires accurate, reliable and efficient means for perceiving roads, obstacles and the dynamic surround of the vehicle. Design of vehicles with advanced intelligent driver assistance features need robust and accurate systems for perceiving and understanding driver state, activities, and driver intent prediction capabilities, in addition to the vehicle surround perception. We will present an overview of a “human-centered” framework for a distributed intelligent system with the driver, vehicle and environment as three key components.

We will emphasize the need and implications of utilizing a holistic approach where driving in a naturalistic context is observed over long periods to learn driving behavior and to predict driver intentions and interactivity patterns. The approach necessitates design and development of novel vision systems for simultaneously “Looking In and Looking Out” (LiLo) of a vehicle. The presentation will include an overview selected ongoing studies and will conclude with some pointers to important outstanding research challenges.

9026-20, Session 5

Optimizing video mosaics for short-term change detection by UAV

Günter Saur, Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung (Germany)

The change detection of e.g. parking cars is a task of short time scale, i. e. the observations are taken in time distances from several minutes up to few hours. Each observation is a short video sequence corresponding to the near-nadir overflight of the UAV over the interesting area. In our previous work, several algorithms for image change detection are compared [1], and an specialized approach for video mosaics has been developed [2].

Our actual work aims on optimizing video mosaics for change detection. The main steps are (a) filtering the metadata by adapting the direct geocoding information to the estimated interframe-transformations, (b) adapting the geometry of the mosaic by a flexible transformation chain and finally (c) adapting the intensity and colour inside the mosaic. These steps have to incorperate all mosaics (two or more) which serve as input for the change detection.

Our results show, that by this method, the area that is covered by the video mosaics and used for change detection, could be enlarged significantly without degradation of the change detection.

[1] G. Saur, W. Krueger, “Short term change detection for UAV video”, SPIE remote sensinge Europe, Edinbourgh 2012

[2] G. Saur, W. Krueger, “Extended images differencing for change detection in UAV video mosaics”, to be presented at SPIE remote sensinge Europe, Dresden 2013

9026-21, Session 5

Real-time traffic jam detection and localization running on a smart camera

Yuriy Lipetski, Gernot Loibner, SLR Engineering GmbH (Austria); Michael Ulm, Wolfgang Ponweiser, AIT Austrian Institute of Technology GmbH (Austria); Oliver Sidla, SLR Engineering GmbH (Austria)

Introduction

Reliable traffic jam detection is of big significance for traffic flow analysis related applications. Still, various outdoor conditions make robust long-term operation a challenging task in terms of both missed detections and false alarms. Intelligent state-of-the-art computer vision algorithms need to be used, resulting in growing complexity of computing. This makes the goal of real-time detection complicated and hard to achieve – especially when the system runs on a smart camera with a low power processing unit.

We present our work aimed at the application of real-time traffic jam detection. The method is designed to be operated on a smart camera, but is also suitable for a standard PC. The camera described here is a custom built smart camera with processing power of 1.6 GHz (single core + hyperthreading). The sensor used has a resolution of 1024x768 pixels. It is sensitive enough for day and night operation; street light during night time is sufficient for the camera to operate well in most situations.

The system is applicable to be used at any common camera mounting position above the road. It can observe either one driving direction or both directions simultaneously. Since the detection algorithm needs only a modest pixel resolution per lane, it can monitor several lanes at once, the method is even able to handle roundabouts. The vehicle resolution in pixels is of no big importance – i.e. the resolution is allowed to be quite poor.

System setup

Before the system can start to operate, road geometry has to be specified once. It is desirable (but not necessary) to calibrate the scene – i.e. to describe mathematically the relationship between camera and world coordinate systems. We use the Tsai calibration algorithm for this purpose. After that each traffic lane will be split automatically on overlapping segments with equal lengths. In such way we are not only able to detect the presence of a traffic jam, but also to localize it by lane(s), start position(s) and end position(s) in world co-ordinates.

Traffic jam detection algorithm

A traffic jam can be roughly described as combination of two states of the traffic:

i) the presence of a large amount of vehicles (vehicle detection algorithm)

ii) the absence of the motion within the road (motion detection algorithm)

Our algorithm tries to measure and combine the two above mentioned properties of traffic and to flag a traffic jam if certain conditions are met:

a) Vehicle detection

Our detection approach is based on appearance based modeling with area based feature computation and a subsequent classification algorithm. For the feature extraction purpose, two different methods are used:

i) Local Binary Patterns (LBP)

ii) Histogram of Oriented Gradients (HOG)

The feature vectors are then classified using Support Vector Machine (SVM) classifiers. The classifier outputs will be merged at the end as AND operation. A classifier itself is one of the following types:

i) Individual pre-trained classifier that was built specifically for the camera geometry

ii) Generic pre-trained classifier that was built using training data from various scenes

iii) Classifier with online learning approach

In the evaluation section we give a comparison of proposed feature extraction methods and classification approaches.

b) Motion detection and analysis

To detect motion, we utilize the well-known Kanade-Lucas-Tomasi (KLT) feature vector approach. We extend it so that we store only trajectories that are valid for at least N successive frames (with N = 5 usually); each trajectory is analyzed to filter noise. Each road segment is considered independent of each other, the average velocity within each segment serves as output value either in km/h if calibration is available or in pixels/sec otherwise.

Achieving real-time operation

Altogether, the described above algorithms are not applicable to run combined in full real-time (continuously with at least 15 frames per second) on the smart camera where the hardware power is limited. Instead, the following technique is used:

i) Acquire N successive frames into a buffer. N has to be large enough to allow a robust decision about presence and localization of a traffic jam as well as motion;

ii) Process acquired frames

a) do vehicle detection per road segment

b) do motion detection per road segment

c) combine results to make decision about traffic jam presence in the individual segments

d) combine all segment results and update global traffic jam information

iii) Skip frames that were lost during processing part

In such a way we obtain traffic jam information which is discrete in time, for a loop time of approximately 12 seconds. This is enough for practical applications, moreover, we are able to handle the “stop and go” behavior pretty exactly. The amount of skipped frames is variable and depends on the capabilities of the hardware: on a modern Intel CPU the algorithm could run fully in real-time and provide continuous output without breaks.

Evaluation results

The automatic visual traffic jam detection has been evaluated with respect to traffic mode estimation using floating car data (FCD) and a model derived from this data:

• the FCD method gathers traffic information from vehicles which send their position in real-time to a server which aggregates this into an estimate of current traffic speeds.

• a new modeling technique then estimates the occurrence and location of traffic jams using the FCD information aggregated over links (street segments).

All three methods, i) visual observation, ii) FCD data and iii) traffic jam modeling are computed and compared with respect to timing and location of traffic jams on the street segment which is observed by the SLR smart camera.

Despite differences in the noise level, all three methods agree very well on their results.

Outlook

Here, the results achieved so will be demonstrated and discussed. We will show evaluation results in the city of Vienna from the operation of two smart cameras over an extended period of time. In addition, we present preliminary results from a traffic monitoring web camera which has low resolution (VGA) and poor video quality. Finally, the known drawbacks and the next steps will be discussed.

9026-22, Session 5

Video-based traffic monitoring and prediction

Shraddha Chaudhary, Indian Institute of Technology Delhi (India); Vikash K. Maurya, Indu Sreedevi, Delhi Technological Univ. (India); Santanu Chaudhury, Indian Institute of Technology Delhi (India)

In this paper a novel approach for the traffic state prediction based on dynamic Bayesian network is presented. Existing methods are using embedded loop detectors for traffic state identification. A camera based system, as it is more accurate, simple and easy to install is proposed in this paper.. The main challenge of camera based system is real time processing, as real time processing of a video is computationally expensive and the techniques available are very complex and involves image segmentation which hence increases the complexity, therefore a novel approach is deduced in which Spatial interest points (SIP) and spatio-temporal interest points (STIP) . SIP are points in spatial domain with significant variation in local intensities whereas STIP are points in space time domain with significant variation in local intensities. It is observed that images of vehicles on a road generate SIP and the images of moving vehicles generate STIP. Hence, the number of SIP is indicative of number of vehicles on a road and ratio of STIP to number of SIP is suggestive percentage of moving vehicles. Traffic on any road can be completely defined by the number of moving vehicles and their average velocity. But these two features depend on each other. Therefore we classify the road states by comparing the no. of vehicles with the number of moving vehicles. If the no. of vehicles in a road link is ? and the no. of moving vehicle is ?, we can classify the road state ?? domain. The major classification considered is: Stopped (S), Heavy congestion (HC), Slight Traffic (ST), and Open (O) .We can extract SIP and STIP from the video frames recorded by the camera. The classification of the traffic state can be done based on the ratio of STIP to SIP as the ratio will give the indication of the state of the road. Using the spatial (?sp) and spatio-temporal interest (?pt) points we can classify the road state by using Gaussian mixture model .Let the feature vector be S = (?sp, ?pt) , no. of classes be k and no. of Gaussian mixtures be n. The conditional probability that belongs to Kth class can be calculated using GMM-EM method.

After simulating this formula and maximizing the expectation, we got 4 Clusters which tells us about the 4 classes that we have specified with different mean and variance values.

After classifying the extracted features in real time, next major concern is traffic prediction .It is a real time process and hence requires system to be self-evolving. One may be interested either in online analysis, where the data arrives in real-time, or in offline analysis, where all the data has already been collected. In online analysis, one common task is to predict future observations, given all the observations up to the present time. Since we will generally be unsure about the future, we would like to compute a best guess. In addition, we might want to know how confident we are of this guess, so we can hedge our bets appropriately. Hence we will try to compute a probability distribution over the possible future observations.

In this paper we are proposing an algorithm for computing the future states of the main road depending on the current and prior states observed on the neighbouring road links and the main road. At every junction state of the main road is changed due to change of the states of the neighbouring road links. Depending on the priors and the likelihood of the particular state , we can predict the future state for the system. After predicting the future state results are validated using the simulator(Synchro studio 8) and even in the real time scenarios and the results are obtained with the accuracy of 95%.

Results:

Open Slight Congestion Heavy Congestion Stopped

Open 97.48 4.83 0 0

Slight Congestion 1.89 94.84 0.71 0

Heavy Congestion 0.63 0.23 96.82 1.33

Stopped 0 0 2.47 98.67

Results obtained are even validated for various ambient conditions (snowy, rainy and dark) by taking videos at random from the YouTube, and with the slight tuning in the parameters of the algorithm .Results thus obtained proved the system highly robust and adaptive.

9026-24, Session 5

A feedback based method for binarization of license plate images

Mandar Sovani, Subhash Challa, The Univ. of Melbourne (Australia) and SenSen Networks Pty Ltd. (Australia); Marimuthu Palaniswami, The Univ. of Melbourne (Australia); Duc Vo, SenSen Networks Pty Ltd. (Australia)

Automatic number plate recognition (ANPR) systems have gained a lot of importance recently because of myriad applications that they are used in, such as traffic infringements, automated toll processing, parking lot management and security related applications. The significant problem that the ANPR systems have to deal with is the variations in the Number plates across the world as well as within the same regions. For e.g., Number Plates from USA and Australia have significant difference in terms of styling, plate dimensions and number of characters and the pattern of the characters. Within Australia itself states such as Victoria and Western Australia have significant differences in the number plates. Thus the ANPR systems have to adapt to the change in the patterns without compromising on the accuracy.

Generally ANPR systems perform the the steps of plate location and cropping, segmentation and Optical Character Recognition(OCR) in a serial order to read the plates. In countries such as Singapore, India where majority of the license plate are non retro-reflective, location and cropping of plates is a hard problem. A good crop relies on the edge information of the plate, since the plate are non-retro this task is done on color images instead of infra-red(IR) illuminated images. As there are many confusing edges in the image, plates are generally under-cropped from the whole image. Correct binarization of such plates for the segmentation is of paramount importance, as any error in this trickles down to the final read.

Generally two types of thresholding methods are considered quite reliable with very few disadvantages: Global and local thresholding. We review the results of global thresholding methods based on Otsu, Kapur, Kittler's algorithms; and that of local thresholding methods based on Niblack, Bersnen, Hysteresis and adaptive threshold. Particularly in the case of under-cropped plate images the global thresholding methods fail to binarize the characters correctly because they work best when the distribution of the pixel intensities is bi-modal. While the local thresholding methods affect the shape of characters and can sometimes produce holes in the character region in the binary image. We

demonstrate these limitations by comparing the binary output images of these methods for a few selected plate images.

Thus we observe that the global thresholding method ensures the uniformity in the characters but is at risk of completely losing characters if an intensity other than the plate region is skewing the histogram. On the other hand local thresholding methods ensure that all the characters are binarized but in this process the character shape can be compromised which affects the final OCR. Based on this knowledge we form a feedback loop where the image binarization is done using adaptive thresholding(local) method. This binary image is passed through the next stages of segmentation and OCR, while storing the OCR likelihood scores for each character. Assuming that the local thresholding method has succeeded in segmenting all the characters, we utilize the spatial information of the characters obtained after segmentation to re-crop the plate image keeping leeways to avoid over-cropping. Now we use one of the global thresholding methods like Otsu on the newly cropped image for binarization.

Even after re-cropping there can be variations in the image due to artifacts like shadows or lights which make the pixels intensities into a multi-modal distribution. In order to avoid this we divide the plate image into multiple regions and calculate the threshold of each region using the Otsu method. If the standard deviation of these thresholds is above a certain threshold T then we binarize the each region independently. In other cases we binarize the whole image using the median value obtained from the thresholds of all regions. Most of the times the second method is preferred as the independent thresholding of regions can produce same errors like local thresholding methods. The output binary image is again passed through the segmentation and OCR modules. In most of the cases the OCR result after this step is fairly accurate, but to accommodate some tolerance we select the read by comparing the OCR score with the one generated from local thresholding.

The results section shows that the proposed method helps in recovering approximately 8% of incorrect final OCR reads which are produced if the global or local thresholding methods are used independently. Repeating the segmentation and OCR modules for the feedback does not affect the timing constraints too much, but the local thresholding methods can be computationally intensive thus affecting the overall read time. We demonstrate the implementation of the adaptive thresholding method on Nvidia based GPU using CUDA architecture based parallel programming which speeds up the operation to produce results within tenths to hundredths of a millisecond. The proposed method is deployed on live sites doing ANPR in Singapore for toll management and vehicle infringements.

9026-25, Session 6

(JEI Invited) Video-based real-time on-street parking occupancy detection system

Orhan Bulan, Robert P. Loce, Wencheng Wu, Yao Rong Wang, Edgar A. Bernal, Zhigang Fan, Xerox Corp. (United States)

Urban parking management is receiving significant attention due to its potential to reduce traffic congestion, fuel consumption, and emissions. Real-time parking occupancy detection is a critical component of on-street parking management systems, where occupancy information is relayed to drivers via smart phone apps, radio, internet, on-road signs, or GPS auxiliary signals. Video-based parking occupancy detection systems can provide a cost-effective solution to the sensing task while providing additional functionality for traffic law enforcement and surveillance. In this paper, we present a video-based on-street parking occupancy detection system that can operate in real-time. Our system accounts for the inherent challenges that exist in on-street parking settings including illumination changes, rain, shadows, occlusions, and camera motion. Our method utilizes several components from video processing and computer vision for motion detection, background subtraction, and vehicle detection. Our experimental results show that the proposed parking occupancy detection method operates at 5 frames per second and achieves better than 90% detection accuracy across several days of videos captured in a busy street block under various weather conditions such as sunny, cloudy, and rainy, among others.

9026-26, Session 6

Automatic parking lot occupancy computation using motion tracking

Francisco Justo, Hari Kalva, Daniel Raviv, Florida Atlantic Univ. (United States)

1. INTRODUCTION

The proposed work is intended to find a way to address the problem of finding parking spots available in a specific parking lot. A simple but expensive way to solve this problem is by using specialized networked sensors that are placed in each parking spot [1], [2], [3]. The main drawbacks of such systems are high cost of installation, long installation times, and high maintenance costs.

Low cost cameras combined with software that allows approximate computation of the number of spots available will provide a cheaper solution to the problem. We propose a low complexity algorithm that relies on camera calibration. We studied the performance of the algorithm at various video resolutions, frame rate, and quality. Our results show that the performance of the system remains high for fairly low quality video. This allows the development of systems that can process large number of video streams.

2. METHODOLOGY

This system will require the use of a camera focusing the target parking lot. Our application will capture images in real time from the camera and will process it using our algorithm applying motion tracking. Internally the proposed algorithm performs blob tracking in which the blobs in the images are classified as cars (base on defined criteria). Then, based on the movement of blobs in the video (transitions in space), we are able to count the number of cars that enter/leave the parking lot and keep track of the number of cars in a given parking lot of interest. The occupancy information is stored in a database for use by external parking management and navigation applications. The entire process of our proposed system is illustrated in figure 1.

The proposed system consists of 4 principal parts: calibration, tracking, classification, and transition process. The system was implemented using OpenCV (Open Source Computer Vision Library) [4]. A Prosilica 2040 camera which has a resolution of 2040x2048 was used for video capture. In contrast with the system proposed by Wu and Zhang [7] that take images of vehicles parked in selected parking spot areas and then analyzed using SVM classification and with [5] that uses a FCM classifier, our proposed algorithm tracks the moving objects in order to detect the vehicle when it goes in or out of a parking lot, which should have lower computational costs than the previously mentioned studies.

3. PERFORMANCE EVALUATION

We recorded a video of an active parking lot on our University campus over a 24 hour period. The area of the parking lot monitored is located outdoors, uncovered, and has a capacity of 20 cars. It is important to mention that the video used was taken from a regularly used parking lot and not staged. The video was recorded in Boca Raton, Florida in springtime. The strong shadows caused by the bright sun, the weak shadows caused by moving clouds, trees moving in the wind, all posed challenges that were successfully addressed by appropriate preprocessing and filtering. A Prosilica GC 2450 camera was used, which gives us a resolution of 2448x2050 full video with 15 fps. In order to fit the desired portion of the parking lot and increasing the frame rate, the resolution of the camera was reduced. The best resolution to capture a large portion of the parking lot was 2000x650. The video was encoded using AVC/H.264 high profile video encoder. From the 24 hour video recorded, approximately 10 hours of daytime video (7:50 AM to 6:00 PM) that contained moving vehicles was used to measure the performance of our algorithm. In order to find the minimal bandwidth and resolution requirements, we also encoded and evaluated the video at lower resolutions and bitrates. The final data set used for performance

evaluation consisted of 56 variations of the 10-hour video (coded at various resolutions and bitrates).

4. EXPERIMENTS AND RESULTS

In order to test the performance of the algorithm, the accuracy at different time instances along the video was measured. The performance was calculated by comparing the occupancy results against the ground truth every 10 minutes. Figure 2 shows results of the algorithm performance for the video with 2000x650 of resolution and a bit rate of 44Kbps (low quality). It can be observed that the average performance is 95.85% and is sufficiently high for parking applications. The performance of the algorithm drops when the lot was fully occupied due to issues such as blob mixing that have not been addressed in this algorithm.

To measure the impact of the video quality, the original video was encoded using different resolutions and bitrates. Based on those results, it can be seen that the algorithm performance remains relatively high even when its quality is decreased to 30db. On the other hand, it was observed that the most influencing factor is the video resolution. Furthermore, the performance drops considerably as the resolution of the image is scaled. The main reasons for the performance degradation was that the thresholds are not dynamically adjusted for the changing resolutions and bitrates. Additional work is necessary to make this system adaptive to varying resolutions and bitrates.

5. PERFORMANCE EVALUATION

The computational performance was calculated based on the time that the algorithm takes to run the 10 hours video. It was observed that computational requirements drop significantly when the resolution of the video is reduced. On the other hand, there is not much difference in computational requirements when the quality/bitrate at the same resolution drops. After analyzing the results we can say that the algorithm performs well even if the video quality is decreased to a very low value. However, reducing the resolution negatively affects the performance. Based on these observations it is clear that we can use relatively cheap cameras and low bandwidth networks to develop large scale parking occupancy applications. Table 1 shows the computation time for 10-hour video at various resolutions. Table 1 shows the capacity increase (complexity reduction) as a result of resolution and bitrate reduction.

6. CONCLUSION

The motion tracking based algorithm presented allows us to determine the parking lot occupancy in a given parking lot. To measure the performance, an evaluation process was developed and applied. The performance was measured by comparing the results to the ground truth. The influence of video resolution and bit rate were also studied. The results observed indicate that motion based tracking can perform very well even with very low quality videos. Most of the issues were because camera angle affects performance. The use of dome cameras would help avoid occlusions and mixing blobs.

Our experimental results showed that the proposed algorithm performs well and that the algorithm behaves well with a low quality video (44kbs). It was observed that the impact of bitrate reduction on the performance was minimal. However, performance quickly drops as resolution is decreased. It was observed that the impact of video resolution reduction affects performance but it was noticed a trend which leads us to think that the thresholds scaling applied is not linear. The results show that relatively inexpensive and low bandwidth networks can be used to develop large scale parking occupancy applications.

REFERENCES

[1] X. Li and U.K. Ranga (2009). “Design and Implementation of a Digital Parking Lot Management System”, the 2009 Technology Interface Journal, Fall 2009.

[2] ST Electronics, “Agilsense Wireless Parking Lot Detection System”, Retrieved on June 13th, 2012, from http://www.stee.com.sg/group/satcom/product/sensor/agilsense_parking_system.html.

[3] MeshNetics, “Parking Lot Gets Smart with ZigBee”, Case study retrieved on June 20th, 2012, from http://www2.ee.ic.ac.uk/t.clarke/projects/Resources/ZDK_v2.0_Complete/Product%20Information/M-253~02-(ZigBee%20Parking%20Automation%20Case%20Study).pdf

[4] Bradski, G. and Kaehler, A., “Learning OpenCV”. First edition, O’Reilly Media, Inc. California, USA, 2008.

[5] Ichihashi, H., Notsu, A., Honda, K., Katada, T., and Fujiyoshi, M., “Vacant Parking Space Detector for Outdoor Parking Lot by Using Surveillance Camera and FCM Classifier”, IEEE International Conference on Fuzzy Systems, Sakai, Japan, 20-24 August 2009, pp 127-134.

[6] Liñál, C., “cvBlob” Retrieved from cvBlob website: http://cvblob.googlecode.com (December 2012).

[7] Wu, Q. and Zhang, Y., “Parking Lots Space Detection”. Machine Learning, Fall 2006, Carnegie Mellon University, 2006.

9026-27, Session 6

Methods for vehicle detection and vehicle presence analysis

Oliver Sidla, Yuriy Lipetski, SLR Engineering GmbH (Austria)

This paper presents our work towards robust vehicle detection in dynamic and static scenes from a historical perspective up to our current state-of-the-art.

The robust detection of vehicles is of great importance for ITS applications,especially traffic monitoring and traffic flow analysis. Still we are some steps away from an optimal, real-time capable algorithm which is able to detect vehicles completely pose and rotation invariant.

Scale invariance can be handled relatively easy by the methods explained in this work, although this may imply considerable more computing time.

We use the vehicle detection methods elaborated in this work for

- speed/flow measurements

- law enforcement systems

- parking lot management/monitoring applications

- verification of vehicle presence in law enforcement and monitoring applications

- vehicle detection for congestion analysis

- vehicle tracking for trajectory and flow analysis

In contrast to the still widely used methods of background modeling and blob analysis for object detection in surveillance systems, our work has early on concentrated on appearance based methods. Although this approach has short term drawbacks because it is more complex, it promises much better detection capabilities in the long term.

This work gives an overview, including examples, of the different algorithms which have been tested and used for vehicle detection. It must be noted here that we mainly focus on detection of vehicles from rear and back frontal views, with a rotation of not more than approximately 30 deg:

- PCA:

the first implementations used PCA with modifications in the sample preprocessing in order to make detection more robust. De-emphasizing the outer regions of a detection window using a Gaussian mask leads to a significant improvement in detection accuracy. In addition to the basic feature design, we have tested several classifiers approaches in this work, with k-NN and linear SVM delivering the best results.

- HOG with features trained using AdaBoost: improving on the classical single stage HOG detector approach, we have implemented a multi-stage HOG detection cascade which has been trained using AdABoost, resulting in a very significant speedup as well as an improvement in recognition quality. The detector resulting from this training can run practically in real-time on a low power Atom processor and has been used in several real-world applications.

- HOG combined with LBP: The addition of new feature dimensions by adding Local Binary Patterns (LBP) improves the recognition performance of our vehicle presence detector still further. Although it needs to be trained for a specific location (geometry), the combination of HOG and LBP can achieve almost 100% accuracy even in adverse and varying outdoor conditions. We briefly present a real-world application (parking lot monitoring) which has been setup very successfully with this approach and was tested in a demanding outdoor environment.

- genetic optimization of HOG stages of a cascade: In order to use the full potential of the combination HOG cascade + linear SVM classifier we are optimizing the features for each stage of the cascade using a genetic algorithm framework (GA). This section describes the GA setup, the optimization procedure and the approach taken for the fitness function.

Special care has to be taken in order to use the right fitness criterion of each HOG detector in the GA process to avoid overfitting on the training samples.

In order to reduce the complexity of the GA training procedure we define the number of cascade stages as well as the number of features to be used in each stage. Still on a multi-core PC an optimization run may take several hours to days of computation.

We consider the latter two methods (cascaded HOG, HOG + LPB) as state-of-the-art and give details into their training and application. An analysis and direct comparison of our AdaBoost trained HOG detector and the GA trained HOG detector will be given.

Long term outlook

The paper will conclude with a discussion of results achieved so far.

The next steps taken in algorithm development will be discussed and outlined, considering currently known drawbacks and known potential for improvement.

Since for some detectors, especially the combined HOG+LBP approach, a camera specific training gives best results, we will emphasize on eliminating this time consuming step.

Fully automatic optimal training with online learning methods as proposed by Roth/Bischof will be discussed in this section.

9026-28, Session 6

Object instance recognition using motion cues and instance specific appearance models

Arne Schumann, Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung (Germany)

Object instance regocnition is a high level task in the image processing chain following image acquisition, image preprocessing, object detection and object classification. The key difference to detection and classification is that the object is already located within the image (i.e. detected) and possibly recognized as belonging to a specific object type or class (i.e. classified) and different occurrences of the same instance of an object must now be recognized as such. The task is a much more challenging one since within-object-class variances are often very small.

Two major areas of application for object instance recognition are the surveillance and security sector where specifically person re-identification is an important task in larger camera networks and the military sector where object instance recognition (e.g. of vehicles) can be applied in aerial videos recorded by unmanned aerial vehicles (UAVs).

As a popular special case of this problem we will in this work investigate instance recognition for persons – or person re-identification – using surveillance video data. To that end we will compare the suitability and discriminative power of a number of color features (most notably fast color histograms, local descriptors and color structure descriptors) which capture an objects principal colors and the way color is distributed on the object's surface as well as a number of texture descriptors that encode information of textural patterns on objects (most notably gabor filters and histograms of local binary patterns). We use these features together with well-known feature matching techniques to establish a baseline system for object instance recognition. We also apply the same approach to aerial images of objects to investigate the versatility of the features for such different kinds of data.

On top of this baseline approach we use object movement to gain robustness to variation in viewpoint. In surveillance camera networks – even uncalibrated ones – the movement direction of a person can yield information about the camera viewpoint relative to the person. Assuming an object always moves in the direction it faces, the relative angle of the camera to the object can be estimated. Based on this information, the instance recognition can be made more robust to viewpoints by weighting matches between object occurrences stronger, if they are close in viewpoint. For example in the case of person re-identification this is useful when a person wears an open jacket or a backpack and their frontal appearance looks different from their backside. While matching scores may in these cases suggest that both instances are not the same person, knowing the difference in viewpoint can help lessen the impact of this difference. In an aerial perspective, the viewpoint does not change much but the rotation of the object might (due to the usually moving camera). Object motion information can in this case help make features more robust to this rotation.

We further improve on the baseline feature matching approach by including a machine learning component. Based on one or more tracks that are known to belong to the same object, an appearance classifier is trained specific to that object. This classifier is then used to reevaluate tracks that were previously (and possibly wrongly) identified as belonging to the specific object. The initial set of tracks that are known to belong to the same object and can thus be used as positive samples for training may only contain one single sample. However, larger sets can be obtained by location constraints given in calibrated surveillance camera networks (i.e. two tracks from different but overlapping cameras of the same person walking in the overlap region can with high confidence be assigned to the same set). Another way to generate such sets in real-world applications is operator feedback in which case an operator is presented with an initial ranked list of possible matches to a query object and selects true and false positive samples. Based on these semi-automatically determined labels, a classifier is trained. The resulting classifier will complement the weaknesses of the baseline approach because it was specifically trained on the false positive samples. To automatically evaluate our system, we use groundtruth to simulate operator feedback. We train multiple types of classifiers (SVMs and Boosting Classifiers) and evaluate their performances after one and more rounds of operator feedback.

Following the two main areas of application, we evaluate our approach on two well known public datasets. The CAVIAR dataset (http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/ second set) provides annotations for persons in a shopping center's surveillance camera network. Actors played out a number of scenarios for this data resulting in many reappearances of persons and thus allowing for the evaluation of person re-identification approaches. Preliminary experiments for our approach show a significant improvement over the baseline configuration.

Secondly, we evaluate our approach on aerial recordings of vehicles from the VIVID dataset (http://vision.cse.psu.edu/data/vividEval/datasets/datasets.html). Since there is not a great number of vehicles in this dataset and they almost never reappear, we cut the annotated tracks into multiple sequences to allow for an instance recognition evaluation. We leave large gaps between the cut out sequences to allow for greater changes in appearance and to get more realistic results in the evaluation. Again, a comparison to our baseline approach shows a significant improvement.

Unfortunately we know of no dataset that contains aerial recordings of reappearing persons on which to evaluate person re-identification from the air. However, we intend to record such data using a UAV at moderate height and perform the corresponding evaluations before the publication of this work.

In conclusion, we present a learning-based object instance recognition approach that yields good reults for person re-identification on surveillance data and shows a degree of versatility when being applied to other, much more challenging data, such as aerial recordings from great height. The performance of the approach is demonstrated on a well known surveillance dataset as well as a public dataset of vehicles recorded from the air.

9026-29, Session 7

Real-time change detection for countering improvised explosive devices

Dennis W. J. M. van de Wouw, Technische Univ. Eindhoven (Netherlands) and ViNotion B.V. (Netherlands); Kris van Rens, Hugo van Lint, Egbert G. T. Jaspers, ViNotion B.V. (Netherlands); Peter H. N. de With, Technische Univ. Eindhoven (Netherlands)

1. INTRODUCTION AND PROBLEM STATEMENT

Improvised Explosive Devices (IEDs) are one of the main causes of casualties amongst NATO troops during transportation of people and materials in conflict zones. In order to reduce casualties, periodical surveillance of the high-risk areas is required where such transportation is planned. One effective method of surveillance is ground-based patrol. During such patrols, potential threats are localized by searching for suspicious patterns in the environment (e.g. suspiciously placed objects, markers etc.) and comparing current and past environment situations. This comparison usually has to be made while the convoy is moving, by military personnel that needs to remember the environment state, as observed during a previous patrol. Any suspicious change in the environment (new objects, moved structures, etc.) may form a potential threat, so the soldier is effectively performing a manual intelligence task focusing on change detection to find possible IEDs. This is a very demanding task for a human, because his ability to concentrate on a task for a longer interval is limited. Furthermore, memories of details about the appearance of a specific environment is hard if the time and distance differenceis significant. This paper proposes a real-time change detection system using automated image analysis. It can aid military personnel in detecting IEDs and thereby help prevent any accidents during surveillance. The development of such a system poses several technological challenges such as spatio-temporal localization, the proper visual interpretation of the scene and image comparisons, where this processing should be performed in real time.

2. RELATED WORK AND CONTRIBUTIONS

Our change detection system combines several existing concepts (i.e. feature extraction, registration, adaptive thresholding, Markov Random Fields), where modifications and/or improvements are applied to each processing stage to improve suitability for the change detection task. Moreover, our system distinguishes itself by applying change detection to high-resolution imagery from a moving vehicle, whereas common change detection algorithms are based on stationary cameras and related imagery. In this, the work of Primdahl et al.[1] most resembles our work, as they propose an approach to detect changes from non-stationary cameras. However, they focus on a small known area of 6 ? 6 meters while our system is designed for and tested on long real-world sequences.

Although we use a similar approach, we use different techniques to obtain our changes.

We also propose a pipelined implementation which allows for real-time processing on a PC with off-the-shelf components. Finally, we discuss the drawbacks of a monocular change detection system and discuss future improvements in the form of a stereo-based change detection system.

3. OUR APPROACH

3.1 Video and Position Capturing & Feature Extraction

We capture images with a state-of-the-art camera with a maximum resolution of 20 Mpixels per image and we employ a GPS-INS system for obtaining position information. While driving, (video) snapshots of the environment are stored in a database alongside their extracted features and GPS location.

3.2 Search & retrieval

Search and retrieval is the process of finding the best matching historic image (i.e. recorded during a previous patrol) with respect to a newly acquired image. Fast image retrieval is facilitated by indexing on the GPS position and driving direction.

3.3 Image registration

The matched features between historic and live images are the basis for image registration, which is necessary to compare historic and present images on a pixel-by-pixel basis. All matched features are assumed to reside on the ground plane. Hence, the ground plane is the 3D plane for which the homography matrix is calculated. This global homography transform is then used to warp the reference frame to the live frame, that is, also the parts that are not on the ground plane. Due to the usage of a single global homography, local warping errors may occur for objects that are not on the ground plane. By orienting the camera slightly downwards, the ground plane is the dominant plane in the scene and warping errors are only moderate.

3.4 Change mask generation

First the difference between the (warped) historic frame and the live frame is computed at the pixel level. The resulting gray-scale image is then thresholded by an adaptive thresholding technique.[2]

3.5 Post processing

The changes resulting from the aforementioned steps are still noisy and inconsistent in space and time. To distinguish relevant, consistent changes from noise, we apply a Markov Random Field (MRF) method, based on energy minimization.[3] After MRF processing, false detections are reduced by investigating their texture in terms of intensities values. This is achieved by employing template matching. Finally, the temporal consistency of the detected changes is evaluated by using a tracking algorithm based on pyramidal optical flow method.[4]

4. EXPERIMENTS AND RESULTS

41 Dataset description

The system has been extensively tested on real-world data, including official live tests at the 2012 NATO SCI-256 trial. During these tests, over 137 videos were recorded by mounting the system on a car (Figure 3) and driving along urban, rural and sand road environments. Ground truth was manually annotated for 60 pairs (historic + current) of these sequences, resulting in more than 30.000 annotated test objects, allowing for automatic batch validation. Our test objects consist of 10x10x10 cm wooden blocks in different colors as shown in Figure 2.

4.2 Preliminary results

Under ideal operating conditions (e.g. similar weather conditions historical and live recordings) the change detection system is able to find all high-contrast test objects of 10 ? 10 ? 10 cm at a 30-m distance as shown in

Table 1 forms a summary of the results. In the full paper we will disclose more experimental results. Furthermore, we will analyze some shortcomings of the visual detection and discuss a possible improvement of the system that enables to extract small objects with low contrast to their background with higher certainty.

REFERENCES

[1] Primdahl, K., Katz, I., Feinstein, O., Mok, Y. L., Dahlkamp, H., Stavens, D., Montemerlo, M., and Thrun, S., “Change detection from multiple camera images extended to non-stationary cameras,” in [In Proceedings of Field and Service Robotics], (2005).

[2] Su, C. and Amer, A., “A Real-time Adaptive Thresholding for Video Change Detection,” in [IEEE Intl Conf. on Image Processing (ICIP)], 157 –160 (2006).

[3] V, K. and R, Z., “What energy functions can be minimized via graph cuts?,” IEEE Transactions on Pattern Analysis And Machine Intelligence 26(2), 147–159 (2004).

[4] J-Y, B., “Pyramidal implementation of the affine Lucas Kanade feature tracker: description of the algorithm,” tech. rep., Intel Corporation (2001).

9026-30, Session 7

Use of automated video analysis for the evaluation of bicycle movement and interaction

Heather A. Twaddle, Tobias Schendzielorz, Technische Univ. München (Germany); Oliver Fakler, TRANSVER GmbH (Germany); Sasan Amini, Technische Universität München (Germany)

In an effort to mediate urban traffic related problems, including congestion, air pollution and space scarcity, governments and urban planners in areas have implemented policies and measures to encourage bicycling. As a result, the volume bicycle traffic in many cities has increased significantly in the last decade and continues to increase today. This increase in volume has made it imperative to include bicycle traffic in microscopic traffic simulation tools, which are frequently used to evaluate and assess transportation measures before they are implemented. The accuracy and reliability of the results of microscopic simulation tools are highly dependent on the validity of the behavior models used within the software. Although the models used to depict the movement and interactions of motorized traffic have been developed extensively, those used to simulate bicycle traffic have received comparatively less attention. Further development of realistic bicycle behavior models is essential for the accurate simulation of urban situations with high volumes of bicycle traffic.

Within the German research initiative UR:BAN (Urban Space: User oriented assistance systems and network management), advanced driver assistance systems (ADAS) and intelligent transportation systems (ITS) are being developed to improve traffic safety, increase efficiency and reduce harmful environmental effects. The goal of a number of specific UR:BAN applications is to improve the safety of vulnerable road users including bicyclists. The accurate evaluation of these applications depends on the development of realistic models of bicycle behavior. This development also takes place within UR:BAN.

In the context of this development, an extensive review of available bicycle models has been carried out. The results indicate that the depiction of the operational behavior of bicyclists, which is the short term (milliseconds and seconds) behavior that takes place on the subconscious level, including lateral and longitudinal spacing between road users, desired speed distributions, passing maneuvers and queuing behavior is possible. There is a need, however, to calibrate and validate these models using data from reality. There are fewer models currently available that depict the tactical behavior of bicyclists, which is the conscious, short to midterm (seconds or minutes) behavior, including the selection of a trajectory to cross an intersection or cooperative behavior with other road users. Due to many factors, including their small size, their ability to ride on different infrastructure types and their possibility of switching between pushing and riding, bicycles are much more flexible that motorized road users. This flexibility has a number of important implications. Firstly, it has a considerable influence on the effect of bicycle traffic on the overall flow of mixed traffic streams. Secondly, the development of effective ITS applications for the protection of vulnerable road users must take this flexibility into consideration. The inclusion of flexible behavior in models takes place mainly on the tactical level, which reinforces the need to develop and extend models on this level.

In urban networks, intersections are not only the most dangerous points for bicyclists, but also present many possibilities for flexible behavior (crossing directly with motorized traffic, using crosswalks with pedestrians, running red lights, riding in the wrong direction, etc.). For both these reasons, the development of bicycle behavior models within UR:BAN focuses on the tactical behavior of bicycles at different types of intersections.

In order to meet these research goals, a large quantity of trajectory data from bicyclists, motor vehicles and pedestrians at a variety of intersections with different geometric and traffic characteristics is required. Video data was selected as a medium to extract trajectory data. Although this method of data collection poses processing challenges, it also makes it possible to gather detailed information about the situation, including the trajectories and type of all road users, current weather conditions and other unique circumstances. However, due to the quantity of required data, the manual processing of video data is infeasible. It is therefore necessary to use an automated or semi-automated video analysis tool. The software “Traffic Intelligence”, which is an open source software developed by Nicolas Saunier and his teams at the Polytechnique Montreal in Canada, was selected because of its accessibility and the potential to extend the software to analyze particular components of the UR:BAN research questions in an automated way.

As a first step of the automated video analysis, a methodology is developed for optimizing the feature detection parameters, the tracking parameters and the object grouping parameters, making it possible to accurately group and track objects at intersections used by large volumes of motor vehicles, bicycles and pedestrians. Different approaches for obtaining valid data are developed and evaluated. An example approach for taking the different characteristics of the various road users, including differences in size, speed and acceleration, into account is to analyze the video data in multiple stages with different parameter settings given in each stage. The trajectories of the different groups are extracted separately during the different stages. The developed approaches for extracting accurate trajectory data from multiple road user groups at busy urban intersections are evaluated and suggestions are provided.

The derived trajectory data from the different road users is then classified depending on the type of maneuver carried out by the road user. For example, a bicycle executing a left hand turn has the possibility of turning directly with the motorized traffic during one signal phase, turning indirectly by crossing the first street during one phase, stopping and waiting to cross the second street during the next phase, using a combined approach, or crossing with the pedestrian traffic. A classification structure for the maneuvers of different road users is presented and a methodology for categorizing the trajectory data using this structure and an automated analysis approach is developed and evaluated.

The derived trajectory data is combined with situational information, including timing data from the signal control at the intersections and other situational data, such as the weather condition and unique events. Finally, a statistical analysis will derive significant relationships between the type of infrastructure, the current situation, including the movement and position of other road users, the signalization and the tactical behavior of bicyclists.

9026-31, Session 7

Mutation detection for inventories of traffic signs from street-level panoramic images

Lykele Hazelhoff, Ivo M. Creusen, CycloMedia Technology B.V. (Netherlands) and Technische Univ. Eindhoven (Netherlands); Peter H. N. de With, Technische Univ. Eindhoven (Netherlands)

Introduction:

Road safety is preserved and improved by both adequate placement and optimal visibility of traffic signs. The visibility of road signs degrades over time due to e.g. aging, vandalism, accidents and vegetation coverage. This implies the need for sign maintenance, in order to preserve a high road safety. This process is usually performed based on accurate and recent inventories of traffic signs, which additionally allow for sign placement analysis. These inventories are commonly acquired by tracking all roads by car, bike or foot and marking all signs manually, which is time-consuming. The efficiency of this process can be improved by exploiting the already available street-level panoramic images, which are nowadays (yearly) captured by several companies in a large number of countries, providing a recent and accurate overview of the road infrastructure. The efficiency can be increased further by employing (semi-)automatic road sign recognition systems, based on computer-vision techniques for sign detection and classification, such that e.g. images not containing road signs can be skipped and projects can be executed at larger scales, while also reducing manual labor costs.

However, automated detection and recognition of traffic signs from images captured from driving vehicles in outside environments is a challenging task. The capturings are taken during all different kinds of weather conditions, including e.g. fog, resulting in significant differences in visual appearance. Furthermore, capturing from a driving vehicle may suffer from motion blur, and the signs are captured from a very wide range of distances and various viewpoints. Next to all variations due to capturing, the road signs themselves also show appearance variations, e.g. due to damage, wearing, besmearing or partial occlusion by e.g. vegetation. Additionally, classification of the traffic signs is complicated, due to the very small differences between some sign types and the possible presence of custom texts and/or symbols. Besides this, many sign-like objects exist, which result in falsely identified road signs.

Relevant literature:

Despite the challenging nature of traffic-sign recognition, multiple systems for road sign detection and classification are described in literature [1] [2] [3] [4] [5] [6] [7]. These systems usually start by detecting the present signs within the individual images. Afterwards, these detections are tracked over consecutive image frames, where the location of the road sign can be identified based on the GPS position of the individual capturings and triangulation of the retrieved detections. Finally, the sign type is retrieved by means of object classification techniques, which are applied either after detection or after tracking. Some of these systems focus on a limited set of traffic signs, i.e. only speed signs4or warning signs[2], while other support numerous different sign types [6] [7] [8] [9].

System description:

Our traffic sign recognition system [7] [10] [11] currently supports over 100 different sign types, and detects around 93% of the signs present within a geographical region, where both the sign type and sign location are returned. This system operates on panoramic images, captured at a 5-m interval. The system consists of three different stages. First, all signs are detected within the individual panoramic images, where we employ a learning-based detector based on Histogram of Oriented Gradients[12], which we have modified to exploit the strong color contrast between the traffic signs and background[13]. For each sign class (i.e. red triangular warning sign or red circular restriction sign) the same detector is employed, only differing in the training samples. Afterwards, the road sign locations are retrieved based on triangulation, where each combination of two detections of the same sign class in subsequent capturing gives a hypothesis of the location of that road sign. These hypotheses are clustered around the real road sign locations, which are therefore extracted using meanshift clustering, resulting in 3D sign locations. Finally, the retrieved road signs are classified to a certain sign type (i.e. prohibited to drive faster than 50 km/hr), based on the popular Bag of Words approach [14]. This process combines all selected detections suitable for positioning the sign for an increased classification robustness. In this stage, also the viewing direction of the sign is estimated [11].Currently, this system supports over 100 sign types from 9 different sign classes, and is able to retrieve about 93% of the signs within a geographical region (i.e. a municipalty, town, rural area or city). This system is employed to perform inventories in a semi-automatic fashion, where the results are checked manually. This stage consists of both correction of errors and addition of extra information (such as i.e. subsign texts).

Problem description and contribution:

Besides performing baseline inventories, road sign inventory systems also allow for the detection of mutations, i.e. locations where signs are removed or added, compared to the baseline inventory. This allows for very efficient updating of existing inventories of road signs, which is a new application of our existing traffic sign recognition system and is of great interest for road maintenance contractors, as they are especially interested in situation changes. Additionally, this enables the retrieval of locations where the visibility of signs could be insufficient (i.e. a missing or barely visible sign, not found by the automated inventory system). By specifically analyzing these situations, e.g. based on the same street-level panoramic images used by the automated inventory system, missing and damaged signs can be efficiently detected, thereby contributing to both a higher road safety and lower maintenance costs. We have performed such a mutation scan for the first time within a large geographical area (about 1,500 km of road), where we have compared the output of our automated traffic sign recognition system to the manually checked results obtained from a year earlier. We should note that this baseline inventory is employed for sign placement analysis, probably resulting in a severe number of changes. In this experiment, a sign is successfully re-identified if the sign type is identical to the baseline type, and the position difference is below 0.5m. This comparison resulted in a re-identification percentage of about 87% (13% of the signs did not match), while about 11% of the baseline sign amount was found as a new sign, resulting in a total difference equal to about 24% of the signs present in the baseline inventory. These changes include both real mutations (i.e. addition/removal of road signs) as mistakes made by our system, such as misclassified signs, falsely detected signs or signs missed, e.g. due to a significantly degraded visibility. At the end, we have found that 6.9% of the signs were newly placed and 8.3% of the signs were removed, such that about 65% of the found differences correspond to real mutations.

Conclusion

Employing automated traffic-sign recognition systems for mutation detection for inventories results in a large efficiency gain, where we have found that the mutation detection can be performed at about 24% of the work of the original semi-automated inventory. When comparing against conventional inventories performed by tracking all roads by car or bike, the efficiency gain is probably orders of magnitude higher. Currently, we explore additional algorithmic options for performing such mutation detection more efficiently. For example, the algorithm for the analysis of the new capturings can be made dependent on the baseline inventory results, e.g. to lower detection thresholds when a sign is expected. The final paper will contain more details about the procedure and used techniques.

References

[1] Lafuente-Arroyo, S., Maldonado-Bascon, S., Gil-Jimenez, P., Acevedo-Rodriguez, J., and Lopez-Sastre, R., “A tracking system for automated inventory of road signs,” in [Intelligent Vehicles Symposium, 2007 IEEE], 166 –171 (june 2007).

[2] Bonaci, I., Kusalic, I., Kovacek, I., Kalafatic, Z., and Segvic, S., “Addressing false alarms and localization inaccuracy in traffic sign detection and recognition,” in [16th computer vision winter workshop], 1–8 (2011).

[3] Ruta, A., Porikli, F., Watanabe, S., and Li, Y., “In-vehicle camera traffic sign detection and recognition,” Mach. Vision Appl. 22(2), 359–375 (2011).

[4] Overett, G. and Petersson, L., “Large scale sign detection using hog feature variants,” in [Intelligent Vehicles Symposium (IV), 2011 IEEE], 326 –331 (june 2011).

[5] Maldonado-Bascon, S., Lafuente-Arroyo, S., Gil-Jimenez, P., Gomez-Moreno, H., and Lopez-Ferreras, F., “Road-sign detection and recognition based on support vector machines,” Intelligent Transportation Systems, IEEE Transactions on 8, 264 –278 (june 2007).

[6] Maldonado-Bascon, S., Lafuente-Arroyo, S., Siegmann, P., Gomez-Moreno, H., and Acevedo-Rodriguez, F., “Traffic sign recognition system for inventory purposes,” in [Intelligent Vehicles Symposium, 2008 IEEE], 590 –595 (june 2008).

[7] Hazelhoff, L., Creusen, I. M., and de With, P. H. N., “Robust detection, classification and positioning of traffic signs from street-level panoramic images for inventory purposes,” in [Applications of computer Vision (WACV), Workshop on], 313 –320 (2012).

[8] Timofte, R., Zimmermann, K., and Van Gool, L., “Multi-view traffic sign detection, recognition, and 3d localisation,” in [Applications of Computer Vision (WACV), 2009 Workshop on], 1 –8 (dec. 2009).

[9] Timofte, R., Zimmermann, K., and Van Gool, L., “Multi-view traffic sign detection, recognition, and 3d localisation,” Machine Vision and Applications (December 2011).

[10] Hazelhoff, L., Creusen, I. M., van de Wouw, D. W. J. M., and de With, P. H. N., “Large-scale classification of trafic signs under real-world conditions,” in [Proc. SPIE 8304B-34], (2012).

[11] Hazelhoff, L., Creusen, I., and de With, P., "Robust classification of traffic signs using multi-view cues," in [Image Processing (ICIP), 2012 19th IEEE International Conference on], 457–460 (2012).

[12] Dalal, N. and Triggs, B., "Histogram of oriented gradients for human detection," in [Proc. IEEE Computer Vision and Pattern Recognition (CVPR)], 1, 886–893 (June 2005).

[13] Creusen, I., Wijnhoven, R. G. J., Herbschleb, E., and de With, P., "Color exploitation in hog-based traffic sign detection," in [Image Processing (ICIP), 2010 17th IEEE International Conference on], 2669–2672 (2010).

[14] Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C., "Visual categorization with bags of keypoints," in [Proc. European Conference on Computer Vision (ECCV)], (May 2004).

9026-32, Session 7

Automatic extraction of direction information in road sign imagery obtained by mobile mapping system

Junhee Youn, Korea Institute of Construction Technology (Korea, Republic of); Gi Hong Kim, Gangneung-Wonju National Univ. (Korea, Republic of); Kyusoo Chong, Korea Institute of Construction Technology (Korea, Republic of)

Road signs are important infrastructures for safe and smooth traffic, since they provide useful information for drivers. For systematic managing the road signs, constructing road sign database and implementing road sign managing system are essential for local government. In this paper, we deals with automatic extraction of direction information in road sign imagery obtained by MMS (Mobile Mapping System). Our approach starts from image pre-processing and binarization. Next, arrow areas are extracted by proposed four-direction consecutive cell measures. From the arrow area, corner points are detected by using good feature to track algorithm, which is extended Newton-Rhapson method. Clearly, some of the detected corner points compose arrow head. Therefore, LSM (Least Square Matching) algorithm is applied to the corner points to extract the direction information. For the LSM algorithm, four direction arrow head shaped templates are used. As a result, we can confirm the possibility of automatic extraction of direction information in road sign imagery.

9026-23, Session PWed

Downhill simplex approach for vehicle headlights detection

Ho-Joong Kang, Mokpo National Univ. (Korea, Republic of); Ho-Kun Kim, SEO Electronics Co., Ltd. (Korea, Republic of); iL-Whan Oh, Mokpo National Univ (Korea, Republic of); Kyoung Ho Choi, Mokpo National Univ. (Korea, Republic of)

Nighttime vehicle detection is an essential problem to be solved in the development of highway surveillance systems that provide information about the vehicle speed, traffic volume, and traffic jams, and so on. In this paper, a novel downhill simplex approach for vehicle headlights detection is presented. In the proposed approach, a rough position of vehicle headlights is detected first. Then, a downhill simplex optimization approach is adopted to find the accurate location of vehicle headlights. For the optimization process, a novel cost function is designed and various headlights are evaluated for possible headlight positions on the detected vehicles, locating an optimal headlight position. Simulation results are provided to show the robustness of the proposed approach for headlights detection.

Downhill simplex is an optimization technique that is based on a simplex and does not require derivative operations during the optimization process. By choosing an initial simplex, i.e. an object with n+1 vertices (n=the number of variables), a downhill simplex algorithm is started, and the vertex with the highest function value is replaced with a new vertex. During the search procedure, reflection, reflection and expansion, contraction, and shrinkage operations are performed until a local minimum point is reached. Locating headlights is a challenging problem, because the shape and size of headlights vary. In addition, headlights may not be extracted easily due to noise pixels from surrounding lighting objects. To locate headlights, a simple headlight detection mask is designed. In addition, it is necessary to determine the size of headlights, the distance between left and right headlights, and the center location of headlights. The detailed procedure of the headlight detection with the downhill simplex optimization will be presented.

The proposed algorithm was tested on highway surveillance video sequences recorded from Korea Expressway Corporation, South Korea. The test video was recorded for more than two hours in different location, covering various types of vehicles and headlights. The resolution and frame rate of the recorded video were 720x480 and 29.9 frames/sec, respectively. For implementation, Matlab R2010a was used with a Windows 7 operating system. The average headlights detection rate was about 92.9%. Headlight detection was considered as a success, if the detected left and right headlight centers were inside of the corresponding true headlights. Although some of headlights were connected with surrounding lighting pixels due to headlights from other vehicles, the proposed headlights detection approach using the downhill simplex method showed good performance for various types of headlights.

9026-33, Session PWed

Template matching based people tracking using a smart camera network

Junzhi Guan, Peter Van Hese, Jorge Oswaldo Niño-Castaneda, Nyan Bo Bo, Sebastian Gruenwedel, Dirk Van Haerenborgh, Dimitri Van Cauwelaert, Peter Veelaert, Wilfried Philips, Univ. Gent (Belgium)

With the advent of low-cost/high-performance smart cameras and super computers, it has become increasingly possible to process video sequences in real-time on affordable tracking system, which makes real-time tracking of people using multiple smart cameras practical. Real-time tracking of people using multiple cameras has many applications, such as surveillance and human behavior analysis. One approach to tracking is tracking by detection: the first step is usually foreground/background segmentation which separates the people to be tracked (foreground) from the background. Next, the resulting binary blobs are linked from one frame to the next. Many different kinds of foreground/background segmentation methods exist, but none of them work very well when illumination change occurs.

Template matching is one of the most popular methods to track objects in a video sequence, after they are found for the first time (e.g., using foreground/background segmentation). The principle of template matching is to track a certain feature or target over time based on the comparison of each frame with a template.

Robust tracking of multiple people is a challenging problem due to frequent occlusions and environment changes such as illumination changes. In the proposed approach, we handle this problem by creating a new template for each frame. This template is computed from the previous frame and from tracking feedback from the server. We propose a correlation coefficient based template matching which is invariant to illumination changes.

First, each smart camera receives from the server the positions of all persons from previous time (feedback). Then it will generate an uncertainty area around the last-known position of each person, assuming that a person cannot have traveled more than a certain distance, which is a good approximation at high frame rate. Then, a template of the person (which contains the silhouette) is extracted from the previous frame based on the feedback from the server. In this process, we assume a rigid person model (e.g. frozen arms legs,

fixed orientation w.r.t the camera; we assume that the silhouette of the person does not change shape but it of course translated to the correct position).

In reality, the silhouette will depend on the position (e.g., because the person will appear smaller when moving away from the camera and will also change because in reality the person is not rigid). In order to solve the aforementioned problem, we restrict the template to part of the human body rather than to a whole person. We experimented with different parts and sizes of the body as template for tracking, and discovered that incorporating only the upper torso of the person in the template reduces the impact of rapid arm and leg motion.

Experiments were also carried on how to solve the issues caused by shape changes of silhouette. We found that correcting the template by slight recursively averaging with earlier templates can handle the aforementioned problem effectively.

Once we have the uncertainty area and the template, the actual tracking is performed using a maximum likelihood approach: for each possible position of the person, a reference image is created by projecting the cuboid fitting the person position onto the background image; within the projected cuboid, the image intensities are replaced by those of the template. Finally the reference image is compared to the current image within a region of interest (ROI) around the projected cuboid, by computing the correlation coefficient in that ROI. The (local) maximum likelihood estimate for that particular camera is then taken as the position which maximizes the correlation coefficient. Each camera then sends the optimal position and the corresponding correlation coefficient to the server.

For occlusion handling, overlapping area is checked between any two persons' cuboid, whenever this is an overlapping between any two persons, the tracking system will ignore information from this certain camera and depends on other cameras without overlapping for tracking of these persons, assuming that there is at least one camera which has a clear view on a specific person, as overlapping cameras are used in the tracking system.

Finally, the server integrates the position estimates of all cameras by computing a weighted average with weight factors related to the correlation coefficients. The fusion center generates a hypothesis testing area around previous position, and it will take the position (on the ground plane) which has the minimum sum distance (weighted by correlation coefficient) between the camera estimate (image domain) and the projected hypothesis testing position (image domain). This fusion method not only takes the relative distance between a person and each camera into account, but also depends on the correlation coefficient of each camera.

For evaluation of the method, we use sequences which are captured by four side-view cameras in a room of 8.6 m by 4.8 m at 20 fps with a resolution of 780 pixels by 580 pixels. First, the method is evaluated on two short sequences which contain frequent occlusion and illumination changes. One of the sequences also has a table inside the tracking area. There is no tracking loss for the proposed tracker, while another state-of-the-art tracker lost persons 8 times for each sequence, good results (no loss, high accuracy) are also obtained when applying the proposed tracker on two long sequences (10 minutes each, two persons walking around in the scene). The proposed tracker also improved the tracking accuracy, so that most of the tracking results could be used directly as ground truth. The results show that we are able to track multiple people in a room using four side cameras. The minimal amount of information exchange between each smart camera and the fusion center makes the system highly scalable with the number of cameras, which means we can add cameras in the system without limitation for better solving occlusion problem. The performance is sufficient to obtain accurate trajectories of people for the tested sequences. The obtained results are promising and will lead to further exploration of our approach.

9026-34, Session PWed

Embedded image enhancement for high-throughput cameras

Stan Geerts, Prodrive B.V. (Netherlands) and Technische Univ. Eindhoven (Netherlands); Dion Cornelissen, Prodrive B.V. (Netherlands); Peter H. N. de With, Technische Univ. Eindhoven (Netherlands)

1. INTRODUCTION

This paper concentrates on embedded image enhancement for a novel ultra-high-resolution video camera with 4K images and higher. This camera is equipped with a highly sensitive image sensor which can also operate at low-light levels. This implies that conventional image enhancement techniques need to be reconsidered for embedding in such cameras. The high resolution and the frame rate of 30 frames per second (fps) result in a very high image processing bandwidth (around 600Mpx/s). In this paper, we study two image processing functions for image enhancement, which are optimized for the special sensor and high throughput, where we evaluate and optimize the algorithms for embedded implementation in programmable logic. The novel camera needs improvements on color and contrast, so that we aim at designing a high-quality auto white balancing and Local Contrast Enhancement (LCE).

2. PROBLEM STATEMENT AND REQUIREMENTS

Besides high image resolution and quality, the camera design should also enable longer operation time by operating it under low-light conditions. These requirements pose serious conditions for the image enhancement functions that are embedded in the camera. The enhancement functions should be more robust than conventional algorithms and at the same time be suited for embedded implementation. This embedding is a second system requirement.

The camera hardware platform is pre-determined, thereby constraining the choice of the algorithms. The camera has a heterogeneous architecture, containing an FPGA (Field Programmable Gate Array) and a general-purpose processor. The FPGA is used as an image buffer and for image processing, while the processor is mainly used for management tasks. To reduce the output bandwidth of the camera, images are encoded with lossy compression prior to transmission. Local contrast enhancement needs to precede the compression stage.

3. METRICS FOR ALGORITHM SELECTION

For both enhancement functions, a suitable algorithm needs to be chosen from a set of possibilities, depending on their suitability for embedded operation and their image quality, using the following three metrics:

a) Objective quality measure: This is a metric for measuring objective reconstruction errors for the white balancing choice and contrast levels for LCE.

b) Implementation complexity: With this metric we measure the intrinsic DSP operations for each algorithm.

c) Subjective performance: This involves a quality evaluation with both customer and image processing experts.

4. AUTO WHITE BALANCING

The Human Visual System (HVS) has the property of color constancy since it can determine the color of an object largely independent of the illumination color [1], while a camera does not have this property. The problem of auto white balancing is how to determine the illuminant color in a scene without any reference. This paper compares algorithms found in literature and analyzes their feasibility within the platform constraints.

After evaluating the results of a set of algorithms, the following subset of algorithms is evaluated for selection:

a) Grey World (GW) [2]

b) Standard deviation weighted grey world (SDWGW) [3]

c) RGB product measure (RGBPM) [4]

d) Color Histogram Stretching (CHS) [5]

Objective measure: The objective error is measured of all four algorithms on a large database of 482 images [9]. The used set contains raw images and white balanced (corrected) images. The raw images are processed with one of the algorithms and are compared with corrected images.

Implementation complexity: All mentioned auto white balancing algorithms contain three parts: 1) Measuring image statistics, 2) Calculating color correction, 3) Applying color correction. Part 1 and 3 are performed in the FPGA. Part 2 can be performed in the processor, since it improves flexibility and the calculations are simple. Since this will be a hybrid implementation, the complexity will be determined in DSP operations per frame. This metric can be mapped between processor implementation and FPGA implementation.

Subjective performance: This is measured by means of a survey containing images from the database, described above, and images from the used image sensor described in Section 1. The question in this survey was: “Which image is perceptually preferred?”.

Considering the results, color histogram stretching is chosen in this camera because it offers a subjective high quality (in over 70% of the images this algorithm is preferred) and it is among the algorithms with the lowest complexity (half compared to SDWGW) while having a small error.

5. LOCAL CONTRAST ENHANCEMENT

The goal with LCE is to improve visibility of details in all regions of the image, particularly in areas where shadows occur or objects are less visible due to local intensity levels. More specifically, details maybe present in the captured signals but are hardly visible on a display due to the threshold of Just Noticeable Differences (JNDs). Therefore dark parts need to be shifted to a visible region and the visibility of details in the total image need to be improved. The platform constraints from Section 2 play an important role in LCE since it requires the use of convolution filters, which are expensive in FPGA implementation.

After evaluating the results of a set of algorithms, the following subset of algorithms is evaluated for selection:

a) Contrast Preserving Gamma (CPG) [6]

b) Single Scale Retinex (SSR) [7]

c) Multi Scale Retinex (MSR) [8]

d) Locally Adaptive Contrast Enhancement (LACE) [6]

Objective measure: To be able to objectively compare different algorithms the Local Contrast (LC) from [6] is used. Since the research of [6] is based on HD images and our camera has a higher resolution, a higher spatial frequency can be captured. The LC metric is changed such that it enables measuring in a multi-layer frequency decomposition by altering the LC measurement kernel. The new LC formula is described in (1), where Lb is the average of the surrounding pixels and Lp the average of the central pixels within the filter kernel.

$$LC^f = |(L_p^f - L_b^f)| / L_b^f \quad (1)$$

For f=1, one central pixel is surrounded by a border of 1 pixel (covering 8 pixels). For f=2, this doubles both for central and surrounding pixels, etc.

The LC is measured for f =1,2,3,4. These four frequencies are chosen in such a way that the LC metric can be compared to the spatial frequencies of the same scene captured with an HD sensor. To be able to rank every algorithm with a single number, the average local contrast gain over the four frequencies is used.

Implementation Complexity: Since LCE needs to be performed before the compression step it requires a full FPGA implementation. The complexity is therefore determined in FPGA building block usage. The used building blocks are: multipliers, adders and line buffers. The platform has a pre-determined number of building blocks available. To measure the building block usage, the algorithms are analyzed and a coarse architecture is determined.

Subjective performance: In an initial experiment, a set of 14 typical images is used, which are all processed by CPG, SSR and LACE. The three results of all images are ranked by the participants from 1 to 10. The participants have a preference for less enhanced images, as they have a more natural appearance.

Considering the results, a combination of CPG and SSR is chosen. The user will have control over the amount of LCE and can choose for an attractive output like CPG up to a more enhanced image like SSR.

6. CONCLUSIONS

We have presented a comparison of auto white balancing and LCE functions that are suited for integration within a ultra-high-definition camera. Both objective and subjective comparisons have shown that color histogram stretching for auto white balancing and contrast preserving gamma combined with single scale Retinex are satisfying the system requirements on quality performance and embedded systems constraints.

In the full paper, we will show how the combination of contrast preserving gamma and single scale Retinex can be exploited for flexible contrast enhancement. Also, we will discuss the mapping of those functions on the FPGA platform to come to a cost-effective implementation while still offering very high picture quality and robustness for low-light conditions.

References

[1] D. A. Forsyth, “A novel algorithm for color constancy,” International Journal of Computer Vision, vol. 5, no. 1, pp. 5–36, 1990

[2] Nikitenko, D.; Wirth, M.; Trudel, K., “White-balancing algorithms in colour photograph restoration,” Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on , vol., no., pp.1037,1042, 7-10 Oct. 2007

[3] Hong-Kwai Lam; Au, O.C.; Chih-Wah Wong, “Automatic white balancing using luminance component and standard deviation of RGB components [image preprocessing],” Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP ‘04). IEEE International Conference on , vol.3, no., pp.iii,493-6 vol.3, 17-21 May 2004

[4] Alcazar, Jose Antonio Mendez. 2012. Auto white balance algorithm using RGB product measure. U.S. Patent 8,339,471, filed Dec. 31, 2009, and issued Dec. 25, 2012.

[5] Su Wang; Yewei Zhang; Peng Deng; Fuqiang Zhou, “Fast automatic white balancing method by color histogram stretching,” Image and Signal Processing (CISP), 2011 4th International Congress on , vol.2, no., pp.979,983, 15-17 Oct. 2011

[6] Cvetkovic, Sasa D.; de With, Peter H.N., “Optimization of video capturing and tone mapping in video camera systems”, PhD thesis for Eindhoven University of Technology, Department of Electrical Engineering, Dec. 2011.

[7] Jobson, D.J.; Rahman, Z.-u.; Woodell, G.A., “Properties and performance of a center/surround retinex,” Image Processing, IEEE Transactions on , vol.6, no.3, pp.451,462, Mar 1997

[8] Parthasarathy, S.; Sankaran, P., “An automated multi Scale Retinex with Color Restoration for image enhancement,” Communications (NCC), 2012 National Conference on , vol., no., pp.1,5, 3-5 Feb. 2012

[9] Image database courtesy of the Vision Group at Microsoft Research Cambridge: http://people.kyb.tuebingen.mpg.de/pgehler/colour/index.html

9026-35, Session PWed

On-road anomaly detection by multimodal sensor analysis and multimedia processing

Fatih Orhan, Anel Research & Development Co. (Turkey); Erhan P. Eren, Middle East Technical Univ. (Turkey)

Intelligent Transportation Systems (ITS) is a promising domain for sensing and computing technologies and many studies are conducted in this field. Companies make important investments on Vehicle-to-Vehicle (V2V) or Vehicle-to-Infrastructure (V2I) communications and the idea of tomorrow’s cars aims to increase the safety of drivers and passengers as well as to lower energy consumption and gas emissions while providing a comfortable, entertaining, and a sociable driving experience.

Although car manufacturers have great interest in built-in vehicular systems, the use of commonly available mobile devices as in-car

application providers is much more practical since the smart phones are widely adopted, present a cost-free solution for the drivers, and already provide a wide range of applications via application stores. Today's typical smart phone includes sensors such as GPS, accelerometer, magnetometer, gyroscope, light, microphone, camera (front and rear), proximity, Bluetooth, Wi-Fi, GSM, touchscreen. The access to device sensors by developers is a de-facto for the mobile application development platforms such as Android and iOS, and the delivery is achieved via application stores.

Although the computing and sensing technology flourishes, and the required infrastructures are setup and ready, there are indeed various challenges to develop smart contextual, sociable applications targeting especially vehicular usage. Due to its dynamic environment and broad requirements, vehicular application development is much more complicated, and further difficulties exist in testing the developed applications. The complications of vehicular sensor-based application development, especially for multimodal analysis, may be listed as (1) difficulties in the collection of sensor values properly; (2) the requirement of integration of diverse libraries for multi-modal sensor analysis including signal and/or image processing techniques and (3) the need to develop and utilize tools for fast and easy machine to machine (M2M) or direct social network communication in order to share the results obtained via sensor analysis.

This study primarily focuses on the challenges of developing vehicular social applications and utilizes the developed multimodal sensor analysis framework in order to enable easy, fast and flexible implementations of such applications. Although any kind of sensor-based application implementation may be developed on this framework, the main goals of the framework are (1) real-time sensing, (2) signal and multimedia processing and (3) effective, fast, easy sharing of results.

This study is an implementation of the framework as a mobile application to be run on the road while driving. The application firstly aims to automatically detect anomalies of the road (hard brake, hazards such as potholes/speed bumps and sudden lane change) using the accelerometer, magnetometer and GPS sensors of a mobile device mounted on the front windshield of the vehicle. For each type of anomaly, a different detector is developed based on distinct pattern of the sensor values and the implementations of the algorithms are performed through common interfaces provided by the framework. All of the sensor-based detectors are utilizing the GPS sensor to access the velocity of the vehicle. The hard brake detector checks the Y axis of the accelerometer for changes in forward-backward motion while the pothole and the speed bump detectors check the motion of the vehicle in Z axis (the gravity axis). On the other hand, the sudden lane change detector utilizes the lateral motion of the vehicle and utilizes X axis as well as Y axis variations. Each detector performs a signal processing algorithm based on the distinct pattern of the event to be detected.

As soon as an anomaly is detected, the video section including the road segment which contains the anomaly is automatically extracted from the recording camera. Thereupon, a second analysis is then started and image processing techniques are applied to the video section in order to identify the anomaly as road hazard (pothole/speed bump)/vehicle/pedestrian or other object. In case of vehicle detection (either because the vehicle is broken or stopped for another reason) the vehicle's plate number is also automatically extracted using a plate number recognition algorithm. All the extracted information including the time, exact location, anomaly type (hard brake/pothole/speed bump/object detection), anomaly severity, detection results and vehicle plate number (if any) along with sample images of the event is uploaded to a central application. The central application receiving such an incident locates nearby online users and provides a visual and/or audio warning including the type, location and image of the incident. The authorities may also be informed in the same way over the central application.

The application is evaluated with several driving tests to demonstrate the usage of the application and anomaly detections of hard brake and road hazard detections are tracked. A total of 25 drives are conducted, each lasted 20 minutes on average. Test drive results are compared with manual inspections and the evaluation is performed in terms of recall and precision values for both sensor-based road anomaly detections and image processing techniques. The obtained results provide a high recall and precision value and show the applicability of the implementation. As an additional output, the images of the detections are also automatically extracted, objects are identified and results are shared with drivers nearby.

As a conclusion, the utilized framework enables performing analysis on real-time synchronized sensor values with multi-modality in mind. The applications developed with this framework are able to effortlessly utilize the sensor values and various libraries, and perform multimodal analysis. The implemented application focuses on the security aspect of on-road travelling and tries to identify the hazardous events by first detecting the vehicle motion and then identifying the cause of the anomaly using image processing techniques. The output of the application is an automatic detector and road hazard warning system for both personal and government use.

As future work, it is planned to focus on more complex analysis using different sensors (such as microphone, ambient light, or proximity), and to explore more social aspects of utilizing the outcomes of the sensor analysis for smart city applications. On the other hand, studies to reduce the energy consumption of the framework are also to be conducted.

9026-36, Session PWed

Modeling dynamics with free-context grammars

Juan M. García-Huerta, Hugo Jiménez-Hernández, Ctr. para el Desarrollo Tecnológico Industrial (Mexico); Ana M. Herrera-Navarro, Univ. Autónoma de Querétaro (Mexico); Teresa Hernández-Díaz, Ctr. para el Desarrollo Tecnológico Industrial (Mexico); Ivan Terol-Villalobos, CIDETEQ (Mexico)

Video surveillance systems becomes more important because the cameras as a source of information represent a great challenge. This is, a camera provide continuously information which a high percentage of it are not representative for decision choices. The information that are representative for the human are based on structured events succeed into the scenarios. However define a general criterion to represent and model motion is not an easy works. Several authors pretend to solve this problem. However the number of variables involved and the difficulty to characterize in synthetic way structures to be more human-tractable make that this sensors become the majority of time sub utilized. Recently with the increment of computational power and the cheaper of camera sensors, results easy to mount distributed system to surveillance big areas.

The amount of information generated by an array of sensors make intractable the store of information in large time stamps. To figure out this situations several criteria are developed to only save events that results of importance in the scenario. This approaches to evaluate efficiently the information requires that the information behave in the same way even there are factors that could affect the quality of acquisition (light intensity, lights sources, reflections, shadows, or technical factors as latency, frame dropping, to mention a few). In this sense, this work presents an approach to model human dynamics in a close environment.

A visual dynamic for our purposes is a structured pattern of motion which preserve certain degree of repeatability. Then, the sequences of a camera to be practical need to catch dynamic primitives of movements performed in a scenario. These primitives start from motion detection. In order to review the state of art, several factors may affect motion detection; but there are a compromise to the quality of motion detected and the computational power needed to model motion. To deal with this situation a model based in temporal differences are presented to detect zones with motion.

Temporal templates represents an economic way to detect motion in outdoors environment and may be tolerant to certain luminance changes. A temporal template is performed as a decay function which preserve temporally, the evidence of motion. For a particular pixel x a temporal template is defined as follows

$$T(x,t-1)=a\,T(x,t)+(1-a)(d(I(x,t),I(x,t-1))$$

The advantage to model motion as a function decay is that all pixels are modelled as a grid, which is stimulated when motion is perceived. To detect motion the difference of successive images in the time is performed by d(I(x,t),I(x,t-1), which is defined as a difference operator which is conformed by the local derivatives of images I(x,t) and I(x,t-1). The local derivatives get information of texture, which make easier to discard luminance changes as global light effects. Further a threshold performed in T(x,t) function is done to determine a movement mask.

Once, motion is detected a set of states need to be defined. These states are represented by a symbol si. The total of states conform an alphabets A={s1,s2,…sn}. Each one symbol has associated a region of image as two tuple (si,r) and r={x| x is a image position}. The selection of regions are estimated in discover state phase. This phase consists on make a cumulative evidence of motion regions by follows expression

M=sum T(x,i) for i=1 to n

Where n is a temporal threshold enough to catch common movement. States are discovering applying a morphological segmentation approach; this is an extended watershed. Extended watershed provide a set of segmented regions based on local maxima. Each region is now associated to each one symbol of the alphabet.

After, to discover grammar structure temporal templates are used to discover when region trigger any symbols. As you can see only a symbol must be triggered. In this case only the proportion of region that present movement is selected as symbol to be triggered. This is

S= argmax(A| such that represent the region of maxima proportion)

A sequence of symbols triggered by the system is used as an information source that be cached by Sequitur algorithm. Sequitur algorithm infers a grammar rules starting from a symbol sequence. The inference of grammar is performed analyzing sequence of symbols produced by the system. States of system and consequently rule inference conforms the two stage of learning process.

At this point the grammar rules inferred are used as symbolic model of motion. This has the advantage that express common rules as a hierarchy, which can be useful to model movements at different abstraction levels.

To test the proposal, a camera is setting up into a building. The human dynamics observed are taken from couple of full working-day. First day is used to develop motion states and second day is used to infer free-context grammar. After several days are used to infer which observed dynamics belongs to the hierarchical structure inferred.

As it can be observed, this proposal might be used in other scenarios with structured motion such as supermarkets, malls, or model object with motion as vehicles, to mention a few.

9026-37, Session PWed

Overtaking vehicles detection and localization for driver assistance

Chung-Lin Huang, Asia Univ. (Taiwan)

The vehicle rear-view detector and the wheel detector based on AdaBoost algorithm are proposed. The former identifies the rear-view of the preceding vehicle and the latter detects the cut-in vehicle which may be partially occluded. Variant scales are computed and a detector containing 3 different aspect ratios is established to overcome the situation that the aspect ratio of the wheels on the vehicle varies with the position and the orientation of the vehicle. Using AdaBoost cascade classifier helps enhance the detection speed, filter out the non-discriminative features so that better detection rate can be achieved.

To obtain the information about the position of the vehicle, further processing is needed to find a pair of detected wheels. First, we find the parallel lines on the road which can be used to find the vanishing point. Then, we connect the center of the wheels and the vanishing point for find the possible pair of wheels. The relative position of the wheels of the same vehicle and its appearance are applied for matching. Finally, the position of the vehicle in the image is found according to the two centers of the matched wheels or when there's only one wheel (front or rear) appearing in the image.

9026-38, Session PWed

License plate location using SIFT and SVM

Roberto M. Souza, Mariana P. Bento, Univ. Estadual de Campinas (Brazil); Rubens C. Machado, Ctr. de Tecnologia da Informacao Renato Archer (Brazil); Roberto A. Lotufo, Univ. Estadual de Campinas (Brazil)

License Plate Location (LPL) is a difficult task, since images can be taken under various conditions, such as changes of illumination (day, night, indoor, outdoor), and weather conditions like rain and snow. Besides, there are differences in plate size, shape and color. The plate can be inclined or occluded, and it is also possible that there are no plates or multiple plates in the image. The vehicular speed, the distance between the vehicle and the camera, complex backgrounds and other objects, such as stickers attached to the car make the localization procedure even more difficult.

This paper proposes a new, simple, robust, and efficient method based on the low-level key-point detector and descriptor SIFT and the SVM classifier to perform LPL. The proposed method is invariant to scale, rotation, the color of the plate and its characters. The proposed LPL method has two stages: the first, which is an off-line stage, consists on training a classifier to distinguish SIFT points in the plate, and SIFT points not in the plate. The second stage, which is a real time task, consists on locating the license plate.

In the first stage of the method, a binary SVM classifier with a Radial Basis Function (RBF) kernel is trained to perform the SIFT points classification. The goal is to detect the plate points, and our idea is that if the classifier is trained with a set of plate points, and another set of points coming from numerous different backgrounds, i.e. numerous samples of classes other than being in the plate. Then, during prediction, when a sample from an unknown background class appears, the chance of it being wrongly classified as being in the plate decreases, since the ratio between the interest class (plate) and the classes that are not of interest (not in the plate) will tend to 0.

The second stage consists in locating the license plate. After receiving an image with a license plate, the first step is to extract its SIFT points. Then, these points are classified as being in the plate or in the background with the classifier trained in the off-line stage. The next step is to create a binary image with the same dimensions of the original image, with 1s in the positions of the points that were classified as being in the plate. Next, a sliding window of dimensions (hx , hy), and with steps (sx , sy) in the axis x and y of the image is slid from left to right, top to bottom through the binary image. The position of the window, which encompasses the highest number of non-zero pixels is considered the location of the plate.

Our LPL method was tested using a benchmark dataset that is consisted of a set of images taken under many different capture conditions, and they are divided in the following groups: blurred, color, grayscale, with shadows in the plate, very close view, difficult dirt shadows, difficult shadows, difficult trucks, and with night flash. They also have a group with images with more than one plate, but this group was not tested here, since our method, as it is today, can only detect one license plate per image.

In order to train the classifier, the off-line stage, it was used 100 random license plate images obtained from the web. The license plates in these images were manually segmented and their SIFT points were extracted. The ratio of SIFT points found in the license plates and SIFT points found in the background is approximately 1: 5, therefore as expected it is an unbalanced classification problem. The parameters of the SVM with a RBF kernel were set through a grid-search technique. In order to access the classification performance, it was used another 100 random images from the web with manually segmented license plates, the classification accuracy obtained was of only 62.54%.

The LPL experiments were performed using a total of 594 images of the dataset, excluding only the images in the group with images with more than one plate. Our method correctly located the license plates in the 93.60% of the cases. The method failed most at the group difficult trucks, probably, due to the presence of other plates and stickers attached to the trucks. Our location results are comparable to other methods in the literature that were tested using the same dataset.

This paper presented a method based on SIFT a binary SVM classifier to perform LPL, which is invariant to scale, rotation, the color of the characters in the plate and is able to achieve good location results even with a poor performance of the points classifier. The method has few parameters to adjust, and they can be easily set by super-estimating the dimensions of the plate. Also, the method performed well even with poor accuracy results of the classifier.

A few possible ideas as future work, is to extend the method to deal with situations where there is no license plate in the image or there are more than one. Implement this method in parallel and evaluate its plate location time, and its viability to be used in real time systems. Also, another interesting idea is to model the classification problem as an Open-Set Classification problem, and replace the binary SVM classifier for an open-set classifier, in order to improve the classification results.

9026-39, Session PWed

An integrated framework for detecting suspicious behaviors in video surveillance

Thi Thi Zin, Pyke Tin, Hiromitsu Hama, Takashi Toriu, Osaka City Univ. (Japan)

In this paper we propose an integrated framework for detecting suspicious activities in public places such as rail way stations, air ports, shopping malls and etc. where video surveillance systems are established. In our framework, a random walk type stochastic model is employed for detecting suspicious behaviors in real time video surveillance systems. First, the proposed framework employs multiple backgrounds modeling technique to detect moving objects. Then the high level motion features of objects of interest such as velocity of and distance between the semantic entities (objects) in the scene are calculated. Based on this record, objects are classified as being either persons or non-persons. Finally the suspicious behaviors are detected by thresholding on the first passage time probabilities for loitering, state interaction probabilities for fighting and time duration spent at a state for detecting abandoned objects. The proposed framework has been tested by using standard public datasets and our own video surveillance scenarios.

1. Introduction

Todays, the public safety and security issues are major concerns in our modern societies. For these purposes video surveillance systems are established in public places such as rail way stations, air ports, shopping malls and etc. for preventing and investigating crimes. In current situation, although video surveillance systems do exist, they have been used mainly for post-event analysis, mostly in the case of crimes investigations and forensics. However, very little has been achieved regarding real-time event recognition. To be more effective the video surveillance systems should be used in crime prevention purposes also. Due to the recent tragic terrorist attacks, the size and complexity of the monitored area are growing fast. In this aspect, a video surveillance system needs to be automatic and real time based for detecting potential criminals such as suspicious abandoned objects, abnormal or loitering behaviors. Therefore in this paper we propose an integrated framework based on a random walk type stochastic model for detecting suspicious activities in video surveillance systems.

2. The Proposed Integrated Framework

The proposed framework integrates a low level image processing technique and a random walk type stochastic modeling technique for analyzing suspicious behaviors in video surveillance systems. First, the proposed framework employs multiple backgrounds modeling technique to detect moving objects. Then the high level motion features of objects of interest such as velocity of and distance between the semantic entities (objects) in the scene are calculated. Based on this record, objects are classified as being either persons or non-persons. Finally the suspicious behaviors are detected by thresholding on the first passage time probabilities for loitering, state interaction probabilities for fighting and time duration spent at a state for detecting abandoned objects.

In particular, we employ the multiple backgrounds model based on short term motion, long term motion histories along with probability concepts. Then the extracted features are used for classifying person and non-person for further behavior analysis. When a new object occurs in the scene, it is classified as unknown (U). When its motion features are sampled, the velocity is used to determine whether it is a person (P) or a non-person or object (O). Using this transition model ensures that a still person is not misclassified as luggage. After classification, the movements of objects are modeled as a random walk type stochastic processes so that we can calculate the following quantities of interests:

(i) the expected first passage time E(T) for crossing boundaries ,

(ii) the expected time duration spent E(D) in a particular state,

(iii) the transition probabilities between the states.

A. Loitering People Detection

Then loitering is defined as “the presence of an individual in an area for a period of time longer than a given time threshold.” This is semantically described as Person is loitering ? In classification process Person ? P ? E(T) ≥ threshold (loitering).....(1)

B. Abandoned Objection Detection

Classification (object1) ? P

Classification (object2) ? O

For all E(D) ≥ t (abandoned) and Distance (object1) and (object2) ≥ d (abandoned). ..... (2)

C. For Fighting Event Detection

Object1 fights Object2 = classification: (object1) ? P and (object2) ? P

Transition probability between object1 and object2 approximately ≥ 0.9 for merging and less than ? for splits ..... (3)

Then the fighting event is detected.

3. Experimental Results

In this paper, carefully selected standard public data sets PETS 2007 and our own video sequences are used to test the proposed framework. Our framework successfully detects all of the events discussed earlier. According to the definition described in the above, a person is loitering if he stays in the field of view for at least 60 sec. This task can be solved rather easily and accurately by considering the expected first passage time for crossing boundary states of the Markov chain and larger probability than going to boundary states. We define the loitering zone as the areas within the camera view. We also use 60 as the threshold which means that the probability of the potential loitering person within the loitering zone is almost certain in specified time duration. On the other hand, the use of second threshold value 0.5 states that the probability of the person crossing the zone boundary is very low compare to the corresponding probability of normal people. These concepts have made our experiment to produce more realistic results. In similar fashion, for abandoned object detection experiment, we use the threshold for the distance between potential abandoned object and associated person 10 meters. In our experimental results, the proposed system can clearly differentiate very still person and abandoned object due to our newly developed background modeling technique. Furthermore, the utilization of the transition probabilities between two moving objects and the threshold values 0.9 for merging and 0.5 for splits give promising results.

9026-40, Session PWed

A novel approach to extract closed foreground object contours in video surveillance

Giounona Tzanidou, Eran A. Edirisinghe, Loughborough Univ. (United Kingdom)

One of the most commonly used approaches for foreground object segmentation in surveillance videos is the method initially proposed by Stauffer and Grimson [7]. According to the authors the background is modelled by a mixture of Gaussian (MoG) distributions; i.e. each pixel in the video frame is modelled by a weighted MoG whose parameters and weights are continuously updated with time. This allows the prediction of the pixel values that belong to the background based on probability. By nature the algorithm is designed to have tolerance to dynamic backgrounds (e.g. waving trees) and slight changes in illumination. However, the initial implementation suffers from inaccuracies when applied to some scenarios such as, shadows, spurious objects, extreme illumination changes, foreground-background similarity, and camera jitter. However this promising method, yet with weaknesses, has motivated a significant number of researchers to invest on the original concept and attempt to address the above weaknesses.

In our attempt to maintain the advantages of using the Gaussian mixture model, but to improve the method's identified weaknesses, we analysed in detail some of the most successful improvements proposed to the original approach of [7]. We used a MATLAB based implementation of the original approach proposed by Stauffer and Grimson [7] as our benchmark algorithm and amended it with shadow detection methods proposed in [3], [6] and [1]. Common approach in the aforementioned methods is the transition to the normalized RnGnI colour space where the I component encodes the intensity information. By observing the changes in the I components the authors were able to detect the shadow areas and the extreme illumination changes. However, in real world this method is rather naïve as sometimes the foreground contains high intensity values (e.g. white clothes on a human or a white car). Therefore, the modern methods for shadow detection take into account texture information, rather than only the colour; the most successful one is proposed by A. Sanin in [5]. In spite of the success of texture base methods in the proposed paper we attempt to approach the matter of shadow detection from a different scope.

Inspired by the work of O. Javed et. al. in [2] we decided to use the magnitude of gradient to isolate the foreground objects. In [2] the authors suggest removing the spurious foreground objects by monitoring the change in gradients. They exclusively perform gradient based segmentation in the case that the colour based detected foreground represents 50% of the total frame. Hence, in our work we attempt to utilise the second part of the method proposed in [2]. However, the gradient feature alone is not sufficient for clear object segmentation. Thus, we combine it with phase congruency as implemented by P. Kovesi in [4]. Having extracted an approximate contour from the above two features we proceed to detect the exact contour edges on the edge image (acquired by applying the Canny edge operator to the current frame.) through the use of an iterative region growing process. As it happens the prominent shadows have a strong gradient and phase congruency; therefore it is inevitable including them in the detected foreground. For this reason we perform detection of edges that belong to shadows with a variation of the method proposed in [8]. Having removed the shadow edges we retain the edges belonging only to the ground truth foreground. However it is likely that the edge contours thus obtained are not always closed. Hence, we propose the use of an edge completion algorithm that is based on pixel interpolation to solve this problem.

Detailed experimental results prove that the proposed method performs well under all of the aforementioned practical scenarios. We provide experimental results to prove the following advantages of the proposed approach as compared to state-of-the-art techniques.

- The spurious object are removed
- The algorithm is resistant to natural illumination changes
- The shadows are removed
- The contour of foreground objects is well defined
- Camera jitter and dynamic background scenarios could be potentially addressed if the system proposed in [8] is trained accordingly.

[1] ELGAMMAL, A., HARWOOD, D., AND DAVIS, L. Non-parametric model for background subtraction. In FRAME-RATE WORKSHOP, IEEE (2000), pp. 751–767.

[2] JAVED, O., AND SHAH, M. Tracking and object classification for automated surveillance. In Proceedings of the 7th European Conference on Computer Vision-Part IV (London, UK, UK, 2002), ECCV '02, Springer-Verlag, pp. 343–357.

[3] KAEWTRAKULPONG, P., AND BOWDEN, R. An improved adaptive background mixture model for real-time tracking with shadow detection. In Video-Based Surveillance Systems, P. Remagnino, G. Jones, N. Paragios, and C. Regazzoni, Eds. Springer US, 2002, pp. 135–144.

[4] KOVESI, P. Image Features from Phase Congruency - Abstract. Videre 1, 3 (1999).

[5] SANIN, A., SANDERSON, C., AND LOVELL, B. Improved shadow removal for robust person tracking in surveillance scenarios. In Pattern Recognition (ICPR), 2010 20th International Conference on (2010), pp. 141–144.

[6] SCHINDLER, K., AND WANG, H. Smooth foreground-background segmentation for video processing. In Proceedings of the 7th Asian conference on Computer Vision - Volume Part II (Berlin, Heidelberg, 2006), ACCV'06, Springer-Verlag, pp. 581–590.

[7] STAUFFER, C., AND GRIMSON, W. Adaptive background mixture models for real-time tracking. In Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on. (1999), vol. 2, pp. 2 vol. (xxiii+637+663).

[8] XIAOYUE JIANG, A. S., AND WYATT, J. Shadow detection based on colour segmentation and estimated illumination. In Proceedings of the British Machine Vision Conference (2011), BMVA Press, pp. 87.1–87.11. http://dx.doi.org/10.5244/C.25.87.

Conference 9027: Imaging and Multimedia Analytics in a Web and Mobile World 2014

Wednesday - Thursday 5 –6 February 2014

Part of Proceedings of SPIE Vol. 9027 Imaging and Multimedia Analytics in a Web and Mobile World 2014

9027-15, Session PWed

Agglomerative clustering using hybrid features for image categorization

Karina Damico, Roxanne L. Canosa, Rochester Institute of Technology (United States)

This research project describes an adaptive agglomerative image clustering technique that is used for the purpose of automating image categorization. Automating the process of image categorization requires that the categories are not pre-defined; rather they must be learned in an unsupervised way from the image collection itself. Adaptive clustering implies that the threshold for establishing new clusters is learned as more images from the collection are evaluated. This adaptive image clustering system is implemented in two steps: feature vector formation, and feature space clustering. The features that we selected are based on texture salience (Gabor filters and a binary pattern descriptor), visual interest points (the Harris corner detector), color (HSV), and several weighted combinations of these features.

For feature vector formation, a binary pattern descriptor based on the census transform is used, along with color and an interest point visual descriptor. An advantage of the census transform is that the global structure of an image is implicitly retained along with local structure. Global properties are encoded via a hierarchical spatial pyramid and local structure is encoded as a bit string, retained via a set of histograms. The algorithm calculates a numeric index from a 3x3 region of intensity values, and hashes it into one of 256 cases, each corresponding to a different type of local structure. The census-transformed output image has enhanced textural features that are used as a robust image descriptor. In addition, the transform can be computed efficiently – it involves only 16 operations (8 comparisons and 8 additions) per 3x3 region; thus, the cost to compute the histogram for a patch is linear in the number of pixels in the region of interest. A disadvantage of the census transform is that it is not invariant to rotation or scale changes; however, the spatial pyramid representing global structure helps to ameliorate this problem.

An agglomerative clustering technique is implemented and a comparison is made between an adaptive and a non-adaptive approach. The goal of the adaptive agglomerative approach is to (1) expand an established category when the rated similarity measure surpasses an adaptively-determined threshold, and (2) create a new category when the similarity measure is lower than another adaptively-determined threshold.

Agglomerative clustering is appropriate for image clustering because it is not necessary to provide the number of clusters in advance, as would be the case with k-means; also, it is efficient for a large number of features. As is the case with any clustering algorithm, agglomerative clustering requires a measure of image similarity; therefore, a variety of similarity metrics are implemented and compared for performance, such as Euclidean, Manhattan, maximum distance, and cosine similarity.

The quality of clustering using our adaptive approach is compared to the quality of clustering using a non-adaptive approach. The comparison metric is the Davies-Bouldin index, which measures clustering quality based on inter- versus intra-cluster feature similarity. A human subjective rating of cluster quality is used in addition to the Davies-Bouldin index. Implementation and initial testing is being conducted on a dataset of 1000 images comprising 10 categories, and will be benchmarked via a standard CBIR dataset such as the Wang dataset.

9027-16, Session PWed

A comparison of histogram distance metrics for content-based image retrieval

Qianwen Zhang, Roxanne L. Canosa, Rochester Institute of Technology (United States)

The type of histogram distance metric selected for a CBIR query varies greatly and will affect the accuracy of the retrieval results. This paper compares the retrieval results of a variety of commonly used CBIR distance metrics: the Euclidean distance, the Manhattan distance, the vector cosine angle distance, histogram intersection distance, ?2 distance, Jensen-Shannon divergence, and the earth-mover's distance. A training set of ground-truth labeled images is used to build a classifier for the CBIR system, where the images were obtained from three commonly used benchmarking datasets: the WANG dataset (http://savvash.blogspot.com/2008/12/benchmark-databases-for-cbir.html), the Corel Subset dataset (http://vision.stanford.edu/resources_links.html), and the CalTech dataset (http://www.vision.caltech.edu/html-files/).

To implement the CBIR system, we use the Tamura texture features of coarseness, contrast, and directionality. Tamura features correspond closely to human visual perception and are commonly used for optimum feature selection and texture analysis. We create texture histograms of the training set and the query images, and then measure the difference between a randomly selected query and the corresponding retrieved image using a k-nearest-neighbors approach. Each query image is classified by a majority vote of its neighbors, with the image being assigned to the class most common among its k nearest neighbors (where k is varied). The various histogram distance metrics define "nearest". Precision and recall is used to evaluate the retrieval performance of the system, given a particular distance metric. Then, given the same query image, the distance metric is changed and performance of the system is evaluated once again.

Several studies have been conducted in the past to determine how different feature descriptors affect the accuracy of retrieval; these studies often apply one specific distance measurement to compare different histograms, while varying the selected feature. However, retrieval results often differ dramatically, depending not only upon the chosen feature, but also upon the chosen distance measure. One current study performs a comparative analysis of histogram distance metrics using color histogram (Vadivel et al., 2013). Color histograms often perform well in image retrieval applications and, since they are simple to calculate, they can be recommended as a simple baseline for many applications. Our study aims to conduct a similar analysis based on texture instead of color.

9027-20, Session PWed

Video salient event classification using audio features

Francesca Gasparini, Gianluigi Ciocca, Silvia Corchs, Univ. degli Studi di Milano-Bicocca (Italy)

Since the amount of multimedia data is constantly increasing, mining and analyzing the content of such data is a challenging task. The integration of computational human attention models within the multimedia data mining techniques can certainly improve their performances.

The visual attention mechanism has been already investigated and incorporated within the data mining methods. Different saliency-based models, such as the well-known by [1] have been considered in the case of static images and also for video (spatio-temporal saliency models) [2], [3]. However, when observing a video sequence, humans are not only

driven by visually salient stimuli but also by auditory salient ones. There exist in the literature several works that model auditory-saliency [4], but the problem of the integration of audio-visual saliencies to generate a multimodal saliency map is not yet solved.

Within this context, the aim of this work is to detect the events in video sequences that are salient with respect to the audio signal. In particular, we focus on the audio-video analysis, with the goal of finding which are the significant features to detect audio-salient events, that should then be integrated into a multimodal saliency map for video analysis.

Keeping this final goal in mind, the database of videos used in our work were chosen from the database collected by "The DIEM Project" [5] freely available online for research and non-commercial use. Within this experiment the eye movements of 250 participants watching 85 different videos were collected. The videos belong to the following categories: advertisements, film clips, real-world scenes, social scenes, film trailers, video game trailers, music videos, documentaries, sports highlights, and news clips.

In our work we have extracted the audio files from videos of 7 different sport events: Cricket, F1, Football, Poker, Basket, Kendo and Tennis. Each audio file has been analyzed using non-overlapping windows (called frames).

For each video, we have manually labeled the salient audio-events using the binary markings (1 for frames where audio events are present, 0 otherwise), for a total of 1442 salient frames and 3930 non-salient frames. On each frame, features in both time and frequency domains have been considered: volume and Low Short Time Energy Ratio (effective in discriminating between speech and music signals) in the temporal domain, and Signal Energy, Sub-Band Energy, Frequency Centroid, Frequency Bandwidth and Spectral Flux in the frequency domain.

The above seven features have been used to train different classifiers. The classifier output is a binary indicator function of 0's (no event) and 1's (audio event), and the classification performance is reported in terms of Recall, Precision and F-measure. In particular for the classification task, we have adopted a Classification and Regression Trees method. The trees obtained by CART are not only used to solve our classification problem but also as a feature selection method to detect the audio features, significant with respect to the detection of salient events. Using CART, it's easy to understand which features are important in making the prediction. In fact, the decision tree generated uses only the features that help to separate the classes, while the others are not considered. The results of the research here presented could be integrated with spatio-temporal saliency maps to design a multimodal saliency strategy to detect video salient events, taking into account both visual and audio information.

[1] L. Itti, C. Koch and E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, 1254-1259, 1998.

[2] S. Corchs, G. Ciocca, R. Schettini, Video summarization using a neurodynamical model of visual attention, In Multimedia Signal Processing, 2004 IEEE 6th Workshop on, IEEE, 71-74, 2004.

[3] Y. Tong, F. Cheikh, F. Guraya, H. Konik and A. Trémeau, A spatiotemporal saliency model for video surveillance, Cognitive Computation vol. 3, 241-263, 2011.

[4] C. Kayser, C. Petkov, M. Lippert and N. Logothetis, Mechanisms for allocating auditory attention: An auditory saliency map. Current Biology, vol. 15, 1943 – 1947, 2005.

[5] http://thediemproject.wordpress.com/

9027-1, Session 1

Representing videos in tangible products *(Invited Paper)*

Reiner Fageth, CeWe Color AG & Co. OHG (Germany); Ralf Weiting, CEWE Stiftung & Co. KGaA (Germany)

Videos can be taken with nearly every camera, digital point and shoot cameras, DSLRs as well as smartphones. The implementation of videos while generating QR codes and relevant pictures out of the video stream was presented in last years' paper.

This year we would like to present first data about what contents is displayed and how the users represent their videos in printed products, e.g. CEWE PHOTOBOOKS and greeting cards. We report the share of the different video formats used, the number of images extracted out of the video in order to represent the video, the positions in the book and different design strategies compared to regular books such as number of text segments, usage of cliparts, pre-defined styles.

9027-2, Session 1

Aesthetic quality inference for online fashion shopping

Ming Chen, Jan P. Allebach, Purdue Univ. (United States)

With the increasing popularity of online fashion shopping, a great number of photos of fashion products have become available online. Among all those photos, some are taken by professional fashion photographer while many others are taken by amateurs. It is desired that photos of the products have higher aesthetic quality to improve users' online shopping experience. In this work, we design features for aesthetic quality inference in the context of online fashion shopping. Psychophysical experiments are conducted to construct a database of the photos' aesthetic evaluation specifically for photos on an online fashion shopping website. We then extract both generic low level features and high level image attributes to represent the aesthetic quality. Metadata such as the product type are also used to further improve the result.

9027-3, Session 1

Smart album: instant photo book creation across multiple platforms

Wiley H. Wang, Mixbook (United States)

Photo automatic organization has been in high demand as more and more photos are taken by digital cameras and mobile phones. Online digital scrap book creation has grown with more advanced features over the last few years. It is still a cumbersome process for the general public. The creation process put much of emphasis on the customization rather than automation. The amount of choices and editing to create a photo book often requires long time engagement from the customers. On the other hands, the needs to preserve and memories has only increased more. In this paper, we present the fast and easy way to create photo books in just a few minutes. The solution deals with not only the photo organization, but also user interaction, multiple platform adaptation, automatic design selection. In the process, we apply computer vision and image processing technology to intelligently learn the content of the photos. Photos are automatically arrange by their relevant relations in the way of natural story telling. We apply various technics to make sure photos fit in well designed page layouts. We integrate artificial intelligent and smooth user interaction design together to achieve complete production creation in a quick and easy way without compromising the quality of the product. The final product creation is instant and only requires minimum post creation adjustment. This product creation process lets average users create beautiful photo books without any scrap booking experience. We demonstrate our product in various platforms and their intuitive user experience design.

9027-4, Session 1

Full-color visibility model using CSF which varies spatially with local luminance

Alastair M. Reed, David Berfanger, Yang Bai, Kristyn Falkenstern, Digimarc Corp. (United States)

A full color visibility model has been developed that uses separate contrast sensitivity functions (CSFs) for contrast variations in luminance and chrominance (red-green and blue-yellow) channels. The width of the CSF in each channel is varied spatially depending on the luminance of the local image content. The CSF is adjusted so that more blurring occurs as the luminance of the local region decreases. The difference between the contrast of the blurred original and marked image is measured using a color difference metric.

This spatially varying CSF performed better than a fixed CSF in the visibility model, approximating subjective measurements of a set of test color patches ranked by human observers for watermark visibility. The effect of using the CIEDE2000 color difference metric compared to CIEDE1976 (i.e., a Euclidean distance in CIELAB) was also compared.

9027-5, Session 2

Text recognition and correction for automated data collection by mobile devices

Suleyman Ozarslan, P. Erhan Eren, Middle East Technical Univ. (Turkey)

In this study, a method is proposed for automatic recognition and enhancement of text in receipt images captured by mobile devices. The proposed method is used to automatically extract information from receipt images, such as product name, price, store name, and purchase date. Our method includes two phases to achieve this. In the first phase, information in store receipts is recognized by image processing methods using Optical Character Recognition (OCR). Compared to document scanner based OCR systems, different problems are encountered in mobile device camera based OCR systems. Insufficient or uneven lighting, text skew, text misalignment, focus loss, grains on the document and perception distortion are examples of these problems. Studies in the literature show that image processing methods cannot completely eliminate these problems. Accordingly, in the second phase of the proposed method, the knowledge based correction algorithm (KBC) developed in the scope of this study is applied on the extracted text in order to address these problems. This algorithm corrects erroneously recognized text pieces. In this second phase, first, segmentation is applied to the text obtained by OCR. After the segmentation, strings in the segments are determined. In order to improve the accuracy, the KBC process corrects errors in store and product names through row-by-row comparison instead of word-by-word comparison. Next, the KBC algorithm is applied to rows (strings) by using different databases which are specific to each segment. We use the Levenshtein distance in order to replace wrong information with the right one. The KBC algorithm also uses frequently encountered OCR errors while correcting character errors. Since the format of the date and time information in store receipts is well defined, this format information and frequently encountered OCR errors are defined as rule templates. In the last step of the second phase, these rule templates are applied to the text. In this study, initial experiments are conducted to evaluate the performance of the proposed method. 45 store receipts are used for our initial experiments. 20 of them are used to train the OCR engine, and the other 25 store receipts are used to measure the accuracy of the proposed method. The resolutions of the images are 5 megapixels, and captured by a mobile phone. At first, OCR engine is used with its default language files. In these experiments, recognition performances appear to be low: 41% for word recognition and 45% for character recognition, since the fonts in the store receipts are not commonly used, and so are not included in the default language files of the OCR engine. Afterwards, the OCR engine is trained. Training process increases the word and character recognition rates to 66% and 71%, respectively. Then, the KBC algorithm is applied to the text extracted by the trained OCR engine. Results show that the KBC algorithm improves the word recognition performance to 90% and the character recognition performance to 97%.

9027-6, Session 2

Text vectorization based on character recognition and character stroke modeling

Zhigang Fan, Xerox Corp (United States); Bingfeng zhou, Peking University (China); Francis Tse, Xerox Corp. (United States)

The shape of a text character can usually be represented in bitmap or outline (vector) forms. In the latter representation, a character is specified with a set of curves describing its outlines. Outline fonts exist extensively in electronically created files. However, they are not native for documents captured by digital cameras or scanners.

Vectorization of text is a process that converts the bitmaps of characters to their vector representations. Specifically, it uses a set of curves to reconstruct the characters. Compared to its bitmap, the vectorized text has the advantages of: 1) Resolution independency: the text can be easily scaled to different output resolutions. This is vital particularly for mobile devices. 2) Better image quality: the vectorized text appears smooth while bitmaps looks often jagged and bumpy. 3) Easy editability: Shape of the text can be edited using standard graphic tools. This enables easy modification of font attributes (size, boldness etc) for repurposing.

Text vectoriztion is typically composed of two procedures, dominant point determination and data fitting. The dominant points partition the character outlines into segments, and each segment is represented by a set of smooth curves by data fitting. The text vectorization is challenging as it often needs to meet conflicting requirements of avoiding outline jaggedness and over-smoothing.

In this paper, a text vectorization method is proposed using OCR (Optical Character Recognition) and Character stroke modeling. This is based on the following observations: 1) For a particular character, its font glyphs may have different shapes, but often share same stroke structures; and 2) The dominant points are usually local extrema points when the text boundary is projected to a certain direction.

In the proposed method, the dominant points are classified as “major” (specifying stroke structures) and “minor” (specifying serif shapes). A set of rules (parameters) are determined offline specifying for each character the number of major and minor dominant points and for each dominant point the detection parameters (projection directions and boundary conditions). For minor points, multiple sets of parameters might be required for different fonts. During operation, OCR is performed and the parameters associated with the recognized character are selected. Both major and minor dominant points are detected as a maximization process as specified by the parameter set. For minor points, an additional step may be required to test the competing hypothesis and detect degenerated cases.

The proposed method shows excellent performances compared to the conventional methods.

9027-7, Session 2

Visual improvement for bad handwriting based on Monte-Carlo method

Cao Shi, Jianguo Xiao, Peking Univ. (China); Canhui Xu, Wenhua Jia, Peking Univ (China)

With increased use of personal computer, handwriting is gradually taken the place of by the American Standard Code for Information Interchange in office. Dramatically, after tablet computer prevailing, people try to

reuse handwriting on touch screen. However, writing on touch screen is totally different with handwriting on a desk or a table. Moreover, in office, people already got used to typing instead of handwriting. Hence, there is an urgent requirement to assist user make handwriting on touch screen and improve user experience in visual effect, because, in the office, people already forgot handwriting skills more or less.

The well designed typefaces are prepared to improve visual effects of handwriting. The size of typeface image is set in algorithm configuration. The size of captured handwriting image is normalized according to the size of typeface, and the format of handwriting image is also converted to the format of typeface image. During this size normalization and format conversion, some drawbacks of handwriting are reduced, such as the letter is too small or too big, one stroke is too long or too short to keep the whole letter beautiful, etc.

The core idea of visual improvement, in this paper, is to use typeface image to approach handwriting image. A series of linear operators for image transformation are defined, including rotating, scaling, and translating, so as to transform typeface image to approach bad handwriting image, using Monte Carlo simulation. Initially, specific parameters of linear operators are randomly generated with Gaussian distribution. Using these parameters to transform typeface image, and match transformed typeface image and handwriting image. Exploiting maximum a posteriori probability estimation, the best matching transformed typeface image is the visual improvement result.

Experiments illustrate the visual improvement result not only reduce bad visual effects because of handwriting habits as mentioned before, such as the letter is too small or too big, one stroke is too long or too short to keep the whole letter beautiful, etc. but also a bad writing condition, just like a unsmooth pen trajectory or a slip of a pen.

In addition, some typefaces have serif of which the design is for publishing to avoid the ink fade out at the end of stroke. The serif in transformed typeface provides vivid effect for visual improvement of handwriting. And some typefaces are designed corresponding to handwriting styles. The random variation of curvature of this kind of typeface let the visual improvement result contain randomicity from handwriting.

The proposed visual improvement algorithm, in this paper, has a huge potential to be applied in tablet computer and Mobile Internet, in order to improve user experience on handwriting.

9027-8, Session 2

Image processing for drawing recognition

Rustem Feyzkhanov, Irina Zhelavskaya, SkTech (Russian Federation)

No Abstract Available

9027-9, Session 3

A web-based video annotation system for crowdsourcing surveillance videos

Neeraj J. Gadgil, Khalid Tahboub, Purdue Univ. (United States); David Kirsh, Univ. of California, San Diego (United States); Edward J. Delp III, Purdue Univ. (United States)

Video surveillance systems used by law enforcement authorities are of a great value to prevent threats and identify/investigate criminal activities. Surveillance cameras can capture, stream and record video data. This data can be analyzed manually or with the help of automated object detection/tracking methods. Manual analysis of a huge amount of data that accounts for feeds from several cameras over a long period of time often becomes impracticable. Automatic detection faces significant challenges when the video contains many objects with complex motion and occlusions. Analysis of videos involving many human actions and behaviors for detecting threats is difficult to a single officer, considering limited time, context and scope of knowledge.

With the rising amount of video data, there have been considerable efforts, particularly from computer vision community, to design generic video annotation tools. These tools, some of which being optimized for high-quality, economical and worker-friendly, are not specifically targeted towards analyzing surveillance videos. Some of them are stand alone tools while the others are embedded into existing crowdsourcing platforms such as Amazon Mechanical Turk. To be able to use crowdsourcing for analyzing surveillance videos, there is a need to develop a video annotation platform as a part of a well-managed system.

Our goal is to develop a web-based system that uses the “Wisdom of the crowd” to help the authorities in the law enforcement process. Our system provides a platform to crowdsource surveillance videos in an organized and controlled way to trained annotators. The system has a hierarchical model for annotators and the law enforcement officers based on level of ability, responsibility, security clearance and experience.

Officers can define labels to specify types of offences such as “assault”, “battery”, “abandoned baggage” and define new roles such as “beginner”, “battery expert”, “assault reporter” with a set of labels. Each annotator can be assigned a role based on the completed trainings. An annotator is restricted to use a certain set of labels corresponding to their role in the annotation process. Officers create and assign a task by dividing a video into parts or by specifying the details of a real-time stream.

Annotators view and annotate video contents reporting any suspicious incident or a potential threat to higher level. The system provides temporal and spatial annotation capabilities for pre-recorded/real-time video content with available labels. It also offers location-based information to annotators to provide a better context of human actions and behaviours. Spatial annotations can be used to improve the performance of object detection and tracking methods.

Officers are able to view the collective annotations and receive email alerts about a newly reported incident. They can also keep track of the annotators’ training performance, manage their activities and reward their success.

By providing this system, the process of video surveillance is made more efficient. This leads to a faster intervention by the law enforcement officers. This system, being managed by the officers, is safer and more reliable than other crowdsourced video annotation tools.

9027-10, Session 3

A Markov chain model for image ranking system in social networks

Thi Thi Zin, Pyke Tin, Takashi Toriu, Hiromitsu Hama, Osaka City Univ. (Japan)

[Introduction]

In today world, many social network services such as Facebook, Twitter, Flickr and many others have been played as independent ecosystem in which users can create and share a number of social activities including social interactions. For example, in Flickr, users can share contents and participate in multiple social communities by submitting photos, by joining groups, by performing actions and interactions such as comments, adding notes, favorite, and so on. In this way the social network contents are created and consumed by the influences of all different social navigation paths that lead to the contents. In such cases, the traditional ranking systems cannot be successful due to their lack of taking into account the properties of navigation paths driven by social connections.

[Proposed Method]

In this paper, we propose a general image ranking system in social networks by establishing a Markov chain model based on social links. In this context, a social link of an image or a photo can be defined as a link to and from a user’s favorite photo collection, from several groups,

art galleries, and via web search engines. In particular, we will rank all of the images or a subset of them in a specific social network domain according to their importance such as popularity, reliability, quality, social value and etc. In social network platforms, our image ranking system will take into account social interactions, internal (i.e. within specific network, say Flickr), and external (i.e. links from outside of Flickr). We then form a user-image relational graph in which nodes represent images and edges represents the relations between the images through user's actions and interactions. From this graph, a Markov chain model is established by forming the normalized number of links between two images as the transition probabilities. Then the following probabilities for the Markov chain are derived.

(i) Time dependent probability distribution,

(ii) Stationary probability distribution,

(iii) First passage time distribution for entering one particular state to another different state.

We then use these probabilities as metric measures for image ranking system in social networks.

Specifically, the image rank for each node in social network is defined as Image-Rank (Image i) = sum of the expected first passage times entering to state i.

[Some Experimental Results]

The proposed method is tested by using self-collected images from the popular social network, Flickr. The retrieved images are ordered by their interestingness value and the corresponding user links are also retrieved. In particular, the co-occurrences of an image and the user who posted are derived and transformed into a Markov chain by using corresponding joint probability density functions. We trained the model for the learned parameters and the information about the interests of the selected users to compute the joint probability function for the top 50 (manually labeled) images in the set. These experimental results show the effectiveness of the proposed method with high precision and recall rates.

9027-11, Session 3

Video quality assessment for web content mirroring

Ye He, Purdue Univ. (United States); Kevin Fei, Gus Fernandez, Google (United States); Edward J. Delp III, Purdue Univ. (United States)

Over the last few years, video streaming traffic has become a significant fraction of internet data traffic. User expectation for high quality streaming video and viewing experience is continuously increasing. Many video streaming devices provide the capability of projecting web video content from the screens in mobile devices to a larger TV screen. We call this feature "web content mirroring". However, fundamental questions, such as whether the mirrored web content quality is good enough in terms of user satisfaction, have not been formally addressed. One of the major challenges lies in the lack of an objective score to quantify the degree of user satisfaction. In this context, it is crucial to develop video quality metrics and understand how these metrics affect user viewing experience in order to best utilize internet resources to optimize user experience.

In this paper, we propose an automated scoring system, geared to Google's Chromecast product, but generalizable to other video streaming services, to quantify user satisfaction. We compare the quality of source videos with the videos transmitted to the TV. To be displayed on a TV, the video data is captured and encoded in Chrome web browser, transmitted over a Wi-Fi network, received and decoded in Chromecast, and finally displayed on the TV screen through HDMI connection. Four video quality metrics, namely Image Quality, Rendering Quality, Freeze Time Ratio and Rate of Freeze Events are used to generate the final User Satisfaction Score (USS). To measure the image quality, we compare the similarity between the original video frames and the available matching video frames in the captured video. Rendering quality is defined as the ratio of the rendered frames per second to the encoded frames per second. Freeze time ratio is defined as the fraction of the total video session time spent in screen freeze. Rate of freeze events is defined as the total number of freeze events divided by the total video session duration.

Several user studies are conducted to evaluate the impact of each individual video quality metric on the real user viewing experience. We scaled the user viewing experience from 1 to 5 corresponding to unwatchable, poor, acceptable, good and perfect. For each video quality metric, we generate corresponding artifacts to videos to such an extent that users score the videos from unwatchable to perfect. In this way, we find the user tolerance threshold for each video quality metric. In particular, we find that users are less sensitive to rendering quality than other quality metrics and the rate of freeze events has the larger impact on the user experience than the freeze time ratio. We use the results from user studies to combine all the video quality metrics and generate the final User Satisfaction Score (USS).

The USS derived from the model is unique in that 1) it is not derived from network-level metrics, such as the bit rate and packet drop ratio, 2) each video quality metric is controlled and measured individually to test user viewing experience, and 3) the objective score is matched to subjective score through user studies.

9027-13, Session 4

Evolving background recovery

Come Genetet, Gady Agam, Illinois Institute of Technology (United States)

The volume of videos containing presentations and lectures is large and is constantly increasing. Indexing such videos allows viewers to search for specific information and support more effective learning.

Automatic indexing of presentation and lecture videos based on their content requires recovering the video background which includes slides and/or whiteboard notes. This video background is evolving due to slides changing and/or the whiteboard being written. The objective of this paper is a method for automatically extracting the evolving background in such videos. In contrast to general background subtraction techniques which aim at extracting foreground objects, the goal here is to extract the background and complete it where the foreground is removed. Specifically, the objective in the context of a presentation or lecture video is to remove the speaker from the video.

In this paper, we propose a novel approach for background extraction. The idea is to detect objects in the foreground by analyzing their motion (assuming that only the foreground and the camera are moving), remove these objects from the video frames, and replace them with the background found in other frames of the video. This approach differs from existing methods in several ways. First, in contrast to most background subtraction techniques, which require prior observation of the background before it can be detected, the approach described in this paper learns the background as the video progresses by relying on motion analysis which takes into account a possible camera motion. Second, by using motion analysis we are able to deal with a dynamic background that is evolving in addition to the motion in the foreground. Third, our algorithm is tuned to the specific case where the evolving background is composed of writing on a whiteboard or a slide that is being changed. Finally, in addition to removing foreground objects, we also fill the holes that are left when removing the objects by using information from subsequent frames. While the proposed approach is described in the context of presentations or video lectures it may also be applied to other types of video where an evolving background needs to be recovered.

To evaluate the proposed approach we perform quantitative and qualitative analysis using a test video collection and compare the results to ones obtained using known algorithms. The test collection is composed of various lecture and presentation scenarios with and without a moving camera, and a synthetically generated videos with evolving background and a moving foreground. The synthetic test videos are generated using a background video (containing text and

handwriting that evolve without occluding foreground), and a foreground video of a person walking and moving in front of a uniform background. The foreground video is segmented and then pasted onto the evolving background to produce the synthetic test video. The synthetic videos are fully labeled meaning that we have accurate knowledge of the exact evolving background that needs to be recovered and the exact moving foreground that needs to be removed. Using these synthetic videos we then measure the difference between the expected and recovered background using RMSE. Initial experimental results demonstrate that the proposed approach is effective and performs better than known techniques.

9027-14, Session 4

HEVC compressed domain content-based video signature for copy detection and video retrieval

Khalid Tahboub, Neeraj J. Gadgil, Mary L. Comer, Edward J. Delp III, Purdue Univ. (United States)

Video sharing platforms and social networks have been growing very rapidly for the past few years. According to the recent figures released by YouTube, over 6 billion hours of video are watched each month, 100 hours of video are uploaded every minute and that's 50% more compared to last year. This exponential growth introduces many challenges in terms of copyright violations detection and video search and retrieval.

Classic video search and retrieval systems are text-based and do not consider the video content. However, to be able to detect copyright violations, a content-based video signature needs to be extracted. This signature can be matched to the signatures database of a video digital library to detect near-duplicate videos. This includes lower quality videos or whenever signal processing techniques are applied. Near-duplicate videos might include a complete or a partial temporal overlap.

Recent studies have focused on extracting video signatures based on the uncompressed domain. A signature that stays very similar when the video is modified is extracted and used to find a match in a large database. However, implementing this method becomes a problem in large systems where computational complexity is a major consideration. Therefore, it is advantageous to extract signatures from the compressed domain. There have been efforts to use DCT coefficients from an MPEG-X bitstream to create video signatures.

With the recent development of High Efficiency Video Coding (HEVC) which may achieve 50% bit rate reduction for equal perceptual video quality, its widespread adaptation is anticipated. One of the key features of HEVC is partitioning a picture into coding tree units (CTUs). HEVC has CTU as the basic processing unit which is further partitioned into coding blocks (CBs) that have associated prediction blocks (PBs) and transform blocks (TBs). HEVC has more Intra prediction modes and a different method for Inter prediction than its predecessors.

We propose a video retrieval system that generates a content-based video signature based on the extracted information from the HEVC-compressed bitstream. In this method, the CTU information, such as partition structure, modes and motion vectors from an HEVC bitstream is extracted, processed and formatted in a standardized way. This accounts for the temporal information which is used as a signature to represent the video or a portion of it. Then, a predefined hashing function in which the extracted information is projected on random matrices is applied. This provides robustness against signal processing techniques , quality changes and rotations and has been widely used in digital watermarking and video retrieval. Finally, the signature is searched in the database for a match. This process is optimized to detect a complete and a partial temporal overlap.

We claim our proposed method to be adequate for video copy detection and video retrieval purposes for HEVC-coded content. Also, by working directly on the compressed domain, the process of creating video signatures is made more efficient by reducing the computational complexity significantly.

9027-17, Session 5

Technology survey on video face tracking *(Invited Paper)*

Tong Zhang, Hewlett-Packard Labs. (United States); Herman Martins Martins Gomes, UFCG (Brazil)

Object tracking is one of the most important topics in computer vision which has been extensively studied in recent years. Among the objects of interest, human face is the most significant one as human activities are essential in many use scenarios. With the prevalence of surveillance cameras in streets, airports, schools, hospitals, data centers, work places and even homes, video face tracking is a crucial technology for applications in the fields of security, education, health care, smart city, smart home, etc.

Similar to the general object tracking problem, face tracking also has challenges such as losing target when there are appearance changes or occlusions, tracking efficiency issue when there are multiple targets, data association across multiple cameras and so on. Moreover, there are additional obstacles to tackle specifically for face tracking. For instance, the tracking model should be robust to quick motions (including in-plane and out-of-plane rotations) of a human head, and it should be able to keep tracking a face even when it turns away from a camera and then comes back. Furthermore, as tracked faces often need to be identified, how to select representative faces from a face sequence for optimal identification performance is another research component.

In this paper, we survey recent literatures on the face tracking topic, and cover the following aspects: 1) the tracking method such as mean shift, CamShift, particle filtering, TLD, etc.; 2) the features of target model such as color histogram, Haar-like features, local features (LBP, HOG, SURF), etc.; 3) integration of online learning and/or online classification into the tracker; 4) parallel computing in the case of tracking multiple faces; 5) data association across multiple face sequences; 6) face identification and face clustering of face sequences; and 7) face tracking across multiple cameras. We also introduce publically available databases for face tracking, and present some experimental results using these databases.

In general, while plenty of researches have been conducted on object tracking, including face tracking, in the last few years, and many promising approaches have been proposed, there is still much to be done to provide a solution that may track a relatively large number of faces simultaneously in high-resolution video frames, as well as identify the faces, all in real time. The goal of this paper is to summarize what are available and what are the state-of-the-art, and then hopefully may reveal potential paths forward.

9027-18, Session 5

Textural discrimination in unconstrained environment

Fatema Albalooshi, Vijayan K. Asari, Univ. of Dayton (United States)

Object region segmentation for object detection and identification in images captured in a complex background environment is one of the most challenging tasks in image processing and computer vision areas especially for objects that have nonhomogeneous body textures. This paper presents an object segmentation technique in an unconstrained environment based on textural descriptors to extract the object region of interest from other surrounding objects and backgrounds in order to get an accurate identification of the segmented area precisely. The proposed segmentation method is developed on a textural based analysis and employs Seeded Region Growing (SRG) segmentation algorithm to accomplish the process.

In our application of obtaining the region of a chosen object for further manipulation through data mining, human input is used to choose the

object of interest through which seed points are identified and employed. User selection of the object of interest could be achieved in different ways, one of which is using mouse based point and click procedure. Therefore, the proposed system provides the user with the choice to select the object of interest that will be segmented out from other background regions and objects. It is important to note that texture information gives better description of objects and plays an important role for the characterization of regions. In region growing segmentation, three key factors are satisfied such as choice of similarity criteria, selection of seed points, and stopping rule. The choice of similarity criteria is accomplished through texture descriptors and connectivity properties. The selection of seed points is determined interactively by the user when they choose the object of interest. The definition of a stopping rule is achieved using a test for homogeneity and connectivity measures, therefore, a region would stop growing when there are no further pixels that satisfy the homogeneity and connectivity criteria. The segmentation region is iteratively grown by comparing all unallocated neighboring pixels to that region. The difference between seed pixels' mean intensity value and the region's textural descriptors is used as a measure of similarity of pixels. The pixel with the smallest difference measured would be allocated to the particular segmentation region. Seeded region growing factors would change interactively according to the intensity levels of the chosen object of interest. The algorithm automatically computes segmentation thresholds based on local feature analysis. The system starts by measuring the intensity level of the selected object and accordingly adapts growing and stopping rules of the segmented region.

The proposed segmentation method has been tested on a relatively large variety of databases with different objects of varying textures. The experimental results show that this simple framework is capable of achieving high quality performance and that this method can better handle the problem of segmenting objects of nonhomogeneous textural bodies and correctly separate those objects from other objects and complex backgrounds. This framework can also be easily adapted to different applications by substituting suitable image feature definitions.

9027-19, Session 5

Image denoising by multiple layer block matching and 3D filtering

Zhigang Fan, Xerox Corp (United States)

No Abstract Available

9027-21, Session 5

Compact binary hashing for music retrieval

Jin Soo Seo, Gangneung-Wonju National Univ. (Korea, Republic of)

With the huge volume of music clips available for protection, browsing, and indexing, there is an increased attention to retrieve the information contents of the music archives. Music-similarity computation is an essential building block for browsing, retrieval, and indexing of digital music archives. For inferring music similarity, there are typically two different approaches; collaborative filtering and content-based approach. In collaborative filtering, the musical preference of one person is predicted by those of other people based on the musical tastes of many people. By the way, content- based approach is based on the perceptual auditory features of a music signal and computes music similarity by measuring the distance between features from two songs. Both approaches have pros and cons. For example, the collaborative filtering cannot be adopted for new songs, and the content-based approach requires perceptually-meaningful feature extraction and computationally-efficient distance measure. This paper deals with the content-based approach and focuses on comparing various binary embedding methods for the efficient music retrieval.

The difficulty in computing the music similarity lies in the fact that the criteria used to determine the level of the similarity between two songs are subjective and are hard to describe quantitatively. For determining content-based music similarity, auditory features representing the music timbre, such as mel-frequency cepstral coefficients or other spectrum descriptors, have been adopted. Inferring song-level representation from the low-level timber features is essential for music similarity estimation. Using the k-means or GMM clusters of the low-level timber features as a song-level representation has been widely adopted for music similarity. However, there are several issues involved in using the cluster models in practice, such as the convergence in iterative cluster construction the high computational cost in model comparison. Recently, to mitigate these problems involved in the cluster-based approaches, the supervector, the Fisher vector, and the centroid model have been applied as a song-level representation. All these approaches do not rely on the iterative modeling and lead to the vector representation for each music clip which can be easily incorporated with various normalization methods and distance measures. Moreover, they performed better than the previous cluster-based approaches in terms of retrieval accuracy. However, both the cluster-based and the supervector approaches have the feature (or model) storage issue. In practice, as the number of songs available for searching and indexing is increased, so the storage cost in retrieval systems is becoming a serious problem. This paper deals with the storage problem by extending the supervector concept with the binary hashing. Especially we utilize the similarity-preserving binary embedding in generating a hash code from the supervector of each music clip. Especially we compare the performance of the various binary hashing methods for music retrieval tasks on the widely-used genre dataset and the in-house singer dataset. Through the evaluation, we will find the optimal way of generating hash codes for music similarity estimation which improves the retrieval performance.

9027-22, Session 6

Efficient eye detection using HOG-PCA descriptor

Andreas E. Savakis, Riti Sharma, Rochester Institute of Technology (United States); Mrityunjay Kumar, RealD (United States)

With the proliferation of webcams and mobile devices, eye detection is becoming increasingly important for human computer interaction. In the context of mobile interfaces, eye detection provides cues on whether or not the user is looking at the screen. Furthermore, eye detection is the starting point for gaze estimation and eye tracking. Following detection, eye locations can be used for face alignment and normalization, pose estimation, or initialization of Active Shape/Active Appearance Models that require good initial placement. While some eye detection systems are relying on infrared light emitting diodes, it cannot be assumed that such devices are generally available. This paper presents a robust and efficient eye detector based on HOG-PCA features obtained from image or video input.

The Histogram of Oriented Gradients (HOG) is a dense descriptor computed on overlapping blocks along a grid of cells. The HOG feature was introduced for people detection where it was combined with an SVM classifier. Following its initial success, the HOG descriptor has been adopted for face recognition and smile detection. HOG-based eye detectors were combined with circular Hough transform and used with adaboost for faster performance.

In this paper, we utilize a HOG-PCA descriptor to obtain an efficient representation of HOG features for robust classifier training. Our HOG-PCA features are computed using Principal Component Analysis on the HOG vectors, which significantly reduces feature dimensionality compared to the original HOG size or the eye image patch size. The HOG-PCA descriptor effectively generates eigen-eyes in HOG space and offers an efficient feature representation for classifier training. In this work, we perform eye detection using HOG-PCA features and a Support Vector Machine (SVM) classifier. Our approach is suitable for real time

implementation, as the reduced dimensionality of HOG-PCA allows for faster computations of the eye detection classifier.

We tested the HOG-PCA eye detector with SVM classifier on two eye datasets generated from the FERET and BioID face databases. Our results demonstrate that eye detection using HOG-PCA outperforms eye detection based on HOG or image patches alone. In conclusion, the HOG-PCA feature is both discriminative and efficient, which makes it suitable for real time object detection.

9027-23, Session 6

Exploiting articulated structure for hand tracking

Prabhu Kaliamoorthi, Ramakrishna Kakarala, Nanyang Technological Univ. (Singapore)

Model based tracking of an articulated hand using RGBD input is an active area of research in the computer vision community. The problem is highly challenging since the state space required to parameterize a human hand is high dimensional. The state of the art methods formulate the problem as a maximum likelihood estimation problem in an analysis by synthesis framework. In this formulation, various hand poses are synthesized and matched with the observed RGBD input, and the pose that results in minimal difference measured by an energy function is considered to be the true pose. The articulated nature of the human hand results in a multimodal likelihood or a non-convex energy function. Furthermore, every evaluation of the function incur a significant overhead as a result of synthesis and matching. Recent studies rely on global optimization methods such as Particle Swarm Optimization to obtain the maximum likelihood estimate.

The formulation we just described does not exploit the structure in the problem. The articulated structure of the human hand enables a number of conditional independence assumptions to be made about the joint likelihood. As a result the joint likelihood could be expressed as the product of several factors of much lower dimension. The factors are the conditional likelihoods for each rigid part given the parameters of the parents in the kinematic chain. The conditional likelihoods further require the observed RGBD input to be decomposed probabilistically into the rigid parts, in order to measure the difference per rigid part. The conditional likelihoods can be fused together to obtain the marginal likelihood for each rigid part using Monte Carlo methods such as importance sampling. This effectively decomposes the high dimensional state space into many subspaces of much lower dimension. Once the state space is decomposed, one could perform stochastic search for the optimal hand pose simultaneously in these subspaces.

Such a decomposition has two advantages. It both reduces the dimensionality of the state space and avoids the combinatorial explosion of the number of hypothesis or modes in the joint likelihood. As a result, the decomposed search would require fewer number of samples for maximum likelihood estimation. In an early study we applied this technique to the problem of human motion reconstruction from multi-view camera input. We showed that using decomposition we are able to achieve state of the art results in human motion reconstruction using less than a sixth of the computational overhead. In this study, we apply the technique to the problem of articulated hand tracking using RGBD data.

We model the hand using a smooth surface mesh. The surface mesh is rigged to a skeleton with 26 degrees of freedom using linear blend skinning. We use annealed particle filter on the decomposed subspaces to search for the optimal pose. Since ground truth is not available for the hand pose, we demonstrate our method using qualitative tracking results. Our initial experiments show that using decomposition, the efficiency of hand tracking is significantly improved.

9027-24, Session 6

Adaptive weighted local textural features for illumination, expression, and occlusion invariant face recognition

Chen Cui, Vijayan K. Asari, Univ. of Dayton (United States)

Biometric features such as fingerprints, iris patterns, and face features help to identify people and restrict access to secure areas by performing advanced pattern analysis and matching. Face recognition is one of the most promising biometric methodologies for human identification in a non-cooperative security environment. However, the recognition results obtained by face recognition systems are affected by several variations that may happen to the patterns in an unrestricted environment. Though several algorithms have been developed for extracting different facial features for face recognition, due to the various possible challenges of data captured at different lighting conditions, viewing angles, facial expressions, and partial occlusions in natural environmental conditions, automatic facial recognition still remains as a difficult issue that needs to be resolved. In this paper, we propose a novel approach to tackling some of these issues by analyzing the local textural descriptions for facial feature representation. The textural information is extracted by an enhanced local binary pattern (ELBP) description of all the local regions in the face. The relationship of each pixel with respect to its neighborhood is extracted and employed to calculate the new representation. ELBP reconstructs a much better textural feature extraction vector from an original gray level image in different lighting conditions. The dimensionality of the texture image is reduced by principal component analysis performed on each local face region. Each low dimensional vector representing a local region is now weighted based on the significance of the sub-region. The weight of each sub-region is determined by employing the local variance estimate of the respective region, which represents the significance of the region. The final facial textural feature vector is obtained by concatenating the reduced dimensional weight sets of all the modules (sub-regions) of the face image. Experiments conducted on various popular face databases show promising performance of the proposed algorithm in varying lighting, expression, and partial occlusion conditions. Four databases were used for testing the performance of the proposed system: Yale Face database, Extended Yale B Face database, Japanese Female Facial Expression database, and CMU AMP Facial Expression database. The experimental results in all four databases show the effectiveness of the proposed system. Also, the computation cost is lower because of the simplified calculation steps. Research work is progressing to investigate the effectiveness of the proposed face recognition method on pose-varying conditions as well. It is envisaged that a multilane approach of trained frameworks at different pose bins and an appropriate voting strategy would lead to a good recognition rate in such situation.

9027-25, Session 6

Research on the face pattern space division in images based on their different views

He Zhixiang, Xiaoqing Ding, Chi Fang, Yanwei Wang, Tsinghua Univ. (China)

The smaller face range of view angle caused less differences of the obtained face images makes the unified processing more convenient, and vice versa. Thus many researches divided the entire face pattern space of multiview faces in images into many subspaces with small range of views based on their view angle. But large number of subspaces needs long time to process all, and different face processing algorithm takes different tolerant of view changing, the research of proper division of face pattern space is needed to ensure its performance.

Different from other researches, this paper proposed an optimal view angle range criterion of face processing algorithms in theory by careful analysis of the structure differences of multiview faces and its influence

to face processing algorithms. Then a face pattern space division method for different face processing algorithms is proposed. Finally, this paper use the proposed criterion and method to divide the face pattern space for face detection and compared with other division results. The final results shows the proposed criterion and method can satisfy the processing performance with minimum number of subspace. The study in this paper can also help other researches which need to divide pattern space of other objects based on their different views.

Conference 9028: Media Watermarking, Security, and Forensics 2014

Monday - Wednesday 3 –5 February 2014

Part of Proceedings of SPIE Vol. 9028 Media Watermarking, Security, and Forensics 2014

9028-1, Session 1

Challenging the doctrines of JPEG steganography

Vojtech Holub, Jessica Fridrich, Binghamton Univ. (United States)

The design of both steganography and steganalysis methods for digital images heavily relies on empirically justified principles. In steganography, the domain in which the embedding changes are executed is usually the preferred domain in which to measure the statistical impact of embedding (to construct the distortion function). Another principle almost exclusively used in steganalysis states that the most accurate detection is obtained when extracting the steganalysis features from the embedding domain. While a substantial body of prior art seems to support these two doctrines, this article challenges both principles when applied to the JPEG format. Through a series of targeted experiments on numerous older as well as current steganographic algorithms, we lay out arguments for why measuring the embedding distortion in the spatial domain can be highly beneficial for JPEG steganography. Moreover, as modern embedding algorithms avoid introducing easily detectable artifacts in the statistics of quantized DCT coefficients, we demonstrate that more accurate detection is obtained when constructing the steganalysis features in the spatial domain where the distortion function is minimized, challenging thus both established doctrines.

9028-2, Session 1

Linguistic steganography on Twitter: personalised language modeling with manual interaction

Alex D. Wilson, Phil Blunsom, Andrew D. Ker, Univ. of Oxford (United Kingdom)

No Abstract Available

9028-3, Session 1

Detection of content adaptive LSB matching: a game theory approach

Tomas Denemark, Jessica Fridrich, Binghamton Univ. (United States)

This paper is an attempt to analyze the interaction between Alice and Warden in Steganography using the Game Theory. We focus on the modern steganographic embedding paradigm based on minimizing an additive distortion function. The strategies of both players comprise of the probabilistic selection channel. The Warden is granted the knowledge of the payload and the embedding costs, and detects embedding using the likelihood ratio. In particular, the Warden is ignorant about the embedding probabilities chosen by Alice. When adopting a simple multivariate Gaussian model for the cover, the payoff function in the form of the Warden's detection error can be numerically evaluated for a mutually independent embedding operation. We demonstrate on the example of a two-pixel cover that the Nash equilibrium is different from the traditional Alice's strategy that minimizes the KL divergence between cover and stego objects under an omnipotent Warden. Practical implications of this case study include computing the loss per pixel of Warden's ability to detect embedding due to her ignorance about the selection channel.

9028-34, Session 1

Further study on security of S-UNIWARD

Tomas Denemark, Jessica Fridrich, Vojtech Holub, Binghamton Univ. (United States)

Recently, a new steganographic method was introduced that utilizes a universal distortion function called UNI-WARD. The distortion between a cover and stego image is computed as a sum of relative changes of wavelet coefficients representing both images. As already pointed out in the original publication, the selection channel of the spatial version of UNIWARD (the version that hides messages in pixel values called S-UNIWARD) exhibits unusual properties – in highly textured and noisy regions the embedding probabilities form interleaved streaks of low and high embedding probability. While the authors of UNIWARD themselves hypothesized that such an artifact in the embedding probabilities may be a security problem, experiments with state-of-the-art rich models did not reveal any weaknesses. In this extended abstract, we present a simple attack on S-UNIWARD that utilizes the fact that the embedding probabilities can be approximately estimated from the stego image. A successful attack appears possible with as few as nine features. The full version of this paper will contain more extensive and optimized experiments as well as recommendations on how to adjust the embedding function of S-UNIWARD to avoid this type of targeted attack.

9028-5, Session 2

Are you threatening me?: Towards smart detectors in watermarking

Mauro Barni, Univ. degli Studi di Siena (Italy); Pedro Comesaña-Alfaro, Fernando Pérez-González, Univ. de Vigo (Spain); Benedetta Tondi, Univ. degli Studi di Siena (Italy)

We revisit the well-known watermarking detection problem, also known as one-bit watermarking. In the absence of an adversary, the design of the detector generally relies on probabilistic formulations (e.g., Neyman-Pearson's lemma) or on ad-hoc solutions. When there is an adversary trying to minimize the probability of correct detection, game-theoretic approaches are possible. However, they usually assume that the attacker cannot learn the secret parameters used in detection. This is no longer the case when the adversary launches an oracle-based attack, which turns out to be extremely effective. In this paper, we discuss how the detector can learn whether it is being subject to such attack, and take proper measures. We present two approaches: one that detects binary line searches that are common in repeated queries, and another based on Afriat's theorem, which is active (i.e., it changes the detector in a smart way) and can be used to determine whether the adversary is trying to maximize a certain utility function.

9028-6, Session 2

On accuracy, robustness, and security of bag-of-word search systems

Svyatoslav V. Voloshynovskiy, Maurits Diephuis, Dimche Kostadinov, Farzad Farhadzadeh, Taras Holotyak, Univ. of Geneva (Switzerland)

No Abstract Available

9028-7, Session 2

An enhanced feature set for pattern recognition based contrast enhancement of contact-less captured latent fingerprints in digitized crime scene forensics

Mario Hildebrandt, Jana Dittmann, Otto-von-Guericke-Univ. Magdeburg (Germany); Claus Vielhauer, Fachhochschule Brandenburg (Germany)

The primary goal of the contact-less acquisition of latent fingerprint traces from crime scenes in digitized forensics is preserving the unaltered trace for more detailed investigations. However, without any physical or chemical preprocessing a distinct ridge pattern is often not visible due to particular characteristics of the substrate it is present on. Hence, either substrate dependent acquisition sensors as described in the Fingerprint Source Book[3] and/or digital preprocessing techniques need to be applied to achieve a sufficient quality of the fingerprint. In biometrics fingerprint segmentation techniques are used to separate the fingerprint pattern from the background noise of the capturing sensor[4]. However, those techniques usually require the fingerprint to be dominant in the image in order to determine local orientations and ridge frequencies for creating optimal Gabor filters. Furthermore, in case of insufficient quality the fingerprint can be easily captured from the finger again. In crime scene forensics fingerprints are found on various substrates with different structural or textural patterns. Hence, such segmentation approaches cannot be directly applied due to the superimposed patterns which are often dominant within the image. Furthermore, the quality of latent fingerprints is usually much lower compared to exemplar fingerprints used in biometrics, often only partial fingerprints are found at a crime scene.

A first block based pattern recognition approach for enhancing the contrast between the latent fingerprint and the substrate within data captured with a chromatic white light sensor is introduced by Makrushin et al.[5] using statistical features. It is extended by Hildebrandt et al. [1] with structural features and features derived from Bedford's law[2]. However, the structural features are directional and thus require a manual alignment of the substrate prior to the acquisition.

Both approaches use blocks of 50 micron which corresponds with a resolution of 500 ppi. Hence, the classification results can be used as an input for additional fingerprint segmentation or enhancement techniques, such as known approaches from biometrics[4]. The features are extracted from the captured intensity and topography data set and from their preprocessed representations. In Hildebrandt et al.[1] this results in a 146 dimensional feature vector. For the classifier training each block is labeled based on an approximated ground truth from binarized differential images. The results indicate very good results of up to 92.7% recognition accuracy on cooperative substrates. However, on structured or textured substrates the results are significantly worse, e.g. brushed stainless steel only 67% of the blocks are correctly classified using the best performing classifier. Hence, further improvements are necessary.

We address this challenge by adding semantic features to the feature set. Such features are determined for an Epsilon-neighborhood of each block allowing for determining fingerprint specific properties. In particular, we apply Gabor filters with a fixed parameter set based on a priori knowledge and different angles in 5 degree steps between 0 and 180 degrees to each data set. Afterwards, the highest filter response is determined by calculating the standard deviation of gray values within an Epsilon-neighborhood of 750 micron around each 50 micron block resulting in a total area of 1550x1550 micron. Thus, each Epsilon-neighborhood should contain multiple ridge lines. The Gabor filter with the matching angle should strengthen the ridge pattern resulting in a higher contrast and thus in a higher standard deviation if a fingerprint is present in the regarded area. Since the Epsilon-neighborhood covers a rather large area which does not allow for distinguishing between fingerprint ridges and valleys, the mean gray value of the 50 micron block is recorded for the determined angle as a feature as well. In addition to those two semantic features we also enhance the approach from Hildebrandt et al.[1] with rotation invariant Hu moments[6] and additional preprocessing techniques. We additionally determine the first derivative of the intensity and topography image from the chromatic white light sensor for the X as well as the Y axis using a Sobel operator as preprocessing. In doing so, the directional pattern of substrates is reduced. This extends the set of preprocessing techniques to Sobel operators in first order for the X, Y, as well as X and Y axis, Sobel operator in second order for X and Y axis and unsharp masking. Since all techniques are applied to the intensity and topography images in total 12 different images for the feature extraction are available.

The preliminary test setup consists of 10 latent fingerprint captured with a chromatic white light sensor with a resolution of 2540 ppi. We investigate 450000 blocks with fingerprint residue and 450000 blocks of the substrate resulting in 900000 feature vectors. We use a two-fold stratified cross-validation using the J48 decision tree classifier from the WEKA data mining suite[7] for determining the recognition accuracy. The preliminary evaluation exclusively based on the two proposed semantic features for the 12 data sets indicates a significantly improved recognition accuracy of 73.7% for fingerprints on brushed stainless steel. In Hildebrandt et al.[1] only 63.2% of the blocks are correctly classified on this particular substrate using a J48 classifier.

In the final paper we investigate the recognition accuracy for the complete feature set for seven substrates: white furniture surface, brushed stainless steel, metallic paint, matte paint, beech veneer, golden oak veneer and aluminum foil. For each substrate ten latent fingerprints are captured using a chromatic white light sensor resulting in 70 latent fingerprints in total. The evaluation is carried out using the same test protocol as used in Hildebrandt et al.[1]: at first the recognition accuracy of various classifiers is determined in a two-fold stratified cross-validation; secondly the impact of the contrast enhancement on the extraction and matching of biometric features is evaluated using the NIST biometric imaging software[8].

The fist evaluation step is necessary to determine the error rates of the approach which addresses the Daubert criterion[9] "known or potential rate of error". The second evaluation step is used in order to show that the suggested approach does not alter biometric features of the fingerprint.

References

[1] Hildebrandt, M., Dittmann, J., and Vielhauer, C., "Statistical latent fingerprint residue recognition in contact-less scans to support fingerprint segmentation," in [Proceedings of 18th International Conference on Digital Signal Processing (DSP)], (2013).

[2] Benford, F., "The law of anomalous numbers," Proceedings of the American Philosophical Society 78(4), 551–572 (1938).

[3] Bleay, S. M., Sears, V. G., Bandey, H. L., Gibson, A. P., Bowman, V. J., Downham, R., Fitzgerald, L., Ciuksza, T., Ramadani, J., and Selway, C., [Fingerprint Source Book], Home Office Centre for Applied Science and Technology (CAST) (2012). [Online]. Available: http://www.homeoffice.gov.uk/publications/science/cast/crime-investigation/fingerprint-source-book-2012/.

[4] Hong, L., Wan, Y., and Jain, A., "Fingerprint image enhancement: algorithm and performance evaluation," Pattern Analysis and Machine Intelligence, IEEE Transactions on 20, 777 –789 (aug 1998).

[5] Makrushin, A., Hildebrandt, M., Fischer, R., Kiertscher, T., Dittmann, J., and Vielhauer, C., "Advanced techniques for latent fingerprint detection and validation using a cwl device," in [Proc. SPIE 8436], (2012).

[6] Hu, M.-K., "Visual pattern recognition by moment invariants," Information Theory, IRE Transactions on 8(2), 179–187 (1962).

[7] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., "The WEKA data mining software: An update," SIGKDD Explorations 11(1), 10 – 18 (2009).

[8] National Institute of Standards and Technology, "NIST Biometric Image Software," (2013).

[9] Dixon, L. and Gill, B., [Changes in the Standards for Admitting Expert Evidence in Federal Civil Cases Since the Daubert Decision], RAND Institute for Civil Justice (2001).

9028-8, Session 2

Robust human face recognition based on locality preserving sparse over complete block approximation

Dimche Kostadinov, Svyatoslav V. Voloshynovskiy, Sohrab Ferdowsi, Univ. of Geneva (Switzerland)

No Abstract Available

9028-9, Session Key1

Digital wallet and mobile payment *(Keynote Presentation)*

Sunil K. Jain, Intel Corp. (United States)

Digitization of the traditional leather wallet goes far beyond payments. It's about identification, access, loyalty, privacy, security, intent, ease, likes, dislikes, trust, and much more. The social, economic and scientific ramifications of digitized wallets could be profound. Digital Wallet opens unprecedented avenues for sophisticated data analytics on all transactions be it as simple as a casual business card exchange or petty purchase from a vending machine, or as complicated as smart watermarking of black money, intelligent deterring of tax evasion maneuvers, or accurate zoning and tracing of crime & terrorism. No wonder the Press and the Analyst communities seem quite excited about this new trend. Businesses are aspiring to ride the Digital Wallet wave.

Realistically, Digital Wallet is in its infancy and currently mired with agonizing teething troubles. Maturity will come, but slow. It'll be several years before the wallet app writers, phone & smartcard makers, ingredient technology suppliers, carriers & operators, trusted service managers, merchants and institutions will arrive mutually agreeable business arrangements. And it'll take many more years further for the Governments and Regulators to start taking full advantage of the efficiency, accuracy and traceability that Digital Wallet innately offers. It's certainly a good news that a worldwide acknowledgement is developing as to how the Digital Wallet impacts the day-to-day lives at very personal, local and contextual levels, and why it requires an unprecedented mix of standardization with customization.

In this talk, the speaker will draw from several key successful and not so successful examples from across the globe, and lead the audience through a systematic thought experiment, motivating them to design a practical digital wallet that feels 'personal, secured & scalable enough' to carry them through 2050.

9028-4, Session 3

Watermarking textures in video games

Huajian Liu, Waldemar Berchtold, Marcel Schäfer, Martin Steinebach, Fraunhofer-Institut für Sichere Informations-Technologie (Germany)

No Abstract Available

9028-10, Session 3

Blind detection for disparity-coherent stereo video watermarking

Cesar Burini, Séverine Baudry, Gwenaël Doërr, Technicolor S.A. (France)

No Abstract Available

9028-11, Session 3

Tuple decoders for traitor tracing schemes

Jan-Jaap Oosterwijk, Technische Univ. Eindhoven (Netherlands) and Irdeto B.V. (Netherlands); Thijs Laarhoven, Technische Univ. Eindhoven (Netherlands); Jeroen Doumen, Irdeto Access B.V. (Netherlands)

No Abstract Available

9028-12, Session 3

Feature-based watermark localization in digital capture systems

Vojtech Holub, Tomas Filler, Digimarc Corp. (United States)

Digital watermarking is an art of embedding auxiliary payload into cover media such that the change remains invisible to humans but allows machines to reliably extract the payload even after common signal-processing operations. We focus on application where watermarks are printed on consumer packages and carry information similar to one found in barcodes. Goal of this paper is to design the very first module in the watermark reading framework that quickly analyzes the image and rejects areas that most likely do not contain watermark – the watermark detector. In this abstract, we present analysis of a specific instance of this problem formulation where each area of captured frame is classified as being watermarked or not. We borrow the basic tools from the field of image steganalysis and, for complexity reasons, deliberately constraint ourselves to image features constructed directly in spatial domain. This abstract shows that the proposed detector is able to reduce the number of images unnecessarily processed by the watermark reader by 60?95% depending on chosen missed-detection rate. We believe the feature extraction and classification process is simple enough to be implemented in software or as part of the camera hardware providing the guidance to the detector in real-time.

9028-13, Session 3

Self-synchronization for spread spectrum audio watermarks after time scale modification

Andrew Nadeau, Gaurav Sharma, Univ. of Rochester (United States)

De-synchronizing operations such as insertion, deletion, and warping, pose significant challenges for watermarking. Because these operations are not typical for classical communications, watermarking techniques such as SS (spread spectrum) can perform poorly. Conversely, specialized synchronization solutions can be challenging to analyze/optimize. This paper addresses desynchronization for blind SS watermarks, detected without reference to the original unmodified signal, by using the robustness properties of short blocks. The proposed self-synchronization scheme detects watermark blocks by correlation using an efficient DTW (dynamic time warping) search for sequences of block detection peaks. This differs from synchronization schemes that must first locate invariant features of the original signal, or estimate and reverse desynchronization before detection. Without these synchronization steps, novel analysis for the proposed scheme can build on classical SS concepts. Specifically, analysis presented here models and optimizes the trade-off between: short watermark blocks for robustness to desynchronizing operations, and long SS sequences that reduce non-desynchronizing interference. A proof-of-concept watermarking scheme is presented to demonstrate simultaneous robustness to 1) MP3 compression, 2) insertion/deletion, and 3) TSM (time-scale modification).

9028-14, Session 3

Drift-free MPEG-4 AVC semi-fragile watermarking

Marwen Hasnaoui, Mihai Mitrea, Télécom SudParis (France)

Introduction:

Intra frame drifting is a major concern for all types of MPEG-4 AVC compressed-domain video processing applications: due to the MPEG-4 AVC intra prediction paradigm, the modification of one block generally results in the alteration of the neighboring blocks. Moreover, such an effect can be propagated several times (according to the visual content, encoding configuration, etc), in an un-controlled way. This drift effect is of particular importance in watermarking applications, where it can depreciate the performances of all the three watermarking properties: data payload, transparency and robustness.

State of the art:

In order to avoid the drift drawbacks, two classes of solutions are currently considered in the literature: compensation-based and selection-based. The former (compensation-based) consists in inserting the mark in the compressed domain, estimating the sub-sequent drift distortions and in compensating them by decoding/re-encoding operations. This way, the side effects can be completely avoided at the expense of increasing the computational complexity and of modifying the compressed stream parameters (hence, of implicitly adding some involuntary attacks). The latter (selection-based) consists in restricting the insertion domain: only the blocks which are not involved in the prediction process can be considered for the mark insertion. This way, no drift effects occur, the computational complexity is kept constant but the data payload is drastically reduced; moreover, the security of the system is also implicitly reduced, making it easier for a brute force attack to be performed.

Paper main contribution:

In order to achieve drift-free watermarking insertion, while keeping the same data payload, robustness, computational complexity, and security constraints, the present paper follows a different approach. It algebraically models the drift behavior and allows the insertion procedure to be customized so as to get rid of this undesired effect.

Method presentation:

When trying to avoid the drift distortions for MPEG-4 AVC watermarking, a two-folded difficulty is encountered. On the one hand, the mark is inserted in the compressed domain; in this respect, several studies demonstrated that the quantized DCT-transformed prediction errors ensure an optimal trade-off among the watermarking functional properties. On the other hand, the prediction modes are established and the prediction itself is performed in the pixel domain. Hence, the dual compressed-uncompressed representations should be considered when eliminating the drift undesired effects.

In order to avoid any decoding – reencoding operation for each watermarked block and for all of its neighbors, our study starts by investigating the analytic expressions of the MPEG-4 AVC encoding operations.

MPEG-4 AVC features 13 prediction modes; despite their peculiarities, all of them act in the pixel domain and compute the predicted blocks based only on the most right column and bottom row of the reference block. Hence, we shall achieve drift-free insertion by constraining the insertion procedure to not change the most right column and bottom row of the host block in the pixel domain. Further on, the MPEG-4 transformations connecting the most right column and bottom row of a pixel block to their corresponding elements in the quantized DCT-transformed prediction errors are expressed by a chain of matrix multiplications. In order to ensure that the most right column and bottom row of the host block are not changed in the pixel domain when the insertion takes place in the compressed domain, we compute an additional matrix by which the watermark elements (represented as 4x4 matrices) are a priori multiplied. This matrix is denoted by DF and its elements are DF = [0 -0.26 1.24 2.57; -0.26 -0.14 1.21 1.82; 1.24 1.21 3.18 4.82; 2.57 1.82 4.82 8.42].

Experimental results:

The experiments consider semi-fragile watermarking applications. They are performed on a video corpus composed of 8 video sequences of about 10 minutes each, downloaded from Internet or recorded under the framework of the SPY (Surveillance imProved sYstem) ITEA2 project. This corpus has a heterogeneous content (city streets, highways, industrial objectives, shopping centers) and arbitrarily changing lighting conditions (indoor and outdoor scenes, natural and artificial lightening). It is encoded in MPEG 4 AVC in Baseline Profile (no B frames, CAVLC entropy encoder) at 512 kbps, 576x576 pixel frames; the GOP size is set to 8.

An m-QIM (multiple-symbols Quantizing Index Modulation) insertion method is adapted according to the advanced drift-free solution. The experiments are devoted to the evaluation of the impact of the new masking model in the watermarking transparency.

The data payload is set to 100 bits/s. The robustness is set at a BER = 0.1± 0.03 after transcoding attacks (down to 50% from the original video stream size). The fragility is set so as to identify content altered blocks of 1/81 from the frame size and with a temporal accuracy of 3s.

The transparency is evaluated according to three types of objective measures: pixel difference-based measures (peak signal to noise ratio – PSNR, absolute average difference – AAD, peak mean square error – PMSE and image fidelity IF), correlation based measures (normalized cross correlation - NCC), and psycho-visual measures (digital video quality - DVQ).

The new drift-free insertion results in average transparency gains of 2 dB in PSNR of 0.01 in IF, of 0.41 in AAD, of 107 in PMSE, 0.01 in NCC, and of 22 in DVQ.

Conclusion:

The present study deals with drift-free semi-fragile watermarking. First, by considering the analytic expressions of the MPEG-4 AVC encoding operations, it algebraically represents the watermark insertion as an optimization problem. Second, it solves this problem under drift-free constraints. This way, the DF= [0 -0.26 1.24 2.57; -0.26 -0.14 1.21 1.82; 1.24 1.21 3.18 4.82; 2.57 1.82 4.82 8.42] matrix which should multiply the mark prior to its insertion is computed. The experiments show significant gains in transparency (e.g. 2dB in PSNR) for fixed data payload, robustness and fragility constraints. Note that the DF matrix is independent with respect to the video content and/or encoding; hence, its use does not represent an additional attack and does not increase the computational complexity.

9028-15, Session 4

Cover estimation and payload location using Markov random fields

Tu-Thach Quach, Sandia National Labs. (United States)

Payload location is an approach to find the message bits hidden in steganographic images, but not necessarily their logical order. Its success relies primarily on the accuracy of the underlying cover estimators and can be improved if more estimators are used. This paper presents an approach based on Markov random field to estimate the cover image given a stego image that has been modified via LSB replacement. It uses pairwise constraints to capture the natural two-dimensional statistics of cover images and forms a basis for more sophisticated models. Experimental results show that it is competitive against current state-of-the-art estimators and can locate payload embedded by simple LSB replacement and group-parity steganography. Furthermore, when combined with existing estimators, payload location accuracy improves significantly.

9028-16, Session 4

A mishmash of methods for mitigating the model mismatch mess

Andrew D. Ker, Univ. of Oxford (United Kingdom); Tomas Pevny, Czech Technical Univ. in Prague (Czech Republic)

Approaching steganalysis as a problem of binary classification has been very successful, but such a scenario assumes that the detector has access to the steganographic embedding method and, crucially, the cover source used by the sender. In reality one cannot normally obtain the exact cover source, and in practice it is necessary to train the classifier on a different one with hopefully-similar characteristics. This induces the model mismatch problem, which has been demonstrated to reduce steganalysis accuracy by very significant, and unpredictable, amounts. It is difficult for a practitioner to trust the output of their steganalysis if its accuracy is unpredictable.

In this paper we describe various methods attempting to attenuate the model mismatch problem.

9028-17, Session 4

Study of cover source mismatch in steganalysis and ways to mitigate its impact

Jan Kodovsky, Vahid Sedighi, Jessica Fridrich, Binghamton Univ. (United States)

When a steganalysis detector trained on one cover source is applied to an image from a different source, generally the detection error increases due to the mismatch between both sources. In steganography, this situation is recognized as the so-called cover source mismatch (CSM). The drop in detection accuracy depends on many factors, including primarily the properties of both sources, the type of the machine learning tool (the classifier construction), the feature space used to represent the covers, and the type of steganographic algorithm. Although well recognized as the single most important factor negatively affecting the performance of steganalyzers in practice, the CSM received surprisingly little attention from researchers. One of the reasons for this is the difficulty of the subject and the diversity with which the CSM can manifest. In this paper, we take the logical first step towards solving this problem: we introduce a methodology for evaluating the severity of the CSM, demonstrate the CSM on a few examples, and investigate various techniques to mitigate the negative impact of the CSM on detection accuracy.

9028-18, Session 4

Implementing the projected spatial rich features on a GPU

Andrew D. Ker, Univ. of Oxford (United Kingdom)

No Abstract Available

9028-35, Session Demo2

Self-verifiable paper documents and automatic content verification

Yibin Tian, Xiaonong Zhan, Chaohong Wu, Wei Ming, Konica Minolta Systems Lab. (United States)

Paper document forgery is a significant issue that governments and businesses face world-wide. Paper document authentication is the process of verifying the authenticity or originality of a paper document [1]. Depending on specific applications, paper document authenticity may refer to media or content authenticity, or the combination of the two. In the case of content authenticity verification, the document content is examined to ensure that it has not been altered. Alterations may occur due to deliberate effort of forgery or accidental events. In this report, we focus on content verification.

Manual verification of paper document content is a common task in various office settings. It is time-consuming and mundane, and prone to human errors. Optical Character Recognition (OCR) has been successfully utilized in various document analysis and management applications [2]. However, the high accuracy of OCR is achievable only on printed text with good image quality or with post-processing human correction [3]. In addition, OCR is language dependent. Though some OCR software can automatically identify languages from document images [4], correct language identification is almost impossible in some multiple-language documents where only a small number of words appear in a different language. While OCR based content verification is constrained by the limitations of OCR software, image based content verification is more versatile. It can verify documents containing handwritings and mixed-languages in a unified approach.

A number of image based paper document content verification methods have been proposed. Unfortunately, most of these methods can only detect whether alterations have occurred as the data hiding or hashing schemes utilized only store limited amount of information of the original document [5-7]. Yang et al. suggested a self-verifiable two-layer binary document that can detect where alterations have occurred (tampering/alteration localization), but there are some limitations on the document, such as content amount and image uniformity [8]. More recently, Sankarasubramaniam et al. used Error Correction Code (ECC) to achieve pixel-level alteration localization. Though its accuracy is high, the method is sensitive to photocopying noise [9].

We have developed a simple solution for the creation and automatic content verification of a low-cost self-verifiable paper document that can achieve symbol-level accuracy of alteration localization. The image of an original document is decomposed to symbol templates and their corresponding bounding boxes. The resulting data is further compressed and encrypted, and encoded in high-capacity color barcodes. The original image and barcodes are printed on the same paper to form a self-verifiable authentic document. During content verification, the paper document is scanned to obtain the barcodes and target image. The original image is reconstructed from data extracted from the barcodes, which is then compared to the target image. The verification is carried out hierarchically from the entire image down to word and symbol levels. The proposed verification method is inexpensive, robust and fast.

Evaluation on 216 character tables and 102 real documents achieved greater than 99% alteration detection rate and less than 1% false positives at the word/symbol level. The average authentication time for a letter-size page scanned at 600dpi (image size 6600x5100) is less than 30s on a regular office laptop computer (Dell Latitude E6520 with Intel Core i7-2720QM

2.20GHz CPU, 6GB RAM, and 64-bit Windows 7 OS) for C++ implementation using OpenCV and OpenMP libraries [10, 11].

The authors' main contributions [12] are:

(1) The JBIG2 approach for document image compression is extended to achieve extremely high compression ratio (average 260:1 for binary input on the single-page real documents evaluated). Topology-preserving down-sampling is applied to classified symbol templates before the data is converted to bit stream and undergoes conventional data compression.

(2) A simple high-capacity color barcode and a method for its robust decoding are developed to achieve data capacity of 4KB/in2. The barcode is created and printed in CMYK color space, scanned as a RGM image, and decoded in HSV space. Both color and spatial references are built into barcode locators and utilized to reduce the impact of color degradations and interferences during decoding.

(3) A hierarchical content authentication scheme is developed to achieve alteration localization to word and symbol levels and to reduce the sensitivity to noises introduced in printing and scanning. For word and symbol level comparisons, cross-correlation, Hausdorff distance,

and a combination of multiple simple image features are utilized, such as zoning profiles, side profiles, lower-order moments, and topology statistics. The authentication scheme allows easy utilization of parallel processing of multi-core CPU as words and symbols can be processed independently.

References

[1] D. Doermann, "The evolution of document authentication", International Conference on Frontiers in Handwriting Recognition, pp.3, 2010.

[2] H. F. Schantz, The History of OCR, Optical Character Recognition. Recognition Technologies Users Association, 1982.

[3] H. Rose, "How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs", D-Lib Magazine, vol. 15, 2009.

[4] G. S. Peake, T. N. Tan, "Script and language identification from document images", Document Image Analysis, pp. 10-17, 1997.

[5] M. Chen, E. K. Wong, N. Memon and S. Adams, "Recent developments in document image watermarking and data hiding", Proceedings of SPIE, vol.4518, pp.166, 2001.

[6] P.M. Hunag, D. C. Wu and W. H. Tsai, "A novel block-based authentication technique for binary images by block pixel rearrangements", International Confernece on Multimedia and Expo, pp. 903-906, 2004.

[7] M. Jiang, E. K. Wong and N. Memon, "Robust document image authentication", International Conference on Multimedia and Expo, pp.1131-1134, 2007.

[8] H. Yang and A. C. Kot, "Two-layer binary image authentication with tampering localization", International Conference on Acoustics, Speech and Signal Processing, pp.309-312, 2006.

[9] Y. Sankarasubramaniam, B. Narayanan, K. Viswanathan and A. Kuchibhotla, "Detecting modifications in paper documents: A coding appraoch", Proceedsings of SPIE, vol.7534, pp.0A1-0A12, 2010.

[10] G. Bradski and A. Kaehler, Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly Media, 2008.

[11] L. Dagum and R Menon, "OpenMP: an industry standard API for shared-memory programming", IEEE Transactions on Computatioanl Science & Engineering, vol.5, pp.46-55, 1998.

[12] The authors have filed patent applications for the core technology disclosed here.

9028-19, Session Key2

Piracy conversion: the role of content protection and forensics *(Keynote Presentation)*

Richard Atkinson, Adobe Systems Inc. (United States)

In this session, Richard Atkinson (Adobe's chief strategist and leader of their piracy conversion effort) will walk us through their unconventional yet very logical approach to viewing pirate users as customers and how they are responding to win their business. Included will be aspects around how they view the areas of Content Protection and Forensics in terms of user-experience, operational value, and direction.

9028-20, Session 5

Estimation of lens distortion correction from single images

Miroslav Goljan, Jessica Fridrich, Binghamton Univ. (United States)

In this paper, we propose a method for estimation of camera lens distortion correction from a single image. The method estimates the parameters of the correction by searching for a maximum energy of the so-called linear pattern introduced into the image during image acquisition prior to lens distortion correction. Potential applications of this technology include camera identification using sensor fingerprint, narrowing down the camera model, estimating the distance between the photographer and the subject, forgery detection, and improving the reliability of image steganalysis (detection of hidden data).

9028-21, Session 5

A reference estimator based on composite sensor pattern noise for source device identification

Ruizhe Li, Chang-Tsun Li, Yu Guan, The Univ. of Warwick (United Kingdom)

No Abstract Available

9028-22, Session 5

Content identification: binary content fingerprinting versus binary content encoding

Sohrab Ferdowsi, Svyatoslav V. Voloshynovskiy, Dimche Kostadinov, Univ. of Geneva (Switzerland)

No Abstract Available

9028-23, Session 5

Additive attacks on speaker recognition

Alireza Farrokh Baroughi, Scott A. Craver, Binghamton Univ. (United States)

No Abstract Available

9028-24, Session 5

Robust hashing for 3D models

Waldemar Berchtold, Marcel Schäfer, Michael Rettig, Martin Steinebach, Fraunhofer-Institut für Sichere Informations-Technologie (Germany)

No Abstract Available

9028-25, Session 6

(JEI Invited) Content authentication for printed images utilizing high capacity data hiding

Orhan Bulan, Xerox Corp. (United States); Gaurav Sharma, Univ. of Rochester (United States)

Authentication of content in printed images poses a challenge that cannot be addressed by conventional digital signature schemes because under the analog transport provided by the printing channel the verifier does not have access to the original digital content in pristine form. In this paper, we present a method for cryptography-based authentication of the content in printed images that also provides the

capability for identifying localized changes made by informed malicious attackerskey functionality that is missing in print scan robust hashes that have traditionally been used for print content authentication. The proposed method operates by embedding, within the printed image, an authentication signature that consists of an encrypted thumbnail of the image using a high capacity data hiding method for halftone images. To authenticate the content, the embedded signature is extracted from a scan of the printed image and, after decryption, compared against the printed content. An implementation of the method that incorporates human or automated verification and identifies potential local tampering by informed malicious attackers is developed and successfully demonstrated.

9028-26, Session 6

A framework for fast and secure packaging identification on mobile phones

Svyatoslav V. Voloshynovskiy, Maurits Diephuis, Taras Holotyak, Nabil Stendardo, Univ. of Geneva (Switzerland); Bruno Keel, U-NICA Group (Switzerland)

Package identification, as a sub-problem of general physical object security, recently emerged in several domains ranging from pharmaceutical products to cosmetics due to the striking numbers and scale of counterfeiting and its spread worldwide. Despite all efforts of the brand owners and the packaging industry, the end-consumers are not well aware of the particularities of the original packaging design which is often changed for various commercial and technical reasons and the entire spectrum of security features applied to protect a particular brand. A similar situation also exists with the protection of banknotes. To protect the consumer, against for example taking a counterfeited drug, and to create an essential element of a global anti-counterfeiting network, it is highly desirable that the users can perform a verification by themselves. Further more, information from these online verifications may be stored and analysed. The latter will contribute towards accurate and fast information on geographical trends in appearance and distribution in counterfeiting as such items enter the market.

The previous attempts of the security industry to introduce security technologies based on special detectors had little success. Moreover, commercially it is highly unattractive to produce a huge amount of hardware devices which need to be distributed and regularly upgraded globally, simply for verifying individual products.

Although there most certainly is a need for fast online authentication of consumer goods, the end consumers shows little interest in using a specialized device for anti-counterfeiting measures.

Therefore, a natural possibility to perform such a verification on their own portable devices looks extremely attractive for both brand owners and consumers. In addition, such a functionality can be considered as a perfect market analysis tool.

It is not needed to mention that the modern smart phones represent powerful platforms with imaging, computational and communication facilities. However, the computational power and memory of these devices represent about tent percent of those of modern desktop computers. Therefore, the corresponding algorithms should be carefully designed to cope with complexity and storage restrictions.

In this paper, we address the problem of packaging identification using a three-layer system: (a) package recognition, (b) package design integrity authentication and (c) package carrier identification based on microstructure images.

In the first stage, the consumer takes a photo of a package with a mobile phone. The product is automatically recognised by the proposed system. It should be pointed out that the problem of package recognition is considered as $(M+1)$-hypothesis testing were M hypothesis correspond to the M enrolled objects and a complementary hypothesis is used for the rejection option.

It is important that the product identity is recognised with a high precision at the first stage for all further processing and decision making. To achieve this goal with the acquisition based on mobile phones one should carefully address several issues related to the particularities of mobile imaging. In particular, it concerns the high variability of lightening conditions, defocusing, affine and projective geometrical distortions and cropping due to the limited field of view in the macro mode. The celebrated recent approaches based on local invariant descriptors and bag-of-features (BOF) can hardly cope with all these distortions. Moreover, most of these methods have been developed for the visual retrieval problem, i.e., k-NN search, while the considered problem requires an 1-NN solution. Therefore, this paper presents a new approach based on local and global descriptors containing geometrical information allowing to perform the product recognition with high precision.

The proposed approach is based on multi-BOFs representation of a package image in a sparse alphabet of features and a new search algorithm based on compressive sensing. We also introduce a concept of feature reliability that makes the construction of multi-BOFs extremely robust to variations in the acquisition process. The search algorithm essentially resembles a hybrid approach that combines the soft fingerprinting decoding and the well-known hierarchical tree.

In the second stage, a set of local micro-features corresponding to the package design is retrieved and the system performs the verification of its authenticity. This problem is formulated as binary hypothesis testing. The package design is represented with high precision by a set of features that include both individual and group descriptors on different hierarchical levels.

The high score of correspondence between the acquired features and the template features leads to a positive decision. Otherwise, the package its authenticity is rejected at this stage.

Although the design of packaging can be reproduced with a certain level of precision, it is surprising that many counterfeited products posses a significant degree of difference which is not always immediately detected by the naked eye or even a trained professional. The level of deviation in package design that can be detected by the developed system is estimated to be around 20-25 mum.

Finally, at the third level, the object identity is verified based on the unique non-clonable features of the package surface or the printed ink. These features are unique for each object and are very difficult to reproduce or clone, even for the brand owners possessing complete knowledge of the entire manufacturing cycle. The main challenge at this stage consists in the selection of distinguishable features and synchronization. We have tested two systems based on explicit synchronization based on graphical design elements and feature extraction from the synchronized images and local descriptors co-occurence matching. The second approach is similar in spirit to the BOF and heavily relies on the geometrical appearance of features. Both approaches are validated on the FAMOS dataset consists of 5'000 microstructure images acquired by two cameras with different resolutions.

{bf Preliminary results}

The first stage consisting in the package recognition was tested on 150 real pharmaceutical and cosmetic packages enrolled by the optical scanner Epson Perfection 4990 with 600 dpi. For the benchmarking purposes, we have tested the SIFT and color descriptors clustered based on the soft-encoding algorithms such as sparse coding and LLC as well as proposed method for different numbers of clusters and decomposition levels. At the recognition stage, the package images have been acquired by Samsung Galaxy S3 in the standard Android imaging application with the full view package mode taken 6 times under different angles and distances and cropped view mode with 10 images. Our best obtained results indicate that SIFT descriptors with 750 final clusters succeeded to reliably, i.e., with the probability 1, recognize all packages in the full view mode, while in the cropped view mode the probability of successful recognition has been reduced to 0.94. The addition of color descriptor enhanced the recognition rate to 0.99. We expect to enhance the system to achieve the perfect recognition and currently we test 250 packages.

The package design integrity authentication was tested on Samsung Galaxy S3. The template of authentic package was tested versus the acquired image for 48 authentic packages and 10 "counterfeited" examples with small deviations in the design. We have received the perfect acceptance of all authentic packages and rejection of all

counterfeited items under approximately 10 cm distance to the item and rotation in the range of 30 degrees.

At the last stage of package carrier identification based on microstructure images, we have tested the package identification based on 5'000 microstructure images acquired 3 times. As the reference implementation we have used the system with the synchronization where we have achieved the probability of correct identification 1. The system based on the co-occurence of SIFT descriptors, i.e, without the synchronization, produced the best result with the probability of correct identification 0.96. In both cases, we have ensured the probability of false acceptance equals to 0.

In the case of the paper's acceptance, we intend to perform an on-line demonstration of the proposed system on several examples of authentic and counterfeited products. In the initial stage, we hope to obtain useful feedback from the academic community and industry as well as to stimulate further research in this direction. For this purpose, we intend to share all acquired databases on-line for public access.

9028-27, Session 6

Printer technology authentication from micrometric scan of a single printed dot

Yves Delignon, Quoc Thong Nguyen, Télécom Lille 1 (France); Lionel Chagas, Institut National Polytechnique de Grenoble (France); François Septier, Télécom Lille 1 (France)

In this paper we are concerned by authentication of printer technologies from microscopic analysis of paper print. At this scale, a print is made of regularly spaced dots whose shape varies from a print to another and also inside the same document. Thus, dot at the microscopic scale can be considered as an intrinsic signature of printer technologies. Modelling and estimating such a signature for the authentication of printer technologies are really challenging.

In this paper, we propose an original modelling of the micrometric scan of document printing. It consists in an extension of the binary response model which takes into account the dot shape. The digital image of a dot is therefore modelled as a set of random pixels distributed following to the so called inverse link function which depends on the center of the dot, its spreading and its shape.

A maximum likelihood estimation algorithm is provided in order to estimate the location, the scale and shape parameter of the dot. From experimental results on three different printer technologies (inkjet, laser and offset), we show that the shape parameter enables to discriminates them and enables to design authentication scheme of paper document.

9028-28, Session Key3

Photo forensics from shadows and shading *(Keynote Presentation)*

Hany Farid, Dartmouth College (United States)

I will describe a method to detect physical inconsistencies in lighting from the shading and shadows in an image. This method extracts a multitude of shading- and shadow-based constraints on the projected location of the illuminating light source. The physical consistency of a collection of such constraints is posed as a linear programming problem. A feasible solution indicates that the combination of shading and shadows is physically plausible, while a failure to find a solution provides evidence of photo tampering. [joint work with Dr. Eric Kee (Dartmouth/Columbia) and Prof. James O'Brien (Berkeley)]

9028-29, Session 7

Digitized locksmith forensics: automated detection and segmentation of toolmarks on highly structured surfaces

Eric Clausing, Claus Vielhauer, Otto-von-Guericke-Univ. Magdeburg (Germany) and Fachhochschule Brandenburg (Germany)

Locksmith forensics is an important area in crime scene forensics. Due to new optical, contactless, nanometer range sensing technology, such traces can be captured, digitized and analyzed more easily allowing a full-fletched digital forensic investigation. In this paper we present a significantly improved approach for the detection and segmentation of toolmarks on surfaces of locking cylinder components (using the example of the locking cylinder component 'key pin') acquired with a 3D Confocal Laser Scanning Microscope. This improved approach is based on our work in [CKD1] using a block-based classification approach with textural features. In [CKD1] we achieve a solid detection rate of 80-85%. Here, in this paper we improve, expand and fuse this prior approach with additional features from acquired surface topography data and an image processing approach using adapted Gabor filters. In particular we are able of raising the detection and segmentation rates above 90% and can provide a precise pixel-based segmentation as opposed to the rather imprecise segmentation of our prior block-based approach. Additionally we present alternatives for a suitable fusion of pixel and block-based approaches on topography, texture and intensity data by discussing our motivations.

1. Application Context

Toolmark investigation is a main aspect of criminal forensics. In this case toolmarks can be everything from clearly visible crowbar impacts to subtle microscopic traces of illegal opening attempts on locking cylinders. Especially for the latter of both a time-consuming, difficult manual process must be performed to detect, acquire and interpret relevant traces in the form of toolmarks. Such an investigation process is nowadays performed manually by forensic experts without the help of technical support except for a classic light microscope. By providing technical support for detection, segmentation and interpretation of toolmarks, one can significantly improve the efficiency and reproducibility of the classic forensic process.

2. Technical Challenge

The surfaces of mechanically fabricated and frequently used metal components (such as the components of a locking cylinder) are naturally cluttered with a vast number of toolmarks of either relevant (toolmarks originating from illegal opening methods) or irrelevant (e.g. toolmarks originating from fabrication) origin. In most cases these toolmarks form complex formations with relevant and irrelevant traces overlapping and distorting each other. Even for the highly experienced and well trained forensic expert, the differentiation of one or the other is not a simple task to solve. With our approach we present a method to automatically differentiate between relevant and irrelevant toolmarks, to segment these relevant toolmarks as precise as possible and to visualize results in an adequate way for forensic experts. This includes the adequate acquisition of the surfaces as proposed in [CKD1], and the processing and preparation of the acquired data (as proposed in [CKD2]) to allow for the application of our segmentation approaches. Following that we need to adapt and specialize the chosen image processing methods and design an adequate method to fuse the gained information of the segmentation approaches for intensity, topography and texture data to one precise segmentation.

3. Our Approach

As basis for our improvements, we use findings from acquisition and pre-processing in [CKD1] and [CKD2]. In [CKD1] we propose an acquisition method which allows for a stepwise partial scanning of the whole key pin surface in about 45 partial scans. In consequence the block-based texture classification method proposed as well in [CKD1] using Gray-Level-Cooccurrence-Matrices (GLCM, see [HSD73]) only works on such partial scans. As the GLCM approach is not rotation invariant and we

introduced a new digital SIFT-based assembling (SIFT, see [Low04]) of the key pin surface as a whole in [CKD2], we have to adapt the texture classification approach to fit this new representation. To compensate for the rotational invariance of the GLCM approach we propose a direction dependent calculation of the GLCM features. For the pixel-based Gabor filtering approach (Gabor filtering, see [Dau85][Dau88]) we use a quite similar concept by utilizing known orientations and patterns of e.g. fabrication toolmarks to specifically dampen or amplify certain trace structures on certain positions of the surface. For the topography approach we use a combination of pixel-based and block-based straightforward thresholding methods along with classification-based methods (tested with a set of fifteen different classifiers)with a feature set including a variety of roughness and topographical texture features. The information gained from these three sub-approaches (for intensity, topography and texture data) are fused in an adequate way to utilize the advantages of each specific approach and optimize the possible segmentation precision.

4. Preliminary Results

Our approach is tested on a preliminary test set of about 1,300 surface sample scans (in 67/33 percentage split of training and test set) of key pins originating from locking cylinders opened with five different opening methods (Normal Key Usage, Single Pin Picking, Raking, Key Bumping and Pick Gun). The surfaces are acquired with the 3D Confocal Laser Scanning Microscope Keyence VK-X 110. Each of the approaches for intensity, topography and the adapted approach for texture alone are able of achieving segmentation rates above 85%. As the fusion of all three approaches is still work in progress and the approaches are still optimized, we cannot give exact results on that but we are positive about achieving segmentation results of above 90% with a segmentation on pixel-level.

5. References

[CKD1] Eric Clausing, Christian Krätzer, Jana Dittmann, and Claus Vielhauer. A First Approach for the Contactless Acquisition and Automated Detection of Toolmarks on Pins of Locking Cylinders Using 3D Confocal Microscopy, In Proceedings of the on Multimedia and security (MM&Sec '12), ACM New York, NY, USA, pages 47-56, 2012.

[CKD2] Eric Clausing, Christian Krätzer, Jana Dittmann, and Claus Vielhauer. A first approach for digital representation and classification of toolmarks on locking cylinders using confocal laser microscopy. In SPIE Security + Defence: Optics and Photonics for Counterterrorism, Crime Fighting and Defence VIII, 854609. SPIE, 2012.

[Dau85] J. Daugman. Uncertainty relations for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. In Journal of the Optical Society of America, volume 2, pages 1160–1169, 1985.

[Dau88] J. Daugman. Complete Discrete 2-D Gabor Transforms by Neural Networks for Image Analysis and Compression. In IEEE Trans on Acoustics, Speech, and Signal Processing, volume 36 (7), pages 1169–1179, 1988.

[HSD73] R. Haralick, K. Shanmugam, and I.Dinstein. Textural features for image classification. In IEEE Transactions on Systems, Man, and Cybernetics SMC, volume 3 (6), pages 610–621, 1973.

[Low04] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. In International Journal of Computer Vision 60 (2), pp. 91-110, 2004.

9028-30, Session 7

Two improved forensic methods of detecting contrast enhancement in digital images

Xufeng Lin, Xingjie Wei, Chang-Tsun Li, The Univ. of Warwick (United Kingdom)

No Abstract Available

9028-31, Session 7

Copy-move forgery detection from printed images

Irene Amerini, Univ. degli Studi di Firenze (Italy); Roberto Caldelli, Univ. degli Studi di Firenze (Italy) and Consorzio Nazionale Interuniversitario per le Telecomunicazioni (Italy); Alberto Del Bimbo, Univ. degli Studi di Firenze (Italy); Andrea Di Fuccia, Anna Paola Rizzo, Luigi Saravo, Prime Minister Office (Italy)

Counterfeiting digital images by means of photo editing tools to alter the original meaning is becoming an immediate and easy practice. Copy-move forgery is the one of the most common ways of manipulating the semantic content of a picture, whereby a portion of the image is copied and pasted once or more times elsewhere into the same image. Image forensics literature offers several examples of detectors for such manipulation [2] and, among them, the most recent and effective ones are those [3][8] based on Zernike moments and [4][5][6] based on Scale Invariant Feature Transform (SIFT) [7]. In particular, the capability of SIFT to discover correspondences among similar visual contents allows the forensic analyst to detect even very accurate and realistic copy-move forgeries.

It could happen, however, instead of a digital image only its analog version may be available. It is interesting to verify whether it is possible to identify tampering from a printed picture rather than its digital counterpart.

Scanned or recaptured (by a digital camera) printed documents are widely used in a number of different scenarios, such as medical imaging, law enforcement, banking and daily consumer use. An example could be a photo published on a newspaper digitally altered by duplicating the crowd, missiles or tanks to amplify the effect, or by hiding some objects or persons.

Dealing with printed images is a novel issue: to the best of our knowledge, the only existing work was proposed in [1] but it takes into account of camera identification issue without considering forgery aspects. In fact, at the state-of-the-art none of the known methods used for digital image tampering discovery has been proven in the printed picture scenario. In this paper, the problem of detecting and localizing copy-move forgeries from a printed picture is focused. The copy-move manipulation is detected by verifying the presence of duplicated patches in the scanned image by using the CMFD (Copy-Move Forgery Detection) method proposed in [6]; such our previous methodology has been adapted in a version tailored for printed image case (e.g. choice of the minimum number of matched keypoints, size of the input image, etc.)

The CMFD algorithm relies on SIFT features matching, then a cluster procedure is performed by a robust clustering based on the J-Linkage algorithm; finally an accurate forgery localization procedure has been set up on the basis of the clusters previously obtained. This is done by resorting at ZNCC (Zero mean Normalized Cross-Correlation) between the checked image and the warped one obtained from the estimated geometric transformation occurred during the tampering attack.

The printing and scanning/recapturing scenario is quite challenging because it involves many diverse kinds of distortions. In fact printing phase can introduce artifacts according to the printer model and its setting, the kind of paper and also the ink type; on the other side, the scanning phase itself affects the recaptured image depending on the type of scanner adopted, acquisition resolution, settings and so on. Also the case of recapturing through a digital camera or a smartphone has been taken into account, thinking that it could be interesting for a real situation in which images have been retaken in an uncontrolled way (e.g. for an on-the-fly analysis on suspicious photos found in a crime scene). Obviously, in this circumstance, the acquisition conditions are more challenging both in terms of illumination and of framing.

The CMFD tool was proved to be robust to a wide range of image processing operations and also to geometrical transformations. It is however worthy wondering whether the tool could still be effective against a digital-analog-digital conversion.

The goal of this paper is to experimentally investigate the requirement set under which reliable copy-move forgery detection is possible. We carry out a series of experiments to pursue, in detail, all the different issues involved in this application scenario. In particular, we have considered two kinds of print circumstances: in the first one, prints have been obtained by our lab printers (laser and inkjet), so this grants a certain level of supervision; in the second case, they have been made by a commercial printing center, thus determining a higher degree of uncertainty. The re-acquisition phase has been carried out by using our lab scanners and smartphones. Experimental results point out that forgery detection is still successful though with reduced performances with respect to the digital case, as expected; the success of detection mainly depends on the printing quality and on the size of the printed picture. Promising results have been obtained in the case of re-acquisition by a digital camera too.

References

[1] M. Goljan and J. Fridrich, Camera identification from printed images, in Proc. SPIE Electronic Imaging, Forensics, Security, Steganography and Watermarking of Multimedia Contents X, San Jose, CA, 2008, vol.6819, pp. OI-1–OI-12.

[2] S. Bayram, H. Sencar, and N. Memon, A survey of copy-move forgery detection techniques. In IEEE Western New York Image Processing Workshop, pp 538–542, 2008.

[3] Seung-Jin Ryu, Min-Jeong Lee, and Heung-Kyu Lee, Detection of copy-rotate-move forgery using Zernike moments. In Proceedings of the 12th International Conference on Information Hiding, Berlin, , pp. 51-65, 2010.

[4] X. Pan and S. Lyu, Region duplication detection using image feature matching. IEEE Transactions on Information Forensics and Security, 5(4):857–867, 2010.

[5] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, and G. Serra, A sift-based forensic method for copy move attack detection and transformation recovery. IEEE Transactions on Information Forensics and Security, 6(3):1099 –1110, 2011.

[6] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, L. Del Tongo, and G. Serra, Copy-Move Forgery Detection and Localization by Means of Robust Clustering with J-Linkage, Signal Processing: Image Communication, vol. 28 (6): 659-669, 2013.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. Int. Journal of Computer Vision, 60(2):91–110, 2004.

[8] S. J. Ryu, M. Kirchner, M. J. Lee and H. K. Lee, Rotation Invariant Localization of Duplicated Image Regions Based on Zernike Moments. IEEE Transactions on Information Forensics and Security, 8(8):1355 –1370, 2013.

9028-32, Session 7

Countering anti-forensics by means of data fusion

Marco Fontani, Mauro Barni, Univ. degli Studi di Siena (Italy)

No Abstract Available

9028-33, Session 7

Image counter-forensics based on feature injection

Massimo Iuliani, Simone Rossetto, Univ. degli Studi di Firenze (Italy); Tiziano Bianchi, Politecnico di Torino (Italy); Alessia De Rosa, Alessandro Piva, Univ. degli Studi di Firenze (Italy); Mauro Barni, Univ. degli Studi di Siena (Italy)

No Abstract Available

Conference 9029: Visual Information Processing and Communication V

Wednesday - Thursday 5 –6 February 2014

Part of Proceedings of SPIE Vol. 9029 Visual Information Processing and Communication V

9029-1, Session 1

An all-zero blocks early detection method for high-efficiency video coding

Zhengguang Lv, Peking Univ. (China); Ronggang Wang, Peking Univ. Shenzhen Graduate School (China)

HEVC, the High Efficiency Video Coding standard, is the most recent joint video project of the ITU-T VCEG and ISO/IEC MPEG standardization organizations, working together in a partnership known as the Joint Collaborative Team on Video Coding (JCT-VC) [1]. HEVC significantly outperforms previous standards such as H.264/AVC in term of coding efficiency. However, it also has a considerable increase in encoder complexity, which is still needed to be decreased. Therefore, reducing the computational complexity of the encoder is vital for this standard.

All-zero blocks (AZBs) are those blocks which all the transformed and quantized coefficients are all zeros. AZBs are quite common in low bit-rate video application [6]. The computations of transform and quantization need not be performed if AZBs can be detected prior to transform and quantization. Thus, early detection of AZBs can effectively reduce the computational complexity. Many works [3]-[10] have been done on H.264/AVC in this filed. For example, the earliest detection method for AZBs proposed in [3] defines a sufficient condition for quantizing all 8?8 floating-point transform coefficients to zero. [4]-[8] propose more precise conditions to detect 4?4 AZBs, in [8] an adaptive method is propose to detect the 4?4 AZBs, in [10] a fast and efficient mode decision algorithm based on all-zero blocks detection method is proposed. All of the zero quantized transform coefficients can effectively reduce the redundant computations.

Methods above are all based on the characteristics of DCT formula and quantization and the simulation platform are all on H.264/AVC. In this paper, all-zero blocks from 4?4 to 32?32 detection method is proposed for inter-prediction mode. Simulation results show that by utilizing our proposed method, the computations of transform and quantization in HEVC can be reduced by about 37% without coding performance loss.

The rest of the paper is organized as follows. Section ? give a review of the existing methods in H.264/AVC. In Section ?, we derive a new AZB early detection method for HEVC through deducing a series of SAD thresholds. Experimental results are presented in Section ?. And Section ? concludes this paper.

9029-2, Session 1

Low-cost multi-hypothesis motion compensation for video coding

Lei Chen, Ronggang Wang, Peking Univ. Shenzhen Graduate School (China); Siwei Ma, Institute of Digital Media, Peking University (China)

In conventional motion compensation for P frame, prediction block is related only with one motion vector. However, this kind of motion compensation does not make good use of spatial and temporal correlation between frames, thus causing the prediction block to be not accurate. In bidirectional motion compensation for B frame, prediction block is derived by two motion vectors, one forward and one backward, thus, making the prediction block more accurate. Generalized B frame[1-2] uses the similar idea to improve the prediction performance of P frame by multi-hypothesis motion compensation (MHMC)[3]. However, in MHMC, at least two motion vectors have to be searched and transmitted in bit stream, thus increase both the encoding complexity and bit-rate.

In this paper, we introduce a new motion compensation technology: low-cost multi-hypothesis motion compensation (LMHMC). This kind of motion compensation is different from conventional motion compensation as prediction block is not only related with motion vector obtained by motion estimation, but also related with motion vectors of neighboring blocks so as to provide higher accuracy for prediction block. Besides, the encoding complexity and bit-rate do not increase since only one motion vector needs to be searched and transmitted in bit-stream.

The proposed method is implemented in the MPEG IVC reference software ITM4.0[4] as an additional mode for P frame. The modified ITM4.0 with our proposed LMHMC method saves about 5% BD rate.

The rest of this paper is organized as follows. Section 2 introduces traditional motion compensation. Section 3 presents our proposed low-cost multi-hypothesis motion compensation in detail. Section 4 gives the experimental results. Finally, we conclude this paper in Section 5.

9029-3, Session 1

An optimized template matching approach to intra coding in video/image compression

Hui Su, Google Inc. (United States) and University of Maryland (United States); Jingning Han, Yaowu Xu, Google (United States)

The template matching prediction is an established approach to intra-frame coding that makes use of previously coded pixels in the same frame for reference. It compares the priorly reconstructed above and left boundaries in searching from the reference area the best matched block for prediction, and hence eliminates the need of sending additional information to reproduce the same prediction at decoder. In viewing the image signal as an auto-regressive model, this work is premised on the fact that pixels closer to the known block boundary are better predicted than those far apart. It significantly extends the scope of the template matching approach, which is typically followed by a conventional discrete cosine transform (DCT) for the prediction residuals, by employing an asymmetric discrete sine transform (ADST), whose basis functions vanish at the prediction boundary and reach maximum magnitude at far end, to fully exploit statistics of the residual signals. It was experimentally shown that the proposed scheme provides substantial coding performance gains on top of the conventional template matching method over the baseline.

9029-4, Session 1

Motion estimation optimization tools for the emerging high efficiency video coding (HEVC)

Abdelrahman Abdelazim, Wassim Masri, Bassam Noaman, The American Univ. of the Middle East (Kuwait)

No Abstract Available

9029-5, Session 2

Efficient determination of intra predictability in H.264/AVC and similar codecs

Seyfullah H. Oguz, Qualcomm Inc. (United States)

An efficient method of determining if a macroblock can be intra predicted

with accuracy according to H.264/AVC intra prediction specification (4x4, 8x8 and 16x16), is introduced. The proposed algorithm can be easily generalized to VP8 and H.265/HEVC including its increased coding unit sizes.

The proposed algorithm is capable of accurately detecting, classifying and reporting directional properties of macroblocks and sub-macroblock regions. It relies on subsampling of pixel domain data on multiple grids rotated with respect to each other with appropriate angular offsets followed by a Hadamard transform based analysis on each subsampling grid. Therefore, only additions, subtractions, absolute value operations and comparisons are utilized by the proposed algorithm resulting in very low computational complexity.

Comparative simulation results demonstrating the detection and R-D performance as well as the computational complexity of the proposal are provided.

9029-6, Session 2

Backwards compatible high dynamic range video compression

Vladimir Dolzhenko, Eran A. Edirisindhe, Loughborough Univ. (United Kingdom); Vyacheslav Chesnokov, Apical Imaging (United Kingdom)

This paper presents a two layer CODEC architecture for high dynamic range video compression.

The research is targeting the ‘compression gap’ in high dynamic range (HDR) imaging: although modern sensor and display technology allow capturing and rendering the scenes with dynamic range exceeding 100dB, the compression algorithms for such data are not wide spread and require specialist equipment. The backwards compatible compression will allow integration of HDR and low dynamic range systems.

The proposed CODEC produces the stream which can be either rendered on the low dynamic range (LDR) devices using conventional equipment or on HDR devices using specific decoder. The CODEC is based on existing video compression algorithms, H.264 or H.265.

To increase usability of the LDR stream the high quality tone mapping operator is required. Experiments shown that the non-uniform tone mapping operators provide better image quality, but make it more challenging to predict HDR content from LDR one. As the tone mapping operators model adaptation process of human visual system, the transform curves are expected to be similar for adjacent pixels and become independent at ? of image size for typical viewing conditions. The model of the tone mapping operator used in this work uses 3D piecewise linear lookup table with nodes along image dimensions and intensity. The research show that although this representation is computationally effective for pixel processing, the adjacent knots are highly correlated. Several techniques were used to compress the data:

The lookup table nodes are predicted from previously decoded frames. Optionally data can be rescaled to accommodate for brightness changes

The nodes are predicted from previously decoded data within frame

The sparse subset of nodes is transmitted first and used as anchors for prediction of remaining nodes.

The residuals are transformed with orthogonal transform, and quantized

The prediction and residual coefficients are arithmetically encoded using the same CABAC as in H.264 algorithms

As the amount of data is relatively small, e.g. 13000 values per frame, non-linear prediction with brute force search can be used to optimize the compressed size.

The HDR data is predicted from the LDR content and effectively only parts that were clipped in the LDR stream are encoded. The clipped parts include out-of-gamut saturated colors and high brightness areas.

The enhancement layer contains the image difference, in perceptually uniform color space, between the result of inverse tone mapped base layer content and the original video stream. Perceptually uniform colorspace enables using standard rate-distortion optimization algorithms, as absolute differences in this colorspace are highly correlated with just noticeable differences.

The encoder uses the standard loops on base and enhancement video layers and on tone mapping operator model, so the prediction or quantization errors are not accumulated.

Different video compression techniques were compared: proposed two-layer, high dynamic range only and simulcast. The quality of high dynamic range content reconstruction was estimated using Mantiuk et al visual difference predictor, HDR-VDP2. The color fidelity was checked using distances in L*a*b colorspace.

9029-7, Session 2

Role-based adaptation for video conferencing in healthcare applications

Oscar Figuerola Salas, Hari Kalva, Antonio Escudero, Ankur Agarwal, Florida Atlantic Univ. (United States)

A large number of health-related applications are being developed using web infrastructure. Video is increasingly used in healthcare applications to enable communications between patients and care providers. We present a video conferencing system designed for healthcare applications. In face of network congestion, the system uses role-based adaptation to ensure seamless service. A new web technology, WebRTC, is used to enable seamless conferencing applications. We present the video conferencing application and demonstrate the usefulness of role based adaptation.

9029-8, Session 2

(JEI Invited) Video compressed sensing using iterative self-similarity modeling and residual reconstruction

Yookyung Kim, Samsung Advanced Institute of Technology (Korea, Republic of); Han Oh, Samsung Electro-Mechanics (Korea, Republic of); Ali Bilgin, The Univ. of Arizona (United States)

Compressed sensing (CS) has great potential for use in video data acquisition and storage because it makes it unnecessary to collect an enormous amount of data and to perform the computationally demanding compression process. We propose an effective CS algorithm for video that consists of two iterative stages. In the first stage, frames containing the dominant structure are estimated. These frames are obtained by thresholding the coefficients of similar blocks. In the second stage, refined residual frames are reconstructed from the original measurements and the measurements corresponding to the frames estimated in the first stage. These two stages are iterated until convergence. The proposed algorithm exhibits superior subjective image quality and significantly improves the peak-signal-to-noise ratio and the structural similarity index measure compared to other state-of-the-art CS algorithms.

9029-9, Session 3

A novel error metric for parametric fitting of point spread functions

Jonathan D. Simpkins, Robert L. Stevenson, Univ. of Notre Dame (United States)

Established work in the literature has demonstrated that with accurate

knowledge of the corresponding blur kernel (or point spread function, PSF), an unblurred prior image can be reliably estimated from one or more blurred observations. It has also been demonstrated, however, that an incorrect PSF specification leads to inaccurate image restoration. In this paper, we present a novel metric which relates the discrepancy between a known PSF and a choice of approximate PSF, and the resulting effect that this discrepancy will have on the reconstruction of an unblurred image. Such a metric is essential to the accurate development and application of a parameterized PSF model.

Several error measures are proposed, which quantify the inaccuracy of image deblurring using a particular incorrect PSF. Using a set of simulation results, it is shown that the desired metric is feasible even without specification of the unblurred prior image or the radiometric response of the camera. It is also shown that the proposed metric accurately and reliably predicts the resulting deblurring error from the use of an approximate PSF in place of an exact PSF.

9029-10, Session 3

Joint deblurring and demosaicking of CFA image data with motion blur

Ruiwen Zhen, Robert L. Stevenson, Univ. of Notre Dame (United States)

No Abstract Available

9029-11, Session 3

Parametric phase information based 2D Cepstrum PSF estimation method for blind de-convolution of ultrasound imaging

Jooyoung Kang, Sung-Chan Park, Jung-ho Kim, Jongkeun Song, SAMSUNG Electronics Co., Ltd. (Korea, Republic of)

2. Problem Statement

Medical ultrasound imaging is a popularly used diagnostic modality in visualization medical tool. In contrast to other medical imaging modalities such as, Magnetic Resonance Imaging, Computed Tomography or X-Ray, ultrasound imaging is considered to be cost effective, easily movable, non-invasive, safety and real time imaging system. However, there are many disadvantages in the ultrasound imaging system such as having a low quality image with speckle noise and blurring in the area of ultrasound image which is not well focused. The reason for the speckle noise is the coherent character of the ultrasound imaging process, and the blurring or low spatial resolution of ultrasound imaging is caused by the finite bandwidth of the ultrasound pulse in the axial and lateral directions. These limitations of ultrasound imaging make depress the ability to detect the clinically important details. For the limitations, the de-convolution can be one of the primary solutions for the low spatial resolution or blurring of ultrasound imaging. The physical phenomenon of blurring in ultrasound imaging is mathematically modeled as convolution of the point spread function (PSF) of the ultrasound probe and the tissue reflectivity function, which indicate the linear operation between acoustical pulse and observed human tissue. Therefore, de-blurring can be achieved by de-convolving the ultrasound images with an estimated of corresponding PSF. However, it is hard to attain an accurate estimation of PSF due to the unknown properties of the tissues of the human body through the ultrasound signal propagates. In this paper, we present a new method for PSF estimation in the Fourier domain (FD) based on parametric minimum phase information, and simultaneously, it performs fast 2D de-convolution in the ultrasound imaging system.

3. Review of prior work

As mention above, ultrasound images can be considered as a convolution model of the PSF of ultrasound probe and the tissue function which characterized linear interaction between acoustic field and observed tissue. One way to improve the spatial resolution of medical ultrasound images is the de-convolution method of the observed radio frequency image by PSF. In the field of image de-convolution, two kinds of approach have been developed depending of capability of prior knowledge of point spread function. These two distinct methods are called blind and non-blind de-convolution. In this paper, we propose one of fast blind de-convolution which able to estimate the unknown properties of the tissues of the human body through the ultrasound signal propagates, more adaptively. In blind de-convolution for ultrasound imaging, the estimation of PSF is the crucial step in the whole process. So far, the homomorphic filtering approach has been mainly applied to find the amplitude and phase of Fourier transform of the PSF in the Fourier domain. Several homomorphic filtering methods used in estimation of the radial PSF are compared in a certain paper. Also, algorithms using the spectral root cepstrum, generalized cepstrum and complex cepstrum methods based on phase unwrapping, logarithmic derivative and polynomial rooting are considered. However, these kinds of methods utilize the logarithmic derivative and phase unwrapping which more computationally demanding. Therefore, the purpose of this paper is two main steps. First, it provides a theoretical framework, establishing that the FD-phase information of the PSF with the parametric weighting factors based on minimum phase assumption and comparing the real PSF from actual transducer being used in offline. A 2D parametric minimum phase signal can be made minimum phase by multiplying with a 2D exponential sequence, having parametric weighting factors. This methods estimate based on minimum phase assumption without considering phase unwrapping and linear phase elimination, thus it is much efficient and less computational. Second, for a de-convolution part with the estimated PSF, we focused on the low complexity issue. So, we are using the Weiner Filter and fast de-convolution technique using hyper-Laplacian priors, which is several orders of magnitude faster than existing techniques that use hyper-Laplacian priors.

4. Proposed approach

The algorithm is based on two-stage reconstruction scheme, in which the estimation of 2D PSF first and image restoration second. In the estimation of 2D PSF block, we obtain the amplitude spectrum PSF based on minimum phase assumption, and the FD-phase spectrum PSF information based on parametric minimum phase assumption and comparing the real PSF from actual transducer being used in offline without considering phase unwrapping methods. In the image restoration block, we use the non-iterative fast de-convolution method which is several orders of magnitude faster than existing techniques that use hyper-Laplacian priors.

Estimation of the PSF

The mathematical model for ultrasound image formation is given by the linear operation equation below (1), where f_s (m,n) is the complex tissue reflectivity function in the IQ segment area, respectively, h_s (m,n) is the spatial invariant PSF function of the imaging process within the IQ segment area, and u_s (m,n) is the measurement noise term. In our algorithm, we classify the each different depth (about 5different depths) for the IQ segment area. And m and n denote the axial and lateral indices of the image samples in the spatial domain, and the operation * stands of 2-D convolution.

g_s (m,n)=f_s (m,n)*h_s (m,n)+u_s (m,n) (1)

The convolution model (1) can be specified alternatively in the frequency domain by taking the Discrete Fourier Transform (DFT) as follow (2):

G_S (w_1,w_2)=F_S (w_1,w_2) ??H?_S (w_1,w_2)+U_S (w_1,w_2) (2)

With the upper-case letters in (2) denoting the DFT of their lower-case counterparts in (1). The magnitude of the DFT of the PSF function is estimated using the linear relationship between the log spectra of the IQ-image, the PSF, and the reflectivity function. The complex logarithmic transformation is employed in both sides of (2) when ignoring the noise term for the sake of simplicity like below (3). And, assuming the minimum phase sequence, we consider only the real parts of resultant relation.

log??|G_S (w_1,w_2)|=log??|F_S (w_1,w_2)|+log?|H_S (w_1,w_2)| ? ? (3)

This substantially simplifies the homomorphic filtering procedure, so that only the logarithm of a real function can be used to map the input radio frequency image data g_s (m,n) to the cepstrum domain:

(g_s) ?(m,n)=?DFT?^(-1) {log?|DFT{g_s (m,n)}| } (4)

where DFT and DFT-1 stands for the forward and inverse 2-D Discrete Fourier Transform, repectively. Having estimated the cepstrum (g_s) ?(m,n) of the input quadrature signal, the cepstrum of the global point spread function is estimated using the cepstrum domain low-pass Butterworth filtering in order to avoid the Gibbs ringing phenomenon:

(h_s) ?(m,n)=(g_s) ?(m,n)?1/(1+(m^2/D_m +n^2/D_n)^r) (5)

where r is the order, and D_m and D_n are the locus parameters in the radial and lateral direction, respectively. For a minimum phase pulse the amplitude and phase spectrum form a Hilbert transform pair. The spectrum of the minimum phase pulse can thus be found by setting the cepstrum for negative time to zero. So, we construct the 2D minimum phase sequence like below (6):

(h_s) ??(m,n)?_min=(h_s) ??(m,n)?_real?w(m,n)_min ,w(m,n)_min={?(1(n =0)@2(n>0)@0(otherwise))? (6)

A derivation of this result for 1D case can be found in certain paper. However, this assumption is not true in general, because it cannot estimate exact the amplitude and phase of the Fourier Transform of the PSF. As the PSF is assumed symmetric, only its amplitude spectrum is estimated and removed in the de-convolution scheme. Therefore, in order to obtain more accurate PSF with assuming that the PSF might be symmetric or asymmetric, we set the cepstrum for negative time to any value like below (7). The computation of the complex cepstrum is greatly simplified like the assumption of minimum phase. Only the norm of the Fourier domain signal must be calculated. No phase unwrapping and linear phase elimination are necessary to obtain the real cepstrum.

(h_s) ??(m,n)?_max=(h_s) ??(m,n)?_real?w(m,n)_max ,w(m,n)_max={?(1(n=0)@2(n<0)@0(otherwise))? (7)

In (6) and (7), we generate the complex cepstrum for the minimum phase sequence (6) and negative time to any value (7) with the window function. In equation (8), we sum these two 2D sequence are with the weighting factor ? and ?. The optimal values of the weighting factors ? and ? are found experimentally with comparing the real PSF from actual transducer being used in offline. We can control the asymmetric shape of PSF with the weight factor values.

(h_s) ?(m,n)=[??(h_s) ?(m,n)_min]+[??(h_s) ?(m,n)_max] (8)

The estimated PSF function is normalized under the sum of its power spectrum values becomes unity in the 2D frequency domain. Consequently, the local normalized PSF function is the cepstrum domain is mapped back to the time domain as follows.

h_s (m,n)=?DFT?^(-1) {exp[DFT{(h_s) ?(m,n)}]} (9)

De-convolution

Restoration is an example of an “ill-posed problem” which can be defined as a problem that does not have a unique solution, or the solution is not a continuous function of the data. Many algorithms have been developed to compute approximate solutions of ill-posed problems. For example, there are several different regularization methods like Weiner Filter. In this paper, we focused on the low complexity for de-convolution part. So, we are using fast de-convolution technique using hyper-Laplacian priors which is several orders of magnitude faster than existing techniques that use hyper-Laplacian priors.

(f_((R))) ?=min?(f_((R)))???y_((R))-x_((R))?f_((R)) ?^2 ?+?|f_((R)) |^? (10)

Blending

In order to prevent discontinuities between the differently restored each depth image regions, we use the piecewise linear interpolation on overlapping regions. For simple explanation, we consider only x-directional case. Region1 and Region2 can have different restoration results. If the center (x value) of Region1 is 0 and the center (x value) of Region2 is T, the equation is like the below:

Blending Value(x)=(T/(T-x))?Restored Region1(x)+(T/x)?Restored Region2(x) (5)

That is, the final blended value of the center of Region1 is only the value of Restored Region1. The middle blended pixel between Region1 and Region2 is the average of the values of Restored Region1 and Restored Region2. Practically, Blending Value is computed with considering x-directional and y-directional case and 3 channels are computed respectively.

4. Experimental Results

In the experimental part, the proposed algorithm was compared using real phantom and tissues ultrasound data. In the figures, a set of RF-images was recorded from phantom CIRS Model-040 and ATS Model-539 with vera-sonic system and commercial ultrasound scanner (Philips C4-2). Its center frequency was measured to be 3Mhz, and the 128 channels used. The speed of sound was chosen as 1540m/s in the vera-sonic system. First, we applied our algorithm to in vitro measurement on a phantom with known speed of sound like Figure 1 (1540m/s) and Figure 2 (1450m/s). Second, we tried our algorithm to in vivo phantoms with speed of sound that are unknown prior like Figure 4. Each figures, we used the de-convolution algorithms like (a) Original Input image, (b) Minimum Phase only with our proposed Hyper-Laplacian filter and (c) Our proposed Parametric Minimum Phase with Hyper-Laplacian filter. In figure 1 and 2, the point targets (a) of the phantom are shown as smaller spots for the de-convolved image (b) and (c) by our proposed Hyper-Laplacian filter. In addition to, the Parametric Minimum Phase of PSF estimation method produced a better restoration result with reduced ripples and ghost effect around point targets, than result by Minimum Phase only PSF estimation.

9029-12, Session 3

Dual tree complex wavelet transform based shadow detection and removal from moving objects

Manish Khare, Rajneesh K. Srivastava, Ashish Khare, Univ. of Allahabad (India)

Presence of shadow degrades the performance of any computer vision system as a number of shadow points are always misclassified as object points. Various algorithms for shadow detection and removal exist for still images but very few algorithms have been developed for moving objects and none of algorithm for shadow detection and removal has been developed in complex wavelet domain. This paper introduces a new method for shadow detection and removal from moving object which is based on Dual tree complex wavelet transform. We have chosen Dual tree complex wavelet transform as it is shift invariant and have a better edge detection property as compared to real valued transform. In the present work, shadow detection and removal has been done by thresholding wavelet coefficients of Dual tree complex wavelet transform of difference of reference frame and the current frame. Standard deviation of wavelet coefficients is used as an optimal threshold. Results after visual and quantitative performance metrics computation shows that the proposed method for shadow detection and removal is better than other state-of-the-art methods.

9029-13, Session 4

Recognition combined human pose tracking using single depth images

Wonjun Kim, ByungIn Yoo, Jae-Joon Han, Changkyu Choi, Samsung Advanced Institute of Technology (Korea, Republic of)

This paper presents a method for tracking human poses in real-time from depth image sequences. The key idea is to adopt recognition for generating the model to be tracked. In contrast to traditional methods utilizing a single-typed 3D body model, we directly define the human body model based on the body part recognition result of the captured depth image, which leads to the reliable tracking regardless of users’ appearances. Moreover, the proposed method has the ability to efficiently reduce the tracking drift by exploiting the joint information inserted into our body model. Experimental results on real-world environments show that the proposed method is effective for estimating various human poses in real-time.

1. INTRODUCTION

In recent years, estimating 3D human poses has drawn considerable attention in the field of computer vision. This growing interest is raised by the depth camera, which enables various advanced applications such as gesture-based TV controls and game plays. For the success of such systems, several approaches have been proposed, which are categorized into two main groups as follows: recognition-based and tracking-based methods. In the former, the basic idea of early work is to utilize several detectors for localizing human body parts, e.g., head, torso, hands, and so on, in depth images [1]. Recently, Shotton et al. [2] have formulated the body part recognition as a per-pixel classification problem. Specifically, they label each pixel of input depth image into a specific body part based on the randomized decision trees. Although these methods perform well in real-time, they still suffer from occlusions and unseen human poses. On the other hand, tracking-based approaches also have been developed. Demirdjian et al. [3] apply the iterative closet point (ICP) algorithm [4] to the articulated body model for 3D human pose tracking. Ganapathi et al. [5] propose to combine the local hill-climbing with body part detections. The resulting distribution of body configurations allows the system to be robust to various human poses. However, most of them still require rendering or preprocessing steps for the 3D body model to be implemented.

In this paper, we focus on improving the model-based tracking method. One important advantage of the proposed method is to efficiently combine tracking with recognition. More specifically, unlike previous approaches, our human body model is directly built on the recognition result of the input depth image and thus becomes robust to variations of users' body shapes. Moreover, the proposed method does not require any rendering process and the manual initialization. This is fairly desirable for achieving 3D human pose estimation even under embedded environments such as smart TVs and phones. The proposed method also leads to the remarkable reduction of tracking drifts by exploiting the joint information inserted into our body model.

2. PROPOSED METHOD

2.1. Body Model Generation

In this work, we employ the body part recognition method based on supervised learning techniques to generate the human body model. More precisely, we utilize the randomized decision trees to train a classifier for labeling each pixel in input depth images. Our training dataset consists of about 311,600 multi-layered synthetic depth images of human poses such as dancing and stretching activities. Each node in the tree contains a simple test that splits the space of features to be classified. Similar to the previous method [2], the difference of depth values between the current and neighborhood pixels is employed as our feature and the optimal threshold for splitting the node is automatically determined by maximizing the information gain. After training, each leaf node contains the posterior distribution over whole classes (i.e., body part IDs). In the test phase, we first conduct random sampling of input depth points captured at the particular frame (e.g., 30th frame) to reduce the computational complexity. Then, sampled points belonging to the user area of the input depth image are fed into K trained decision trees. When they reach a leaf node, the probability belonging to one of 10 body parts (i.e., torso (T), head-neck (HN), left upper/lower arms (LUA/LLA), right upper/lower arms (RUA/RLA), left upper/lower legs (LUL/LLL), right upper/lower legs (RUL/RLL)) is computed according to the distribution stored during the training phase. Probabilities computed from K trees are finally combined to achieve better recognition performance. For the skeletal representation, positions of 17 joints (i.e., head, neck, three spines, left(right) shoulder/elbow/hand, left(right) sacrum/knee/foot) are subsequently determined by using spatial means of point distributions [2]. Now, our human body model is generated by inserting each joint into the point cloud of the corresponding body part. It should be emphasized that the joint defined between two adjacent body parts is concurrently inserted into both body parts. For example, the joint of the left elbow is inserted into both the left upper arm and the left lower arm. This is because connecting such joints efficiently enforces body parts to be located within allowable ranges, which reduces tracking drifts yielding high-level failures in traditional methods. Our human body model is shown in Fig. 1(a).

2.2. Tracking with the Proposed Body Model

Once the body model is built at the beginning of the video as mentioned, body part tracking is performed for estimating human poses from depth image sequences. Based on the assumption that each body part is rigid, we update the pose of individual body parts using the iterative closet point (ICP) algorithm [4] with the simple articulated constraints. The proposed method goes further by introducing the strategy of connecting adjacent body parts, which is defined based on joints concurrently inserted into our body model. The overall procedure of tracking can be summarized as follows: Given two clouds of 3D points (i.e., our body model and observed 3D data at the current frame), we update the pose of individual body parts using the iterative closet point (ICP) method and apply articulated constraints. More specifically, ICP finds the corresponding points and then iteratively estimates the deformation parameters of each body part, i.e., rotation and translation matrices, by minimizing the Euclidean distance between matched points. For a complete description of how to compute the optimal translation and rotation, see [6]. To consider articulated constraints, we propose to apply the forward kinematic chain with the body part connection. Since the deformation of the torso model is propagated to the end body parts (e.g., left lower arm, right lower leg, etc.) due to the chain structure, we update our body model from the torso to the end body parts. For instance, in order to update the left upper leg as shown in Fig. 1(b), the whole parts of our body model are firstly updated according to deformation of the torso, i.e., R(T) and T(T). Then, we conduct ICP for the left upper leg model, which finds R(LUL) and T(LUL), followed by connecting that part to the torso model. It is worth noting that such connection of body parts is efficiently achieved by using the direction vector defined between joints concurrently inserted into both sides in our body model. This leads to the remarkable reduction of the tracking drift by increasing the robustness of the global optimization of the ICP cost function as mentioned. It is important to note again that end body parts (i.e., LLA, RLA, LLL, and RLL) need to be preveniently updated by deformation parameters of the torso model, followed by the relevant upper body model (i.e., LUA, RUA, LUL, and RUL), before estimating their current positions. The overall procedure of the proposed method is shown in Fig. 1.

3. EXPERIMENTAL RESULTS

We demonstrate experimental results on tracking human poses using single depth images, which are captured by the PrimeSense depth camera in real-world environments. Figure 2 shows some results of human pose tracking by the proposed method. Note that we show skeletal models estimated from tracked joints of our body model with the user area of the input depth image for better view. We can see that the proposed method provides reliable tracking results even in variations of users' body shapes as well as significant occlusions. To show the efficiency and robustness of the proposed method, we compare ours with the competitive method proposed in the literature [2]. As can be seen in Fig. 3, the previous method often fails to correctly yield human poses in occlusion cases and unfamiliar poses whereas the proposed one successfully estimates human poses even in such situations. Moreover, we demonstrate the effect of our strategy for connecting adjacent body parts as shown in Fig. 4. We can see that end body parts highly tend to stick to other ones and be also separated without body part connection (see Fig. 4(b)). Therefore, it is thought that our body model is desirable to robustly track human poses (see Fig. 4(c)). The framework of the proposed method has been implemented by using Visual Studio 2009 (C++) on a single low-end PC (3.0 GB RAM without parallel processing). The processing time for tracking human poses with our algorithm takes about 20 fps on average, which can be applied to various real-time applications.

4. CONCLUSION

A novel method for tracking human poses from single depth images has been proposed in this paper. The basic idea behind the proposed approach is to track human poses with the recognition-based body model, which is robust to variations of users' appearances. To this end, we define the body model by efficiently combining labels of body parts and corresponding joints derived from the body part recognition. The proposed strategy for connecting adjacent body parts allows our body model to be robust to the tracking drift yielding high-level failures in traditional methods. Experimental results show that the proposed

method provides the reliable tracking results even in occlusions as well as unusual poses.

REFERENCES

[1] Y. Zhu, B. Dariush, and K. Fujimura, "Controlled human pose estimation from depth image streams," in Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1-8, 2008.

[2] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1297-1304, 2011.

[3] D. Demirdjian, T. Ko, and R. Darrell, "Constraining human body tracking," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2, pp. 1071-1078, 2003.

[4] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," IEEE Transactions on Pattern Analysis and Machine Intelligence 14(2), pp. 239-256, 1992.

[5] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 755-762, 2010.

[6] R. M. Haralick, H. Joo, C.-N. Lee, X. Zhuang, V. G. Vaidya, and M. B. Kim, "Pose estimation from corresponding point data," IEEE Transactions on Systems, Man, and Cybernetics 19(6), pp. 1426-1446, 1989.

9029-14, Session 4

Video-based facial discomfort analysis for infants

Eleni Fotiadou, Svitlana Zinger, Technische Univ. Eindhoven (Netherlands); Walter Tjon a Ten, Sidarto Bambang Oetomo, Maxima Medisch Centrum (Netherlands); Peter H. N. de With, Technische Univ. Eindhoven (Netherlands)

1. SUMMARY

Prematurely born infants are admitted to the Neonatal Intensive Care Unit (NICU) to receive special care. Various physiological parameters, such as their heart rate, oxygen saturation and temperature are continuously monitored, while there is no system for monitoring and interpreting their facial expressions, the most prominent discomfort indicator. In this paper, we present a monitoring system for automatic discomfort detection in infants' faces based on the analysis of their facial expression. The proposed system uses an Active Appearance Model (AAM) to robustly track both the global motion of the newborn's face, as well as its inner features. The system detects the presence of discomfort by employing the AAM representations of the face on a frame-by-frame basis, using a Support Vector Machine (SVM) classifier. We contribute in two ways. First, we increase the accuracy of the system by extracting several histogram-based texture descriptors to improve the AAM appearance representations. Second, the AAM operation is improved by the fusion of different features. The proposed system is evaluated in video recordings of 6 infants, yielding 0.977 area under ROC curve (AUC). The system offers the benefit of monitoring the infant's expressions when it is left unattended and it also provides objective judgments.

2. INTRODUCTION AND PROBLEM STATEMENT

Pain and discomfort are major indicators of infant illnesses, while persistent unrelieved pain can cause severe complications, such as nervous system changes and delayed development. Despite the significance of pain recognition, most neonatal intensive care units do not have sufficient resources for identifying it. Medical staff is highly responsible for assessing infant pain and discomfort, as infants are unable to verbally communicate their experiences. However, neonatal pains are often brief and may pass unnoticed, since healthcare professionals cannot provide continuous surveillance of all infants at risk.

Facial expressions play a major role in discomfort assessment, as they are the most specific and frequent discomfort indicators. Brahnam et al. [1] was the first to use various face classification techniques, including PCA, LDA, SVMs and NNSOA, to classify the facial expressions of neonates into pain and no-pain classes. The experiments were conducted in the infant COPE image database, demonstrating that such classification techniques can achieve reasonable accuracy. However, the authors did not examine the possibility of implementing a pain recognition system within a hospital environment. In the pilot system designed by Han et al. [2], important facial features, such as eyes, mouth and eyebrows, are automatically extracted and analyzed for the purpose of discomfort detection. To further adapt this system to a real hospital setting, non-ideal situations, such as changes in lighting conditions and viewpoint are taken into consideration.

Lucey et al. [3] used the UNBC-McMaster Shoulder Pain Archive to classify video sequences into pain and no-pain segments. An Active Appearance Model was employed to track the face and derive features, by decoupling the face into rigid and non-rigid shape and appearance parameters. A Support Vector Machine (SVM) classifier with linear kernel was then utilized for pain classification. In their work, they found that the fusion of all AAM representations produces higher recognition rate, revealing that there exists complementary information in AAM representations.

Our work adopts the approach of [3] for the problem of neonatal discomfort detection, but differentiates from it in the following aspects. First, a robust initialization and recovery technique is incorporated to the AAM face tracker, providing face recovery after partial or total face occlusion that may occur due to infant movements or external objects. Instead of using the coarse AAM appearance representations for classification, descriptors of high discriminative power are extracted to boost the performance of the system. Finally, to deal with temporal aspects of discomfort display, an averaging filter is applied to the classification outputs.

3. PROPOSED APPROACH

3.1 AAM-based Face Tracking

The first and most challenging step of the discomfort detection system is the tracking of the infant's face across the successive frames of the video by using an Active Appearance Model (AAM) [4]. In the proposed system, an AAM is fitted to each video frame by using the inverse compositional algorithm [4].

The AAM fitting algorithm requires a suitable initial estimation of the face shape and position to find a proper landmark matching. In the first video frame, face detection, followed by eye and mouth detection and rough pose estimation, are performed for the purpose of AAM initialization. Afterwards, the estimated facial shape can be used to initialize the AAM in the next frame. However, if the fitting error exceeds a predefined threshold, then tracking recovery is employed by providing an initial AAM estimate like in the first frame. Thus, the algorithm is able to recover the face in cases where the tracking is temporarily lost due to occlusions, obstructions or sudden movements.

3.2 Feature Extraction

Once the face is tracked with an AAM by estimating the shape and the appearance parameters, facial features can be obtained based on this information. According to [3], three features can be extracted based on AAM parameters: (1) the similarity-normalized shape (SPTS), (2) the similarity-normalized appearance (SAPP) and (3) the canonical-normalized appearance (CAPP). The similarity-normalized shape (SPTS) and appearance (SAPP) refer to the shape and appearance of the fitted AAM, respectively, when all rigid geometric variation has been removed. The canonical-normalized appearance (CAPP) refers to the appearance of the fitted AAM when all non-rigid shape variation has been normalized with respect to the base shape of the AAM. To improve the performance of the system, highly discriminative histogram-based texture descriptors are extracted from the AAM appearance representations. The following descriptors are extracted and compared: Local Binary Patterns (LBP) [5], Completed LBP (CLBP) [6], Local Phase Quantization (LPQ) [7] and Histogram of Oriented Gradients (HoG) [8]. These descriptors have shown a high performance in facial expression recognition tasks. Prior to the feature extraction, the image is divided into regions and the computed histograms of each region are concatenated.

3.3 Discomfort Classification

For the discomfort classification, an SVM classifier with RBF kernel has been selected because it is suited for many binary classification tasks. The SVM outputs of successive frames are post-processed, with an averaging of the SVM results, to include temporal information. The temporal depth of the averaging filter has been set in such a way that it maximizes the area under the receiver-operator characteristic (ROC) curve of the cross-validation.

3.4 Fusion of SVM scores

In SVM classification, the decision is based on an output score, which is the distance from the optimal hyperplane. In order to combine the outputs of different SVMs, calibration of the scores in a common domain is needed, because of the difference in their nature. For the purpose of calibration and in order to examine the effect of fusing more features together, Logistic Linear Regression [9] is adopted.

4. EXPERIMENTAL RESULTS

4.1 Dataset and training

The facial expressions of 10 infants, experiencing heel puncture, diaper change, hunger, resting or sleeping, were monitored in the neonatal intensive care unit. In order to train the SVM classifier, 156 images are selected from the video recordings of the 10 neonates, half of them displaying discomfort and half of them comfort. The regularization parameter, the sigma of the RBF kernel, and the number of regions for the histogram-based descriptors are selected such that they maximize the accuracy using leave-one-subject-out cross-validation.

4.2 Preliminary results

For the evaluation of the overall system, leave-one-subject-out cross-validation is used due to the limited amount of data. As the number of frames displaying discomfort is quite different from the one with comfort, the accuracy cannot be used as a reliable metric. Instead, the area under the ROC curve (AUC) is adopted as a more reliable performance metric.

Preliminary tests have been conducted to 13 videos of 6 neonates, 8 showing discomfort while the remaining not. To improve tracking performance and robustness, a person-specific and grayscale AAM was constructed for each infant. The results of face tracking in the videos are presented in Table 1. The AAM loses the face during tracking mostly when the hands of the infants or blankets cause partial or total occlusion and in cases that the out-of-plane face rotation is such that both eyes are not visible. However, after these difficult situations have normalized, the position of the face is recovered.

?he performance of the overall system is illustrated by the AUC figures in Table 2 for the various features selected. Only the frames that are correctly tracked are used for the performance estimation. For the individual features, the maximum area under ROC curve (AUC) is 0.955 and this number is achieved by the LBP features computed on the canonical-normalized appearance (CAPP). When the CAPP representation is used, a higher AUC is obtained. This is expected since in many frames the infants are not in frontal position, while good face registration is of high importance for all situations. However, the fusion of the shape and the LBP computed on both AAM appearance representations has achieved an AUC of 0.977, confirming that there is complementary quality in AAM representations.

4. CONCLUSIONS

With the increasing evidence of the distractive pain effect on the mental and physical health of prematurely born infants, it is of vital importance to provide an automatic monitoring system for discomfort detection. We have proposed an AAM-based facial analysis system with an improved accuracy by using selected histogram-based descriptors for texture features, while the AAM has been improved with fusion of these features. In the full paper, these features will be explained in detail, while more experimental results will be provided.

References

[1] S. Brahnam, C. Chuang, R. S. Sexton, F. Y. Shih, “Machine assessment of neonatal facial expressions of acute pain,” Special Issue on Decision Support in Medicine in Decision Support Systems, 43, 1247–1254 (2007).

[2] J. Han, L. Hazelhoff, P. H. N. de With, “Neonatal Monitoring Based on Facial Expression Analysis,” Neonatal Monitoring Technologies: Design for Integrated Solutions, IGI Global, 303-323 (2012).

[3] P. Lucey, J.F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, K.M. Prkachin, “Automatically Detecting Pain in Video Through Facial Action Units,” Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Trans., 41, 3, 664-674 (2011).

[4] I.Matthews and S. Baker, “Active Appearance Models Revisited,” Int. Journal of Computer Vision, 60, 2, 135 – 164 (2004).

[5] T. Ojala, M. Pietikainen, T. Maeenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Transactions on Pattern Analysis Machine and Intelligence, 24, 7, 971–987 (2002).

[6] G. Zhenhua, Z. Lei, Z. David, “A completed modeling of local binary pattern operator for texture classification,” Trans. Img. Proc. 19, 6, 1657-1663 (2010).

[7] V. Ojansivu and J. Heikkila, “ Blur insensitive texture classification using local phase quantization,” In In Proc. Int. Conf. on Image and Signal Processing, 5099, 236–243 (2008).

[8] O. Ludwig, D. Delgado, V. Goncalves, and U. Nunes, “Trainable Classifier-Fusion Schemes: An Application To Pedestrian Detection,” In: 12th International IEEE Conference On Intelligent Transportation Systems, St. Louis, 1, 432-437 (2009).

[9] N. Brummer and J. du Preez, “Application-Independent Evaluation of Speaker Detection,” Computer Speech and Language, 20, 2-3, 230-275 (2005).

9029-15, Session 4

Onomatopoeia characters extraction from comic images using constrained Delaunay triangulation

Xiangping Liu, Kenji Shoji, Hiroshi Mori, Fubito Toyama, Utsunomiya Univ. (Japan)

The paper is aiming at onomatopoeia characters extraction from a comic image. The method we proposed was developed based on stroke width feature of the characters.

A comic image was segmented with the constrained Delaunay triangulation firstly. In all generated triangles, they are classified into 3 types, that is, we call the triangle without any sides on the edge line as T0 triangle, that with one side on the edge line as T1 triangle, that with two sides on the edge line as T2 triangle. The constrained condition for the constrained Delaunay triangulation is to avoid a side in one triangle intersecting an edge line, because it’s better to generate more T1 triangles for stroke width calculation rather than T0 or T2 triangle in our study.

Two neighboring triangles sharing a side were grouped together if they have similar average colors. A blob composed of all grouped triangles is one of character candidate areas.

Then we employed a set of fairly flexible rules parameters of which were learned by the training set of comic images, like the ratio between the total area of all T1 triangles and the total area of all triangles, the variance of the stroke width, candidate size.

The experimental results proved the effectiveness of the proposed method.

9029-16, Session 4

Improved global-sampling matting using sequential pair-selection strategy

Ahmad F. Al-Kabbany, Eric Dubois, Univ. of Ottawa (Canada)

No Abstract Available

9029-17, Session 4

Register multimodal images of range information

Yong Li, Beijing Univ. of Posts and Telecommunications (China); Robert L. Stevenson, Univ. of Notre Dame (United States)

No Abstract Available

9029-18, Session 5

Overview and importance of heterogeneous computing systems for imaging applications *(Invited Paper)*

Michael Frank, LG Electronics Inc. (United States)

No Abstract Available

9029-19, Session 5

HSA overview and how it impacts image processing *(Keynote Presentation)*

Phil Rogers, AMD (United States) and HSA Foundation (United States)

No Abstract Available

9029-20, Session 5

Heterogeneous compute via OpenCL in OpenCV 3.0 *(Invited Paper)*

Harris Gasparakis, AMD (United States)

No Abstract Available

9029-21, Session 5

Introduction to android RenderScript *(Invited Paper)*

Tim Murray, Google (United States)

No Abstract Available

9029-22, Session 6

Deeper look into HSA architecture and its runtime *(Invited Paper)*

Ben Sander, AMD (United States)

No Abstract Available

9029-23, Session 6

Image processing on mobile SOC using ARM Mali GPU *(Invited Paper)*

Akshay Agarwal, ARM Inc. (United States)

No Abstract Available

9029-24, Session 6

Imaging algorithms and implementations using RenderScript in mobile systems *(Invited Paper)*

Shereef Shehata, Mahesh Renduchintala, LG Electronics Inc. (United States)

No Abstract Available

Conference 9030: Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications 2014

Monday - Wednesday 3 –5 February 2014

Part of Proceedings of SPIE Vol. 9030 Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications 2014

9030-1, Session 1

Conception of a course for professional training and education in the field of computer and mobile forensics, Part III: network forensics and penetration testing

Knut Kröger, Reiner Creutzburg, Fachhochschule Brandenburg (Germany)

IT security and computer forensics are important components in the information technology. From year to year, incidents and crimes increase that target IT systems or were done with their help. More and more companies and authorities have security problems in their own IT infrastructure. To respond to these incidents professionally, it is important to have well trained staff. The fact that many agencies and companies work with very sensitive data makes it necessary to further train the own employees in the field of network forensics and penetration testing. Motivated by these facts, this paper - a continuation of a paper of January 2012 [1] which showed the conception of a course for professional training and education in the field of computer and mobile forensics - addresses the practical implementation important relationships of network forensic and penetration testing.

9030-2, Session 1

Remote laboratory content delivery in radio-communications

Achot Matevossyan, Rudresh Gandhinagar Ekanthappa, Rodrigo Escobar, David Akopian, The Univ. of Texas at San Antonio (United States)

Labs provide valuable hands-on experience of the real-world aspects of theories, concepts, and ideas and allow students to observe a range of phenomena and to explore the effects of various factors. However, costly equipment is not always available, maintenance and servicing costs are often very high, and laboratory class scheduling may constrain other course selections. At the same time, lab equipment is often underutilized because of a limited number of classes and servicing costs. One possible solution to this issue is to consolidate lab equipment usage by offering remote laboratories over the Internet as it has become a common networking medium and is increasingly used to enhance education and training.

This paper is a description of a new flexible remote experimentation concept “eComLab using USRP N2920” for experiments in radio-communication. Remote labs can be offered at flexible times by allowing students to schedule different times for different experiments by matching their availability and reserve there different time slots for accessing different experiment with the single hardware board, So the students can access the lab on his scheduled day at the particular time for different experiments from remote locations so equipment can be accessed individually and shared by different students. This will also provide more opportunities for underserved areas and for students with disabilities, thus closing educational gaps. The system has been tested in engineering labs and student assessment is provided.

Remote labs can be accessed without the effort of any software installation on the user side. The users should register to remote gateway website labreservation.com to get access to the remote labs. The registered student is added to the particular course to get access to those labs under that course. Under the given period of time for different courses the students can schedule their available time slots for different labs under that particular course like the way students can schedule different time slots for more number of courses. On the particular scheduled time student can work on his lab experiment independently from remote place. The rest of the students in that particular course will get their chance to work independently only on their reserved time slot and also these students have a chance to only view the labs operation controlled by the student who is working on his reserved time.

The USRP N2920 (Universal Software Radio Peripheral) is a software reconfigurable RF hardware from National Instruments used in Wireless Communication Remote lab to build a digital communication system. This hardware can be easily configured using LabVIEW. LabVIEW is a graphical programming language developed by National Instruments. In the Remote lab setup USRP N2920 connected to a PC (running LabVIEW). This PC controls the USRP through the gigabit ethernet cable connecting the two together. Through the remote desktop students can control RF hardware USRP N2920 and work on their labs.

9030-3, Session 2

Accessing multimedia content from mobile applications using semantic web technologies

Joern Kreutel, Andrea Gerlach, Beuth Hochschule für Technik Berlin (Germany); Stefanie Klekamp, Kristin Schulz, Humboldt-Univ. zu Berlin (Germany)

Our paper will describe the ideas and results of an applied research project aiming at leveraging the expressive power of Semantic Web technologies (RDF, OWL, SPARQL) as a server-side backend for mobile applications. In particular, we are dealing with access to location and multimedia data for providing a rich user experience for mobile application use cases, ranging from city and museum guides to multimedia enhancements of any kind of “narrative” content, e.g. E-Book applications. For these purposes, we will develop a range of reusable software components for both server-side functionality and native mobile platforms that are aimed at significantly decreasing the effort required for developing particular applications of that kind. It will allow to focus application development on the aspects of visual design and interface design, on the one hand, and content modelling, on the other. From our perspective, concrete applications will serve as customised explorers that offer particularly mediated paths for the user to experience a wide web of interrelated content, where access to that content and exploration of its interrelations is enabled by our core technology components.

The primary use case for verifying the power and limitations of our approach is the development of a mobile application for the iOS platform that allows users to experience work, life and personality of the East German writer Heiner Müller (1929-1995), who spent most of his life time living in the city of Berlin. This work is carried out in cooperation between the Department of Informatics and Media at the University of Applied Sciences of Brandenburg and the Department of German Literature at Humboldt University of Berlin. Beyonds its innovative usage of semantic web technologies in the field of mobile applications, it understands itself as an interdisciplinary contribution to the field of “digital humanities” aiming to bridge the gap between IT experts, media designers and domain experts.

The paper will provide an outline of the ideas underlying our project and will propose a generic system architecture as part of our research efforts. We will further give an insight into the core technology components that will have been implemented by the time of final submission.

9030-4, Session 2

Real-time global illumination on mobile device

Minsu Ahn, Inwoo Ha, SAMSUNG Electronics Co., Ltd. (Korea, Republic of); Hyong-Euk Lee, Samsung Advanced Institute of Technology (Korea, Republic of); Dokyoon Kim, SAMSUNG Electronics Co., Ltd. (Korea, Republic of)

The mobile devices, such as Smartphones, and Tablet PCs, are the most widespread devices with rendering capabilities and can support the interactive rendering with the recent advances in processing power and memory capacity,

A number of global illumination methods have been developed with high performance device such as PCs, workstations.

Unfortunately, the rendering techniques on mobile devices only focus on the direct illumination and global illumination is still considered quite burdensome for real-time rendering on mobile.

Because of the resources on such devices are extremely limited; small amount of memory, little bandwidth, and low processing power.

In this paper, we introduce a novel method for real-time global illumination on mobile devices.

Our approach is based on the reflective shadow map and splatting methods of Dachsbacher et al.

We use the hybrid strategy, which collaboratively combines the CPUs and GPUs available in a mobile SoC and enables real-time global illumination on mobile devices.

That is, a CPU and a GPU in mobile SoC generate the direct illumination with shadow and indirect lights simultaneously in order to increase rendering performance. During this stage, two processors share deferred shading buffers for indirect lights through memory. Finally the GPU generates global effects with indirect lights.

Thus, our hybrid approach takes advantage of the whole computing power available in mobile devices and further reduces the processing time.

With limited computing resources in mobile SoC, a small number of indirect lights for global effects are allowed for real-time rendering and the sampling for indirect light is particularly important in order to represent the global effects more efficiently and accurately.

Our sampling method is based on multi-resolution approach using the camera, 3D geometry and attributes such as normal, color.

Due to the low-frequency nature of the indirect illumination such as instant radiosity, many pixels receive radiance quite similar to their neighbors and can be processed as a group. Using hierarchical structure, we detect the discontinuities of geometry, normal, and color and group virtual lights, which are similar to each other.

Furthermore, this paper also proposes an indirect rendering method, with a small number of secondary bounce lights.

Most of rendering techniques with a reflective shadow map distribute one-bounce indirect lights from a primary light source efficiently, but do not provide a solution for secondary bounce lights, which is similar to photon mapping.

With this restriction, to accurately represent the global affects, a much larger number of one-bounce indirect lights are required.

First, one-bounce indirect lights are distributed from the primary light source. Second, we use multi-resolution space partitioning tree and each sub-space includes the whole geometry or simplified one with the attributes as well as one-bounce indirect lights. Only sub-spaces, which do not include one-bounce light, represent the secondary bounce lights. Finally the colors and intensities of secondary bounce lights are determined by the attributes in each sub-space and nearest one-bounce lights.

Experimental results demonstrate the global illumination performance of the proposed method.

9030-5, Session 2

Micro modules for mobile shape, color, and spectral imaging with smartpads in industry, biology, and medicine

Dietrich Hofmann, Paul-Gerald Dittrich, Eric Düntsch, Daniel Kraus, SpectroNet (Germany); Nicolaus Hettler, Angelika Murr, CDA GmbH (Germany)

Aim of the paper is the demonstration of a paradigm shift in shape, color and spectral measurements in industry, biology and medicine as well as in measurement education and training. Innovative hardware apps (hwapps) and software apps (swapps) with smartpads are fundamental enablers for the transformation from conventional stationary working places towards innovative mobile working places with in-field measurements and point-of-care (POC) diagnostics. Mobile open online courses (MOOCs) are transforming the study habits. Practical examples for the application of innovative optodigital micro shape meters, color meters and spectrometers will be given. The innovative approach opens so far untapped enormous markets for measurement science, engineering and training. These innovative working conditions will be fast accepted due to their convenience, reliability and affordability. A highly visible advantage of smartpads is the huge number of their distribution, their worldwide connectivity via Internet and cloud services, the standardized interfaces like USB and the experienced capabilities of their users for practical operations.

9030-6, Session 2

A mobile phone user interface for image-based dietary assessment

Ziad Ahmad, Purdue Univ. (United States); Nitin Khanna, Graphic Era Univ. (India); Deborah A. Kerr, Curtin Univ. (Australia); Carol J. Boushey, Cancer Research Ctr. of Hawai'i (United States); Edward J. Delp III, Purdue Univ. (United States)

Many chronic diseases, including obesity and cancer, are related to diet. Such diseases may be prevented and/or successfully treated by accurately monitoring and assessing food intake. Existing dietary assessment methods such as the 24-hour dietary recall, in which a dietitian interviews an individual about foods consumed in the last 24 hours; and the food record, which relies on the individual to record on paper the dietary information of the food consumed, tend to be burdensome and not generally accurate due to misrecording and underreporting. With the burden and inaccuracy associated with existing dietary assessment methods and with the enormous advances in technology, the need for new methods that rely on technology seem inevitable. Indeed, it has been shown that using a mobile telephone as a food record promises to be a valuable tool for dietitians and researchers in diet assessment. Today, mobile telephones are bundled with many features. These features include a built-in camera, network access, location information retrieval, an accelerometer, and many more. Such advancements in the mobile telephone technology allow for new dietary assessment methods.

The mobile telephone user interface described in this paper was developed as part of a larger complete image-based dietary assessment system. In this system, user begins the process by capturing images of foods using the built-in camera on the mobile telephone, and then the images, along with information about the images, are sent to the server for food analysis. The analysis part is performed on the server due to its computational resource requirements. The analysis process consists of two main parts: food identification and volume estimation. The analysis results are then sent back to the mobile telephone for the user to review and confirm. After the server obtains the user confirmation, food consumption is stored in a specific database on the server that is used for finding the nutrient information using the FNDDS database.

The FNDDS database contains the most common foods consumed in the U.S., their nutrient values, and weights for typical food portions. Finally, the results can be made available to the research community or to healthcare professionals. In this paper we focus on the design and implementation of the client side for such a system. We have implemented the user interface on the Apple iPhone device.

In this paper, we will focus on the development of a user interface for mobile telephones for dietary assessment that uses food images, which are captured via the built-in camera, as the primary method of recording. We will describe in the paper various unique features of the user interface and system that allow a user to capture food images and have them used for evaluating their diet. The user interface we have developed has been tested in several user studies that we will describe and has shown to be a very robust way of capturing dietary information.

9030-7, Session 2

Interactive real-time media streaming with reliable communication

Xunyu Pan, Kevin M. Free, Frostburg State Univ. (United States)

Streaming media is a recent technique for delivering multimedia information from a source provider to an end-user over the Internet. The major advantage of this technique is that the media player can start playing an audio or video file even before the entire file is transmitted. Most streaming media applications are currently implemented with a client-server model, where a server system hosts the video file while a client system connects to this server system to view the file. While the client-server model is great in many situations, it may not be ideal to rely on a third party system to provide the streaming service. For example, a user may want to watch a video simultaneously with a friend in another part of the world over the internet. Although streaming media applications with dedicated server support are very popular, similar applications implemented with the peer-to-peer (P2P) model are not uncommon. In a peer-to-peer model, streaming video is shared directly between end-users, called peers, with minimal or no reliance on a dedicated server. One of the drawbacks of client-server model is that users are required to be a member of the domain in order to use their streaming service. In addition, with constant traffic to and from the server resulting in network congestion and slower video streaming, these dedicated servers are not as robust as a P2P connection. In this paper, we develop a new media streaming application based on the peer-to-peer model to overcome these challenges in mobile circumstance. With the proposed software ????? (pronounced [rév?ma]), named for the Greek word meaning stream, we can host a video file on any computer and directly stream it to a connection partner. To accomplish this, ????? utilizes the Microsoft .NET Framework and Windows Presentation Framework (WPF). The application uses custom File-Buffers and the Universal Datagram Protocol (UDP) to stream HD video at speeds upwards of 20 Mbps. In order to do this, the application implements special threads denoted hostMediaThread and partnerMediaThread, this mechanism are specially designed to transfer video/audio information at extremely high speeds allowing users to view a full length HD movie on their screens simultaneously. In addition to the media threading technology, the ????? also features an Instant Messenger (IM) that allows users to communicate with each other while they enjoy a video. The ultimate goal of the application is to offer users a hassle free way to watch a movie or other video file over long distances without having to upload any of their personal information into a third party database. In order to accomplish this, the users can communicate with each other and stream media directly from one mobile device to another while maintaining an independence from any form of software installation or sign up.

9030-8, Session 3

Efficient burst image compression using H.265/HEVC

Hoda Roodaki-Lavasani, Tampere Univ. of Technology (Finland); Jani Lainema, Nokia Corp. (Finland)

There are number of emerging use cases relying on capturing and editing of multiple correlated pictures. Such cases include e.g. capturing bursts of photos instead of a single one, either to be able to select the most desirable one, be able to add context to the photos, be able to apply special effects or be able to combine a series of photos into an enhanced representation. Burst images are typically captured at a relatively high speed suggesting that there is often inherent correlation between the images. That seems to imply that the temporal prediction approach in hybrid video compression schemes can be used to compress such sequences efficiently.

One important aspect with such image sequences is that those are typically not simply to be played back in temporal order, but often the user is expected to browse such bursts back and forth or jump to an arbitrary position in the sequence. Also in the case of computational photography use cases a random access to arbitrary pictures in the sequence is of importance to be able to allow algorithms to analyze and process the data freely. Thus, it appears evident that a viable image burst format would support low latency access to all the pictures in the sequence.

In this paper we will propose coding structures useful in compressing image bursts and we will study applicability of those in the context of H.265/HEVC video coding standard. We analyze compression performance of such structures as comparison to H.265/HEVC intra only and predictive coding as well as JPEG and JPEG 2000 image codecs. We use wide range of test material to analyze the compression performance for various categories of content, including traditional burst photographs, focal stacks, exposure stacks and short user generated animations.

The provided experimental results show that H.265/HEVC codec with the proposed coding structures for image bursts provides promising bitrate savings compared to the previous still image coding solutions while enabling quick access to any picture in the coded sequence. We are further showing how the coding efficiency is affected by choosing the common reference frames for the coded bursts with different strategies.

9030-9, Session 3

MPEG-4 solutions for virtualizing RDP-based applications

Bojan Joveski, Mihai Mitrea, Rama Rao Ganji, Télécom SudParis (France)

Introduction:

According to the nowadays user exigencies, no functional differences should be encountered between users accessing a cloud application from a mobile thin client or from a fixed desktop PC. Under this framework, defining virtualization mechanism suitable for mobile thin clients remains a challenging research topic: specifying a high performance compression algorithm for heterogeneous content and ensuring versatile, user-friendly and real time interaction are issues to be jointly dealt with. The underlying technical deadlocks are mainly connected to the network and to the terminal.

State-of-the-art:

Application virtualization can be implicitly ensured by remote display technologies. Currently, such technologies convert all types of graphical content into sequences of images rendered by the client. Consequently, important information concerning the content semantics is lost. Our previous study brought to light the possibility of designing MPEG-4 multimedia-based remote display technologies for X11 (Unix) applications.

Paper main contribution:

This paper provides the proof-of-concepts for the use of the MPEG-4 multimedia scene representations (BiFS and LASeR) as a virtualization tool for RDP-based applications. Two main applicative benefits are thus granted. First, any legacy application can be virtualized without additional programming effort. Second, heterogeneous mobile devices can collaboratively enjoy full multimedia experiences. The main methodological novelty consists in (1) designing an architecture allowing the conversion of the RDP content into a semantic multimedia scene-graph and its subsequent rendering on the client and (2) providing the underlying scene graph management and interactivity tools.

Some details about the new architecture:

The paper reconsiders and extends our previous architecture (devoted to X11) so as to cope with the RDP peculiarities. The main components are:

RDP interpreter: captures the graphical content generated by the application and identifies its basic semantic components;

Scene-graph creation: converts the basic components and aggregates them into an MPEG-4 hierarchical, semantic multimedia scene-graph;

Scene-graph manager: optimizes in real-time the scene-graph under network bandwidth/latency constraints;

Scene-graph compression and transmission

Interactivity manager: captures the MPEG-4 events generated by the user and posts them back to the RDP application.

Experimental results:

Experiments consider 5 users and two RDP applications (MS Word and Internet Explorer). For text editing, each user was typing for 5 minutes the beginning of the Plato's Republica. For the Internet browsing, each user performed a complex 9 step scenario (searching, scrolling, watching multimedia, using the top menu bar, …).

The device is connected to the server through a Wi-Fi 802.11 connection.

The new architecture was benchmarked against VNC and FreeRDP. The visual quality is evaluated by seven objective measures (e.g. PSNR>37dB, SSIM>0.99). The network traffic evaluation shows that: (1) for text editing, the MPEG solutions outperforms the VNC by a factor 1.8 while being 2 times heavier then the FreeRDP; (2) for Internet browsing, the MPEG solutions outperforms both VNC and FreeRDP by factors of 1.9 and 1.5, respectively. The average round-trip time (less than 40ms) copes with real-time applications. All these results are computed with 95% relative errors lower than 0.01.

9030-10, Session 3

Evaluation of the emerging scalable high efficiency video coding (SHVC) standard for video stream adaptation in lossy mobile networks

James M. Nightingale, Qi Wang, Christos Grecos, Univ. of the West of Scotland (United Kingdom); Sergio R. Goma, Qualcomm Inc. (United States)

The recently standardised High Efficiency Video Coding (HEVC, H.265) compression scheme can deliver video streams of comparable quality to the current H.264 Advanced Video Coding (H.264/AVC) standard while consuming up to 50% less bandwidth. Modern Internet Protocol (IP)-based video transmission systems must be capable of delivering video content to a wide variety of user devices, many of which are mobile devices with different constraints in display capabilities, available memory and processing power. Furthermore, video delivery is increasingly taking place over heterogeneous mobile networks characterised by dynamically changing network path conditions in which available bandwidth, end-to-end delay and packet loss ratios can vary over time. Previously the scalable extension (H.264/SVC) to the H.264/AVC standard had commonly been proposed as a mechanism to adapt video streams to the challenges of both a diverse mix of user devices and dynamically changing network environments. Work on the development of a scalable extension (SHVC) to the new HEVC standard, which will also address the scalability issues yet at substantially higher compression efficiency, is essential to realise the next-generation adaptive, scalable video delivery. Currently, the SHVC standardisation is underway within the Joint Collaborative Team for Video Coding (JCT-VC). However, research into SHVC especially in lossy mobile networks is still in its infancy.

One important area for investigation is whether and how, given the significantly greater compression ratio of HEVC (and SHVC), the loss of packets containing video content would have a more notable impact on the quality of delivered video than is the case with H.264/AVC or H.264/SVC. In this work, we have empirically implemented and evaluated real-time in-network adaptation of video streams encoded using SHVC. We have considered the two major categories of adaptations. In the first situation, we study the means of adapting SHVC streams to meet a mobile device constraint by exploring and comparing the relative merits of spatial, temporal and quality scalability in SHVC. In the second scenario, we investigate adaptation to meet changing network conditions focussing on the effects of varying rates of packet losses of between 0% and 5% on the quality of received SHVC video streams. In each scenario we evaluate relative performance using both the objective Peak Signal to Noise Ratio (PSNR) and the pseudo-subjective Video Quality Metric (VQM) based on a range of experiments conducted on a realistic testing platform.

Through extensive experimentation, we establish a comprehensive set of numerical results for SHVC delivery in loss-prone network environments such as those commonly found in mobile networks. Our experimental evaluation reveals in-depth advantages and disadvantages for the spatial, quality and combined temporal/quality adaptation of SHVC encoded content. Empirical quantitative comparisons are provided with previously proposed adaptation schemes for both HEVC and the scalable extension (H.264/SVC) to the current standard. Among the highlighted numerical results reported, we show that packet losses of only 1% can lead to a substantial reduction in PSNR of over 3dB and error propagation in over 50 pictures following the one in which the loss occurred. To the best of the authors' knowledge, this work would be one of the earliest studies in this cutting-edge area that offers empirical and analytical insights into SHVC adaptation to lossy, mobile networking conditions.

9030-11, Session 3

Spatial domain entertainment audio decompression/compression

Yiu Keung Chan, Consultant (Hong Kong, China); Ka Him K. Tam, Hong Kong Aircraft Engineering Co. Ltd. (Hong Kong, China)

MP3 and its variants have their roots on "polyphase filter" and "Modified Discrete Cosine Transform(MDCT). In layman context, "polyphase filter" is very much like "divide and rule" and generates a set of subband frequencies from the original audio. MDCT is renown for its excellent energy compaction characteristics and generates MDCT coefficients. The coefficients are then further distilled to factor out non-contributing subband coefficients through frequency masking without affecting decompressed audio quality. The coefficients are then compactly coded for high compression efficiency. HE-AAC takes it a step further and targets low bit rate application with Spectral Band Replication and Parametric Stereo technology.

The computation complexity for MP3 decoding is high. An ARM7 NEON processor with 128bit SIMD hardware accelerator requires a peak performance of 13.99 Mega Cycles Per Second[1] for MP3 decoding.

The PCT application entitled SYSTEM AND METHOD FOR SPATIAL DOMAIN AUDIO DECOMPRESSION AND COMPRESSION[2] is characterized by extremely low decoding complexity.

Reference [2]

Title: SYSTEM AND METHOD FOR SPATIAL DOMAIN AUDIO DECOMPRESSION AND COMPRESSION

Applicants/Inventors: TAM Ka Him Kevin, CHAN Yiu Keung

PCT Application Number: PCT/GB2012/052107

Date of receipt 28 August 2012

Other References will be listed in Full paper if this submission were accepted

The utility of the invention as submitted in [2] uses two sources of high audio complexity as proof of concept. The first source is a male Speech downloaded at random from the Web and seemed as specifically designed to challenge MDCT energy compaction capability by saturating low/mid frequency range with very high Sound Pressure Level energy.

The second source is truly entertainment[3]. The female artist is very famous in the Chinese speaking community, versatile in both English and Chinese, with her formative years spent in Canada. The audio characteristics is frequency and energy rich with very high dynamic range. With many entertainment audio evaluated, only funeral for a friend[4] by Elton John is more complex. These audio sources does not pose much of a challenge to the innovation as submitted in the PCT.

For compression, a mono audio source can be decomposed into frames and dealt with one at a time. An MP3 frame consists of 1,152 audio samples. The PCT application operates on a frame of 2,028 samples as a sequence of interleaving Maximum magnitude(Max) and minimum magnitude(min). Each of these “min to Max” can be said as a rising segment and the succeeding “Max to Min” can be said as a falling segment, and so forth. Number of interior samples bound by “min to Max” or “Max to min” can be {0|1|…|N}. The magnitude of these interior samples, if exist, are irregular and contributes to frequency and energy variation in an audio source.

The differentiating feature of PCT application with respect to existing audio compression/decompression systems is that, in the decode audio, the interior sample magnitude between “min to Max” or “Max to min” are literally distributed. As an example, falling segment or rising segment with 4 interior samples(thus, 6 samples in total) is distributed as 6 constants from a literal table defined as {0.0, 0.1, 0.35, 0.65, 0.9, 1.0}.

Thus, decoding computational complexity is governed by decoding the set of “sample positions where magnitude is a Max or a min”. This is trivial and followed by decoding the set of “Max” or “min” magnitudes. With Max and min magnitudes decoded at their respective sample positions, the operation of filling in interior samples with appropriate magnitude governed by literals then follow and completes the decoding operation.

The PCT application demonstrates how the set of “sample positions” and the set of “Max/min magnitude at their respective sample positions” can be coded with very high compression efficiency through the steps of “aliasing”,”structurally compaction” and “multi-huffman tree decoding”. A single huffman position symbol decoded contains multiple “sample position” and a single huffman magnitude symbol contains multiple magnitudes.

The efficient decoding computational complexity, depending on table look up approach or compute bound approach, can easily approach two order of magnitude better than vanilla MP3 decoding. Apart from the processor cycle advantage, the hardware requirement can also be two order of magnitude better; the data path width requirement will also be several times better. Power consumption advantage is way ahead of current decoding schemes given the above stated advantage.

In any decompression scheme, quantization is a mandatory requirement. A small quantizer will adversely affect compression efficiency while as highly quantized approach will affect decompressed audio quality. The PCT application illustrates a very simple scalar quantization criteria which is based on energy of the audio frame.

If this submission were accepted as an oral presentation, the innovation will be presented with 4 frames from [3] at 0th second, 5th second, 10th second and 1st second. If it were judged that presentation time might not be adequate, example at 0th, 10th and 1st second will be presented. The scalar Quantizer used for 0th, 5th, 10th, 1st second are respectively 100, 500, 500, 300.

As a glimpse of what is to come, 10th second comparison is illustrated here. Constant bit rate of 128kbps is used for both LAME[5] and OGG[6] over a 20 second period stereo. Because of complexity of source audio, LAME seemed to have band limited source audio to 32kHz instead of 44,100Hz. As PCT application was intend to compared with MP3, the input to PCT was band limited to 31kHz in conformance to 15.5kHz human audio perception Bark Critical Band limit.

As the result is a table form and can not be reproduced in here with clarity, the digest comparison is that for the same given bit rate. PCT is superior to OGG[6] in terms of linear phase error between source and decompressed audio. PCT is also superior for Sound Pressure Level Energy comparison between source and decompressed audio. Both PCT and OGG is much superior to LAME[5] which is considered as the best in the pack, especially when both compression and decompression are in floating point format.

9030-12, Session 4

Power efficient imaging for mobile displays

Chihao Xu, Tobias Jung, Daniel Schaefer, Univ. des Saarlandes (Germany)

Power saving technologies like local dimming of LCD backlight offer a huge potential for mobile devices. On the other hand, mobile displays pose specific challenge due to its specific LED arrangement and application. The LEDs are just placed at one edge of the panel, so that the backlight of the opposite part is contributed by many LED strings. The state of the art local dimming method, derived for TVs, cannot properly be applied.

Our model starts from the physical viewing direction. The backlight is composed by every LED string with an individual PWM value. The backlight distribution has to cover the image requirement. The local dimming process is considered as an integer linear optimization problem. The sum of LED PWM values has to be minimized. Due to the wide distribution of the backlight emitted by an LED, a global optimization approach was developed.

The next challenge is called pixel compensation. Due to the dimmed backlight, which is no more uniform, the transmission of every tft-pixel has to be adapted. Through a higher transmission, a gray value may be achieved at a lower backlight. This is the key of power saving. However, every pixel has to be processed at video rate (e.g. 60 Hz). Effectively, the input image is brightened. Such an image processing algorithm has to be efficient, since the power consumption of the computation on a mobile device has to be low.

Another issue correlated with pixel compensation is the clipping artifact. If the image is properly filtered, a considerably higher power saving rate may be achieved. For this purpose, the high resolution image is condensed to a low resolution image. The condensed image is the input for the optimization core. However, clipping may occur that image details get lost. A specific method was introduced which virtually increases local backlight luminance to suppress visible clipping artifact, while the overall brightness of the image is nearly preserved.

The ambient environment of a mobile device is a further challenge for a reasonable visual quality and needs to be considered for power saving. In a dark ambient, human eye is sensitive to image details, while the LED power is low anyway. At daylight, the LED has to be much brighter to ensure viewablity. The image condensation function may consider this environment. Furthermore, proven method to enhance the daylight viewablity like CLAHE may be amended to this local dimming algorithm.

Another issue for mobile devices is the diversity of the image content. It may include artificial menu, gaming or natural content. These contents have different characteristics which need to be preserved. This may be realized in the image condensation and/or pixel compensation stage.

The algorithm has been implemented on an FPGA. Validation on a 6 –inch panel (720RGB X 1280 with 7 strings à 2 LEDs) and a 10-inch panel (1280RGB X 800 with 12 strings à 3 LEDs) were performed. One processor with different control parameters and LUTs can control these two (and other) panels. The power saving rate for IEC 62087 is above 38%. Over 300mW or 800mW power is saved.

9030-13, Session 4

Combining spherical harmonics and point lights for real-time photorealistic rendering

Inwoo Ha, James D. Kim, Hyungwook Lee, SAMSUNG Electronics Co., Ltd. (Korea, Republic of)

For photorealistic rendering, we need to simulate real-world lights and materials. In the real world, lights arrive at surfaces with all frequency from all directions, usually represented with environment map, and materials have complex properties. Therefore, rendering a scene with complex lights and materials in real time is a difficult problem.

The environment lights and complex materials can be approximated with spherical harmonics defined in spherical Fourier domain. Then, low frequency components of complex environment lights and materials, such as large smooth light and diffuse reflection, are projected on just a few bases of spherical harmonics, which makes real-time rendering possible in low dimensional space. However, high frequency components, such as small bright light and specular reflection, are filtered out during the spherical harmonics projection. Some approaches, such as wavelet rendering, are developed to render all frequency lights and materials, but limited to offline rendering.

Our approach is based on combining spherical harmonics and point lights in real time. Spherical harmonics is efficient to represent low frequency lights and materials, while point lights are efficient to high frequency lights. Based on additive property of lights, our approach can separate input environment map image to high frequency layers and low frequency layers. Spherical harmonics projection is applied to the input image. Then we can get spherical harmonics coefficients for low frequency components. Subtracting low frequency components from the input image, high frequency components are computed.

For the low frequency lights, preprocessing of visibility computation is needed. The computed visibility is projected on spherical harmonics. In real-time, dynamic lights are projected on spherical harmonics, and the dot product of the lights coefficients and visibility coefficients are computed. For the high frequency lights, virtual point lights (VPL) are sampled from the separated high frequency environment map. For each VPL, shadow map is computed for each frame. With VPLs and their shadow maps, the result for high frequency and view-dependent components is rendered.

Finally, adding the images rendered from high and low frequency components, we can render a scene under all frequency lights and materials. Our experiment results show all frequency dynamic lights and materials can be rendered in real time on mobile devices.

9030-14, Session 4

Fast ice image retrieval based on a multilayer system

Guoyu Lu, Chandra Kambhamettu, Univ. of Delaware (United States)

Image retrieval is an important topic for the multimedia community. With the development of mobile devices, image retrieval achieves an increasing interest of application. There are mainly two existing approaches for retrieving images. One is the text based image retrieval method, which extracts images based on the manually labelled notations. With the expansion of available images, a vast amount of labour work has to be involved for labelling the images. To solve this issue, content based image retrieval method was proposed. Depending on the different contents in the images, various methods are used for searching relevant images. As Arctic plays an important role in maintaining the global climate and contains large amount of nature resources, Arctic sea ice is a critical research area for both atmosphere and biology scientists. We collect several hundred thousands ice images during a trip to Arctic. Among this large amount of images, locating a query image is an extreme time-consuming task. Ice images contains several characteristics that differ from most other images. The largest difference is that ice images are quite texture-less and the contents of ice images are not distinguishable. For this reason, ice images are difficult to retrieve. For achieving high retrieval accuracy, high level local features (e.g. SIFT) are usually employed. This may be applicable for small image datasets. However, for large-scale image retrieval tasks, this could be extremely expensive as high level local features are usually high dimensional and slow to compute the similarity between descriptors. To overcome this problem, we develop a multilayer searching system for fast locating a query image. In each layer, we largely reduce the searching scope for the next layer. As the ice location and state differ, the appearance of sea ice differs significantly. During the process of ice melting and freezing, the images' color will gradually change from deep blue to pure white. In the first layer, we use image color as a holistic feature to filter out large amount of irrelevant images. The color information will be stored in the memory when reading images. This will save us time for extracting the features. The color space is quantized into several bins. Each bin contains a certain range of color space. Based on the quantized space, a Hash table is built for increasing the searching speed. The query image is cast into a certain color category. Though ice images are texture-less, the breaking ices contain abundant edge information. In the second layer, we detect edges existing in the images by Canny edge detection algorithm and describe the detected edges by a gradient histogram. We select the first 100 nearest neighbours based on the edge descriptors which will be passed into the third layer. As Canny edge detection is fast compared with high-level feature detection, the second layer consumes much less time in finding the nearest neighbours than most methods based on local features. For the top 100 extracted images, we build a new descriptor describing the salient points' graph structure to find the best matched image. We detect Laplacian of Gaussian (LoG) feature on the original images. 20 points with highest LoG value are selected as the salient features. We sort the 20 features from low to high value and then connect the sorted features' image coordinates. This gives us a graph with 20 vertices and 19 edges. We then calculate the length scale for each connected edge pair, which outputs 18 scaling numbers. These 18 numbers are the first part of our new descriptor. The angle between each connect edges will contribute as the next 18 dimensions of our descriptor. So the whole image will be described as a 36 dimensional descriptor. As we are computing the length scale of edge pairs, our descriptor is invariant to image scaling change. Meanwhile, by maintaining the properties of graph matching, our new descriptor is also rotation and transformation invariant. We captured the data using 2 cameras. The images captured by the right camera are used as the training data. And a random selected subset of images captured by the left camera are used as the testing set. Experiments show that our image retrieval system achieve comparable accuracy compared with high-level local features, while the retrieving speed is large improved. As the descriptors we use are all small, the system can be employed in mobile devices which provides convenience for fast locating an image of interest. The system can also be adjusted into other texture-less image retrieval problems.

9030-15, Session 4

Multi-frame knowledge based text enhancement for mobile phone captured videos

Suleyman Ozarslan, P. Erhan Eren, Middle East Technical Univ. (Turkey)

Recent developments in mobile device technology have increased digital imaging capabilities of these devices, thereby enabling the emergence of various new approaches in solving problems, such as automated data collection using mobile phone cameras. While the manual collection and organization of the data take a long time, the same data can be collected in a shorter amount of time by extracting information automatically from images captured by mobile phone cameras. In this study, we explore automated text recognition and enhancement on mobile phone captured

images and videos of store receipts. We propose two methods in order to address text recognition, enhancement and erroneous character correction problems related to mobile phone camera based Optical Character Recognition (OCR) systems. Most of these problems are caused by challenges that are specific to mobile phone-captured images, such as uneven lighting, text skew, text misalignment, and focus loss. The first method is the image based scheme which includes OCR and our proposed Knowledge Based Correction (KBC) algorithm. In this method, first, the trained OCR engine is used for recognition; then, the proposed KBC algorithm is applied to the output of the OCR. The KBC algorithm used in support of the OCR process addresses erroneous character recognition problems, which cannot be completely handled by common image processing methods. The KBC algorithm significantly improves accurate detection rate, but it cannot correct all deficiencies in the single image of a receipt. However, a video capture of the same receipt has hundreds frames, and these problems affect different parts in different frames. Accordingly, we propose the video based scheme as the second method that includes standard OCR, our proposed Row Based Video Frame Integration (RB-VFI), and our proposed KBC algorithm. The RB-VFI method determines the most accurate rows of the text outputs of the video frames which are extracted by OCR. After identification of the most accurate rows, the KBC process is applied to these rows to correct erroneous characters. Therefore, the video based approach integrates the most accurate rows of the frames, and also corrects their erroneous characters by using the KBC process to enhance character recognition rate of the single image based approach. In this study, initial experiments are conducted to evaluate the performance of the proposed methods. 40 classic store receipts are used for the initial experiments. 20 of them are used to train the OCR engine, and the other 20 store receipts are used to measure the accuracy of the proposed methods. According to the results, the word and character recognition rates of the trained OCR engine are 33% and 53%, respectively. Applying the KBC after the OCR process improves the word and character recognition rates to 69% and 73%. Finally, experiments for the video-based scheme are conducted. These results show that the video-based scheme that includes the RB-MFI and the KBC algorithm increases the word character recognition rate to 95%, and the character recognition rate to 98%.

9030-16, Session PWed

Possibilities for retracing of copyright violations on current video game consoles by optical disk analysis

Frank Irmler, Reiner Creutzburg, Fachhochschule Brandenburg (Germany)

This paper deals with the possibilities of retracing copyright violations on current video game consoles (e.g. Microsoft Xbox, Sony PlayStation,...) by studying the corresponding optical storage media DVD and

Blu-ray.

The possibilities of forensic investigation of DVD and Blu-ray discs are presented. It is shown which information can be read by using freeware and commercial software for forensic examination. A detailed analysis is given on the visualization of hidden content and the possibility to find out information about the burning hardware used for writing.

In connection with a forensic analysis of the Windows registry of a suspects PC a detailed overview of the crime scene for forged DVD and Blu-ray discs can be obtained.

Optical discs are examined under forensic aspects and the obtained results are implemented into automatic analysis scripts for the commercial forensics program EnCase Forensic. It is shown that for the optical storage media a possibility of identification of the drive used for writing can be obtained. In particular Blu-ray discs contain the serial number of the burner. These and other findings were incorporated into the creation of various EnCase scripts for the professional forensic investigation with EnCase Forensic.

Furthermore, a case study for a forensics training program for investigators is developed.

9030-17, Session PWed

Fault tolerant position fingerprinting algorithm

Mohammadhafez Bazrafshan, The Univ. of Texas at San Antonio (United States)

Position Fingerprinting algorithm using Received Signal Strength Indicator (RSSI) measurements in Wireless Local Area Networks has been suggested to be very effective in providing accurate location estimates in indoor environments. This algorithm comprises of two stages: the offline phase and the online phase. In the offline phase, also known as the training stage, a set of RSSI measurements from surrounding Access Points (AP) are collected for various reference locations. Due to the availability of many APs in recent years and the innate complex structure of the indoor environments, these RSSI measurements tend to be unique for each reference location and hence can be thought of location fingerprints. In the online phase or testing stage, a mobile device receives a set of RSSI measurements from surrounding APs and is to infer its location by comparing these measurements to the fingerprints. A reference location for which the online measurements is most similar to the fingerprints is provided as an estimate of the mobile device location. This similarity check is often done through algorithms such as nearest neighbor, maximum likelihood, or multiclass classification algorithms such as logistic regression and Support Vector Machines. However, in many scenarios, an alteration in the environment, an unprecedented failure or a malicious attack can cause faults in AP measurements. These faults can result in a no measurement, a slight deviation from the actual value or a significant change in the value from the values already stored in the database.Therefore, in this paper we have investigated the susceptibility of conventional location fingerprinting algorithms to these various faulty scenarios via simulation and experiments. After selecting the most tolerant algorithm to the various faulty scenarios we have also proposed a simple algorithm that performs a sanity check on the online measurements from APs to see whether the received values are consistent. This sanity check procedure helps to detect most likely faulty measurements in the online data and suggests a set of reliable APs. The set of reliable APs are once again used to perform fault tolerant positioning. Simulation and experimental results indicate that the proposed approach has superior performance in various faulty environments.

9030-18, Session PWed

Indoor positioning system using fingerprinting of WLAN multipath signals for mobile devices

Anirban Saha, David Akopian, The Univ. of Texas at San Antonio (United States)

The Global Positioning System (GPS) has emerged as a popular approach in the outdoor location based services but the satellite signals have yet to overcome many obstacles to provide meter accuracy for indoor positioning. Meanwhile, due to the wide deployment of Wireless Local Area Networks (WLAN) in recent years, WLAN position fingerprinting algorithms has become popular for the positioning of mobile devices in indoor environments with very good coverage and accuracy at the expense of an additional effort of pre-operation surveying. This algorithm consists of two stages. In the first stage, called the offline phase (training), the Received Signal Strength Indicator (RSSI) from a set of available Access Points (AP) in the area is measured for a number of reference locations and stored in a database. Due to the availability of many APs and the complex structure of indoor environments, this information is distinctive for each reference location and thus is called a position fingerprint. In the second stage, called the online phase (testing), a mobile device receives RSSI values from the set of APs in the area. These measurements are compared to the fingerprints that are stored in the database using the nearest neighbor or probabilistic

methods and the reference location for which the fingerprint is most similar to the test measurements is selected as the mobile device's position. Position Fingerprinting can provide high accuracy in indoor environments when there are many APs available in the area because the complex structure can impose a unique fingerprint for different reference locations. However, there are certain scenarios that there is limited number of access points available for positioning and received signal measurements by mobile devices can not provide distinctively separable fingerprints for unambiguous positioning. Therefore, other useful fingerprint measurements for mobile devices ought to be sought for. One of the many possible measurements that mobile devices can provide is the delay profile channel data in the WLAN networks. Additionally multipath is one of the crucial problem in indoor environments as typically line-of-sight (LOS) is not available. Adopted these fluctuated signals may lead to inaccurate results. To mitigate these problem, this paper provides a novel approach to distinguish between various user locations based on the multipath pattern observed at particular location from the WLAN signals. More concretely, we investigate WLAN multipath signals as a possible fingerprint for positioning algorithms with limited number of access points. In the offline phase, we record multipath delay profile measurements from only three access points using Beetle Bang Wi-Fi optimizer device. In the online phase, upon receiving multipath measurements we show that by performing conventional nearest neighbor or maximum likelihood algorithms the position of the mobile device can be derived using only three access points. Experimental results show comparable accuracy performance for RSSI based fingerprint positioning using many access points and the proposed multipath based fingerprint positioning using only three access points.

9030-19, Session PWed

Human activity recognition by smartphones regardless of device orientation

Jafet A. Morales, David Akopian, Sos Agaian, The Univ. of Texas at San Antonio (United States)

A new method for activity recognition using smartphones is proposed. Using the three-axis accelerometer and three-axis gyroscope embedded in a smartphone, a system that uses the proposed method for activity recognition is able to identify exercise activities such as walking, jogging, squatting, executing jumping jacks, and resting. The system works regardless of the orientation of the smartphone with respect to the body part to which it is attached. Most researchers calculate features only from three-axis acceleration data in the device's coordinate system, but such methods are not robust by themselves because the device must be oriented the same way during the training and testing phases. This may not happen in a real life scenario were the orientation of the smartphone can be inverted if the user decides to place the screen facing in or out and also because the orientation is highly dependent on clothing, in which case the smartphone can have an offset in orientation up close to 90 degrees if he is wearing shorts. Such problem can be solved by using any coordinate system that always generates the same acceleration signature for a particular activity, regardless of the orientation of the device with respect to the user (as long as the device is worn in the same on-body location.) This requirement does not mean that the coordinate system will be oriented the same way with respect to the user for all activities. Different coordinate systems can be used for every activity, as long as two coordinate systems do not output the same acceleration signature for different activities. In fact, due to the restricted freedom of motion of the limbs in the human body, and the small number of patterns that are usually executed, we believe that it is very unlikely that the same acceleration signature will be generated for two different activities of interest, even when using a different coordinate system for every activity. The proposed system applies a signal transformation before calculating features. This transformation makes both, acceleration and rotation signals invariant to the orientation of the device inside the user's pocket. As opposed to other methods which are part of the background for this research, which have used only the principal component signal from the accelerometer as an orientation invariant signal for feature extraction, the proposed method uses all three components of acceleration and all three components of the gyroscope in the principal component analysis (PCA) transformation, which makes all motion tracking signals orientation-invariant. All signals in the PCA transformation are needed, not just the principal component because most low level activities require the user to accelerate and rotate the phone in patterns that cannot be fully represented with one dimension. After having transformed all six motion tracking signals to their orientation-invariant counterparts, any other fixed orientation techniques can be used for classification. By using a mutual information based feature selection method, a small set of features have been selected for orientation-invariant algorithms from a large pool of features proposed by other researchers for the fixed orientation problem. Other features which were also chosen by the feature selector are proposed by the authors. As suspected, the final set of features includes both acceleration and rotation standardized features from all three axis, not just from the principal component of acceleration. In the classification stage, two different classifiers were tested. In the first case, a k-nearest neighbor classifier in Euclidean feature space is used. In the second case, a multilayer perceptron with one hidden layer trained by backpropagation is used. When gathering data for this research the users were asked to place the phone in their front right pocket as they would typically do it in a normal situation but it is proposed that the same method be used for other on-body locations. The system achieves a high level of accuracy when tailored to a specific user, but also when trained on a small set of users and tested on an unknown user.

9030-20, Session PWed

Conception, implementation, and test of a Windows registry forensic tool

Knut Kröger, Reiner Creutzburg, Christian Leube, Fachhochschule Brandenburg (Germany)

This paper describes the design and prototypic implementation of a forensic tool for the automated analysis of the Windows registry. The concept provides a complete object-oriented analysis of functional requirements as well as detailed descriptions of the program components and the software architecture of the tool. The prototypical implementation of the tool on basis of the developed concept shows its consistency. The implementation is partially described as object-oriented design. The subsequently defined tests prove the consistency of the concept and the implementation.

9030-21, Session PWed

Virtual tutorials, Wikipedia books, and multimedia-based teaching for blended learning support in a course on algorithms and data structures

Jenny Knackmuss, Reiner Creutzburg, Fachhochschule Brandenburg (Germany)

The aim of this paper is to describe the benefit and support of virtual tutorials, Wikipedia books and multimedia-based teaching in a course on Algorithms and Data Structures.

We describe our work and experiences gained from using virtual tutorials held in Netucate iLinc sessions and the use of various multimedia and animation elements for the support of deeper understanding of the ordinary lectures held in the standard classroom on Algorithms and Data Structures for undergraduate computer sciences students. We will describe the benefits, form, style and contents of those virtual tutorials.

Furthermore, we mention the advantage of Wikipedia books to support the blended learning process using modern mobile devices.

Finally, we give some first statistical measures of improved student's scores after introducing this new form of teaching support.

9030-22, Session PWed

Hacking and securing the AR.Drone 2.0 quadcopter: investigations for improving the security of a toy

Ricardo Band, Johann-Sebastian Pleban, Reiner Creutzburg, Fachhochschule Brandenburg (Germany)

In this article we describe the security problems of the Parrot AR.Drone 2.0 quadcopter. Due to the fact that it is promoted as a toy with low acquisition costs, it may end up being used by many indivuduals which makes it a target for harmful attacks. In addition, the videostream of the drone could be of interest for a potential attacker due to its ability of revealing confidential information. Therefore, we will perform a security threat analysis on this particular drone. We will set the focus mainly on obvious security vulnerabilities like the unencrypted WiFi connection or the user management of the GNU/Linux operating system which runs on the drone. We will show how the drone can be hacked in order to hijack the AR.Drone 2.0. Our aim is to sensitize the end-user of AR.Drones by describing the security vulnerabilities and to show how the AR.Drone 2.0 could be secured from unauthorized access. We will provide instructions to secure the drones WiFi connection and its operation with the official Smartphone App and third party PC software.

9030-23, Session PWed

A new 1D parameter-control chaotic framework

Zhongyun Hua, Yicong Zhou, Chi-Man Pun, C. L. Philip Chen, Univ. of Macau (Macao, China)

This paper introduces a novel parameter-control framework to produce many new one-dimensional (1D) chaotic maps. It has a simple structure and consists of two 1D chaotic maps, in which one is used as a seed map while the other acts as a control map that controls the parameter of the seed map. Examples and analysis results show that these newly generated chaotic maps have more complex structures and better chaos performance than their corresponding seed and control maps.

9030-24, Session PWed

A new collage steganographic algorithm using cartoon design

Shuang Yi, Yicong Zhou, Chi-Man Pun, Univ. of Macau (Macao, China); C. L. Philip Chen, Univ. of Macao (Macao, China)

Existing collage steganographic methods suffer from low payload of embedding messages. To improve their payload while providing high security protection to messages, this paper introduces a new collage steganographic algorithm using cartoon design. It embeds messages into the least significant bits (LSBs) of color cartoon objects, applies different permutations to each object, and adds objects to a cartoon cover image to obtain the stego image. Computer simulations and comparisons demonstrate that the proposed algorithm shows significantly higher capacity of embedding messages compared with traditional collage steganographic methods.

9030-25, Session PWed

Fixed tile rate codec for bandwidth saving in video processors

Vladimir Lachine, Chon-Tam Le Dinh, Dinh Kha Le, Jeffrey Wong, Qualcomm Inc. (Canada)

The paper presents an image compression circuit for bandwidth saving in video display processors. This is intra frame tile based compression algorithm offering visually lossless quality for compression rates between 1.5 and 2.5. RGB and YCbCr (4:4:4, 4:2:2 and 4:2:0) video formats are supported for 8/10 bits video signals.

The Band Width Compressor (BWC) consists of Lossless Compressor (LC) and Quantization Compressor (QC) that generate output bit streams for tiles of pixels. Size of output bit stream generated for a tile by the LC may be less or greater than a required size of output memory block. The QC generates bit stream that always fits output memory block of the required size. The output bit stream generated by the LC is transmitted if its size is less than the required size of output memory block. Otherwise, the output bit stream generated by the QC is transmitted.

The LC works on pixel basis. A difference between original and predicted pixel's values for each pixel of a tile is encoded as prefix and suffix. The prefix is encoded by means of variable length code, and suffix is encoded as is.

The QC divides a tile of pixels on a set of blocks and quantizes pixels of each block independently of the other blocks. The number of quantization bits for all pixels of a block depends on standard deviation calculated over the block. A difference between pixel's value and average value over the block is quantized and transmitted.

2014 Electronic Imaging

SCIENCE AND TECHNOLOGY

Conferences and Courses
2–6 February 2014

Location
Hilton San Francisco, Union Square
San Francisco, California, USA

Technologies for digital imaging systems, 3D display, image quality, multimedia, and mobile applications

Technical Summaries

www.electronicimaging.org

IS&T imaging.org

SPIE®