Using covariance improves computer detection and tracking of humans

Fatih Porikli

Understanding the relationships between statistical variables allows them to be used more effectively.

The object detection and tracking our eyes do so innately are among the most essential components of computer vision applications, from consumer electronics to smart weapons. In video surveillance, these methods facilitate understanding of motion patterns to uncover suspicious events. Navigation systems employ them to keep vehicles in lanes and prevent collisions. In traffic management systems, they control the flow of vehicles to keep traffic moving smoothly. Video broadcasting makes use of them to better compress data, improving download speeds for users accessing video files on the Internet. In medical fields, they are used to analyze tumors and cellular entities to obtain accurate diagnoses. Still, robust detection and tracking of deforming, non-rigid, and fast moving objects—such as human bodies—presents a challenge.

Computers use many different object descriptors, from aggregated statistics to appearance models, to translate images of real world into the world of numbers. Histograms are among the most popular representations, but these descriptors disregard the spatial arrangement of features, and do not scale to higher dimensions. Another popular type of descriptor, the appearance model, is highly sensitive to noise and shape distortions. To overcome these shortcomings, we have developed a novel object descriptor, a bag of covariance matrices, to represent an image window. We use this representation to automatically detect and track any target object in video images.\(^1\)

The concept of covariance is essentially a measure of how much two variables vary together. By constructing the covariance of different features of an image window such as coordinate, color, gradient, edge, texture, and motion—as illustrated in Figure 1—we capture the information embodied in both histograms and appearance models. By using a bag of such covariance matrices, we improve robustness in the detection of pose and shape changes. The bag of covariance matrix descriptors also provides a natural method of fusing multiple features: it has a very low dimensionality; it is scale and illumination independent; and noise that corrupts individual samples is largely filtered out during the covariance computation. We are able to compute the covariance matrix very quickly using integral images,\(^2\) which significantly accelerates the computation time by taking advantage of the spatial arrangement of the points.

To detect a specific object—for example, a human—in a given image, we first train a boosted classifier. This is done offline using covariance descriptors of positive and negative training samples, representing humans and non-human objects in an image, respectively (see Figure 2). We then apply the classifier online at each candidate image window to determine whether the target object is present. To track a given object, we compare the covariance descriptor of object and candidate windows in consecutive video frames using an eigenvector based distance metric.\(^3\) We select the window that has the minimum distance and assign it as the estimated location. The covariance tracker does not make any assumptions about an object’s motion; in other words, it keeps track of objects even if the motion is erratic and fast. Sample results are given in Figure 3.

Continued on next page
Figure 2. A classifier is trained with positive (depicting humans) and negative (non-human) examples. Each weak classifier makes its estimation based on a single matrix from the bag of covariance matrices.

Figure 3. Detection initiates objects, while tracking resolves identity correspondence problems.

We use a cascade of rejectors and a boosting framework to increase the speed of the detection process. Each rejector is a strong classifier, and consists of a set of weighted linear weak classifiers. The number of such classifiers at each rejector is determined by the target true and false positive rates. Each weak classifier corresponds to a region in the training window and splits the high-dimensional input space with a hyperplane. The boosting framework then sequentially fits weak classifiers to reweighted versions of the training data. We fit an additive logistic regression model by stage-wise optimization of the Bernoulli log-likelihood.

To address the fact that objects undergo appearance changes in time, we construct and update a temporal kernel of covariance descriptors corresponding to previously estimated object regions. From this set, we compute an intrinsic mean matrix that blends all the descriptors in the kernel. Since the space of covariance descriptors is not Euclidean space, we transform their manifold onto its tangent space, where the relation between the vectors on the tangent space and the geodesics on the manifold are given by an exponential map. This enables us to define the dissimilarity between covariance descriptors by the sum of the squared logarithms of their generalized eigenvalues.4

We are currently integrating this promising technology into surveillance products for multi-camera setups. We also plan to further improve the computational complexity of the method by hardware implementation. Finally, while we have mainly highlighted our success in detecting and tracking human targets here, our method can be used on any object type, and expect it to be applied to other target detection and tracking tasks as well.

Author Information

Fatih Porikli
Mitsubishi Electric Research Labs - MERL
Cambridge, MA

Fatih Porikli is leading research in the areas of computer vision, pattern recognition, data-mining, robust optimization for multi-camera surveillance, and intelligent transportation systems at MERL. He serves as associate editor for Journal of Real-Time Image Processing, has authored over 50 papers, and is the holder over 30 patents. He is a senior IEEE, ACM, and SPIE member. In addition, he has served among the program committees of the Real-Time Imaging and Video Communication and Image Processing conferences, and chaired several sessions at these meetings over the past three years.

References


© 2006 SPIE—The International Society for Optical Engineering