KERNEL-BASED HIGH-DIMENSIONAL HISTOGRAM ESTIMATION FOR VISUAL TRACKING

Peter Karasev *James Malcolm* *Allen Tannenbaum*

pkarasev@gatech.edu, {malcolm,tannenba}@ece.gatech.edu
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, Georgia

ABSTRACT

We propose an approach for non-rigid tracking that represents objects by their set of distribution parameters. Compared to joint histogram representations, a set of parameters such as mixed moments provides a significantly reduced size representation. The discriminating power is comparable to that of the corresponding full high-dimensional histogram yet at far less spatial and computational complexity. The proposed method is robust in the presence of noise and illumination changes, and provides a natural extension to the use of mixture models. Experiments demonstrate that the proposed method outperforms both full color mean-shift and global covariance searches.

Index Terms— Object tracking, mean-shift, region covariance, kernel density estimation

1. INTRODUCTION

The goal of visual tracking is to follow the movement of a target through a video sequence. As objects encounter clutter, occlusion, changes in illumination, or changes in view, this becomes a difficult problem.

One popular technique called kernel tracking represents each object as a joint probability density function (pdf). In each frame, a local mean-shift computation is performed to minimize a statistical similarity metric between the target and reference densities [1].

An alternate algorithm is that of covariance tracking [2]. In this approach, a target is described by the covariance of its features. Faster than constructing a full histogram, this method allows efficient models of more complex feature spaces than simply color, for example image derivatives or spatial characteristics.

We propose to represent objects as a set of their distribution parameters that approximate the underlying joint density function. In this work, we used the covariance and mean feature vector as the parameter set. After defining a distance function with respect to this parameter set, we employ a variational approach to find the target. This method is fast, uses a compact representation, does not require an expensive exhaustive search, and takes into account a potentially large number of feature dimensions without becoming computationally intractable. In contrast, the computational cost of mean-shift grows exponentially with the number of histogram dimensions, making it impractical for much more than color intensity. Further, the proposed method outperforms the approach using covariance alone by incorporating knowledge of the distribution mean and naturally extending to mixture models.

Fig. 1. One frame of the SUBWAY sequence showing target location using color mean-shift, global covariance search, and the proposed method *(left to right)*.

2. RELATED WORK

This note draws upon two areas of research, namely kernel tracking and covariance tracking.

Kernel tracking, popularized by Comaniciu *et al.* [1], is a simple and robust technique with many extensions. Two extensions deserve particular note for their approaches to object representation and comparison. First, to incorporate the spatial arrangement of an object's color, Birchfield and Rangarajan [3] introduced "spatiograms", joint histograms of both color and position. Second, Yang *et al.* [4] demonstrated direct evaluation of the similarity metric on the density estimates. In addition, they employed the Improved Fast Gauss Transform to deal efficiently with the higher dimensional feature spaces. While powerful, both techniques still suffer from exponential time and space complexity as the density estimates grow. Thus, in practice mean-shift typically uses color without spatial information.

Region covariance was introduced as an alternative compact representation of joint feature spaces [2]. For each region, only the covariance of the feature vectors at each pixel describe the object, and tracking is performed in a global search to match rectangular regions. Since covariance matrices lie in a Riemannian space, Porikli *et al.* [2] are careful to use an appropriate distance metric when comparing regions, a costly computation since tracking is performed via global search. Further, since this uses global detection, covariance tracking can recover after prolonged occlusion or large movements. In the context of high-dimensional density estimation, characterizing regions by their covariance implies each region fits a Gaussian distribution with identical mean. The proposed method improves upon this approach by incorporating knowledge of the means and extending to arbitrary parameters. Extension to Gaussian mixtures further increases the representational power of the technique incurring only a linear complexity increase.

3. PROPOSED METHOD

We propose to represent objects by estimating a parameterized distribution to approximate the underlying joint density containing features such as color, spatial coordinates, and image derivatives. Kernel-weighted averages are computed for a region of interest. To represent the target object, we construct a vector $\mathbf{q} = \{q_1 \dots q_n\}$ consisting of these parameter elements. If we take $\{q_i\}$ to be the set of all mixed central moments for the feature vector $\boldsymbol{z}$, we arrive at the covariance descriptor. We extend this to include both the covariance and the mean of the feature vector. Thus, for an n-dimensional choice of the feature vector, we define $\mathbf{q}$ as a vector containing the $\frac{n^2+n}{2}$ unique elements in the covariance matrix, in addition to the n means.

Given an image $\mathcal{I}$ over spatial domain Ω, a parameter set for the reference target is constructed by computing $\mathbf{q}$ for a kernel-weighted subset $X \subseteq \Omega$ centered at x. For a Gaussian distribution, we estimate the kernel-weighted covariance and mean:

$$\boldsymbol{\mu}(x) = \frac{1}{C} \sum_{x_k \in X} K(x_k - x)\boldsymbol{z}_k \tag{1}$$

$$\boldsymbol{\Sigma}(x) = \frac{1}{C} \sum_{x_k \in X} K(x_k - x)(\boldsymbol{z}_k - \boldsymbol{\mu}(\boldsymbol{x}))(\boldsymbol{z}_k - \boldsymbol{\mu}(\boldsymbol{x}))^T \tag{2}$$

where $C = \sum_{x_k \in X} K(x_k - x)$ is a normalization constant and $K(\cdot)$ is a kernel function.

Given a reference target characterized by $\mathbf{q}$, we seek a region centered at position x whose candidate parameter vector $\mathbf{p}(x)$ minimizes a distance to $\mathbf{q}$. We must now choose an appropriate measure of distance in the parameter space. While this space is often Riemannian, if we employ a variational solution, we may assume the manifold to be locally Euclidean and so use a standard L_2 distance. Our optimization problem is now:

$$\underset{x}{\operatorname{argmin}}\, d(\mathbf{q}, \mathbf{p}(x)) = \underset{x}{\operatorname{argmin}} \|\mathbf{q} - \mathbf{p}(x)\|^2_W \tag{3}$$

where W is a matrix used in a weighted L_2 distance that affects the influence of the different distribution parameters on the distance computation.

Expanding (3) as $(\mathbf{q} - \mathbf{p}(x))^T W (\mathbf{q} - \mathbf{p}(x))$ and taking the derivative with respect to x, we arrive at

$$\frac{\partial d(\mathbf{q}, \mathbf{p}(x))}{\partial x} = (\mathbf{q} - \mathbf{p}(x))^T W \frac{\partial \mathbf{p}(x)}{\partial x} \tag{4}$$

Computation of $\frac{\partial \mathbf{p}(x)}{\partial x}$ is element-wise for each $p_i \in \mathbf{p}$. For example, for the Gaussian parameter set, the spatial derivatives of $\mathbf{p}$ are:

$$\frac{\partial p_1}{\partial x} = \frac{\partial \boldsymbol{\mu}(x)}{\partial x} = \frac{1}{C} \sum_{x_k \in X} K'(x_k - x)\boldsymbol{z}_k \tag{5}$$

$$\frac{\partial p_2}{\partial x} = \frac{\partial \boldsymbol{\Sigma}(x)}{\partial x} = \frac{1}{C} \sum_{x_k \in X} K'(x_k - x)(\boldsymbol{z}_k - \boldsymbol{\mu})(\boldsymbol{z}_k - \boldsymbol{\mu})^T \tag{6}$$

where we simplify the computation in (6) by approximating $\boldsymbol{\mu}$ as constant, since its derivative is small relative to the rest of the expression.

In mean-shift, the kernel is often taken to be Epanechnikov, so we can solve directly for x [1]; however, in general $K(\cdot)$ may not have closed form. The use of arbitrary kernels (see Section 4) requires nonlinear optimization. The solution to the minimization problem (3) is found by using gradient descent:

$$x_{t+1} = x_t - h \frac{\partial d(\mathbf{q}, \mathbf{p}(x))}{\partial x} \tag{7}$$

where is h is an appropriate time step.

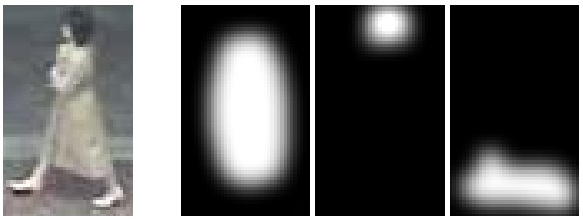

Fig. 2. Reference image from SUBWAY sequence and kernels corresponding to mixture model components. Each region has its own mean, covariance, and kernel.

4. KERNEL SELECTION FOR GAUSSIAN ESTIMATE

A simple design choice is for $K(\cdot)$ to be isotropic and decreasing with distance from the center so as to weight pixels as the edge of the window less, as they are more likely to be subjected to occlusion and background clutter. In this section we describe how more powerful kernels improve tracking by better approximating the often multimodal pdf.

If we take the choice of target distribution to be an n-dimensional Gaussian, we will have the number of parameters B equal to $(n^2 + 3n)/2$, to include the mean and unique elements of the covariance matrix. Using these distribution parameters, the estimate of the target's joint pdf is characteristically a unimodal normal random variable:

$$\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K). \tag{8}$$

While the density estimate in (8) gives an improvement over characterization using the means or covariance alone, it still describes the object as a single Gaussian distribution. If the object has light and dark areas that are both used to compute the means, the resulting Gaussian estimate would not look like the underlying bimodal distribution. The proposed framework naturally allows the characterization of a target's joint density as a superposition of M Gaussian functions, each created with a different spatial kernel:

$$\sum_{m=1}^{M} \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_{K_m}, \boldsymbol{\Sigma}_{K_m}) \tag{9}$$

thus using MB parameters to estimate the density function. The additional parameters are concatenated into a larger parameter vector; the form of the optimization solution (7) is unchanged.

The task of kernel creation can be automated as a segmentation or clustering task to separate areas with different means, hence allowing the description to capture the multi-modal distribution. For the SUBWAY sequence, we illustrate a simple example of this in Figure 2 using a thresholding procedure. Several strategies for capturing important features with multiple kernels are described in [5]. Sophisticated kernel creation methods employ expectation maximization [6], and work has been done to collaborate the movement of multiple kernels [7]. Optimal kernels for tracking are described in [8].

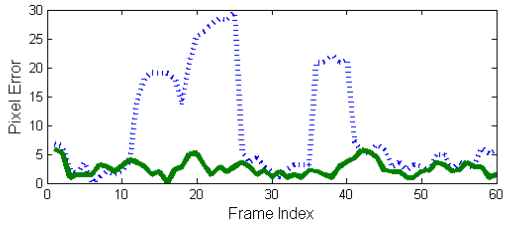

Fig. 3. Track point distance from ground truth for 60 frames of SUBWAY sequence using global covariance search *(dotted)* and proposed method *(solid)*. Notice that the covariance search loses track when it drifts due to clutter three times.

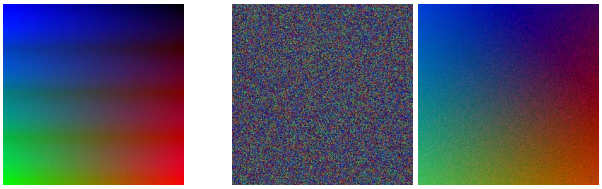

Fig. 4. Sampling the distribution of a synthetic image *(left)*. Images created by sampling the color histogram *(center)* and the joint density approximated by the proposed method's parameter vector $\mathbf{q}$ *(right)*.

5. RESULTS

Tracking was performed on four video sequences with representative examples chosen to exhibit clutter, changes in illumination, and changes in view. The system was prototyped in Matlab running on a 2.0 GHz laptop, and for 640x480 resolution video and multiple kernels it runs at roughly 2 Hz (500ms per frame). All sequences assumed Gaussian distributions using means and covariances to describe objects. The CROWD sequence assumed unimodal distribution (one kernel) while the DOG, SUBWAY, and CHASE sequences used a mixture model each employing three kernels.

The first experiment demonstrates the expressibility of the proposed method. From a synthetic image we calculated both the color histogram and density estimate described by $\mathbf{q}$, and from these we generated sample images. Figure 4 illustrates the loss of spatial information in the joint color histogram; the resulting sample does not resemble the original. In contrast, we are able to retain the global intensity gradient of the synthetic image when generating values from the Gaussian parameterized by $\mathbf{q}$ conditioned on the pixel locations.

The CROWD sequence shows tracking among many objects of indistinct color and little texture. While the covariance method obtains very high detection rates, Figure 5 shows selected frames where the global covariance method wandered. Here, the proposed method gave robust results with only one kernel.

The SUBWAY sequence involves a woman walking among other pedestrians that share similar covariance. Figure 1 shows that both color mean-shift and global covariance search are distracted by this clutter. To illustrate how the covariance method can frequently drift in the presence of clutter, Figure 3 plots the distance between ground truth and the track points reported in both the covariance and proposed methods. While the proposed method smoothly maintains track, notice that in this span of 60 frames, the covariance method pulls away three times. Its global nature enables it to eventually recover. For this sequence of frames, Figure 6 shows the covariance method pulling away.

In the DOG sequence, the dog changes statistics significantly. The covariance method with its update scheme is able to maintain track for roughly 92% of the frames; however, the proposed technique maintains track for the entire sequence.

As a final sequence, the CHASE demonstrates significant changes involving color, illumination, and view as the car weaves through the countryside. Even without updating the reference parameters, the proposed method is able to keep a steady track throughout the entire sequence (see Figure 8).

6. DISCUSSION

We have described a variational tracking technique that minimizes the distance between distribution parameters of the target and candidate objects. In addition to being more discriminative, the proposed method is more efficient than mean-shift and more robust than covariance tracking. The extension to mixture models and multiple kernels proved to increase performance.

Future work on this method might focus on incorporating optimal or collaborative kernel techniques [8, 7].

Acknowledgments

We would like to thank Dr. Fatih Porikli for providing the videos used in the results section.

This work was supported in part by grants from NSF, AFOSR, ARO, MURI, as well as by a grant from NIH (NAC P41 RR-13218) through Brigham and Women's Hospital. This work is part of the National Alliance for Medical Image Computing (NAMIC), funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant U54 EB005149. Information on the National Centers for Biomedical Computing can be obtained from `nihroadmap.nih.gov/bioinformatics`.

References

[1] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, 2003.

[2] F. Porikli, O. Tuzel, and P. Meer, "Covariance tracking using model update based on means on Riemannian manifolds," in *Computer Vision and Pattern Recognition*, 2006.

[3] S. Birchfield and S. Rangarajan, "Spatiograms versus histograms for region-based tracking," in *Computer Vision and Pattern Recognition*, 2005.

[4] C. Yang, R. Duraiswami, and L. Davis, "Efficient mean-shift tracking via a new similarity metric," in *Computer Vision and Pattern Recognition*, 2005.

[5] G. Hager, M. Dewan, and C. Stewart, "Multiple kernel tracking with ssd," in *Computer Vision and Pattern Recognition*, 2004.

[6] Z. Zivkovic and B. Kroese, "An EM-like algorithm for color-histogram-based object tracking," in *Computer Vision and Pattern Recognition*, 2004.

[7] Z. Fan, M. Yang, and Y. Wu, "Multiple collaborative kernel tracking," *Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 7, 2007.

[8] M. Dewan and G. Hager, "Toward optimal kernel-based tracking," in *Computer Vision and Pattern Recognition*, 2006.

Fig. 5. CROWD sequence (70 frames) using global covariance search *(dashed)* and proposed method with one kernel *(solid)*. Notice the covariance method picking up incorrect objects with similar covariance.

Fig. 6. SUBWAY sequence (60 frames) using global covariance search *(dashed)* and proposed method with three kernels *(solid)*.

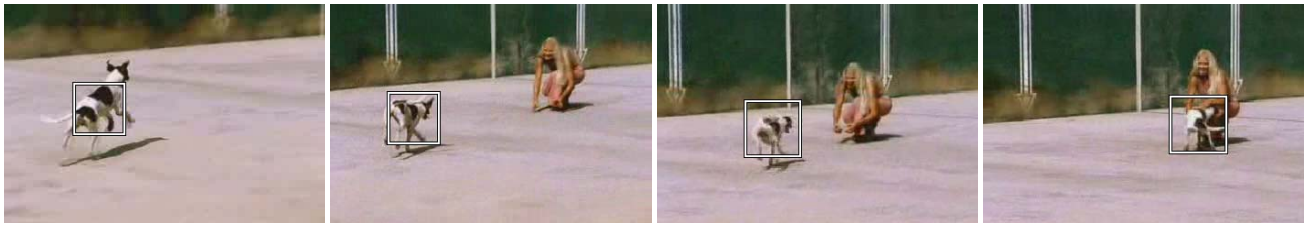

Fig. 7. DOG sequence (100 frames) using proposed method with three kernels. The target undergoes signficant changes in shape, scale, and intensity.

Fig. 8. CHASE sequence (411 frames) using proposed method with three kernels. The target undergoes significant scale and illumination changes.