Real-Time Tracking of Non-Rigid Objects using Mean Shift

Dorin Comaniciu Visvanathan Ramesh
Imaging & Visualization Department
Siemens Corporate Research
755 College Road East, Princeton, NJ 08540

Peter Meer
Electrical & Computer Engineering Department
Rutgers University
94 Brett Road, Piscataway, NJ 08855

Abstract

A new method for real-time tracking of non-rigid objects seen from a moving camera is proposed. The central computational module is based on the mean shift iterations and finds the most probable target position in the current frame. The dissimilarity between the target model (its color distribution) and the target candidates is expressed by a metric derived from the Bhattacharyya coefficient. The theoretical analysis of the approach shows that it relates to the Bayesian framework while providing a practical, fast and efficient solution. The capability of the tracker to handle in real-time partial occlusions, significant clutter, and target scale variations, is demonstrated for several image sequences.

1 Introduction

The efficient tracking of visual features in complex environments is a challenging task for the vision community. Real-time applications such as surveillance and monitoring [10], perceptual user interfaces [4], smart rooms [16, 28], and video compression [12] all require the ability to track moving objects. The computational complexity of the tracker is critical for most applications, only a small percentage of a system resources being allocated for tracking, while the rest is assigned to preprocessing stages or to high-level tasks such as recognition, trajectory interpretation, and reasoning [24].

This paper presents a new approach to the real-time tracking of non-rigid objects based on visual features such as color and/or texture, whose statistical distributions characterize the object of interest. The proposed tracking is appropriate for a large variety of objects with different color/texture patterns, being robust to partial occlusions, clutter, rotation in depth, and changes in camera position. It is a natural application to motion analysis of the mean shift procedure introduced earlier [6, 7]. The mean shift iterations are employed to find the target candidate that is the most similar to a given target model, with the similarity being expressed by a metric based on the Bhattacharyya coefficient. Various test sequences showed the superior tracking performance, obtained with low computational complexity.

The paper is organized as follows. Section 2 presents and extends the mean shift property. Section 3 introduces the metric derived from the Bhattacharyya coefficient. The tracking algorithm is developed and analyzed in Section 4. Experiments and comparisons are given in Section 5, and the discussions are in Section 6.

2 Mean Shift Analysis

We define next the sample mean shift, introduce the iterative mean shift procedure, and present a new theorem showing the convergence for kernels with convex and monotonic profiles. For applications of the mean shift property in low level vision (filtering, segmentation) see [6].

2.1 Sample Mean Shift

Given a set $\{\mathbf{x}_i\}_{i=1\ldots n}$ of n points in the d-dimensional space R^d, the *multivariate kernel density estimate* with kernel $K(\mathbf{x})$ and window radius (bandwidth) h, computed in the point $\mathbf{x}$ is given by

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right). \tag{1}$$

The minimization of the average global error between the estimate and the true density yields the multivariate Epanechnikov kernel [25, p.139]

$$K_E(\mathbf{x}) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1-\|\mathbf{x}\|^2) & \text{if } \|\mathbf{x}\|<1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where c_d is the volume of the unit d-dimensional sphere. Another commonly used kernel is the multivariate normal

$$K_N(\mathbf{x}) = (2\pi)^{-d/2}\exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right). \tag{3}$$

Let us introduce the *profile* of a kernel K as a function $k:[0,\infty)\rightarrow R$ such that $K(\mathbf{x}) = k(\|\mathbf{x}\|^2)$. For example, according to (2) the Epanechnikov profile is

$$k_E(x) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1-x) & \text{if } x<1 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

and from (3) the normal profile is given by

$$k_N(x) = (2\pi)^{-d/2}\exp\left(-\frac{1}{2}x\right). \tag{5}$$

Employing the profile notation we can write the density estimate (1) as

$$\hat{f}_K(\mathbf{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} k\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right). \tag{6}$$

We denote

$$g(x) = -k'(x)\ , \tag{7}$$

assuming that the derivative of k exists for all $x\in[0,\infty)$, except for a finite set of points. A kernel G can be defined as

$$G(\mathbf{x}) = Cg(\|\mathbf{x}\|^2), \tag{8}$$

where C is a normalization constant. Then, by taking the estimate of the density gradient as the gradient of the density estimate we have

$$\hat{\nabla} f_K(\mathbf{x}) \equiv \nabla \hat{f}_K(\mathbf{x}) = \frac{2}{nh^{d+2}} \sum_{i=1}^{n} (\mathbf{x} - \mathbf{x}_i)\, k'\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right)$$

$$= \frac{2}{nh^{d+2}} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{x})\, g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right) = \frac{2}{nh^{d+2}} \times$$

$$\times \left[\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right) \right] \left[\frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_i}{h} \right\|^2 \right)} - \mathbf{x} \right], \quad (9)$$

where $\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_i}{h} \right\|^2 \right)$ can be assumed to be nonzero. Note that the derivative of the Epanechnikov profile is the uniform profile, while the derivative of the normal profile remains a normal.

The last bracketin (9) contains the sample mean shift vector

$$M_{h,G}(\mathbf{x}) \equiv \frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_i}{h} \right\|^2 \right)} - \mathbf{x} \quad (10)$$

and the density estimate at $\mathbf{x}$

$$\hat{f}_G(\mathbf{x}) \equiv \frac{C}{nh^d} \sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right) \quad (11)$$

computed with kernel G. Using now (10) and (11), (9) becomes

$$\hat{\nabla} f_K(\mathbf{x}) = \hat{f}_G(\mathbf{x}) \frac{2/C}{h^2} M_{h,G}(\mathbf{x}) \quad (12)$$

from where it follows that

$$M_{h,G}(\mathbf{x}) = \frac{h^2}{2/C} \frac{\hat{\nabla} f_K(\mathbf{x})}{\hat{f}_G(\mathbf{x})}. \quad (13)$$

Expression (13) shows that the sample mean shift vector obtained with kernel G is an estimate of the normalized density gradient obtained with kernel K. This is a more general formulation of the property first remarked by Fukunaga [15, p. 535].

2.2 A Sufficient Convergence Condition

The *mean shift procedure* is defined recursively by computing the mean shift vector $M_{h,G}(\mathbf{x})$ and translating the center of kernel G by $M_{h,G}(\mathbf{x})$.

Let us denote by $\{\mathbf{y}_j\}_{j=1,2\ldots}$ the sequence of successive locations of the kernel G, where

$$\mathbf{y}_{j+1} = \frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\left\| \frac{\mathbf{y}_j - \mathbf{x}_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{y}_j - \mathbf{x}_i}{h} \right\|^2 \right)}, \quad j = 1, 2, \ldots \quad (14)$$

is the weighted mean at $\mathbf{y}_j$ computed with kernel G and $\mathbf{y}_1$ is the center of the initialk ernel. The density estimates computed with kernel K in the points (14) are

$$\hat{f}_K = \left\{ \hat{f}_K(j) \right\}_{j=1,2\ldots} \equiv \left\{ \hat{f}_K(\mathbf{y}_j) \right\}_{j=1,2\ldots} . \quad (15)$$

These densities are only implicitly defined to obtain $\hat{\nabla} f_K$. How ever we need them to prove the convergence of the sequences (14) and (15).

Theorem 1 *If the kernel K has a convex and monotonic decreasing profile and the kernel G is defined according to (7) and (8), the sequences (14) and (15) are convergent.*

The Theorem 1 generalizes the con vergence shown in [6], where K was the Epanechnikov kernel, and G the uniform kernel. Its proof is given in the Appendix. Note that Theorem 1 is also valid when we associate to each data point $\mathbf{x}_i$ a positive weight w_i.

3 Bhattacharyya Coefficient Based Metric for Target Localization

The task of finding the target location in the current frame is formulated as follows. The feature $\mathbf{z}$ representing the color and/or texture of the target model is assumed to have a density function $q_{\mathbf{z}}$, while the target candidate centered at location $\mathbf{y}$ has the feature distributed according to $p_{\mathbf{z}}(\mathbf{y})$. The problem is then to find the discrete location $\mathbf{y}$ whose associated density $p_{\mathbf{z}}(\mathbf{y})$ is the most similar to the target density $q_{\mathbf{z}}$.

T o define the similarity measure we take into account that the probability of classification error in statistical hypothesis testing is directly related to the similarity of the t w o distributions.The larger the probability of error, the more similar the distributions. Therefore, (con trary to thehypothesis testing), we formulate the target location estimation problem as the derivation of the estimate that *maximizes* the Bayes error associated with the model and candidate distributions. For the moment, w e assume that the target has equal prior probability to be present at any location $\mathbf{y}$ in the neighborhood of the previously estimated location.

An entity closely related to the Bayes error is the Bhattacharyya coefficient, whose general form is defined by [19]

$$\rho(\mathbf{y}) \equiv \rho[p(\mathbf{y}), q] = \int \sqrt{p_{\mathbf{z}}(\mathbf{y}) q_{\mathbf{z}}}\, d\mathbf{z} . \quad (16)$$

Properties of the Bhattacharyya coefficient such as its relation to the Fisher measure of information, quality of the sample estimate, and explicit forms for various distributions are given in [11, 19].

Our interest in expression (16) is, how ever, motivated by its near optimality given by the relationship to the Bayes error. Indeed, let us denote by $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ tw o sets of parameters for the distributions p and q and by $\boldsymbol{\pi} = (\pi_p, \pi_q)$ a set of prior probabilities. If the value of (16) is smaller for the set $\boldsymbol{\alpha}$ than for the set $\boldsymbol{\beta}$, it

can be proved [19] that, there exists a set of priors π^* for which the error probability for the set α is less than the error probability for the set β. In addition, starting from (16) upper and lower error bounds can be derived for the probability of error.

The derivation of the Bhattacharyya coefficient from sample data involves the estimation of the densities p and q, for which we employ the histogram formulation. Although not the best nonparametric density estimate [25], the histogram satisfies the low computational cost imposed by real-time processing. We estimate the discrete density $\hat{\mathbf{q}} = \{\hat{q}_u\}_{u=1\ldots m}$ (with $\sum_{u=1}^m \hat{q}_u = 1$) from the m-bin histogram of the target model, while $\hat{\mathbf{p}}(\mathbf{y}) = \{\hat{p}_u(\mathbf{y})\}_{u=1\ldots m}$ (with $\sum_{u=1}^m \hat{p}_u = 1$) is estimated at a given location $\mathbf{y}$ from the m-bin histogram of the target candidate. Hence, the sample estimate of the Bhattacharyya coefficient is given by

$$\hat{\rho}(\mathbf{y}) \equiv \rho\left[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}\right] = \sum_{u=1}^{m} \sqrt{\hat{p}_u(\mathbf{y})\hat{q}_u}. \tag{17}$$

The geometric interpretation of (17) is the cosine of the angle between the m-dimensional, unit vectors $\left(\sqrt{\hat{p}_1}, \ldots, \sqrt{\hat{p}_m}\right)^\top$ and $\left(\sqrt{\hat{q}_1}, \ldots, \sqrt{\hat{q}_m}\right)^\top$.

Using now (17) the distance between two distributions can be defined as

$$d(\mathbf{y}) = \sqrt{1 - \rho\left[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}\right]}\ . \tag{18}$$

The statistical measure (18) is well suited for the task of target localization since:

1. It is nearly optimal, due to its link to the Bayes error. Note that the widely used histogram intersection technique [26] has no such theoretical foundation.
2. It imposes a metric structure (see Appendix). The Bhattacharyya distance [15, p.99] or Kullback divergence [8, p.18] are not metrics since they violate at least one of the distance axioms.
3. Using discrete densities, (18) is invariant to the scale of the target (up to quantization effects). Histogram intersection is scale variant [26].
4. Being valid for arbitrary distributions, the distance (18) is superior to the Fisher linear discriminant, which yields useful results only for distributions that are separated by the mean-difference [15, p.132].

Similar measures were already used in computer vision. The Chernoff and Bhattacharyya bounds have been employed in [20] to determine the effectiveness of edge detectors. The Kullback divergence has been used in [27] for finding the pose of an object in an image.

The next section shows how to minimize (18) as a function of $\mathbf{y}$ in the neighborhood of a given location, by exploiting the mean shift iterations. Only the distribution of the object colors will be considered, although the texture distribution can be integrated into the same framework.

4 Tracking Algorithm

We assume in the sequel the support of two modules which should provide (a) detection and localization in the initial frame of the objects to track (targets) [21, 23], and (b) periodic analysis of each object to account for possible updates of the target models due to significant changes in color [22].

4.1 Color Representation

Target Model Let $\{\mathbf{x}_i^\star\}_{i=1\ldots n}$ be the pixel locations of the target model, centered at $\mathbf{0}$. We define a function $b : R^2 \to \{1 \ldots m\}$ which associates to the pixel at location $\mathbf{x}_i^\star$ the index $b(\mathbf{x}_i^\star)$ of the histogram bin corresponding to the color of that pixel. The probability of the color u in the target model is derived by employing a convex and monotonic decreasing kernel profile k which assigns a smaller weight to the locations that are farther from the center of the target. The weighting increases the robustness of the estimation, since the peripheral pixels are the least reliable, being often affected by occlusions (clutter) or background. The radius of the kernel profile is taken equal to one, by assuming that the generic coordinates x and y are normalized with h_x and h_y, respectively. Hence, we can write

$$\hat{q}_u = C \sum_{i=1}^{n} k(\|\mathbf{x}_i^\star\|^2)\delta\left[b(\mathbf{x}_i^\star) - u\right]\ , \tag{19}$$

where δ is the Kronecker delta function. The normalization constant C is derived by imposing the condition $\sum_{u=1}^m \hat{q}_u = 1$, from where

$$C = \frac{1}{\sum_{i=1}^{n} k(\|\mathbf{x}_i^\star\|^2)}\ , \tag{20}$$

since the summation of delta functions for $u = 1 \ldots m$ is equal to one.

Target Candidates Let $\{\mathbf{x}_i\}_{i=1\ldots n_h}$ be the pixel locations of the target candidate, centered at $\mathbf{y}$ in the current frame. Using the same kernel profile k, but with radius h, the probability of the color u in the target candidate is given by

$$\hat{p}_u(\mathbf{y}) = C_h \sum_{i=1}^{n_h} k\left(\left\|\frac{\mathbf{y} - \mathbf{x}_i}{h}\right\|^2\right)\delta\left[b(\mathbf{x}_i) - u\right]\ , \tag{21}$$

where C_h is the normalization constant. The radius of the kernel profile determines the number of pixels (i.e., the scale) of the target candidate. By imposing the condition that $\sum_{u=1}^m \hat{p}_u = 1$ we obtain

$$C_h = \frac{1}{\sum_{i=1}^{n_h} k(\|\frac{\mathbf{y}-\mathbf{x}_i}{h}\|^2)}. \tag{22}$$

Note that C_h does not depend on $\mathbf{y}$, since the pixel locations $\mathbf{x}_i$ are organized in a regular lattice, $\mathbf{y}$ being one of the lattice nodes. Therefore, C_h can be precalculated for a given kernel and different values of h.

4.2 Distance Minimization

According to Section 3, the most probable location $\mathbf{y}$ of the target in the current frame is obtained by minimizing the distance (18), which is equivalent to maximizing the Bhattacharyya coefficient $\hat{\rho}(\mathbf{y})$. The search for the new target location in the current frame starts at the estimated location $\hat{\mathbf{y}}_0$ of the target in the previous frame. Thus, the color probabilities $\{\hat{p}_u(\hat{\mathbf{y}}_0)\}_{u=1\dots m}$ of the target candidate at location $\hat{\mathbf{y}}_0$ in the current frame have to be computed first. Using Taylor expansion around the values $\hat{p}_u(\hat{\mathbf{y}}_0)$, the Bhattacharyya coefficient (17) is approximated as (after some manipulations)

$$\rho\left[\hat{\mathbf{p}}(\mathbf{y}),\hat{\mathbf{q}}\right] \approx \frac{1}{2}\sum_{u=1}^{m}\sqrt{\hat{p}_u(\hat{\mathbf{y}}_0)\hat{q}_u} + \frac{1}{2}\sum_{u=1}^{m}\hat{p}_u(\mathbf{y})\sqrt{\frac{\hat{q}_u}{\hat{p}_u(\hat{\mathbf{y}}_0)}} \tag{23}$$

where it is assumed that the target candidate $\{\hat{p}_u(\mathbf{y})\}_{u=1\dots m}$ does not change drastically from the initial $\{\hat{p}_u(\hat{\mathbf{y}}_0)\}_{u=1\dots m}$, and that $\hat{p}_u(\hat{\mathbf{y}}_0) > 0$ for all $u = 1\dots m$. Introducing now (21) in (23) we obtain

$$\rho\left[\hat{\mathbf{p}}(\mathbf{y}),\hat{\mathbf{q}}\right] \approx \frac{1}{2}\sum_{u=1}^{m}\sqrt{\hat{p}_u(\hat{\mathbf{y}}_0)\hat{q}_u} + \frac{C_h}{2}\sum_{i=1}^{n_h} w_i k\left(\left\|\frac{\mathbf{y}-\mathbf{x}_i}{h}\right\|^2\right) \tag{24}$$

where

$$w_i = \sum_{u=1}^{m}\delta\left[b(\mathbf{x}_i)-u\right]\sqrt{\frac{\hat{q}_u}{\hat{p}_u(\hat{\mathbf{y}}_0)}}. \tag{25}$$

Thus, to minimize the distance (18), the second term in equation (24) has to be maximized, the first term being independent of $\mathbf{y}$. The second term represents the density estimate computed with kernel profile k at $\mathbf{y}$ in the current frame, with the data being weighted by w_i (25). The maximization can be efficiently achieved based on the mean shift iterations, using the following algorithm.

Bhattacharyya Coefficient $\rho\left[\hat{\mathbf{p}}(\mathbf{y}),\hat{\mathbf{q}}\right]$ Maximization

Given the distribution $\{\hat{q}_u\}_{u=1\dots m}$ of the target model and the estimated location $\hat{\mathbf{y}}_0$ of the target in the previous frame:

1. Initialize the location of the target in the current frame with $\hat{\mathbf{y}}_0$, compute the distribution $\{\hat{p}_u(\hat{\mathbf{y}}_0)\}_{u=1\dots m}$, and evaluate
$$\rho\left[\hat{\mathbf{p}}(\hat{\mathbf{y}}_0),\hat{\mathbf{q}}\right] = \sum_{u=1}^{m}\sqrt{\hat{p}_u(\mathbf{y}_0)\hat{q}_u}\ .$$

2. Derive the weights $\{w_i\}_{i=1\dots n_h}$ according to (25).

3. Based on the mean shift vector, derive the new location of the target (14)
$$\hat{\mathbf{y}}_1 = \frac{\sum_{i=1}^{n_h}\mathbf{x}_i w_i g\left(\left\|\frac{\hat{\mathbf{y}}_0-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n_h} w_i g\left(\left\|\frac{\hat{\mathbf{y}}_0-\mathbf{x}_i}{h}\right\|^2\right)}\ . \tag{26}$$
Update $\{\hat{p}_u(\hat{\mathbf{y}}_1)\}_{u=1\dots m}$, and evaluate
$$\rho\left[\hat{\mathbf{p}}(\hat{\mathbf{y}}_1),\hat{\mathbf{q}}\right] = \sum_{u=1}^{m}\sqrt{\hat{p}_u(\mathbf{y}_1)\hat{q}_u}\ .$$

4. While $\rho\left[\hat{\mathbf{p}}(\hat{\mathbf{y}}_1),\hat{\mathbf{q}}\right] < \rho\left[\hat{\mathbf{p}}(\hat{\mathbf{y}}_0),\hat{\mathbf{q}}\right]$
Do $\hat{\mathbf{y}}_1 \leftarrow \frac{1}{2}\left(\hat{\mathbf{y}}_0+\hat{\mathbf{y}}_1\right)$.

5. If $\|\hat{\mathbf{y}}_1-\hat{\mathbf{y}}_0\| < \epsilon$ Stop.
Otherwise Set $\hat{\mathbf{y}}_0 \leftarrow \hat{\mathbf{y}}_1$ and go to Step 1.

The proposed optimization employs the mean shift vector in Step 3 to increase the value of the approximated Bhattacharyya coefficient expressed by (24). Since this operation does not necessarily increase the value of $\rho\left[\hat{\mathbf{p}}(\mathbf{y}),\hat{\mathbf{q}}\right]$, the test included in Step 4 is needed to validate the new location of the target. However, practical experiments (tracking different objects, for long periods of time) showed that the Bhattacharyya coefficient computed at the location defined by equation (26) was almost always larger than the coefficient corresponding to $\hat{\mathbf{y}}_0$. Less than 0.1% of the performed maximizations yielded cases where the Step 4 iterations were necessary. The termination threshold ϵ used in Step 5 is derived by constraining the vectors representing $\hat{\mathbf{y}}_0$ and $\hat{\mathbf{y}}_1$ to be within the same pixel in image coordinates.

The tracking consists in running for each frame the optimization algorithm described above. Thus, given the target model, the new location of the target in the current frame minimizes the distance (18) in the neighborhood of the previous location estimate.

4.3 Scale Adaptation

The scale adaptation scheme exploits the property of the distance (18) to be invariant to changes in the object scale. We simply modify the radius h of the kernel profile with a certain fraction (we used $\pm 10\%$), let the tracking algorithm to converge again, and choose the radius yielding the largest decrease in the distance (18). An IIR filter is used to derive the new radius based on the current measurements and old radius.

5 Experiments

The proposed method has been applied to the task of tracking a football player marked by a hand-drawn ellipsoidal region (first image of Figure 1). The sequence has 154 frames of 352×240 pixels each and the initial normalization constants (determined from the size of the target model) were $(h_x, h_y) = (71, 53)$. The Epanechnikov profile (4) has been used for histogram computation, therefore, the mean shift iterations were computed with the uniform profile. The target histogram has been derived in the RGB space with $32 \times 32 \times 32$ bins. The algorithm runs comfortably at 30 fps on a 600 MHz PC, Java implementation.

The tracking results are presented in Figure 1. The mean shift based tracker proved to be robust to partial occlusion, clutter, distractors (frame 140 in Figure 1),

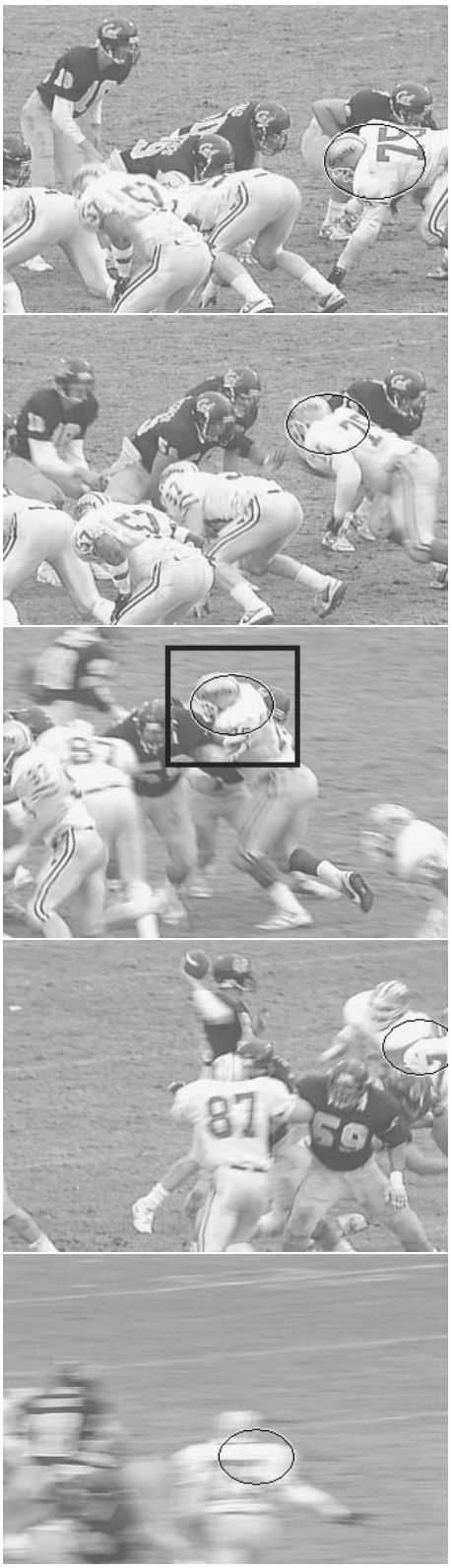

Figure 1: *Football* sequence: Tracking the player no. 75 with initial window of 71×53 pixels. The frames 30, 75, 105, 140, and 150 are shown.

and camera motion. Since no motion model has been assumed, the tracker adapted well to the nonstationary character of the player's movements, which alternates abruptly between slow and fast action. In addition, the intense blurring present in some frames and due to the camera motion, did not influence the tracker performance (frame 150 in Figure 1). The same effect, however, can largely perturb contour based trackers.

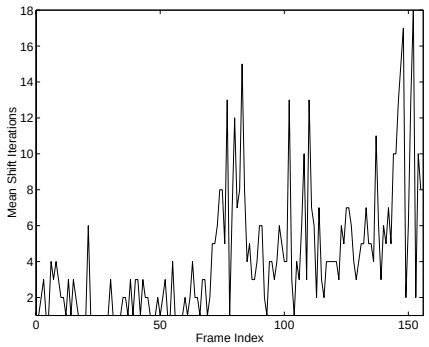

Figure 2: The number of mean shift iterations function of the frame index for the *Football* sequence. The mean number of iterations is 4.19 per frame.

The number of mean shift iterations necessary for each frame (one scale) in the *Football* sequence is shown in Figure 2. One can identify two central peaks, corresponding to the movement of the player to the center of the image and back to the left side. The last and largest peak is due to the fast movement from the left to the right side.

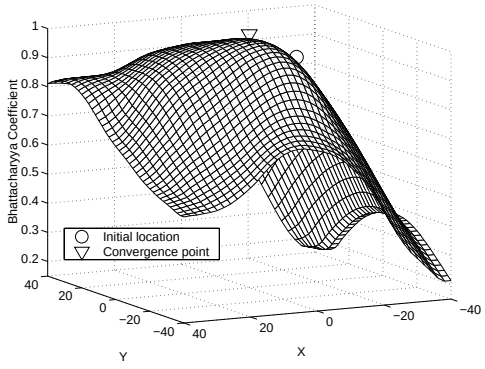

Figure 3: Values of the Bhattacharyya coefficient corresponding to the marked region (81×81 pixels) in frame 105 from Figure 1. The surface is asymmetric, due to the player colors that are similar to the target. Four mean shift iterations were necessary for the algorithm to converge from the initial location (circle).

To demonstrate the efficiency of our approach, Figure 3 presents the surface obtained by computing the Bhattacharyya coefficient for the rectangle marked in Figure 1, frame 105. The target model (the selected elliptical region in frame 30) has been compared with the target candidates obtained by sweeping the elliptical region in frame 105 inside the rectangle. While most of the tracking approaches based on regions [3, 14, 21]

must perform an exhaustive search in the rectangle to find the maximum, our algorithm converged in four iterations as shown in Figure 3. Note that since the basin of attraction of the mode covers the entire window, the correct location of the target would have been reached also from farther initial points. An optimized computation of the exhaustive search of the mode [13] has a much larger arithmetic complexity, depending on the chosen search area.

The new method has been applied to track people on subway platforms. The camera being fixed, additional geometric constraints and also background subtraction can be exploited to improve the tracking process. The following sequences, however, have been processed with the algorithm unchanged.

A first example is shown in Figure 4, demonstrating the capability of the tracker to adapt to scale changes. The sequence has 187 frames of 320×240 pixels each and the initial normalization constants were $(h_x, h_y) = (23, 37)$.

Figure 5 presents six frames from a 2 minute sequence showing the tracking of a person from the moment she enters the subway platform till she gets on the train (≈ 3600 frames). The tracking performance is remarkable, taking into account the low quality of the processed sequence, due to the compression artifacts. A thorough evaluation of the tracker, however, is subject to our current work.

The minimum value of the distance (18) for each frame is shown in Figure 6. The compression noise determined the distance to increase from 0 (perfect match) to a stationary value of about 0.3. Significant deviations from this value correspond to occlusions generated by other persons or rotations in depth of the target. The large distance increase at the end signals the complete occlusion of the target.

6 Discussion

By exploiting the spatial gradient of the statistical measure (18) the new method achieves real-time tracking performance, while effectively rejecting background clutter and partial occlusions.

Note that the same technique can be employed to derive the measurement vector for optimal prediction schemes such as the (Extended) Kalman filter [1, p.56, 106], or multiple hypothesis tracking approaches [5, 9, 17, 18]. In return, the prediction can determine the priors (defining the presence of the target in a given neighborhood) assumed equal in this paper. This connection is however beyond the scope of this paper. A patent application has been filed covering the tracking algorithm together with the Kalman extension and various applications [29].

We finally observe that the idea of centroid computation is also employed in [22]. The mean shift was used for tracking human faces [4], by projecting the

Figure 4: *Subway1* sequence: The frames 500, 529, 600, 633, and 686 are shown (left-right, top-down).

histogram of a face model onto the incoming frame. However, the direct projection of the model histogram onto the new frame can introduce a large bias in the estimated location of the target, and the resulting measure is scale variant. Gradient based region tracking has been formulated in [2] by minimizing the energy of the deformable region, but no real-time claims were made.

APPENDIX

Proof of Theorem 1

Since n is finite the sequence $\hat{f}_K$ is bounded, therefore, it is sufficient to show that $\hat{f}_K$ is strictly monotonic increasing, i.e., if $\mathbf{y}_j \neq \mathbf{y}_{j+1}$ then $\hat{f}_K(j) < \hat{f}_K(j+1)$, for all $j = 1, 2 \ldots$.

By assuming without loss of generality that $\mathbf{y}_j = \mathbf{0}$ we can write

$$\hat{f}_K(j+1) - \hat{f}_K(j) = \\ = \frac{1}{nh^d} \sum_{i=1}^{n} \left[k\left(\left\| \frac{\mathbf{y}_{j+1} - \mathbf{x}_i}{h} \right\|^2 \right) - k\left(\left\| \frac{\mathbf{x}_i}{h} \right\|^2 \right) \right] \quad \text{(A.1)}$$

Figure 5: *Subway2* sequence: The frames 3140, 3516, 3697, 5440, 6081, and 6681 are shown (left-right, top-down).

The convexity of the profile k implies that

$$k(x_2) \geq k(x_1) + k'(x_1)(x_2 - x_1) \tag{A.2}$$

for all $x_1, x_2 \in [0, \infty)$, $x_1 \neq x_2$, and since $k' = -g$, the inequality (A.2) becomes

$$k(x_2) - k(x_1) \geq g(x_1)(x_1 - x_2). \tag{A.3}$$

Using now (A.1) and (A.3) we obtain

$$\hat{f}_K(j+1) - \hat{f}_K(j) \geq$$

$$\geq \frac{1}{nh^{d+2}} \sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right) \left[\|\mathbf{x}_i\|^2 - \|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2\right]$$

$$= \frac{1}{nh^{d+2}} \sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right) \left[2\mathbf{y}_{j+1}^\top \mathbf{x}_i - \|\mathbf{y}_{j+1}\|^2\right] = \frac{1}{nh^{d+2}}$$

$$\times \left[2\mathbf{y}_{j+1}^\top \sum_{i=1}^{n} \mathbf{x}_i g\left(\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right) - \|\mathbf{y}_{j+1}\|^2 \sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right)\right] \tag{A.4}$$

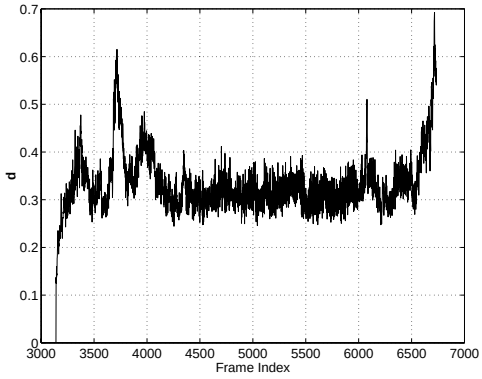

Figure 6: The detected minimum value of distance d function of the frame index for the 2 minute *Subway2* sequence. The peaks in the graph correspond to occlusions or rotations in depth of the target. For example, the peak of value $d \approx 0.6$ corresponds to the partial occlusion in frame 3697, shown in Figure 5. At the end of the sequence, the person being tracked gets on the train, which produces a complete occlusion.

and by employing (14) it results that

$$\hat{f}_K(j+1) - \hat{f}_K(j) \geq \frac{1}{nh^{d+2}} \|\mathbf{y}_{j+1}\|^2 \sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right). \tag{A.5}$$

Since k is monotonic decreasing we have $-k'(x) \equiv g(x) \geq 0$ for all $x \in [0, \infty)$. The sum $\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}_i}{h}\right\|^2\right)$ is strictly positive, since it was assumed to be nonzero in the definition of the mean shift vector (10). Thus, as long as $\mathbf{y}_{j+1} \neq \mathbf{y}_j = \mathbf{0}$, the right term of (A.5) is strictly positive, i.e., $\hat{f}_K(j+1) - \hat{f}_K(j) > 0$. Consequently, the sequence $\hat{f}_K$ is convergent.

To prove the convergence of the sequence $\{\mathbf{y}_j\}_{j=1,2...}$ we rewrite (A.5) but without assuming that $\mathbf{y}_j = \mathbf{0}$. After some algebra we have

$$\hat{f}_K(j+1) - \hat{f}_K(j) \geq \frac{1}{nh^{d+2}} \|\mathbf{y}_{j+1} - \mathbf{y}_j\|^2 \sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{y}_j - \mathbf{x}_i}{h}\right\|^2\right) \tag{A.6}$$

Since $\hat{f}_K(j+1) - \hat{f}_K(j)$ converges to zero, (A.6) implies that $\|\mathbf{y}_{j+1} - \mathbf{y}_j\|$ also converges to zero, i.e., $\{\mathbf{y}_j\}_{j=1,2...}$ is a Cauchy sequence. This completes the proof, since any Cauchy sequence is convergent in the Euclidean space.

Proof that the distance $d(\hat{\mathbf{p}}, \hat{\mathbf{q}}) = \sqrt{1 - \rho(\hat{\mathbf{p}}, \hat{\mathbf{q}})}$ is a metric

The proof is based on the properties of the Bhattacharyya coefficient (17). According to the Jensen's inequality [8, p.25] we have

$$\rho(\hat{\mathbf{p}}, \hat{\mathbf{q}}) = \sum_{u=1}^{m} \sqrt{\hat{p}_u \hat{q}_u} = \sum_{u=1}^{m} \hat{p}_u \sqrt{\frac{\hat{q}_u}{\hat{p}_u}} \leq \sqrt{\sum_{u=1}^{m} \hat{q}_u} = 1, \tag{A.7}$$

with equality iff $\hat{\mathbf{p}} = \hat{\mathbf{q}}$. Therefore, $d(\hat{\mathbf{p}}, \hat{\mathbf{q}}) = \sqrt{1 - \rho(\hat{\mathbf{p}}, \hat{\mathbf{q}})}$ exists for all discrete distributions $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$, is positive, symmetric, and is equal to zero iff $\hat{\mathbf{p}} = \hat{\mathbf{q}}$.

The triangle inequality can be proven as follows. Let us consider the discrete distributions $\hat{\mathbf{p}}$, $\hat{\mathbf{q}}$, and $\hat{\mathbf{r}}$, and define the associated m-dimensional points $\boldsymbol{\xi}_p = \left(\sqrt{\hat{p}_1}, \ldots, \sqrt{\hat{p}_m}\right)^\top$, $\boldsymbol{\xi}_q = \left(\sqrt{\hat{q}_1}, \ldots, \sqrt{\hat{q}_m}\right)^\top$, and $\boldsymbol{\xi}_r = \left(\sqrt{\hat{r}_1}, \ldots, \sqrt{\hat{r}_m}\right)^\top$ on the unit hypersphere, centered at the origin. By taking into account the geometric interpretation of the Bhattacharyya coefficient, the triangle inequality

$$d(\hat{\mathbf{p}}, \hat{\mathbf{r}}) + d(\hat{\mathbf{q}}, \hat{\mathbf{r}}) \geq d(\hat{\mathbf{p}}, \hat{\mathbf{q}}) \tag{A.8}$$

is equivalent to

$$\sqrt{1 - cos(\boldsymbol{\xi}_p, \boldsymbol{\xi}_r)} + \sqrt{1 - cos(\boldsymbol{\xi}_q, \boldsymbol{\xi}_r)} \geq \sqrt{1 - cos(\boldsymbol{\xi}_p, \boldsymbol{\xi}_q)}. \tag{A.9}$$

If we fix the points $\boldsymbol{\xi}_p$ and $\boldsymbol{\xi}_q$, and the angle between $\boldsymbol{\xi}_p$ and $\boldsymbol{\xi}_r$, the left side of inequality (A.9) is minimized when the vectors $\boldsymbol{\xi}_p$, $\boldsymbol{\xi}_q$, and $\boldsymbol{\xi}_r$ lie in the same plane. Thus, the inequality (A.9) can be reduced to a 2-dimensional problem that can be easily demonstrated by employing the half-angle sinus formula and a few trigonometric manipulations.

Acknowledgment

Peter Meer was supported by the NSF under the grant IRI 99-87695.

References

[1] Y. Bar-Shalom, T. Fortmann, *Tracking and Data Association*, Academic Press, London, 1988.

[2] B. Bascle, R. Deriche, "Region Tracking through Image Sequences," *IEEE Int'l Conf. Comp. Vis.*, Cambridge, Massachusetts, 302–307, 1995.

[3] S. Birchfield, "Elliptical Head Tracking using intensity Gradients and Color Histograms," *IEEE Conf. on Comp. Vis. and Pat. Rec.*, Santa Barbara, 232–237, 1998.

[4] G.R. Bradski, "Computer Vision Face Tracking as a Component of a Perceptual User Interface," *IEEE Work. on Applic. Comp. Vis.*, Princeton, 214–219, 1998.

[5] T.J. Cham, J.M. Rehg, "A Multiple Hypothesis Approach to Figure Tracking," *IEEE Conf. on Comp. Vis. and Pat. Rec.*, Fort Collins, vol. 2, 239–245, 1999.

[6] D. Comaniciu, P. Meer, "Mean Shift Analysis and Applications," *IEEE Int'l Conf. Comp. Vis.*, Kerkyra, Greece, 1197–1203, 1999.

[7] D. Comaniciu, P. Meer, "Distribution Free Decomposition of Multivariate Data," *Pattern Anal. and Applic.*, 2:22–30, 1999.

[8] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 1991.

[9] I.J. Cox, S.L. Hingorani, "An Efficient Implementation of Reid's Multiple Hypothesis Tracking Algorithm and Its Evaluation for the Purpose of Visual Tracking," *IEEE Trans. Pattern Analysis Machine Intell.*, 18:138–150, 1996.

[10] Y. Cui, S. Samarasekera, Q. Huang, M Greiffenhagen, "Indoor Monitoring Via the Collaboration Between a Peripheral Senson and a Foveal Sensor," *IEEE Workshop on Visual Surveillance*, Bombay, India, 2–9, 1998.

[11] A. Djouadi, O. Snorrason, F.D. Garber, "The Quality of Training-Sample Estimates of the Bhattacharyya Coefficient," *IEEE Trans. Pattern Analysis Machine Intell.*, 12:92–97, 1990.

[12] A. Eleftheriadis, A. Jacquin, "Automatic Face Location Detection and Tracking for Model-Assisted Coding of Video Teleconference Sequences at Low Bit Rates," *Signal Processing - Image Communication*, 7(3): 231–248, 1995.

[13] F. Ennesser, G. Medioni, "Finding Waldo, or Focus of Attention Using Local Color Information," *IEEE Trans. Pattern Anal. Machine Intell.*, 17(8):805–809, 1995.

[14] P. Fieguth, D. Terzopoulos, "Color-Based Tracking of Heads and Other Mobile Objects at Video Frame Rates," *IEEE Conf. on Comp. Vis. and Pat. Rec.*, Puerto Rico, 21–27, 1997.

[15] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Second Ed., Academic Press, Boston, 1990.

[16] S.S. Intille, J.W. Davis, A.F. Bobick, "Real-Time Closed-World Tracking," *IEEE Conf. on Comp. Vis. and Pat. Rec.*, Puerto Rico, 697–703, 1997.

[17] M. Isard, A. Blake, "Condensation - Conditional Density Propagation for Visual Tracking," *Intern. J. Comp. Vis.*, 29(1):5–28, 1998.

[18] M. Isard, A. Blake, "ICondensation: Unifying Low-Level and High-Level Tracking in a Stochastic Framework," *European Conf. Comp. Vision*, Freiburg, Germany, 893–908, 1998.

[19] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," *IEEE Trans. Commun. Tech.*, COM-15:52–60, 1967.

[20] S. Konishi, A.L. Yuille, J. Coughlan, S.C. Zhu, "Fundamental Bounds on Edge Detection: An Information Theoretic Evaluation of Different Edge Cues," *IEEE Conf. on Comp. Vis. and Pat. Rec.*, Fort Collins, 573–579, 1999.

[21] A.J. Lipton, H. Fujiyoshi, R.S. Patil, "Moving Target Classification and Tracking from Real-Time Video," *IEEE Workshop on Applications of Computer Vision*, Princeton, 8–14, 1998.

[22] S.J. McKenna, Y. Raja, S. Gong, "Tracking Colour Objects using Adaptive Mixture Models," *Image and Vision Computing*, 17:223–229, 1999.

[23] N. Paragios, R. Deriche, "Geodesic Active Regions for Motion Estimation and Tracking," *IEEE Int'l Conf. Comp. Vis.*, Kerkyra, Greece, 688–674, 1999.

[24] R. Rosales, S. Sclaroff, "3D trajectory Recovery for Tracking Multiple Objects and Trajectory Guided Recognition of Actions," *IEEE Conf. on Comp. Vis. and Pat. Rec.*, Fort Collins, vol. 2, 117–123, 1999.

[25] D.W. Scott, *Multivariate Density Estimation*, New York: Wiley, 1992.

[26] M.J. Swain, D.H. Ballard, "Color Indexing," *Intern. J. Comp. Vis.*, 7(1):11–32, 1991.

[27] P. Viola, W.M. Wells III, "Alignment by Maximization of Mutual Information," *IEEE Int'l Conf. Comp. Vis.*, Cambridge, Massachusetts, 16–23, 1995.

[28] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans. Pattern Analysis Machine Intell.*, 19:780–785, 1997.

[29] "Real-Time Tracking of Non-Rigid Objects using Mean Shift," US patent pending.