Probabilistic Tracking in Joint Feature-Spatial Spaces

Ahmed Elgammal
Department of Computer Science
Rutgers University
Piscataway, NJ
elgammal@cs.rutgers.edu

Ramani Duraiswami
UMIACS
University of Maryland
College Park, MD
ramani@umiacs.umd.edu

Larry S. Davis
Department of Computer Science
University of Maryland
College Park, MD
lsd@cs.umd.edu

Abstract

In this paper we present a probabilistic framework for tracking regions based on their appearance. We exploit the feature-spatial distribution of a region representing an object as a probabilistic constraint to track that region over time. The tracking is achieved by maximizing a similarity-based objective function over transformation space given a nonparametric representation of the joint feature-spatial distribution. Such a representation imposes a probabilistic constraint on the region feature distribution coupled with the region structure which yields an appearance tracker that is robust to small local deformations and partial occlusion. We present the approach for the general form of joint feature-spatial distributions and apply it to tracking with different types of image features including row intensity, color and image gradient.

1 Introduction

The problem of region tracking can be defined as: given a region in an image find the global transformation of the region in successive frames where the region motion is constrained to be chosen from a given class of transformations. Region tracking can be achieved by tracking the boundary (contour) of the region [12, 13], the interior of the region [6, 16, 10, 14, 3, 8, 11, 1], or both [2].

Appearance-based approaches for region tracking vary from approaches that strictly preserve region structure - by imposing rigidity constraints on the region, to approaches that totally ignore the region structure and track based on feature distributions. Approaches that preserve region structure typically establishes correspondences at the pixel level (or sparse feature level) between the region model and any target region by minimizing some metric in the feature space to obtain the global transformation, for example [10, 18]. Such approaches perform well for tracking rigid bodies and have been generalized to track articulated

and even deformable objects if the articulation or the deformation model are known. On the other end, approaches that ignore region structure establish correspondence at the region level based on feature distribution such as histogram-based approaches, for example [6, 8]. These approaches have great flexibility to track deformable and non-rigid objects as well as being robust to partial occlusion, but they can lose the tracked region to another region with similar feature distribution.

In this paper we present a probabilistic framework for tracking regions based on their appearance. In our formulation we consider both the feature value and the feature location to be probabilistic random variables. We exploit the feature-spatial distribution of a region representing a non-rigid object as a constraint to track that region over time. Given a sample from a region representing the object, we estimate the feature-spatial joint distribution using kernel density estimation. The tracked object is located in each new frame by searching a transformation space for the transformation that maximizes a similarity-based objective function. Representing the joint feature-spatial distribution of the region imposes a probabilistic constraint on the tracked region that facilitate tracking even under small local deformations or measurement uncertainties. The framework we present is general and we will show that rigid tracking approaches based on minimizing SSD in the feature space are special cases of this framework. On the other end, non-rigid tracking approaches based on feature distributions (histogram trackers) are also special cases of the proposed framework. We present the approach for general form joint feature-spatial distributions and apply it to tracking with different types of image features including intensity, color and edge features.

Tracking using joint feature-spatial distributions involves some problematic issues: For example, the type of transformations involved in the spatial domain (e.g., geometric or deformation) are different from the type of transformation that might be applicable in the feature domain. The feature-spatial distributions are nonstandard in shape

1

and can be high dimensional, therefore they require a general approaches to handle the density estimation and an efficient computational framework. Also, typically, the features are correlated with their locations. The approach presented in this paper addresses these issues.

The structure of the paper is as follows: Section 2 presents the joint space representation and the tracking objective function and it's asymptotic behavior. Section 3 presents the region localization approach. Section 4 shows a tracking example using a synthetic target and studies some tracking issues. Section 5 shows example experimental results with different feature-spatial spaces. Finally, section 6 conclude the paper with a discussion about some important issues.

2 Representation

2.1 Joint Distribution Representation

Our objective is to model the feature-spatial joint probability distribution of a region. In rigid body tracking, the relative locations of the features are deterministic and the region undergoes certain geometric transformation, where the objective of the tracking is to find such transformation under the assumption of feature invariance (for example brightness invariance). In our formulation, we consider both the feature value and the feature location to be probabilistic random variables.

Let $u(x)$ be a d-dimensional feature vector at image location x. We use image features in a general sense, so u can be image intensity, color, texture, edge features, wavelet filter response, etc. We use R to denote the region or a sample from the region interchangeably.

Let $S = \{y_i = (x_i, u_i)\}_{i=1..N}$ be a sample from the target region R where x_i is the sample 2-dimensional location, and $u_i = u(x_i)$ is a d-dimensional feature vector at x_i. S is a sample from the joint spatial-feature probability distribution $P(X, U)$ of the region. Given, S, we can obtain an estimate of the probability of any point, (x, u), in the joint space (assuming a discrete distribution) using multivariate kernel density estimation [17] as

$$\hat{P}(x, u) = \frac{1}{N} \sum_{i=1}^{N} K_\sigma(x - x_i) G_\kappa(u - u_i) \quad (1)$$

where K_σ is a 2-dimensional kernel with a bandwidth σ and G_κ is a d-dimensional kernel with a bandwidth κ. The bandwidth in the spatial dimensions represents the variability in feature location due to the local deformation or measurement uncertainty while the bandwidth in the feature dimensions represent the variability in the value of the feature. We use the notation $K_\lambda(t)$ to denote the d dimensional kernel $\frac{1}{\lambda_1 \cdots \lambda_d} K(\frac{t_1}{\lambda_1}, \cdots, \frac{t_d}{\lambda_d})$. Equation 1 gives an estimate of the joint probability density at any point in the $d + 2$ dimensional spatial-feature space.

If the region R undergoes some geometric transformation to a new region R' where $T(\cdot; a)$ is the inverse transformation from R' to R with parameters a such that $R = T(R'; a)$, then we can obtain an estimate for observing a certain feature, u, at a certain location x using

$$\hat{P}_a(x, u) = \frac{1}{N} \sum_{i=1}^{N} K_\sigma(T(x; a) - x_i) G_\kappa(u - u_i) \quad (2)$$

We can estimate the likelihood $L(a)$ of the transformation $T(\cdot; a)$ by sampling from the new region and integrating the joint probability estimates for P from (2)

$$L(a) \equiv L(R') = \int_{R'} \log \hat{P}_a(x, u) \quad (3)$$

2.2 Similarity Based Tracking

Let $P(y)$ and $Q(y)$ be the probability distributions for R and R' respectively in the joint spatial-feature space. We can measure the similarity between the region R and the transformed region R' by measuring the similarity between the distributions P and Q using the Kullback-Leibler (cross entropy) information distance

$$D(Q\|P) = \int_{-\infty}^{\infty} Q(y) \log \frac{Q(y)}{P(y)} dy \quad (4)$$

where $D(Q\|P) \geq 0$ and equality holds when the two distributions are identical. This can be written as

$$D(Q\|P) = \int Q(y) \log Q(y) dy - \int Q(y) log P(y) dy \quad (5)$$

Since we sample from the transformed region, i.e., the samples y are drawn according to the joint density Q, the first term in equation 5 is (-1) times the entropy $H(Q)$ of the transformed region. By the law of large number, the second term is the expectation given the sample Q of the log likelihood of the sample under the density P, i.e,

$$\int Q(y) log P(y) dy \approx E_Q(log P(y))$$

Therefore the second term is the likelihood function $L(R')$ defined in equation 3.

We can define a similarity-based objective function

$$\psi(R') = -H(R') - L(R') \quad (6)$$

Equivalently we will write the objective function in terms of the transformation parameter vector a, since the choice of R' depends only on a

$$\psi(a) = -H(a) - L(a) \quad (7)$$

Given this objective function, we can formalize region tracking as a search over the transformation parameter space

for the parameter a_t^* that maximizes the similarity of the transformed region at time t

$$a_t^* = arg_a min \psi(a)$$

The optimal hypothesis represents a local maxima in the parameter search space. The likelihood function as defined in equation 3 as well as the probability estimate in equation 2 are continuous and differentiable and, therefore, a gradient based optimization technique would converge to that solution as long as the initial guess is within a small neighborhood of that maxima.

2.3 Asymptotic Behavior

The bandwidth of the spatial kernel represents the variability in feature location (for example, due to uncertainty in the measurement or due to local deformation). The spatial kernel acts as a weighting mechanism such that the expected feature value in a certain location depends on other feature in the neighborhood. To see the effect of changing the bandwidth of the spatial kernel K on the tracking formulation we consider here the extreme cases where the bandwidth is set to 0 or ∞

2.3.1 Conversion to SSD Tracking

Consider the case where $\sigma = 0$. In this case The spatial kernel reduced to a kronecker delta function

$$k_\sigma(x - x_i) = \begin{cases} 1 & x = x_i \\ 0 & o.w. \end{cases}$$

In this case the location of the features are deterministic and we have a collection of random variables representing the feature value at each of the locations x_i's. Let us denote the feature value at location x_i by $u(x_i)$. Then, given the estimate of equation 1, the probability density function of the feature at x_i reduces to

$$p_{x_i}(u) = \hat{P}(u|x_i) = \frac{\hat{P}(u, x_i)}{\hat{P}(x_i)} = G_\kappa(u - u_i)$$

If we used a Gaussian kernel for G, the likelihood function of a certain transformation in 3 reduces to sum of squared distance (SSD)

$$L(a) = \sum_i \|u(T(x_i; a)) - u_i\|^2$$

Therefore, in this case, the region tracking is mainly a search for the transformation that minimizes the sum of squared error given feature consistency constraints. This is similar to many rigid body tracking approaches such as [10, 18]

2.3.2 Conversion to Histogram Tracking

Now let us consider the other extreme where the bandwidth of the spatial kernel tends to ∞. In this case K becomes a flat kernel with infinite support, i.e., $K(\cdot) = c$ where c is a constant and the joint probability estimate of equation 1 reduces to

$$\hat{P}(x, u) = \frac{c}{N} \sum_{i=1}^{N} G_\kappa(u - u_i) = c\hat{P}(u)$$

which is an estimate the feature probability density function $\hat{P}(u)$ using the kernel G. we can think of this as a histogram of the feature values smoothed with G. Therefore the tracking reduces histogram tracking where the objective is to find the transformation that yields a region with similar feature distribution and no rigidity constraint is enforced and therefore no structure is preserved such as the work of [6, 8]

In conclusion, the formalization provided in this paper is general. Changing the spatial kernel bandwidth provides a continuum between treating the locations of the features as deterministic (imposing rigidity constraint), i.e., template matching, and totally ignoring the spatial domain by tracking based on feature distribution only, i.e., histogram matching. Instead, the structure of the region is preserved in a relaxed manner based on the joint spatial-feature distribution.

3 Region localization

3.1 Likelihood Maximization

Assume that the tracked region R undergoes a geometric transformation to region R' where T is the inverse transformation such that $R = T(R'; a)$. Let $y' = [x' \quad u']^t$ be samples from R' represented as $2 + d$ dimensional vector and, similarly, let $y = [x \quad u]^t = [T(x'; a) \quad u']^t$, i.e., the sample transformed with T. Given the transformation $T(x; a)$, the Jacobian matrix of T is a $k \times 2$ matrix where k is the number of the parameters in the transformation

$$\nabla_a T(x; a) = \left[\frac{\partial T(x; a)}{\partial a_1} \Big| \frac{\partial T(x; a)}{\partial a_2} \Big| \cdots \Big| \frac{\partial T(x; a)}{\partial a_k} \right]^t . \quad (8)$$

The gradient of the likelihood function $L(a)$ as defined in equation 3 with respect to the parameter a is

$$\nabla_a L(a) = \sum_{(x', u') \in R'} \nabla_a T(x'; a) \cdot \frac{\nabla_x \hat{P}_a(x', u')}{\hat{P}_a(x', u')} \quad (9)$$

where $\nabla_x \hat{P}_a(x', u')$ is the gradient of the density estimate with respect to the location, x, at sample point (x', u').

For estimating the density gradient, we will consider the general case where the feature is considered to be a continuous function over the spatial domain $u(x) =$

$[u_1(x), u_2(x), \cdots u_d(x)]$, then feature gradient with respect to the location x is the $2 \times d$ matrix

$$\nabla_x u(x) = \left[\frac{\partial u_1(x)}{\partial x} \Big| \frac{\partial u_2(x)}{\partial x} \Big| \cdots \Big| \frac{\partial u_d(x)}{\partial x} \right] \qquad (10)$$

Given the estimate for the joint density $\hat{P}_a(x, u)$ as defined by equation 2 where we use a Gaussian kernels[1] for both the spatial and feature kernels. For a Gaussian kernel $K_\sigma(x)$ the derivative is $\acute{K}_\sigma(x) = -\frac{x}{\sigma^2} K_\sigma(x)$. Therefore, the gradient of the density estimate is

$$\nabla_x \hat{P}_a(x', u') = \frac{-1}{N} \cdot$$
$$\left(\frac{1}{\sigma^2} \sum_{i=1}^{N} (x - x_i) \cdot K_\sigma(x - x_i) G_\kappa(u - u_i) \right.$$
$$+ \quad \nabla_x u \cdot \frac{1}{\kappa^2} \sum_{i=1}^{N} (u - u_i) \cdot K_\sigma(x - x_i) G_\kappa(u - u_i) \Bigg)$$
$$= \quad \frac{-1}{N} \cdot \left(\sum_{i=1}^{N} K_\sigma(x - x_i) G_\kappa(u - u_i) \right) \cdot$$
$$\cdot \left[\frac{1}{\sigma^2} \left(x - \frac{\sum_{i=1}^{N} x_i K_\sigma(x - x_i) G_\kappa(u - u_i)}{\sum_{i=1}^{N} K_\sigma(x - x_i) G_\kappa(u - u_i)} \right) \right.$$
$$+ \quad \nabla_x u \cdot \frac{1}{\kappa^2} \left(u - \frac{\sum_{i=1}^{N} u_i K_\sigma(x - x_i) G_\kappa(u - u_i)}{\sum_{i=1}^{N} K_\sigma(x - x_i) G_\kappa(u - u_i)} \right) \Bigg]$$

The first term in the last equation is the density estimate and therefore

$$\frac{\nabla_x \hat{P}_a(x', u')}{\hat{P}_a(x', u')} =$$
$$\left[\frac{1}{\sigma^2} \left(\frac{\sum_{i=1}^{N} x_i K_\sigma(x - x_i) G_\kappa(u - u_i)}{\sum_{i=1}^{N} K_\sigma(x - x_i) G_\kappa(u - u_i)} - x \right) \right.$$
$$+ \quad \nabla_x u \cdot \frac{1}{\kappa^2} \left(\frac{\sum_{i=1}^{N} u_i K_\sigma(x - x_i) G_\kappa(u - u_i)}{\sum_{i=1}^{N} K_\sigma(x - x_i) G_\kappa(u - u_i)} - u \right) \Bigg] \quad (11)$$

This can be written in the matrix form as

$$\frac{\nabla_x \hat{P}_a(x', u')}{\hat{P}_a(x', u')} = \Gamma \cdot \Sigma \cdot M(y) \qquad (12)$$

where

$$\Gamma = \left[\begin{array}{cc} 1 & 1 \\ 1 & 1 \end{array} \Big| \nabla_x u \right] \qquad (13)$$

and Σ is a $(d+2) \times (d+2)$ diagonal matrix with kernel

---

[1] Similar formulas can be derived with other kernels.

bandwidths in each spatial-feature dimensions

$$\Sigma = \left(\begin{array}{cccccc} \frac{1}{\sigma_1^2} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{\kappa_1^2} & \cdots & \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & & & \frac{1}{\kappa_d^2} \end{array} \right) \qquad (14)$$

and $M(y)$ is $(d+2) \times 1$ vector

$$M(y) = m(y) - y$$

$M(y)$ is the mean shift vector as defined in [9, 4, 6] and $m(y)$ is the $(2+d)$ vector representing the sample mean using kernel $K \cdot G$ [4]

$$m(y) = m([x \quad u]^t) = \left(\begin{array}{c} \frac{\sum_{i=1}^{N} x_i K_\sigma(x - x_i) G_\kappa(u - u_i)}{\sum_{j=1}^{N} K_\sigma(x - x_i) G_\kappa(u - u_i)} \\ \frac{\sum_{i=1}^{N} u_i K_\sigma(x - x_i) G_\kappa(u - u_i)}{\sum_{i=1}^{N} K_\sigma(x - x_i) G_\kappa(u - u_i)} \end{array} \right)$$
$$(15)$$

Mean shift is a steepest-ascent like procedure but with variable steps that leads to fast convergence. It was shown in [6] that if the density gradient is on the form of $m(y) - y$ where a convex kernel is used then this optimization procedure is guaranteed to converge to a stationary point satisfying $m(y) = y$ which is the density mode.

Using the density gradient from equation 12, the likelihood function gradient as defined by equation 9 will be,

$$\nabla_a L(a) = \sum_{(x', u') \in R'} \nabla_a T(x'; a) \cdot \Gamma \cdot \Sigma \cdot M(y) \qquad (16)$$

This means that the sum of the mean shift vectors over the region R' is an estimate of the gradient of the likelihood function $L(\cdot)$. This formula shows how we can shift the parameters a in the parameter space to maximizes the likelihood by linear transformation from the joint feature-spatial space. It can easily be verified that if a sequence of successive steps maximizing the density, it will also maximizes the log-likelihood of the density and therefore would lead to the local maxima of the likelihood function.

3.2 Entropy Maximization

As pointed out in section 2.2, the similarity between region R and R' defined by using the Kullback Leibler information distance depends on the entropy term H and a likelihood term L. For the likelihood term $L(\cdot)$ in the objective function, it is sufficient to build a representation of the joint density of the original region once and use this representation to evaluate the likelihood of any new region. Unfortunately, for the entropy term, the entropy of any hypothesis region R' depends on the spatial-feature joint distribution Q

of that region. Therefore it is required to compute a representation of such distribution for any new region hypothesis R'.

One way to estimate the entropy of a region is using empirical entropy estimation [19]. Two samples A and B are used, where A is used to estimate the density function and sample B is used to estimate the entropy using the density estimate obtained by A. Unfortunately, such an approach is computationally expensive if we consider that the entropy has to be computed for each new hypothesis region. Therefore, we use an approximation for the entropy as follows:

Let $H(X, U)$ be the entropy of the spatial-feature joint distribution. By the chain rule $H(X, U) = H(X) + H(U|X)$. For the case of dense features (intensity, color, etc.) we can assume that the samples from any region are taken uniformally over the spatial domain. $H(X)$ will then be the same for any region hypothesis. Therefore, only $H(U|X)$ is the relevant term in the maximization. To approximate $H(U|X)$, if we assume that the local distribution of the feature $P(U|x)$ at point x can be approximated as a Gaussian distribution with mean $\mu(x)$ and variance $\gamma(x)$, then the entropy will be

$$H(U|x) = \frac{1}{2}\log(2e\pi\gamma(x))$$

i.e, the local entropy depends on the local feature variance, which can be efficiently computed. Therefore the entropy of any region R can be computed as

$$
\begin{aligned}
H(U|X) &= \sum_{x \in R} \hat{P}(x) H(U|X = x) \\
&= \frac{1}{2|R|} \sum_{x \in R} \log(2e\pi\gamma(x))
\end{aligned}
$$

and the gradient of the entropy using variance gradient is

$$\nabla_x H(X, U) = \nabla_x H(U|X) = \frac{1}{2|R|} \sum_{x \in R} \frac{\nabla_x \gamma(x)}{2e\pi\gamma(x)}$$

where the variance gradient can be computed from the feature gradient over a neighborhood $\mathsf{N}(x)$ around x as

$$\nabla_x \gamma(x) = \frac{2}{|\mathsf{N}|} \left(\sum_{y \in \mathsf{N}(x)} u(y)\nabla_x u(y) - \mu(x) \sum_{y \in \mathsf{N}(x)} \nabla_x u(y) \right)$$

The feature variance and its gradient need to be computed only once per frame which can be implemented efficiently.

4 Tracking Example

In this section we present an example for tracking using the proposed framework to discuss some of the related issues. In this example we used grayscale image intensity as the feature. To study the tracker performance we generated
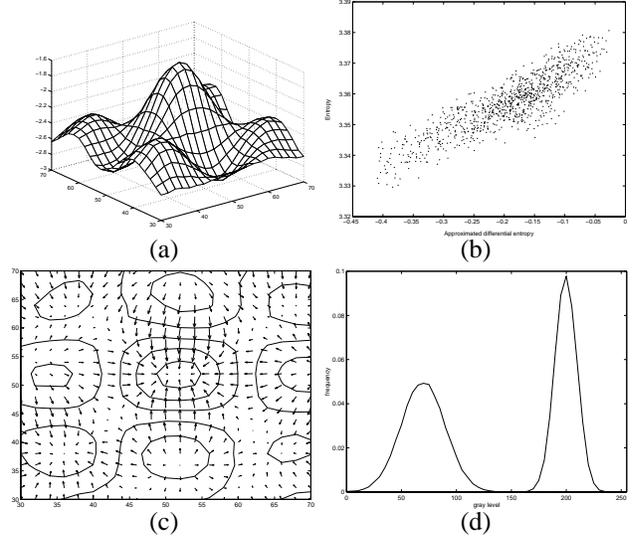

(a) (b)

(c) (d)

Figure 1. Synthetic example: (a) Objective function surface. (b) Scatter plot for the entropy and the approximated differential entropy. (c) Objective function gradient. (d) pdf for both the background and synthetic target.

a synthetic target that is similar in the feature distribution to the background. Figure 2 shows the tracked target which consists of two disks. All the pixels in the background as well as the target pixels are drawn from the same probability distribution which is shown in figure 1-d except that structure is imposed on the target. So, basically, trackers based on feature distribution only will fail to track such target. The intensity of each pixel in the target as well as the background is changing by randomly generating new intensities according to the mixture at each new frame. To emulate local deformation, the size of the internal disk is random at each frame which leads to large changes in the appearance of the object as can be seen from figure 2.

Figure 1-a shows the objective function surface which shows the local maxima corresponding to the target image location. Note that, because of the geometric symmetry of this particular synthetic target, there are four local maximums around the target which is not typically the case in real images. Figure 1-c shows the objective function gradient vectors. Figure 1-b shows the scatter plot between the target entropy and the approximated differential entropy as was presented in section 3.2. The scatter plot shows the correlation between the approximation using image variance and the actual entropy (which is calculated using 3D histogram in this case). Typically, the entropy approximation using image variance leads to a smooth entropy surface. Figure 2 shows some frames from the tracking results where the appearance of the target is changing as mentioned above. The error in the tracking was 1.667 pixels on aver-

age with spatial kernel bandwidth set to 2 and feature kernel bandwidth set to 1% of the feature spectrum.
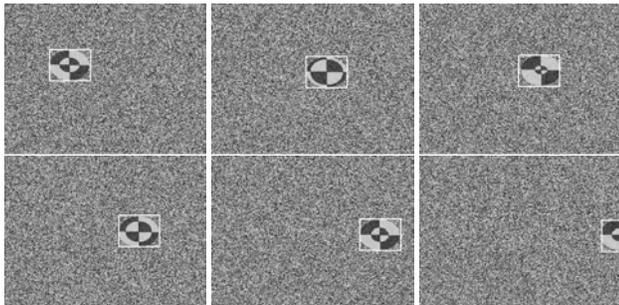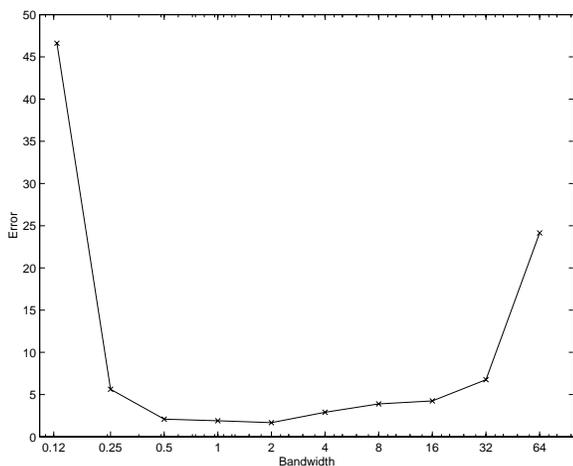

Figure 2. Tracking synthetic target

Figure 3. Effect of the spatial kernel bandwidth on the tracker. The tracking breaks as $\sigma \to 0$ (rigidity imposed) and as $\sigma \to \infty$ (structure ignored)

To study the effect of the spatial kernel bandwidth, σ, on the tracking, we used different values for the bandwidth to track the synthetic target shown in figure 2. Figure 3 illustrates the relation between the selected spatial kernel bandwidth and the error in tracking for the synthetic target. The figure shows two important issues. First the asymptotic behavior: As was discussed in section 2.3, as the bandwidth decreases and approaches zero, the framework converges to a rigid body tracker that minimizes the SSD in the feature space and since our target is not rigid the tracker will lose the target under this bandwidth. In this experiment, with $\sigma < 0.125$ the tracker started to lose the target. On the other hand, as we increase the σ the tracker will converge to a feature distribution tracker (histogram tracker) that ignores the region structure. Since both the target and the background have the same feature distribution in this example, the tracker loses the target as well. In this experiment, the tracker lost the target with $\sigma > 64$. As a conclu-

sion, both rigid body tracker and histogram tracker will lose such target. The second issue that the figure shows is the tracker insensitivity to the choice of an optimal bandwidth. In the experiment, the tracker performed very well as σ varied from 0.5 to 8 with error within 5 pixels. This is because kernel density estimation yields, generally, good estimates of the density even if we varied the bandwidth within reasonable range around the optimal value which is typically unknown.

5 Experimented Results

In this section we show some tracking results using different feature-spatial spaces.

Figure 4. Tracking in gray-scale from a moving camera

Figure 4 shows some tracking results using raw image intensity. In this case we used three dimensional feature-spatial space (2D location and 1D intensity). The target was tracked over a 500 frame span taken at about 10 frames per second (320x240 images) with kernel bandwidths set to (2,2,1%). The tracker failed when the target changes his appearance. To test the robustness of the tracker under low quality images, we compressed this video using MPEG compression with compression ratios of 52:1, 105:1, 178:1 and 198:1. The tracker performance (in terms of time to failure) was 91%, 85%, 83% and 79% respectively compared to the uncompressed video. In this experiment, as well as all other experiments presented here, only one frame samples were used to represent the joint space.

Figure 5. Tracking under occlusion

Figure 5 shows an example of tracking using color. In this case we used a four dimensional feature-spatial space (2D location and 2D chromaticity) with bandwidths (2,2,1% 1%) respectively. The sequence shows tracking where the target is partially occluded. We get similar tracking results when we used gray-scale as well as image gradient instead of color.

Figure 6,7 show tracking using edges represented as raw image gradient. In this case the feature-spatial space was four dimensional (2D image location and 2D image gradient) and therefore the invariant is the edge probabilistic structure of the region. In both cases, the joint space representation was obtained using only one frame with bandwidths = (2,2 10%, 10%) for each dimension respectively. Figure 6 shows tracking of a face using image gradient over about 200 frame span with 320x240 images. Some of the traced faces are also shown in the figure which shows that the tracked face undergoes different variations in the pose and the tracker was successful in locating the target at each new frame even under this changes. Figure 7 shows tracking of a human at a distance using the same space. Here, also, the edge probabilistic structure is the invariant represented as a four dimensional feature-spatial space as the previous example. The tracker also performs well in this case in spite of the deformation in the appearance due to the walking. Note that the camera was moving in this sequence.
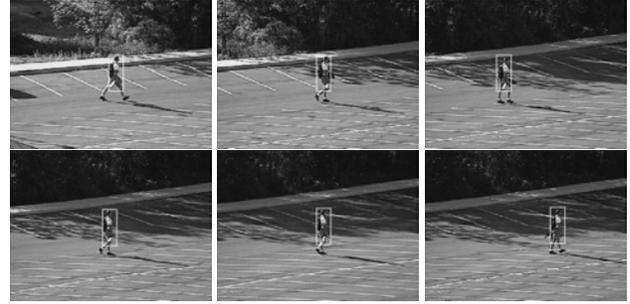

Figure 6. Tracking based on image gradient

Figure 7. Tracking based on image gradient - moving camera

6 Discussion

The framework presented in this paper is a general framework for estimating the geometric transformation that a region undergoes in a sequence of images by imposing probabilistic feature and spatial constraints. The geometric transformation represents a global constraint on the region motion while the feature values and feature locations represent local constraints on the region appearance and structure. We showed that the approach is general and its asymptotic behavior converges to either a rigid body tracker that preserves region structure. At the other end, it converges to feature distribution tracker that totally ignores the region structure. Therefore, the approach preserves the region structure constraint as it is tracked and, on the mean time, facilitates tracking under small local deformations. We presented some results of tracking using different features including image intensity, color and intensity gradient. The results shows that the approach is robust to partial occlusions and local deformations and small changes in target's pose. In general, it was found that the likelihood term in the objective function is the dominant term and the entropy variations is small with respect to the likelihood term.

One of the main advantages of the framework we present is that the objective function used is continuous and differentiable. The optimal target hypothesis represents a local maxima in the feature-spatial space as well as local maxima in the search space (by linear transformation). Therefore a gradient based optimization technique would converge to that solution as long as the initial guess is within a sufficiently small neighborhood of that maxima. We showed that maximizing the likelihood term, which is the dominant term, follows the mean shift procedure which guaranteed the conversion. The approach converges to the objective function maxima in few iteration (typically 2-5 iteration per frame.)

The framework presented in this paper can efficiently be implemented to run in real-time with low dimensional feature-spatial spaces. At each new frame, the computa-

tion involves evaluation of the likelihood and the entropy functions. The likelihood computation as well as it's gradient can be efficiently computed using lookup tables since most of the computations are repeated from frame to frame. Also, in case of Gaussian kernels, fast multipole methods can be used for efficient computation as was shown in [7]. The computation of the entropy term involves computing the image variance and it's gradient only once per frame.

This framework is an approach to localize the tracked region and can be used in conjunction with any tracker formalization. For example, a Kalman filter or a particle filter tracker can be used to predict a new location for the target at each new frame. This prediction can be used as an initial guess for the search for the optimal target hypothesis. This would provide measurements for the tracker at each new frame.

The framework presented in this paper can handle both dense features such as intensity, color, etc. and sparse features such as edges, corners, etc. The application of the framework presented in this paper was mainly focused on dense features which imposes connectivity constraint on the region. Applying the framework to sparse features and handling the issues related to this remain as part of our future work. Current paper does not address the topic of how to choose the optimal kernel bandwidth. As was shown in the examples, the tracker is insensitive to this choice within some range. This remains a topic that we will address in our future work.

Related work includes [6]. In their work, the color distribution of the target region was represented by a color histogram, i.e., no spatial information is preserved. The Bhattacharyya coefficient was used as a metric for similarity between the model and candidate region histogram, where each pixel is weighted using an external spatial kernel. The external kernel was necessary for the tracking. The mean shift algorithm was then used to obtain the location that maximizes the similarity metric. Our approach differs in three fundamental aspects: First, we preserve the spatial information by representing the joint feature-spatial distribution of the region. Second, our approach does not need an external spatial kernel imposed over the region to drive the tracker. The tracking in our case is driven by both the region structure as well as the region appearance. Third, our representation of the distribution is based on kernel density estimation which provides smooth estimates. Another related work include the work of [19, 15] for feature registration by maximizing mutual information. Related work also includes the work of [1], [3], and [5].

References

[1] B. Bascle and R. Deriche. Region tracking through image sequences. In *Proc. IEEE International Conferance on Computer Vision*, pages 302–307, 1995.

[2] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Jun 1998.

[3] G. Bradski. Computer vision face tracking as a component of a perceptual user interface. In *Proc. IEEE Workshop on Application of Computer Vision*, pages 214–219, 1998.

[4] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 17(8):790–799, Aug 1995.

[5] D. Comaniciu. Bayesian kernel tracking. In *Annual Conf. of the German Society for Pattern Recognition (DAGM'02), Zurich, Switzerland*, pages 438–445, 2002.

[6] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 142–149, Jun 2000.

[7] A. Elgammal, R. Duraiswami, and L. S. Davis. Efficient computation of kernel density estimation using fast gauss transform with applications for segmentation and tracking. In *Proc. of IEEE 2nd Int. workshop on Statistical and Computational Theories of Vision, Vancouver, CA, July 2001*, 2001.

[8] P. Fieguth and D. Terzopoulos. Color-based tracking of heads and other objects at video frame rates. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Jun 1997.

[9] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with application in pattern recognition. *IEEE Transaction on Information Theory*, 21:32–40, 1975.

[10] G. D. Hager and P. N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(10):1025–1039, 1998.

[11] B. Heisele, U. Kerssel, and W.Ritter. Tracking non-rigid, moving objects based on color cluster flow. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1997.

[12] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *ECCV (1)*, pages 343–356, 1996.

[13] D. G. Lowe. Robust model-based motion tracking through the integration of search and estimation. *International Journal of Computer Vision*, (2):113–122, 1992.

[14] J. Martin, V. Devin, and J. Crowley. Active hand tracking. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, 1998.

[15] C. F. Olson. Image registeration by aligning entropies. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume II, pages 331–336, 2001.

[16] Y. Raja, S. J. Mckenna, and S. Gong. Tracking colour objects using adaptive mixture models. *Image Vision Computing*, (17):225–231, 1999.

[17] D. W. Scott. *Mulivariate Density Estimation*. Wiley-Interscience, 1992.

[18] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *Int. J. of Comp. Vision*, 9(2):137–154, 1992.

[19] P. A. Viola. *Alignment by Maximization of Mutual Information*. PhD thesis, M.I.T., June 1995.