Object Tracking by Asymmetric Kernel Mean Shift with Automatic Scale and Orientation Selection

Alper Yilmaz
Ohio State University
Photogrammetric Computer Vision Laboratory
yilmaz.15@osu.edu

Abstract

Tracking objects using the mean shift method is performed by iteratively translating a kernel in the image space such that the past and current object observations are similar. Traditional mean shift method requires a symmetric kernel, such as a circle or an ellipse, and assumes constancy of the object scale and orientation during the course of tracking. In a tracking scenario, it is not uncommon to observe objects with complex shapes whose scale and orientation constantly change due to the camera and object motions. In this paper, we present an object tracking method based on the asymmetric kernel mean shift, in which the scale and orientation of the kernel adaptively change depending on the observations at each iteration. Proposed method extends the traditional mean shift tracking, which is performed in the image coordinates, by including the scale and orientation as additional dimensions and simultaneously estimates all the unknowns in a few number of mean shift iterations. The experimental results show that the proposed method is superior to the traditional mean shift tracking in the following aspects: 1) it provides consistent object tracking throughout the video; 2) it is not effected by the scale and orientation changes of the tracked objects; 3) it is less prone to the background clutter.

1. Introduction

In computer vision, tracking refers to the task of generating the trajectories of the moving objects by computing its motion in a sequence of images. Object motion recorded in the form of a trajectory commonly contains translational motion. Numerous approaches have been dedicated to computing the translation of an object in consecutive frames, among which the mean shift method is one of the most common methods which is also used in the commercial applications. The popularity of the mean shift method is due its ease of implementation, real time response and robust tracking performance.

Mean shift is a nonparametric density estimator which iteratively computes the nearest mode of a sample distribution. After its introduction in the literature [7], it has been adopted to solve various computer vision problems, such as line fitting [2], segmentation [5] and object tracking [6]. Despite its promising performance, as discussed in various papers [4], [3] and [12], the traditional mean shift method has two main limitations, the first of which is the constancy of the kernel bandwidth. The changes in the object scale requires an adjustment of the kernel bandwidth in order to consistently track the object. An intuitive approach to estimate the object scale is to search for the best scale by testing different kernel bandwidths and selecting the bandwidth which maximizes the appearance similarity [6]. Alternatively, after the object center is estimated, a mean shift procedure can compute the bandwidth of the kernel in the scale space, which is formed by convolving the image with a set of Gaussian kernels at various scales [3].

The second limitation of the traditional mean shift method is the use of radially symmetric kernels which are isotropic in shape. In the view of often anisotropic structure of the object, radially symmetric kernels inhibit robust image segmentation [12] and object tracking. For example, representing an elongated object with a circular kernel will bias the position estimation due to the non-object regions residing inside the kernel. This object can be better represented by an anisotropic symmetric kernel, such as an ellipse. The scale and orientation of a kernel representing an object can be computed by evaluating the second order moments from the object silhouette [1] or the posterior appearance probabilities [10]. Both of these methods, however, are computationally expensive compared to the mean shift tracking method. This observation resulted in the introduction of anisotropic symmetric kernels to the mean shift analysis. In particular, Wang et al. [12] has shown the improved performance of the mean shift segmentation when a circular kernel is replaced with an elliptical kernel. In their approach, in contrast to analyzing the local data distribution,

1-4244-1180-7/07/$25.00 ©2007 IEEE

the authors estimated the orientation and the scale of the elliptical kernel from images.

In this paper, we extend the traditional mean shift method by introducing the use of *asymmetric kernels* in the density estimation process. Asymmetric kernels have been used in the area of statistics for over a decade and have been shown to improve the density estimation [8]. This paper, however, does not use asymmetric kernels which have parametric profiles, due to their inability to represent the shape of the tracked objects. We introduce the level set kernel to nonparametric density estimation whose profile is also nonparametric. The second contribution of our paper is the online estimation of the kernel *scale* and *orientation* at each mean shift iteration by analyzing the local data statistics. In our approach, the kernel scale and orientation are introduced as additional dimensions in the estimation process. Proposed approach evaluates the mean shift vector simultaneously in all dimensions. This approach overcomes local solutions observed due to the estimation of the kernel translation, scale and orientation sequentially. In order to demonstrate the advantages of the proposed method, we chose the object tracking application, where the objects usually have asymmetric shapes.

We organized the paper as follows. The next section provides a brief overview of the traditional mean shift tracking for completeness. We introduce the asymmetric kernel mean shift, its motivation and formulation within the tracking framework in section 3. Section 4 provides insight to the automated scale and orientation estimation for both symmetric and asymmetric kernels. Experimental results using the proposed method are given in section 6. Finally, we discuss the future directions and conclude in section 7.

2. Mean Shift Tracking: An overview

The mean shift method iteratively computes the closest mode of a sample distribution starting from a hypothesized mode. In the context of tracking, a sample corresponds to the color observed at a pixel $\mathbf{x}$ and has an associated sample weight $w(\mathbf{x})$, which defines how likely the pixel color $I(\mathbf{x})$ belongs to an object model $\mathbf{q}$.

The traditional mean shift tracking method evaluates the displacement (translation) of the object centroid by computing the mean shift vector $\Delta\mathbf{x}$. Let the initial object position be $\hat{\mathbf{x}}$, and we want to compute the new object position $\hat{\mathbf{x}}'$, such that $\hat{\mathbf{x}}' = \hat{\mathbf{x}} + \Delta\mathbf{x}$. The mean shift vector is computed using the following:

$$\Delta\mathbf{x} = \frac{\sum_i K(\mathbf{x}_i - \hat{\mathbf{x}}) w(\mathbf{x}_i)(\mathbf{x}_i - \hat{\mathbf{x}})}{\sum_i K(\mathbf{x}_i - \hat{\mathbf{x}}) w(\mathbf{x}_i)}, \qquad (1)$$

where the denominator serves as a normalization term, and $K(.)$ is a radially symmetric kernel with bandwidth h defining the tracked object region. Given the color distribution functions q and p generated from the model and candidate object regions, the weight at pixel $\mathbf{x}$ is derived from the Bhattacharyya measure and is given by:

$$w(\mathbf{x}) = \sqrt{q(I(\mathbf{x}))/p(I(\mathbf{x}))}. \qquad (2)$$

3. Asymmetric Kernel Mean Shift

At each mean shift iteration, the samples that lie within the kernel effect the amount of kernel shift, hence, contribute equally to the selection of the mode. Although, this property is one of the strengths of the mean shift method [7], it also poses a limitation due to the method's sensitivity to the kernel selection [12]. Especially in the context of object tracking, where the kernel represents the object shape, this limitation becomes more evident. Take a look at Figure 1, where we show four different kernels defining the tracked object region. The traditional mean shift requires radially symmetric kernels in the shape of a circle given in part (a). These kernels have promising performance for object segmentation, however, their use dramatically effects the performance of object tracking. The main reason for the performance loss stems from considering the non-object regions residing inside the kernel as a part of the object. The use of anisotropic symmetric kernels, such as the ones shown in parts (b) and (c), has shown to improve segmentation [12], and object tracking [3].

The main advantage of anisotropic symmetric kernels is that most of the non-object region resides outside of the kernel. These kernels, however, do not represent the object shape and still contain non-object regions as part of the object. An ideal kernel has the shape of the tracked object which may be asymmetric as shown in Figure 1 d. Asymmetric kernels, however, may not have an analytical form. This observation suggests the use of implicit functions for defining the profile of arbitrarily shaped kernels. Implicit functions are commonly employed in fluid dynamics, where an evolving fluid is modeled by either the volume fluid representation or the level set representation. Between the two approaches, level set is more popular among vision researchers for its simplicity and robustness and is often used for contour based region segmentation or object tracking [13]. In this paper, we adopt a modified form of the implicit level set function as the asymmetric kernel required for the mean shift method.

Implicit level set function $\phi(\mathbf{x})$ encodes the signed distances of the pixels $\mathbf{x}$ (inside positive, and outside negative) from the object boundary, such that the object boundary resides on the zero level set. This property provides a smooth and differentiable function [9], and satisfies the regularity requirement of a kernel [11, p. 95]. The level set function, however, does not provide a compact support which is a required in order to be used as a density estimator in the mean shift framework. Hence, we introduce an indicator function

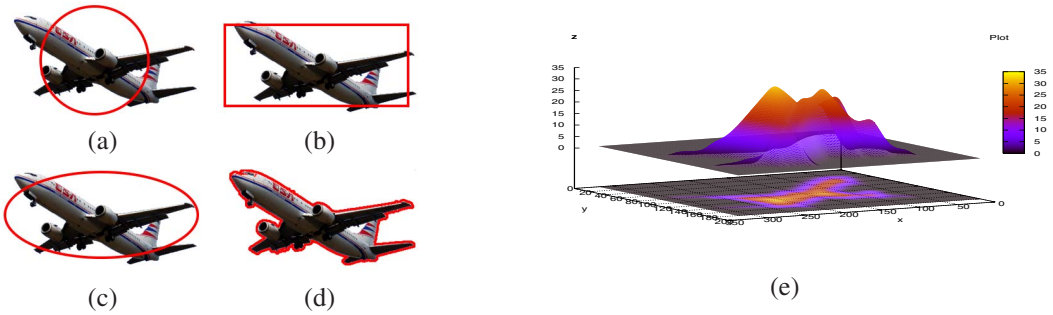


Figure 1. (a) Radially symmetric object kernel. (b, c) Symmetric object kernels which are anisotropic in shape. These kernels are commonly used for object tracking. (d) Asymmetric object kernel tightly fitting the object shape. (e) Non parametric level set kernel associated with the shape given in (d); the bottom view shows the projection of the kernel weights into the image coordinates.

$I(\mathbf{x})$ to bound the level set by the zero level set, such that outside of the object boundary is set to 0. In this formalism, the level set kernel is computed by:

$$\hat{\phi}(\mathbf{x}) = C\phi(\mathbf{x})I(\mathbf{x}),$$

where C is the normalization term: $\frac{1}{\sum \phi(\mathbf{x})I(\mathbf{x})}$. Using the level set kernel, the density estimator in the mean shift framework becomes:

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\phi}_{\mathbf{x}}(x_i). \tag{3}$$

The level set kernel generated from this equation for the object shown in Figure 1 d is given in part (e).

4. Automated Scale and Orientation Selection

A radially symmetric kernel is an unbiased density estimator compared to an anisotropic symmetric kernel or an asymmetric kernel. For instance, an elliptical kernel has more data along its major axis compared to its minor axis, which is not the case for a circular kernel. If used appropriately, such as by controlling the orientation of the kernel, the estimation bias can be controlled by the underlying distribution and produce a more accurate shift of the mean resulting in better mode estimation.

An intuitive approach to search for the kernel scale and orientation is to evaluate the Bhattacharyya distance over a set of predefined scales and orientations. A brute-force search of predefined scales and orientations, however, is not practical. In addition, a brute force search requires that the object translation has already been computed. Performing the search independent of the computing the object translation −mean shift in image coordinates−, however, may result in loss of the tracked object. Alternatively, the mean shift iterations can be carried out simultaneously for the object scale and orientation as well as the object translation. This approach requires the mean shift iterations to be performed in a higher dimensional space which corresponds to scale, orientation, apses and ordinate.

In the following discussion, we provide details of the proposed scale and orientation selection method which is applicable to both the asymmetric and anisotropic kernels. In order to be consistent with the experiments given later in the text, we will structure the discussion around object tracking.

4.1. Scale Dimension

Before providing the details, we will first define some of the variables used throughout the discussion. Let the spatial object center be denoted by $\mathbf{o} = (o_x, o_y)$ where x and y are the image coordinates. The scale and orientation dimensions are defined by σ and θ respectively. The kernel bandwidth at an angle $\alpha \in [0, 2\pi)$ is denoted by $r(\alpha)$. The distance of a pixel $\mathbf{x}_i = (x_i, y_i)$ inside the object from $\mathbf{o}$ is referred to as $\delta(\mathbf{x}_i)$.

The scale of an object pixel is represented by a linear transformation of the image coordinates to the scale dimension. This transformation is computed by the ratio between $\delta(\mathbf{x}_i)$ of point $\mathbf{x}_i$ and the bandwidth observed at angle θ_i:

$$\sigma_i = \frac{\delta(\mathbf{x}_i)}{r(\theta_i)} = \frac{||\mathbf{x}_i - \mathbf{o}||}{r(\theta_i)}. \tag{4}$$

Considering that the object is represented by a closed curve, the scale dimension becomes an irregular grid in contrast to the regular image coordinates. This can be exemplified by the variation in the number of observations with an increase in the scale; for instance, at $\sigma = 0.1$ there are less number of pixels on the circumference of a circle than there are at $\sigma = 0.7$. This property also results in an important observation regarding the constancy of the sample mean, $\hat{\sigma}$, in the scale dimension. The scale-mean satisfies the equilibrium of the sum of the pixel-scales on both sides of the scale-mean:

$$\int_0^{2\pi}\int_0^{\hat{\sigma}r(\alpha)} \frac{\delta}{r(\alpha)} d\delta d\alpha = \int_0^{2\pi}\int_{\hat{\sigma}r(\alpha)}^{r(\alpha)} \frac{\delta}{r(\alpha)} d\delta d\alpha. \tag{5}$$

Solving this equation for the unknown $\hat{\sigma}$ provides us with $2\hat{\sigma}^2 - 1 = 0$, and the solution of the scale-mean is given

by $\hat{\sigma} = \sqrt{2}^{-1}$. The scale-mean for any type of object represented with a closed curve is not dependent on the shape of the object and is constant throughout the sequence. This property differentiates the scale dimension from the image coordinates when conducting the mean shift iterations. An update in the scale dimension should take relation between the bandwidth and the scale into consideration. Let the mean shift iteration provide a scale update of $\Delta\sigma$. Substituting the new scale, $\hat{\sigma} + \Delta\sigma$ in equation (5) and replacing the upper bound of the bandwidth to its prospective value $kr(\alpha)$ will result in the bandwidth update factor k given by $k = 1 + \sqrt{2}\Delta\sigma$.

4.2. Orientation Dimension

There is an analogy between the scale and orientation dimensions in terms of their differences from the image coordinates. Similar to the scale dimension the number of pixels at each angle is different, however, the number of observations does not increase with an increase in the angle. To generate samples in the orientation space, the object pixels $\mathbf{x}_i = (x_i, y_i)$ are nonlinearly transformed to the orientation space by computing:

$$\theta_i = \arctan \frac{y_i - o_y}{x_i - o_x}, \tag{6}$$

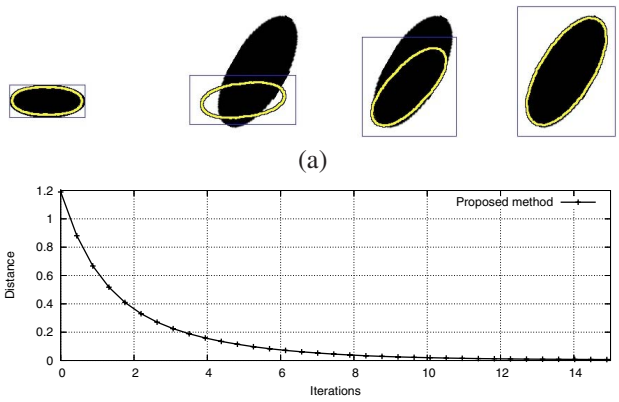
where $\mathbf{o} = (o_x, o_y)$ is the object centroid. An important observation arising from equation (6) is that the orientation is not directly dependent on the object shape which is expressed by the bandwidth $r(\theta_i)$. We should, however, note that there is an indirect dependence between the shape and orientation, which can be exemplified by the difference in the number of observations between the minor and major axis of an elliptical kernel. This dependence results in computing the orientation-mean $\hat{\theta}$ of the object region:

$$\int_0^{\hat{\theta}} \int_0^{r(\alpha)} \alpha d\delta d\alpha = \int_{\hat{\theta}}^{2\pi} \int_0^{r(\alpha)} \alpha d\delta d\alpha. \tag{7}$$

solving for the inner integral and rearranging terms, we have:

$$2\int_0^{\hat{\theta}} \alpha r(\alpha) d\alpha = \int_0^{2\pi} \alpha r(\alpha) d\alpha. \tag{8}$$

In contrast to the scale-mean, the orientation-mean is not constant and requires an update with every new frame. This update has the form of the mean shift iterations performed the image coordinates to find the object centroid: $\hat{\theta}_{new} = \hat{\theta}_{old} + \Delta\theta$. An exception to the orientation update is observed when the kernel is an isotropic kernel, where $r(\alpha) = r$. In this case, similar to the scale-mean, orientation-mean becomes a constant, $\hat{\theta} = \sqrt{2}\pi$, such that $\Delta\theta = 0$.



(a)

(b)

Figure 2. Mean shift iterations in the joint space. Yellow contour represents the tracked object. (a) Initial orientation and intermediate mean shift iterations. (b) Plot showing the distance between the model histogram and the candidate histogram computed using L^1 norm as a function of the iteration number.

5. Object Tracking in (σ, θ, x, y) Space

Estimating the object position in the joint space refers to finding the spatial position as well as the scale and orientation of the object. To reduce the estimation bias of one dimension over the other dimension, the mean shift iterations are required to be conducted simultaneously on all dimensions. Let $\Gamma = (\sigma, \theta, x, y)$ denote the joint space. The density estimator in Γ is given by

$$\hat{f}(\Gamma) = \frac{1}{n}\sum_{i=1}^{n} K(\Gamma - \Gamma_i) \tag{9}$$

Since the orientation and scale of the object is independent of its centroid, the 4D kernel $K(.)$ can be broken into a product of three different kernels:

$$K(x, y, \sigma, \theta) = K(x, y)K_E(\sigma)K_E(\theta), \tag{10}$$

where $K_E(.)$ is a one-dimensional Epanechnikov kernel with a profile $k_E(z) = 1 - |z|$ if $|z| < 1$ and 0 otherwise, and $K(x, y)$ is the spatial kernel with the profile as discussed in Section 3. Using the estimator given in equation (10), the mean shift vector which maximizes the density iteratively is computed by:

$$\Delta\Gamma = \frac{\sum_i K(\Gamma_i - \hat{\Gamma})w(\mathbf{x}_i)(\Gamma_i - \hat{\Gamma})}{\sum_i K(\Gamma_i - \hat{\Gamma})w(\mathbf{x}_i)}, \tag{11}$$

where $\Delta\Gamma = (\Delta x, \Delta y, \Delta\sigma, \Delta\theta)$ denotes the update of the mode. The mean shift in the joint space always points in the direction of the true density mode due to the monotonic decreasing profiles of the kernels in each dimension; hence, the mean shift procedure convergences to a local solution in the joint space.

In contrast to estimating the object translation in the image coordinates, simultaneous estimation in the joint space requires a slightly higher number of iterations. This observation is due to increase in the number of parameters to estimate. Nevertheless, the tracking performance is real time. Empirically, the number iterations ranges from three to nine depending on the amount of motion and the distinctiveness of the object appearance from its background. In Figure 2 (a) we show a synthetic sequence where the object undergoes a scaling, translation and rotation from one frame to the next. As shown in part (b) of the figure, the Bhattacharyya distance decreases exponentially with each iteration.

6. Experiments

Given the object outline in the first frame, proposed tracking algorithm is initialized by computing the object centroid and orientation. The scale-mean is not computed as it is independent of the shape and is constant. In order to represent the asymmetric shape of an object, we create an r-table which encodes the bandwidth observed at each angle. The r-table is utilized for generating the modified level set kernel for conducting the mean shift iterations. We should emphasize that the level set kernel is computed only for once and used throughout the sequence. The appearance of the object is modeled by the kernel density estimated (KDE) in the RGB space. For practical purposes, we have used the weighted histogram approximation of the KDE using the Epanechnikov kernel with 8 bins along each dimension. The bandwidth of the Epanechnikov kernels used in the scale and the orientation dimensions are set to $h_\sigma = .4$ and $h_\theta = 10^\circ$ respectively. In addition, we also generate the KDE of the immediate outside region of the tracked object to suppress the bins of the model histogram which also appear in the background histogram. This step improves the tracking considerably due to reducing the ambiguity between the background and object appearances. Once the initialization is achieved, tracking is performed iteratively computing the mean shift vector in the search space following the steps given in Figure 3.

In order to evaluate the performance of the proposed method, we present two sequences in which the objects undergo scaling, translation and in-plane rotations. The video sequences of the results given in this section are provided as a media attachment to this submission. The proposed approach successfully computes the translation, orientation and scale on the average of 40fps where the estimation in every frame took on the average five iterations. For the same sequences, the traditional mean shift tracker took longer time due to fluctuations of the object center around multiple modes in the weight surface.

The first sequence, which is given in Figure 4 a, includes a phone handle. The outline of the handle is marked by

Tracking()

1 Compute $\mathbf{o}_0$, θ_0 (equation 8)
2 Generate object model $\mathbf{q}$
3 Generate r-table and level set kernel $\phi(\mathbf{x})$ in spatial dimensions
4 for (all frames)
5 Loop until convergence
6 Generate candidate prior $\mathbf{p}$
7 Perform mean shift iterations in Γ (equation (11)
8 Update object centroid: $\mathbf{o}_{new} = \mathbf{o}_{old} + \Delta\mathbf{o}$
9 Update object orientation: $\theta_{new} = \theta_{old} + \Delta\theta$
10 Update object scale: $rTable_{new} = rTable_{old}(1 + \sqrt{2}\Delta\sigma)$
11 endLoop
12 endFor

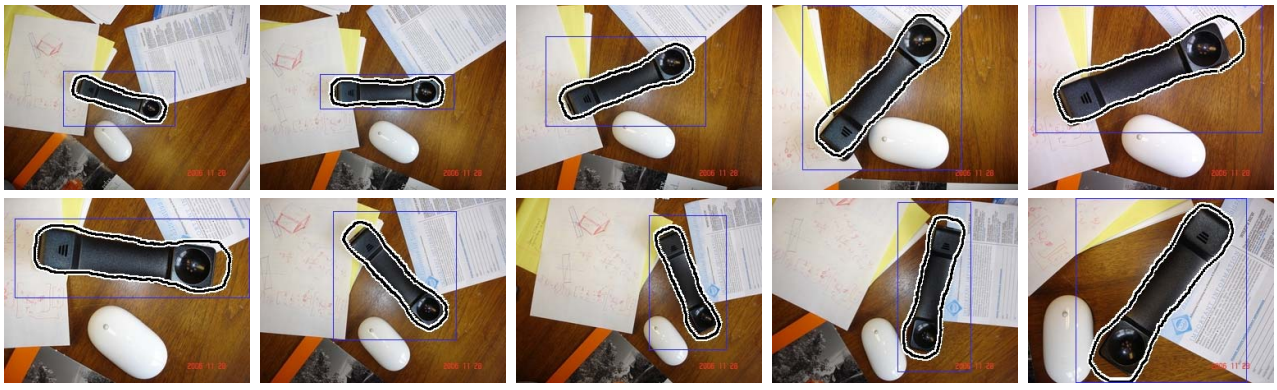
Figure 3. Mean shift algorithm for tracking objects by automated scale and orientation selection.

hand in the smallest object scale which provides a rough estimate of the object shape. This initialization results in a slightly imperfect tracking for higher object scales as shown in the second row of part (a). The second sequence is a zooming-in sequence of a cruiseliner. Towards the end of the sequence, out-of-plane rotation of the object due to the camera motion results in the tracked shape not to perfectly fit on the ship outline, especially around the foremast of the ship. This effect will occur regardless of the initial shape due to the change in the object view. Regardless, the tracked outline provides the minimum Bhattacharyya distance in the mean shift tracking space.

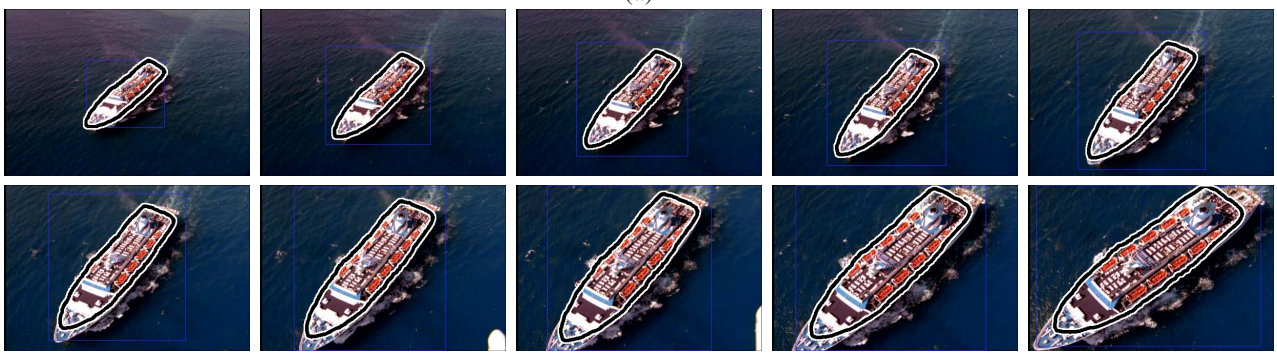
7. Conclusions and Future Work

This paper introduces the use asymmetric kernels to the mean shift tracking framework. The complex object shapes which may not have an analytical form are implicitly represented by a modified level set formalism. Introducing the asymmetric kernels reduce the estimation bias due to a better representation of the underlying density. Another advantage of the asymmetric kernel is that it is a generalization of the previous kernels, such that both radially symmetric and anisotropic kernels are forms of an asymmetric kernel. Tracking is achieved by automatically estimating the kernel scale and orientation in addition to its translation in the image coordinates. Proposed method's simultaneous estimation of all parameters using the kernel mean shift contrasts with most kernel tracking methods which sequentially estimates the parameters (first, the translation parameters, then the scale parameter without the orientation parameter).

Despite the advantages of the proposed tracking method, our method suffers from the constancy of the kernel shape which is also a drawback for all other kernel trackers. In computer vision literature, this problem has been addressed by contour trackers which are capable of tracking the complete object boundary. We are working on techniques to merge the kernel trackers, especially the mean shift tracker, with contour trackers which would overcome the computational complexity problem of the contour trackers. This proposal would require a better contour handling technique



(a)



(b)

Figure 4. Tracking result using the proposed tracking method. For video files, please look at the attached media file.

compared to the currently used r-table representation.

References

[1] G. Bradski. Computer vision face tracking for use in a perceptual user interface. In *IEEE Workshop on Applications of Computer Vision*, 1998.

[2] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17:790–799, 1995.

[3] R. Collins. Mean-shift blob tracking through scale space. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

[4] D. Comaniciu. An algorithm for data-driven bandwidth selection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(2):281–288, 2003.

[5] D. Comaniciu and P. Meer. Mean shift analysis and applications. In *IEEE Int. Conf. on Computer Vision*, volume 2, pages 1197–1203, 1999.

[6] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25:564–575, 2003.

[7] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE IT*, 21(1):32–40, 1975.

[8] P. Michels. Asymmetric kernel functions in non-parametric regression analysis and prediction. *Statistician*, 41(4):439–454, 1992.

[9] J. Sethian. *Level set methods: evolving interfaces in geometry,fluid mechanics computer vision and material sciences*. Cambridge University Press, 1999.

[10] H. Tao, H. Sawhney, and R. Kumar. Object tracking with bayesian estimation of dynamic layer representations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(1):75–89, 2002.

[11] M. Wand and M. Jones. *Kernel smoothing*. Chapman and Hall, 1995.

[12] J. Wang, B. Thiesson, Y. Xu, and M. Cohen. Image and video segmentation by anisotropic mean shift. In *European Conf. on Computer Vision*, 2004.

[13] A. Yilmaz, X. Li, and M. Shah. Contour based object tracking with occlusion handling in video acquired using mobile cameras. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(11):1531–1536, 2004.