Improved Kernel-Based Object Tracking Under Occluded Scenarios

Vinay P. Namboodiri, Amit Ghorawat, and Subhasis Chaudhuri

Department of Electrical Engineering
Indian Institute of Technology, Bombay
Mumbai, India
{vinaypn, sc}@ee.iitb.ac.in, amit.ghorawat@gmail.com

Abstract. A successful approach for object tracking has been kernel based object tracking [1] by Comaniciu *et al.*. The method provides an effective solution to the problems of representation and localization in tracking. The method involves representation of an object by a feature histogram with an isotropic kernel and performing a gradient based mean shift optimization for localizing the kernel. Though robust, this technique fails under cases of occlusion. We improve the kernel based object tracking by performing the localization using a *generalized (bidirectional) mean shift* based optimization. This makes the method resilient to occlusions. Another aspect related to the localization step is handling of scale changes by varying the bandwidth of the kernel. Here, we suggest a technique based on SIFT features [2] by Lowe to enable change of bandwidth of the kernel even in the presence of occlusion. We demonstrate the effectiveness of the techniques proposed through extensive experimentation on a number of challenging data sets.

1 Introduction

Real-time object tracking is indispensable to a vast number of computer vision applications like video surveillance and security, driver assistance, video abstraction, traffic management and video editing. Segmenting and tracking objects accurately with low computational complexity is a challenge.

A method which has been quite successful in handling this task is the kernel based object tracking algorithm [1]. In this method the target is spatially masked with an isotropic kernel. A spatially-smooth similarity function is defined and the target localization problem is then done by a gradient based optimization method, based on mean shift filter [3]. This method has been demonstrated to successfully work for non-rigid motion and in the presence of significant clutter. While in some cases it does handle partial occlusion, it unfortunately fails in a large number of cases. The reason for this can be traced to the mean shift based approach used for target localization. It is an effective method for clustering when the modes are distinct. However if there are multiple modes which are nearby then the gradient based optimization step can often converge to a local mode which is not necessarily the "true" mode. In this paper we address this issue effectively by considering a *generalized mean shift* based approach. This

P. Kalra and S. Peleg (Eds.): ICVGIP 2006, LNCS 4338, pp. 504–515, 2006.

method effectively handles the problem of partial occlusion and in some cases total occlusion in a more robust manner. A recent work which addresses the same problem is by Babu *et al.* [4] in which they consider the problem of improving the kernel object tracker. However, they address this problem by considering multiple tracking systems, that is they combine the mean shift filter with an SSD based tracking system. This affects the real time performance of the system and besides it does not actually address the core issue of the mean shift procedure which we have considered. Another approach [5], has been based on combination of particle filtering with blob tracking and is very successful in handling the occlusion problem, however, the method is computationally expensive.

The other aspect which is of interest has been that of adapting the bandwidth of the kernel to account for a change in scale of the object of interest. There have been a few approaches for data driven bandwidth selection [6,7] and a scale space based approach [8] for the mean shift procedure to account for the scale as well. However, while these approaches work well to account for a scale change when there is no occlusion, they fail when the scale changes with partial occlusion. To handle this aspect we consider a approach where we compute the SIFT [2] based features and compute the matches of key-points over the frames. Using this technique we are able to handle scale change even in the presence of occlusion.

In the next section we discuss the original kernel object tracker. In section 3 we discuss the procedure of generalized mean shift. Next, in section 4 we formulate a tracker based on generalized mean shift. The technique for scale change is presented in section 5. The experimental results are presented in section 6 and we conclude in section. 7.

2 Kernel-Based Object Tracking

The main contribution of the kernel based object tracking algorithm [1] has been in the target representation and localization aspects of tracking. The other aspects of tracking like initial object segmentation can be addressed using methods like background subtraction. Further, to make it more robust it can be associated with a prediction filter like Kalman filter. The target representation and localization is a bottom up process and has to handle changes in the appearance of the object. We now briefly discuss these aspects of the object tracker.

2.1 Target Representation

The reference *target model* is represented by its probability distribution function (p.d.f.) q in the feature space. Here the p.d.f.s are represented using m-bin histograms due to the low computational cost involved and the real-time processing restrictions. A target is represented by an ellipsoidal region in the image. Let $x_i^*, i = 1 \dots n$ be the normalized pixel locations in the region defined as the target model. The region is locally centered at 0. An isotropic kernel with a convex and monotonic decreasing kernel profile $k(x)$, assigns smaller weights to pixels farther from the center. The function b associates to the pixel at location x_i^*

the index $b(x_i^*)$ of its bin in the quantized feature space. The probability of the feature $u = 1 \ldots m$ in the target model is then computed as

$$\hat{q}_u = C \sum_{i=1}^{n} k(||x_i^*||^2)\delta[b(x_i^*) - u] \tag{1}$$

where δ is the Kronecker delta function and C is the normalization constant and is given by

$$C = \frac{1}{\sum_{i=1}^{n} k(||x_i^*||^2)}. \tag{2}$$

The target model can be considered as centered at the spatial location 0. In the subsequent frame, a *target candidate* is defined at location y and is characterized by the pdf $p(y)$. Let $x_i, i = 1 \ldots n_h$ be the normalized pixel locations of the target candidate, centered at y in the current frame. Using the same kernel profile $k(x)$, but with bandwidth h, the probability of the feature $u = 1 \ldots m$ in the target candidate is given by

$$\hat{p}_u(y) = C_h \sum_{i=1}^{n_h} k(||\frac{y - x_i}{h}||^2)\delta[b(x_i) - u], \tag{3}$$

where

$$C_h = \frac{1}{\sum_{i=1}^{n_h} k(||\frac{y-x_i}{h}||^2)}. \tag{4}$$

A similarity function is defined that defines the distance among target model and candidates as

$$d(y) = \sqrt{1 - \rho|\hat{p}(y), \hat{q}|}, \tag{5}$$

where

$$\hat{\rho}(y) = \rho|\hat{p}(y), \hat{q}| = \sum_{u=1}^{m} \sqrt{\hat{p}_u(y)\hat{q}_u}, \tag{6}$$

is the sample estimate of the Bhattacharyya coefficient between p and q.

2.2 Localization

In the localization phase the distance measure between the target model and target candidates is minimized. Minimizing the distance given in eqn.(5) is equivalent to maximizing the Bhattacharyya coefficient $\hat{\rho}(y)$. The search for the new target location in the current frame starts at the location $\hat{y}_0$ of the target in the previous frame. The linear approximation of the Bhattacharyya coefficient in eqn.(6) is

$$\hat{\rho}(y) \approx \frac{1}{2} \sum_{u=1}^{m} \sqrt{\hat{p}_u(\hat{y}_0)\hat{q}_u} + \frac{1}{2} \sum_{u=1}^{m} \hat{p}_u(y) \sqrt{\frac{\hat{q}_u}{\hat{p}_u(\hat{y}}} \tag{7}$$

The resultant expression considering eqn.(3) is

$$\hat{\rho}(y) \approx \frac{1}{2} \sum_{u=1}^{m} \sqrt{\hat{p}_u(\hat{y}_0)\hat{q}_u} + \frac{C_h}{2} \sum_{u=1}^{n_h} w_i k(||\frac{y - x_i}{h}||^2), \tag{8}$$

where

$$w_i = \sum u = 1^m \sqrt{\frac{\hat{q}_u}{\hat{p}_u(\hat{y}_0)}}\delta[b(x_i) - u]. \tag{9}$$

To minimize the distance, the second term in eqn.(8) has to be maximized. The second term represents the density estimate with kernel profile $k(x)$ at y in the current frame. The mode of this density in the neighborhood is the sought maximum that can be found employing the mean shift procedure. In this procedure, the kernel is recursively moved from the current location $\hat{y}_0$ to the new location $\hat{y}_1$ according to the mean shift procedure with the relation being

$$\hat{\mathbf{y}}_1 = \frac{\sum_{i=1}^{n_h} \mathbf{x}_i w_i g(||\frac{\hat{y}_0 - x_i}{h}||^2)}{\sum_{i=1}^{n_k} w_i g(||\frac{\hat{y}_0 .- x_i}{h}||^2)} \tag{10}$$

where $g(x) = -k'(x)$. In the next section we discuss the generalized mean shift procedure which can be used to find the modes more robustly.

3 Generalized Mean Shift

The mean shift procedure ([3,9]) when applied on a set of points explicitly moves the points towards their modes. The mean shift procedure has been extended in [10] to perform reverse mean shift which moves the points away from their modes. The generalized mean shift procedure combines forward and reverse mean shift methods so as to move the points to their correct modes without getting stuck in the local mode.

3.1 Generalized Mean Shift

The forward and reverse mean shift based methods move the points towards and away from the mode of the cluster respectively. However, when there are multiple modes close by it is possible that the point may be clustered to an incorrect mode away from its "true" mode. In order to handle this case, we formulate the notion of generalized mean shift where the points are perturbed away from their mode by the reverse mean shift and then clustered again using forward mean shift. This is not a purely convex optimization and hence it is able to move away from local minima and converge to the global minima provided that the global minima is near. The threshold for the global minima being nearer is decided by a dissimilarity factor and is discussed in section 6. The process of generalized mean shift is illustrated in fig. 1. It involves combining the forward and reverse mean shift procedures in an iterative manner with the switching between forward and reverse mean shift being decided using an automatic switching criterion. The reverse mean shift is a divergent procedure and tends to move the cluster values away from their mode in the direction of the gradient and the forward mean shift is a convergent procedure. Hence, in case of mixed clusters (that correspond to occluded scenarios), the generalized mean shift mixes the forward and reverse mean shift, ensuring that it is a convergent procedure, by switching the direction based on the dissimilarity factor.

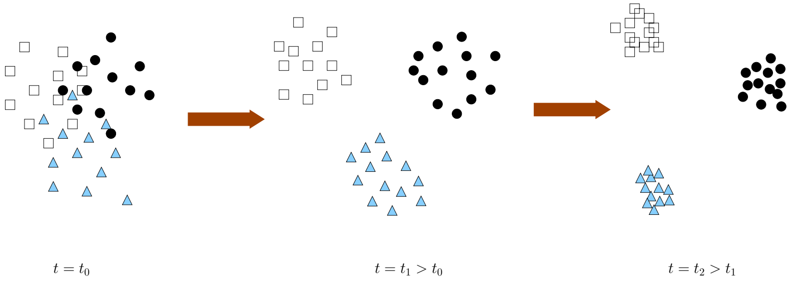


Fig. 1. Illustration of mixed diffusion in the feature space. The inverse diffusion results in the mixed clusters being separated and the individual elements of clusters coming closer together due to forward diffusion.

4 Tracking Using Generalized Mean Shift

The application of generalized mean shift optimization for tracking becomes relevant in the case of partial or total occlusion of tracked objects. In this case the modes of the histogram are affected and the kernel tends to be attached to the false mode, i.e. the occluding object. By using adaptive forward and reverse mean shift, i.e. the generalized mean shift, one can recover the true mode even after partial or total occlusion. The generalized mean shift is then given by:

$$\hat{\mathbf{y}}_1 = \mathrm{sgn(y)} \frac{\sum_{i=1}^{n_h} \mathbf{x}_i w_i g(|| \frac{\hat{y}_0 - x_i}{h} ||^2)}{\sum_{i=1}^{n_k} w_i g(|| \frac{\hat{y}_0 . - x_i}{h} ||^2)} \tag{11}$$

where sgn(y) is a sign function and is determined by a dissimilarity factor threshold θ.

$$\begin{aligned} d(y) > \theta &\Rightarrow \mathrm{sgn(y)} = -\gamma \\ d(y) <= \theta &\Rightarrow \mathrm{sgn(y)} = +1. \end{aligned} \tag{12}$$

Here γ is the reverse mean shift coefficient such that $0 < \gamma \ < 1$. The value of γ is generally less than 1 since the reverse mean shift procedure is divergent and hence it is required to dampen the divergent procedure. The value of θ is determined based on the distance measure between the target model and the candidate model and is fixed for a wide class of tracking scenarios. However, since the reverse mean shift is a divergent procedure, if the distance function during reverse mean shift increases beyond the value θ, then the sgn function is again made positive and the forward mean shift procedure is used. Thereby, one ensures that the generalized mean shift procedure is always convergent. The algorithm for the generalized mean shift procedure for Bhattacharyya coefficient is given in Algorithm 1.

Algorithm 1. Bhattacharyya Coefficient Maximization using **Generalized mean shift**

```
Input: The target model $\{\hat{q}_u\}_{u=1\ldots m}$ and its location $\hat{y_0}$ in the previous frame
 1: Initialize the location of the target in the current frame with $\hat{y_0}$, compute
    $\{\hat{p_u}(\hat{y_0})\}_{u=1\ldots m}$, and evaluate
        $$\rho[\hat{p}(y_0), \hat{q}] = \sum_{u=1}^{m} \sqrt{\hat{p_u}(\hat{y_0})\hat{q_u}}.$$
 2: Derive the weights $\{w_i\}_{1=1\ldots n_h}$.
 3: if $d[\hat{p}(y_0), \hat{q}] > \theta$ then
 4:     $\mathrm{sgn(y)} = -\gamma, 0 < r < 1$ and $a = 0$.
 5: end if
 6: Find the next location of the target candidate according to eqn.(11).
 7: Compute $d(\hat{y}_1)$
 8: if $d(\hat{y}_1) < d(\hat{y}_0)$ then
 9:     $\hat{y_1} \leftarrow \frac{1}{2}(\hat{y_0} + \hat{y_1})$
10:     Evaluate $\rho[\hat{p}(\hat{y_1}), \hat{q}]$
11: else
12:     reinitialize sgn(y) = 1 and go to Step 6.
13: end if
14: if $||\hat{y_1} - \hat{y_0}|| < \epsilon$ then
15:     Stop.
16: else
17:     Set $\hat{y_0} \leftarrow \hat{y_1}$
18:     go to Step 2.
19: end if
```

5 Scale Adaptation

While there have been works related to adapting the kernel bandwidth h based on the scale [6,7,8], the methods assume that there will be no occlusion during scale change or relatively no occlusion. In order to consider real world scenarios where there may be scale change while there is occlusion we consider a different approach based on Scale Invariant Feature Transform (SIFT) based features [2] proposed by Lowe.

5.1 SIFT Features

The SIFT features [2] are highly robust and are invariant to image scale and rotation and provide robust matching across a substantial range of affine distortion, change in 3D viewpoint and change in illumination which are pervasive in tracking. The scale of the key-points are computed by a search over all scales and image locations using a difference of Gaussian function and the interest points selected are invariant to scale and orientation. At each key-point location a detailed model is fit to determine the location and scale and it is ensured that the key-points selected are stable.

5.2 Scale Adaptation Using SIFT

Given a kernel at the current location and a kernel from an earlier location, the key-points are selected using the SIFT operator and key-point matches are calculated between the key-points in the kernels from the selected frames. Then the average change in the matched key-points is calculated. The scale of the kernel, i.e. the bandwidth factor h is then resized using the change in the scale as indicated by the matched key-points. Let S_m be the average scale of the matched key-points in the target model and S_c be the average scale of the matched key-points in the target candidate. Then we obtain the new value for the bandwidth parameter h as

$$\hat{h} = h * \frac{S_c + \alpha S_m}{(1+\alpha) S_m} \tag{13}$$

where α is a weight factor which denotes the weight given to the scale of the target model key-points as compared to the scale of the target candidate key-points. We have used a value of $\alpha = 4$ in our experiments. Since even matches of a few key-points are sufficient to indicate the scale change, this method is able to adaptively change the size of the kernel even under severe partial occlusion thus making the kernel tracker more robust. In case there are no matches between the kernels as can happen in certain cases of total occlusion, the scale of the kernel is chosen to be the same as in the previous frame. This method of adapting to scale is more robust as compared to the scale space based approach advocated in [8] where the author proposes a scale space based mean shift approach. While, the idea of Gaussian scale space is similar, since these are considered for key-points instead of the whole kernel they are more resilient in case of occlusion.

6 Experimental Results

The proposed algorithm has been extensively tested on numerous videos from the Caviar [11], Jojic [12], and Karl-Wilhelm-Strasse [13] datasets. The generalized mean-shift tracker performed well for almost all the test cases under partial as well as full occlusion and in real time.

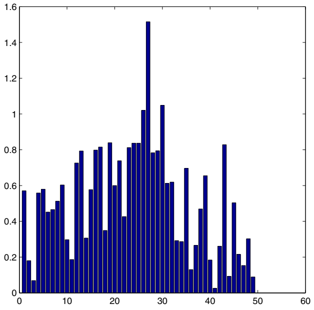


Fig. 2. Plot of Dissimilarity Factor (Y-axis) vs image frames (X-axis)

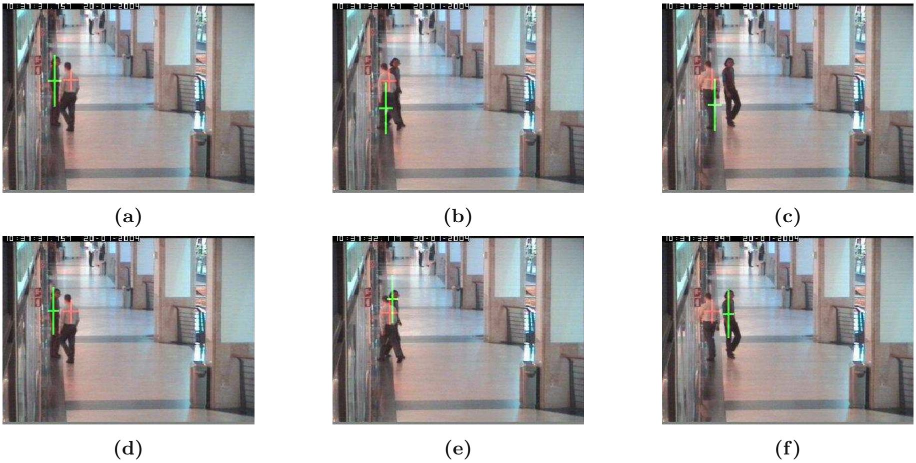


Fig. 3. *EnterExit* sequence: Tracking with occlusion.(a), (b), (c) show the results of the plain kernel tracker while (d), (e), and (f) show that of the proposed method on the Caviar data set.

The targets have been initialized by a manually chosen ellipse in all the video sequences. However, colored crosses have been used to indicate the kernel positions, they represent the minor and major axes of the tracking ellipse. We now discuss the results.

The *EnterExit* sequence is a set of 50, 384 x 288 pixel frames taken from the Caviar dataset [11]. It is a scene from a mall where one person enters a shop and another person exits it resulting in the two people crossing each other. Thus one observes partial occlusion. When we use the forward mean shift tracker [1], then the tracker fails to track the person entering the shop correctly and latches onto the person leaving the shop. This is due to the partial occlusion. However, as can be seen in Fig. 3, the proposed method is able to successfully track the person entering as well as the person leaving correctly even in case of partial occlusion.

Next we consider a close range sequence used by Jojic and Frey in [12]. The sequence consists of 40 frames with each frame 320x240 pixels in size. Here one can observe that there is full occlusion present. The results for the forward mean shift tracker and the proposed method are presented in Fig. 6. The interesting part is that the two close range observations are quite similar in terms of skin color. The forward mean shift tracker fails to track the two persons when there is full occlusion. However, due to the improvements proposed in terms of generalized mean shift optimization we are able to track the two persons even in case of full occlusion.

The algorithm presented is scalable and hence can be extended to higher number of objects being tracked simultaneously with occlusion. It has been tested on three objects and can be scaled up with ease. For this purpose another Caviar sequence has been used: *ThreePastShop* sequences (Fig. 4) which consists of 100

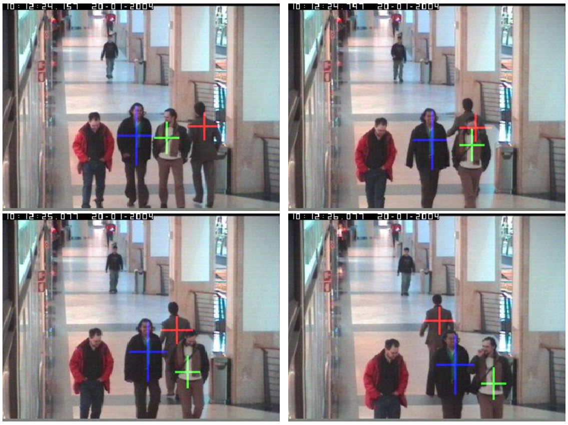


Fig. 4. *ThreePastShop* sequence: Tracking 3 targets with occlusion on the Caviar data set

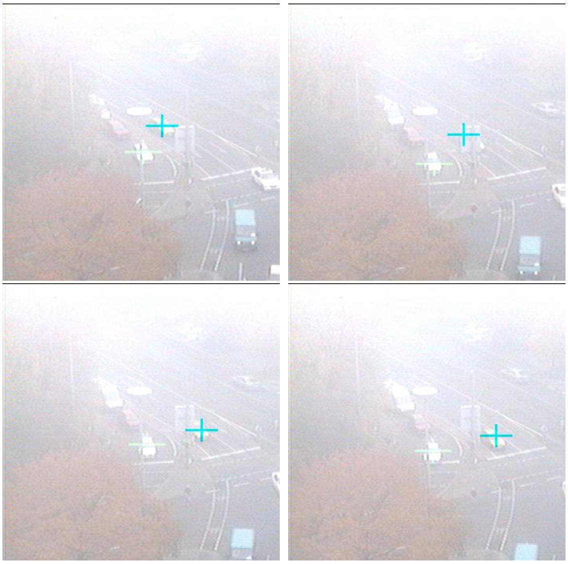
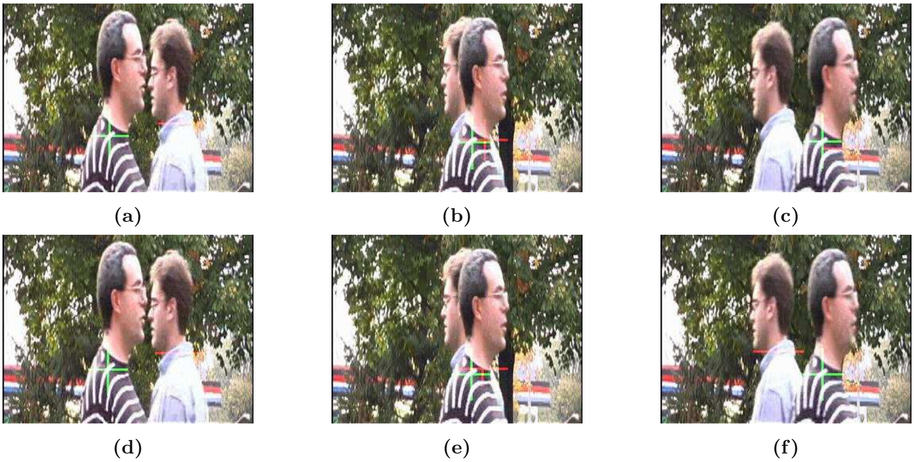


Fig. 5. *Karl-Wilhelm-Strass* sequence: Tracking vehicle targets with occlusion and fog

384 x 288 pixel frames shot at a corridor of a shopping mall. In this case we are able to successfully track three people and multiple occlusions. The results shown in Fig. 4 demonstrates this.

The proposed method has also been successfully tested on traffic videos. We ran experiments on the Karl-Wilhelm-Strass data set [13] (60 frames 350 x 350 pixels) with considerable fog and occlusion (Figure 5). Here we are able to track a car under severe fog and also occlusion when it passes under a billboard. This demonstrates the robustness of our approach.


(a) (b) (c)

(d) (e) (f)

Fig. 6. *Jojic* sequence: Tracking with occlusion.(a), (b), (c) show the results of the plain kernel tracker while (d), (e), and (f) show that of the proposed method on the Jojic data set.

Further, we have demonstrated the effectiveness of the SIFT based technique to handle scale change in tracking videos. The results of the same are shown on a video clip from the movie "Breakfast at Tiffany's", which shows a scene where two persons are climbing down while being occluded by a passing motor-vehicle. The SIFT based technique enables effective handling of scale change as can be seen from the results in Fig. 7. The results can be better considered from the result videos available at *http://vinaypn.googlepages.com/tracking*. We now discuss the parameters used in our experiments. The generalized mean-shift approach has two parameters that need to be initialized manually. The first is a similarity factor threshold θ in eqn. (12) which is used to determine the sign of the sgn function. It can be seen from Figure 2 that at areas of partial or total occlusion the distance factor $d(y)$ is quite high and this can be used to determine a threshold value θ. In the Caviar, Karl-Wilhelm-Strass, "Breakfast at Tiffany's"sequences we used a threshold value of 0.4. While, in the sequence of Jojic we had to use a value of 0.15. This change can be attributed to the fact that this sequence was a close range video while all the others were shot from a considerable distance. Hence, we believe that if the range of the scene is known approximately, we need to initialize these parameters just once.

The other coefficient we used is the reverse mean-shift weight γ in eqn. (12) and we used a value of 0.4 for all the test cases. We found that these parameters are fairly global and were not changed in most of the test sequences.

There are a few areas where the tracker might fail. Whenever the foreground object is considerably bigger than the tracking kernel of the background object, the tracker will not be able to locate the object once the occlusion frames are over. This situation may be handled by using a much wider search window when the tracker fails to locate a match after a certain number of iterations.

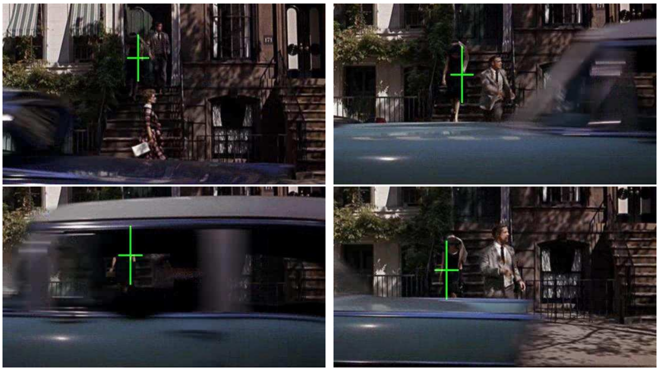


Fig. 7. *Breakfast at Tiffany's* sequence: Tracking target with occlusion and scale change in a clip from the movie

The SIFT based technique to handle scale change requires relatively high resolution videos to be able to generate adequate number of match points to work effectively. Often, the tracked object's orientation changes in a video sequence. We have taken care to update our matching image to take care of these situations.

7 Conclusion

In this paper we address the problem of occlusion while tracking multiple objects using a kernel based tracker. We identify the problem as incorrect mode estimation due to the convex optimization method of mean shift based optimization used for localization. Hence, we suggest a modification based on generalized mean shift based optimization which is able to escape problems of local minima in a neighborhood. We further consider the problem of scale adaptation and propose a solution based on identifying scale change in key-points computed using SIFT. This method of scale change works well even in case of occlusion. The improved kernel tracker thus developed is robust and also processes the data in real time. The tracker's efficiency has been proved by extensive testing on various popular data sets.

There are certain cases where in case there is prolonged total occlusion, the errors are propagated. We intend to explore solutions based on a global search paradigm in such cases to handle the problems which are inherent due to the local nature of the approach considered.

Acknowledgments

Financial assistantship under the *Swarnajayanti* fellowship scheme from the Department of Science and Technology, India is gratefully acknowledged.

References

1. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **25** (2003) 564–575
2. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
3. Comaniciu, D., Meer, P.: Mean Shift: A Robust Approach toward Feature Space Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002) 603–619
4. Babu, V., Perez, P., Bouthemy, P.: Kernel-based robust tracking for objects. In: Proc. Asian Conference on Computer Vision, Hyderabad India, Part II. (2006) 353–362
5. Isard, M., MacCormick, J.: Bramble: A bayesian multiple-blob tracker. In: Proc.IEEE International Conf. on Computer Vision (ICCV), vol. 2. (2001) 34–41
6. Comaniciu, D., Ramesh, V., Meer, P.: The variable bandwidth mean shift and data-driven scale selection. In: Proceedings of IEEE International Conference on Computer Vision, Vol. 1. (2001) 438–445 held in Vancouver, Canada.
7. Comaniciu, D.: An algorithm for data-driven bandwidth selection. IEEE Transactions on Pattern Analysis and Machine Intelligence **25** (2003) 281–288
8. Collins, R.T.: Mean shift blob tracking through scale space. In: CVPR 2003 Conference Proceedings. (2003) 234–240 held in Madison, Wisconsin, June.
9. Fukunaga, K., Hostetler, L.D.: The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory **21** (1975) 32–40
10. Namboodiri, V.P., Chaudhuri, S.: Shock filters based on implicit cluster separation. In: Proc. Conference on Computer Vision and Pattern Recognition (CVPR 2005),20-26 June 2005, San Diego, CA, USA. (2005) 82–87
11. Fisher, R.B.: Pets04 surveillance ground truth data set. In: Proc. Sixth IEEE Int. Work. on Performance Evaluation of Tracking and Surveillance. (2004) 1–5
12. Jojic, N., Frey, B.: Learning flexible sprites in video layers. In: Proc.IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 1. (2001) 199–206
13. Haag, M., Nagel, H.H.: Tracking of complex driving manoeuvres in traffic image sequences. Image and Vision Computing **16** (1998) 517–527