Multi-Model Component-Based Tracking using Robust Information Fusion

Bogdan Georgescu[1], Dorin Comaniciu[1], Tony X. Han[2], Xiang Sean Zhou[1]

[1]Real-Time Vision and Modeling Department, Siemens Corporate Research
755 College Road East, Princeton, NJ 08540, USA
bogdan.georgescu, dorin.comaniciu, xiang.zhou@scr.siemens.com
[2]Beckman Institute and ECE Department, University of Illinois at Urbana-Champaign
405 N. Mathews Ave., Urbana, IL 61801, USA
xuhan@ifp.uiuc.edu

Abstract. One of the most difficult aspects of visual object tracking is the handling of occlusions and target appearance changes due to variations in illumination and viewing direction. To address these challenges we introduce a novel tracking technique that relies on component-based target representations and on robust fusion to integrate model information across frames. More specifically, we maintain a set of component-based models of the target, acquired at different time instances, and combine robustly the estimated motion suggested by each component to determine the next position of the target. In this paper we allow the target to undergo similarity transformations, although the framework is general enough to be applied to more complex ones. We pay particular attention to uncertainty handling and propagation, for component motion estimation, robust fusion across time and estimation of the similarity transform. The theory is tested on very difficult real tracking scenarios with promising results.

1 Introduction

One of the problems of visual tracking of objects is to maintain a representation of target appearance that has to be robust enough to cope with inherent changes due to target movement and/or camera movement. Methods based on template matching have to adapt the model template in order to successfully track the target. Without adaptation, tracking is reliable only over short periods of time when the appearance does not change significantly. However, in most applications, for long time periods the target appearance undergoes considerable changes in structure due to change of viewpoint, illumination or it can be occluded. Methods based on motion tracking [1], [2], where the model is adapted to the previous frame, can deal with such appearance changes. However accumulated motion error and rapid visual changes make the model to drift away from the tracked target. Tracking performance can be improved by imposing object specific subspace constraints [3], [4] or maintaining a statistical representation of the model [5], [6], [7]. This representation can be determined a priori or computed on line. The appearance variability can be modeled as a probability distribution function which ideally is learned on line. Previous work approximated this p.d.f. as a normal distribution in which case the mean represent the most likely model template. Updating the distribution parameters can be done using EM based algorithms. Also adaptive mixture models have been proposed to cope with outliers and sudden appearance changes [5].

We propose a method where the appearance variability is simply modeled by maintaining several models over time. This amounts for a nonparametric representation of the probability density function that characterizes the object appearance. We also adopt a component based approach and divide the target into several regions which are processed separately. Tracking is performed by obtaining independently from each model a motion estimate and its uncertainty through optical flow. A recently proposed robust fusion technique [8] is used to compute the final estimate for each component. The method, named Variable-Bandwidth Density-based Fusion (VBDF), computes the location of the most significant mode of the displacements density function while taking into account their uncertainty. The VBDF method manages the multiple data sources and outliers in the motion estimates. In this framework, occlusions are naturally handled through the estimate uncertainty for large residual errors. The alignment error is used to compute the scale of the covariance matrix of the estimate, therefore reducing the influence of the unreliable displacements.

The paper is organized as follows. Section 2 contains previous work on appearance modeling related to our approach. The multi-model component based tracking method is presented in Section 3. Experiments on real sequences under considerable occlusions are in Section 4 and we conclude in Section 5.

2 Related Work

An intrinsic characteristic of the vision based tracking is that the appearance of the tracking target and the background are inevitably changing, albeit gradually. Since the general invariant features for robust tracking are hard to find, most of the current methods need to handle the appearance variation of the tracking target and/or background. Every tracking scheme involve a certain representation of the 2D image appearance of the object, even though this is not mentioned explicitly.

Fleet et al. [5] proposed a generative model containing 3 components: the stable component, the wandering component, and the occlusion component. The stable component identifying the most reliable structure for motion estimation and the wandering component representing the variation of the appearance are two Gaussian distributions. The occlusion component accounting for data outliers is uniformly distributed on the possible intensity level. Using the phase parts of the steerable wavelet coefficients [9] as feature, this algorithm achieves satisfactory tracking results. It needs a relative long time for the stable component to gain confidence in appearance estimation. Since the stable component, as the tracking template, is modeled as an unimodal Gaussian, it needs to restart from time to time to accommodate the natural multimodal case.

Based on the hypothesis that most promising features for tracking are the same features that best discriminate between object and background classes, Collins et al. [10] empirically evaluate all candidate features to estimate the distributions and a new image composed of the log likelihood ratio of these distributions is used to track. However some salient feature that distinguish the tracking object from background may change drastically which imperils the validity of the hypothesis used in [10].

Using Earth Movers Distance as feature, Sharma et al. [11] present a complete object appearance learning approach, dealing with 3D surfaces. The problem is modeled by a continuous Markov Random Field with clique potentials defined as energy function. This approach safely maps and maintains object appearance on the 3D model surface. An online adapted appearance model is proposed in [12] using a Markov Random Field of the color distributions over a 3D model. Appearance driven cue confidences are used to balance the contribution for model update.

3 Multi-Model Component-Based Tracker

Object tracking challenges due to occlusions and appearance variations are handled in our framework through a multi-model component-based approach. Maintaining several representatives for the 2D appearance model does not restrict it to a unimodal distribution and the VBDF fusion mechanism robustly integrates multiple estimates to determine the most dominant motion for each component. These key ideas are introduced in Subsection 3.1 followed by the VBDF algorithm in Subsection 3.2. Details about our method and implementation issues are in Subsection 3.3.

3.1 Main Ideas

The steps of our proposed method are outlined in Figure 1. To model the changes during tracking we propose to maintain several exemplars of the object appearance over time. This is in contrast to the approach adopted in [5] where the stable appearance component is modeled using a Gaussian distribution for each pixel. Maintaining explicitly the intensities is equivalent to a nonparametric representation of the appearance distribution.

The top row in Figure 1 illustrates the current exemplars in the model set, each having associated a set of overlapping components. A component-based approach is more robust that a global representation, being less sensitive to illumination changes and pose. Another advantage is that partial occlusion can be handled at the component level by analyzing the matching likelihood.

Each component is processed independently, its location and covariance matrix is estimated in the current image with respect to all of the model templates. For example, one of the components is illustrated by the gray rectangle in Figure 1 and its location and uncertainty with respect to each model is shown in I_{new}. The VBDF robust fusion procedure is applied to determine the most dominant motion (mode) with the associated uncertainty (Figure 1, center bottom row). Note the variance in the estimated location of each component due to occlusion or appearance change.

The location of the components in the current frame is further constrained by a global parametric motion model. We assume a similarity transformation model and its parameters are estimated using the confidence in each component location. Therefore the reliable components contribute more to the global motion estimation.

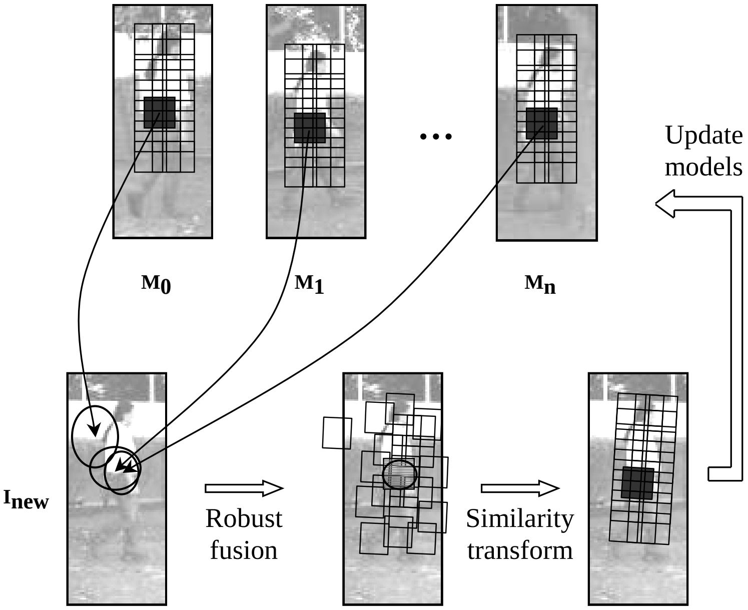


Fig. 1. The steps of the multi-model component based tracker.

The current frame is added to the model set if the residual error to the reference appearances is relatively low. The threshold is chosen such that we do not add the images where the target has significant occlusion. The number of templates in our model set is fixed, therefore the oldest one is discarded.

3.2 Variable-Bandwidth Density-based Fusion

The VDBF estimator is based on nonparametric density estimation with adaptive kernel bandwidths. It was introduced in [8] with an application to robust optical flow computation. The choice of the VDBF estimator is motivated by its good performance in the presence of outliers in the input data when compared to previously proposed methods such as *Covariance Intersection* [13] or BLUE estimate assuming single source, statistically independent data [14]. The robustness with respect to outliers of the VDBF technique comes from the nonparametric estimation of the initial data distribution while exploiting its uncertainty. The VBDF estimator is defined as the location of the most significant mode of the density function. The mode computation is based on the variable-bandwidth mean shift technique in a multiscale optimization framework.

Let $\boldsymbol{x}_i \in \mathbb{R}^d$, $i = 1 \ldots n$ be the available d-dimensional estimates, each having an associated uncertainty given by the covariance matrix C_i. The most significant mode of the their density function is determined iteratively in a multiscale fashion. A bandwidth matrix $\mathrm{H}_i = \mathrm{C}_i + \alpha^2\mathrm{I}$ is associated with each point $\boldsymbol{x}_i$, where I is the identity matrix and the parameter α determines the scale of the analysis. The sample point density estimator at location $\boldsymbol{x}$ is defined by

$$\hat{f}(\boldsymbol{x}) = \frac{1}{n(2\pi)^{d/2}} \sum_{i=1}^{n} exp\left(-\frac{1}{2}D^2(\boldsymbol{x}, \boldsymbol{x}_i, \mathrm{H}_i)\right) \tag{1}$$

where D represents the Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{x}_i$

$$D^2(\boldsymbol{x}, \boldsymbol{x}_i, \mathrm{H}_i) = (\boldsymbol{x} - \boldsymbol{x}_i)^\top \mathrm{H}_i^{-1}(\boldsymbol{x} - \boldsymbol{x}_i) \tag{2}$$

The variable bandwidth mean shift vector at location $\boldsymbol{x}$ is given by

$$\boldsymbol{m}(\boldsymbol{x}) = \mathrm{H}_h(\boldsymbol{x}) \sum_{i=1}^{n} \omega_i(\boldsymbol{x})\mathrm{H}_i^{-1}\boldsymbol{x}_i - \boldsymbol{x} \tag{3}$$

where H_h represents the harmonic mean of the bandwidth matrices weighted by the data-dependent weights $\omega_i(\boldsymbol{x})$

$$\mathrm{H}_h(\boldsymbol{x}) = \left(\sum_{i=1}^{n} \omega_i(\boldsymbol{x})\mathrm{H}_i^{-1}\right)^{-1} . \tag{4}$$

The data dependent weights computed at the current location $\boldsymbol{x}$ have the expression

$$\omega_i(\boldsymbol{x}) = \frac{\frac{1}{|\mathrm{H}_i|^{1/2}} exp\left(-\frac{1}{2}D^2(\boldsymbol{x}, \boldsymbol{x}_i, \mathrm{H}_i)\right)}{\sum_{i=1}^{n} \frac{1}{|\mathrm{H}_i|^{1/2}} exp\left(-\frac{1}{2}D^2(\boldsymbol{x}, \boldsymbol{x}_i, \mathrm{H}_i)\right)} \tag{5}$$

and note that they satisfy $\sum_{i=1}^{n} \omega_i(\boldsymbol{x}) = 1$.

It can be shown that the density corresponding to the point $\boldsymbol{x} + \boldsymbol{m}(\boldsymbol{x})$ is always higher or equal to the one corresponding to $\boldsymbol{x}$. Therefore iteratively updating the current location using the mean shift vector yields a hill-climbing procedure which converges to a stationary point of the underlying density.

The VBDF estimator finds the most important mode by iteratively applying the adaptive mean shift procedure at several scales. It starts from a large scale by choosing the parameter α large with respect to the spread of the points $\boldsymbol{x}_i$. In this case the density surface is unimodal therefore the determined mode will correspond to the globally densest region. The procedure is repeated while reducing the value of the parameter α and starting the the mean shift iterations from the mode determined at the previous scale. For the final step the bandwidth matrix associated to each point is equal to the covariance matrix, i.e. $\mathrm{H}_i = \mathrm{C}_i$.

The VBDF estimator is a powerful tool for information fusion with the ability to deal with multiple source models. This is important for motion estimation as points in a local neighborhood may exhibit multiple motions. The most significant mode corresponds to the most relevant motion.

3.3 Tracking Multiple Component Models

Consider that we have n models $\mathrm{M}_0, \mathrm{M}_1, \ldots, \mathrm{M}_n$. For each image we maintain c components with their location denoted by $\boldsymbol{x}_{ij}$, $i = 1 \ldots n$, $j = 1 \ldots c$. When a new image is available we estimate the location and the uncertainty for each component and for each model. This step can be done using several techniques such as ones based on image correlation, spatial gradient or regularization of spatio-temporal energy. Based on the image brightness constancy, one of the most popular optical flow techniques has been developed by Lucas and Kanade [15]. For a small image patch the pixels flow estimates are combined assuming a translational model by solving a weighted least squares problem. However they neglect the uncertainty of the initial estimates, therefore we adopt the robust optical flow technique proposed in [8] which is also an application of the VBDF technique. The result is the motion estimate $\hat{\boldsymbol{x}}_{ij}$ for each component and its uncertainty $\hat{\mathrm{C}}_{ij}$. Thus $\hat{\boldsymbol{x}}_{ij}$ represents the location estimate of component j with respect to model i. The scale of the covariance matrix is also estimated from the matching residual errors. This will increase the size of the covariance matrix when the respective component is occluded therefore we naturally handle occlusions at the component level.

The VBDF robust fusion technique presented in the previous subsection is applied to determine the most relevant location $\hat{\boldsymbol{x}}_j$for component j in the current frame. The mode tracking across scales results in

$$\begin{aligned}\hat{\boldsymbol{x}}_j &= \mathrm{C}(\hat{\boldsymbol{x}}_j)\sum_{i=1}^{n}\omega_i(\hat{\boldsymbol{x}}_j)\hat{\mathrm{C}}_{ij}^{-1}\hat{\boldsymbol{x}}_{ij} \\ \mathrm{C}(\hat{\boldsymbol{x}}_j) &= \left(\sum_{i=1}^{n}\omega_i(\hat{\boldsymbol{x}}_j)\hat{\mathrm{C}}_{ij}^{-1}\right)^{-1} .\end{aligned} \tag{6}$$

with the weights ω_i defined as in (5).

Following the location computation of each component, a weighted rectangle fitting is carried out with the weights given by the covariance matrix of the estimates. We assume that the image patches are related by a similarity transform T defined by 4 parameters. The similarity transform of the dynamic component location $\boldsymbol{x}$ is characterized by the following equations.

$$T(\boldsymbol{x}) = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}\boldsymbol{x} + \begin{pmatrix} t_x \\ t_y \end{pmatrix} \tag{7}$$

where t_x, t_y are the translational parameters and a, b parametrize the 2D rotation and scaling.

The minimized criterion is the sum of Mahalanobis distances between the reference location $\boldsymbol{x}_j^0$ and the estimated ones $\hat{\boldsymbol{x}}_j$ (j^{th} component location in the current frame).

$$\mathcal{J} = \sum_{j=1}^{c}(\hat{\boldsymbol{x}}_j - T(\boldsymbol{x}_j^0))^T\mathrm{C}(\hat{\boldsymbol{x}}_j)^{-1}(\hat{\boldsymbol{x}}_j - T(\boldsymbol{x}_j^0)) . \tag{8}$$

Minimization is done through standard weighted least squares. Note that because we use the covariance matrix for each component the influence of points with high uncertainty is reduced.

After the rectangle is fitted to the tracked components, we uniformly resample the dynamic component candidate inside the rectangle. We assume the relative position of each component with respect to the rectangle does not change a lot. If the distance of the resample position and the track position computed by the optical flow of a certain component is larger than a tolerable threshold, we regard the tracked position as an outlier and replace it with the resampled point. The current image is added to the model set if sufficient components have low residual error. The median residual error between the models and the current frame is compared with a pre-determined threshold T_h.

Given a set of models $M_0, M_1, \ldots, M_n$ in which the component j has location $\boldsymbol{x}_{ij}$ in frame i, our object tracking algorithm can be summarized by the following steps:

1. Given a new image I_f compute $\hat{\boldsymbol{x}}_{ij}^{(f)}$ through robust optical flow [8] starting from $\hat{\boldsymbol{x}}_j^{(f-1)}$, the location estimated in the previous frame;
2. For $j = 1 \ldots c$ estimate the location $\hat{\boldsymbol{x}}_j^{(f)}$ of component j using the VBDF estimator (Subsection 3.2) resulting in (6);
3. Constrain the component location using the transform computed by minimizing (8);
4. Add the new appearance to the model set if its median residual error is less that T_h.

The proposed multi-template framework can be directly applied in the context of shape tracking. If the tracked points represent the control points of a shape modeled by splines, the use of the robust fusion of multiple position estimates increases the reliability of the location estimate of the shape. It also results in smaller corrections when the shape space is limited by learned subspace constraints. If the contour is available, the model templates used for tracking can be selected online from the model set based on the distance between shapes.

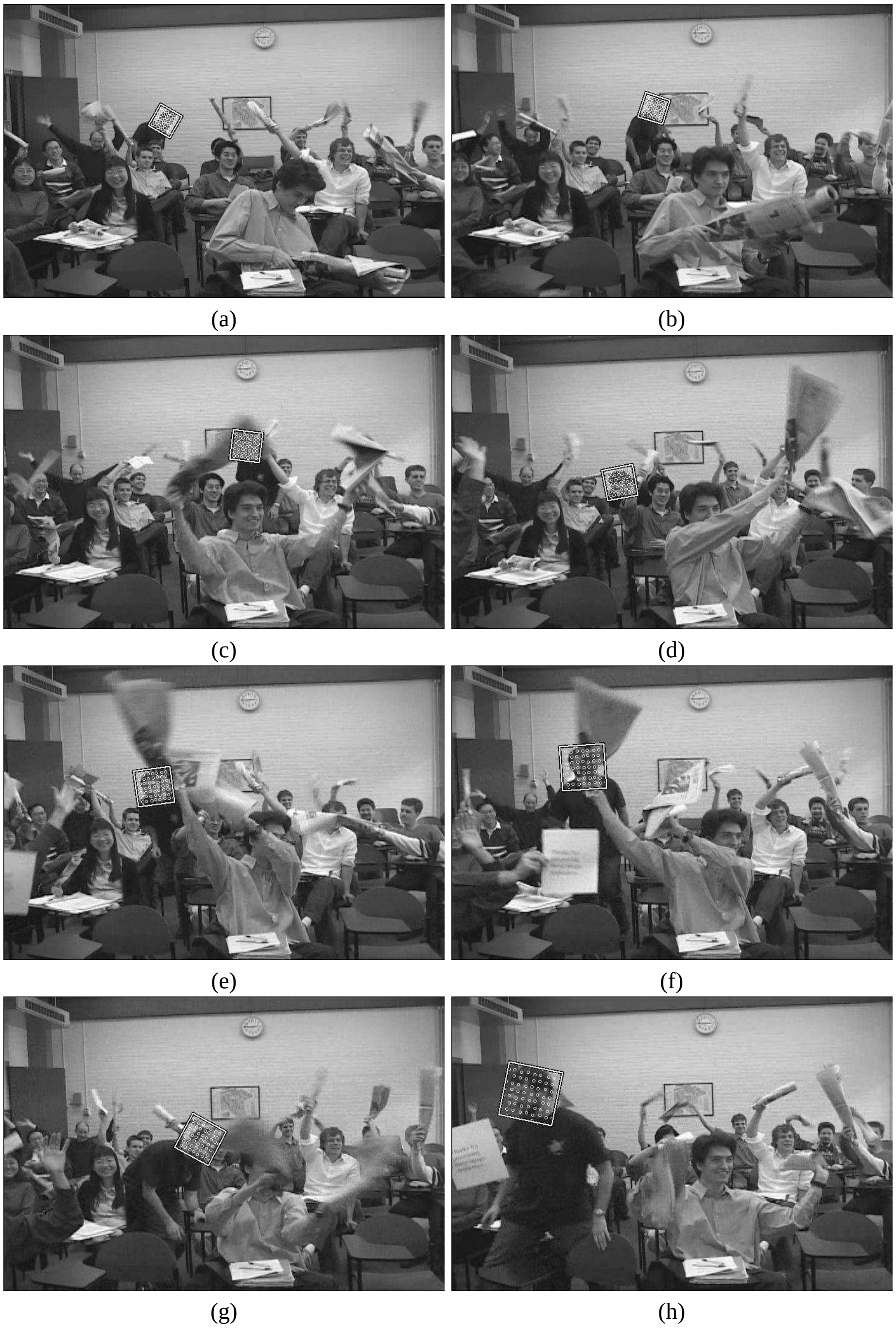
4 Experiments

The proposed method was applied for object tracking in real videos with significant clutter and occlusions. We used $n = 20$ model templates and the components are at 5 pixels distance with their number c determined by the bounding rectangle. The threshold T_h for a new image to be added to the model set was $1/8$ of the intensity range. The value was learned from the data such that occlusions are detected.

We successfully tested the method on different image sequences including medical videos. We present results on only two sequences that illustrate the advantages of our approach.

The results for tracking a person's face is presented in Figure 2. This is a very challenging sequence where the scene has significant clutter with several faces and multiple occlusions affect the tracked region. Figure 3 shows the median residual error over time which is used for model update. The peaks in the graph correspond to frames where the target is completely occluded. As mentioned earlier, the model update occurs when the error passes the threshold $T_h = 32$ which is the horizontal line in Figure 3.

Figure 4 shows the results of tracking of a human body. The method is able to cope with the appearance changes such as the arm moving and it is able to recover the



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 2. Face tracking results; the white rectangle represents the target. (a) Frame 0; (b) Frame 49; (c) Frame 137; (d) Frame 249; (e) Frame 281; (f) Frame 312; (g) Frame 375; (h) Frame 527.

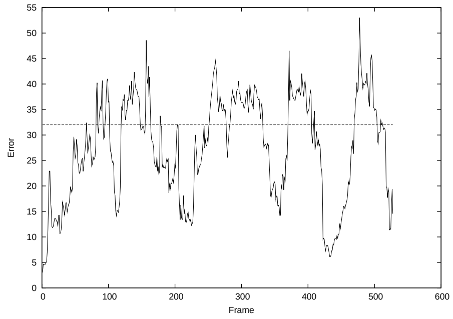


Fig. 3. Residual error over time for face tracking sequence.Horizontal line represents the model update threshold.

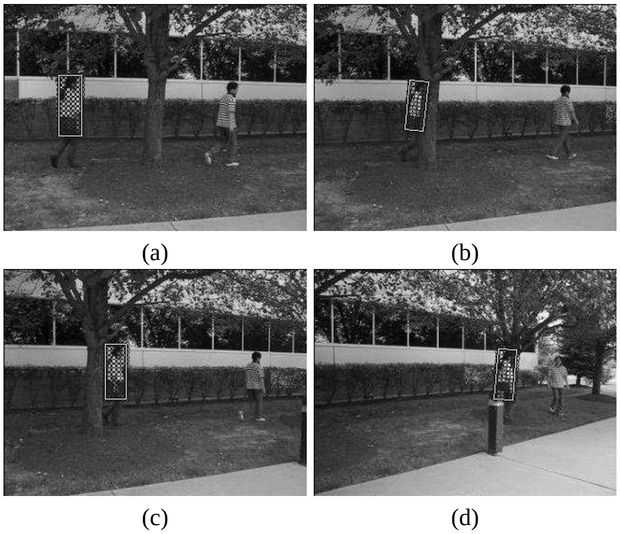


Fig. 4. Human body tracking results.(a) Frame 0; (b) Frame 27; (c) Frame 38; (d) Frame 106.

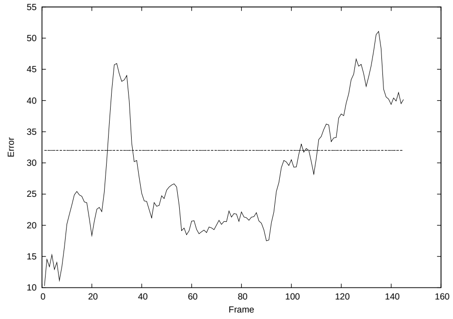


Fig. 5. Residual error over time for human body tracking sequence. Horizontal line represents the model update threshold.

tracking target after the tree occlusion. Figure 5 plots the residual error over time. The first peak corresponds to the target being occluded by the tree while toward the end the error is due to the person turning and its image size becoming smaller with respect to the fixed component size.

5 Conclusion

This paper introduced an object tracking method based on multiple appearance models and VBDF-based fusion of estimates. We showed the ability of the proposed approach to deal with significant occlusions, clutter and appearance changes on real image sequences. Although we used image templates as models, our approach is general enough to integrate information from different model representations such as color distributions or filter responses. Further work include solving for the global motion through a robust approach and use of multiple hypothesis for tracking.

References

1. Shi, J., Tomasi, C.: Good features to track. In: 1994 IEEE Conf. on Computer Vision and Pattern Recog., San Juan, Puerto Rico (1994) 593–600
2. Sidenbladh, H., Black, M.J., Fleet, D.J.: Stochastic tracking of 3d human figures using 2d image motion. In: 2000 European Conf. on Computer Vision. Volume 2., Dublin, Ireland (2000) 702–718
3. Black, M.J., Jepson, A.D.: Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. International J. of Computer Vision **26** (1998) 63–84
4. Edwards, G.J., Cootes, T.F., Taylor, C.J.: Face recognition using active appearance models. In: 1998 European Conf. on Computer Vision, Freiburg, Germany (1998) 581–595
5. Jepson, A.D., Fleet, D.J., El-Maraghi, T.F.: Robust online appearance models for visual tracking. IEEE Trans. Pattern Anal. Machine Intell. **25** (2003) 1296–1311
6. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: 1999 IEEE Conf. on Computer Vision and Pattern Recog. Volume 2. (1999) 246–252
7. Tao, H., Sawhney, H.S., Kumar, R.: Dynamic layer representation with application to tracking. In: 2000 IEEE Conf. on Computer Vision and Pattern Recog. Volume 2. (2000) 134–141
8. Comaniciu, D.: Nonparametric information fusion for motion estimation. In: 2003 IEEE Conf. on Computer Vision and Pattern Recog. Volume I., Madison, WI (2003) 59–66
9. Freeman, W.T., Adelson, E.H.: The design and use of steerable filters. IEEE Trans. Pattern Anal. Machine Intell. **13** (1991) 891–906
10. Collins, R., Liu, Y.: On-line selection of discriminative tracking features. In: 2003 International Conf. on Computer Vision. (2003)
11. Krahnstoever, N., Sharma, R.: Robust probabilistic estimation of uncertain appearance for model based tracking. In: IEEE Workshop on Motion and Video Computing. (2002)
12. Krahnstoever, N., Sharma, R.: Appearance management and cue fusion for 3d model-based tracking. In: 2003 IEEE Conf. on Computer Vision and Pattern Recog., Madison, WI (2003)
13. Julier, S., Uhlmann, J.: A non-divergent extimation algorithm in the presence of unknown correlations. In: Proc. American Control Conf., Alberqueque, NM (1997)
14. Singh, A., Allen, P.: Image-flow computation: An estimation-theoretic framework and a unified perspective. CVGIP: Image Understanding **56** (1992) 152–177
15. Lucas, B., Kanade, T.: An iterative image registration technique with application to stereo vision. In: International Joint Conf. on Artificial Intelligence, Vancouver, Canada (1981) 674–679