AUTOMATIC MULTI-HEAD DETECTION AND TRACKING SYSTEM USING A NOVEL DETECTION-BASED PARTICLE FILTER AND DATA FUSION

Wei Qu, Nidhal Bouaynaya and Dan Schonfeld

Multimedia Communications Laboratory,
ECE Department, University of Illinois at Chicago, Chicago, IL 60607.
E-mail: {wqu, nbouayna, ds}@ece.uic.edu

ABSTRACT

We present a novel automatic system integrating head detection with particle filter for realtime multi-head tracking (MHT) in video. Distinct with the conventional particle filter which gets particles from the prior density, we propose a novel importance function based on the up to date detection and motion observation which makes the particles more effective and helps us to achieve stable tracking by using much less particles. We also propose a general likelihood model in the context of MHT. Different information can be fused in a principle manner to make the tracker more stable. The proposed approach can handle not only the changes of scale, lighting, zooming, and pose, but also fast motion, appearance, and hard multi-head occlusion.

1. INTRODUCTION

Head detection and tracking has been an intensive research area due to its wide applications. However, because of lacking effective representation and good scheme to handle occlusion, robust and efficient head tracking especially MHT in complex environment is still an open research problem.

Particle filter has received much attention in recent years. Consider a dynamic system presented by the continuous-time *Hidden Markov Model*. The tracking problem is to estimate the posterior $p(x_t|Z_{1:t})$ by the Bayesian inference

$$p(x_t|Z_{1:t}) \propto kp(z_t|x_t)p(x_t|Z_{1:t-1}) \qquad (1)$$

where k is the normalization constant and

$$p(x_t|Z_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|Z_{1:t-1})\,dx_{t-1}. \qquad (2)$$

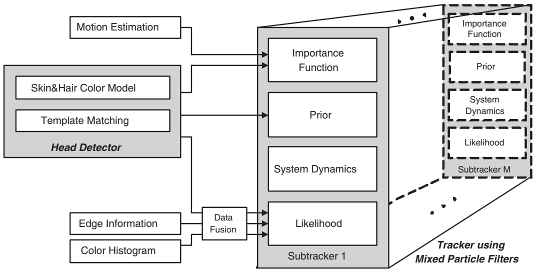

Fig. 1. The structure of MHT system.

Because the likelihood $p(z_t|x_t)$ is usually nonlinear, non-Gaussian which makes the integral unfeasible, the posterior density can be approximated by properly weighted particles sampled from any proposal distribution q (also called the importance function) [1]. Since particles are sampled from $p(x_t|Z_{1:t-1})$ and weights are only computed by likelihood, the standard particle filter is vulnerable to the degeneracy problem [2]. [3] has proved that the optimal importance function which minimizes the variance of weights is

$$q_{opt} = p(x_t|x_{t-1}, Z_t) \qquad (3)$$

where Z_t is the newest observation. But no realization is given in this paper and most discussion in this context limits to single object tracking. In this paper, we propose a novel optimal importance function suitable for MHT.

The paper is organized as follows. Section 2 presents the system structure. Section 3 and 4 describe the detection and motion estimation methods we used respectively. Section 5 discusses the detection and motion-based MHT framework we proposed. Experiment results are given in Section 6. In Section 7, we conclude this paper.

2. SYSTEM STRUCTURE

Fig. 1 shows the system structure. The tracker is composed of mixed particle filters where each keeps tracking one head. Up to date motion and detection information has been used to make the importance function optimal. The detector is also used for initialization. Different information has been fused together to make the tracker more stable. To handle multi-head occlusion where the particle filters are dependent, the likelihood is associated with corresponding head under the following assumptions: *I.* One head can produce zero or one observation at one time. *II.* One observation can

0-7803-8874-7/05/$20.00 ©2005 IEEE

originate from the clutter or from several heads. The latter situation occurs during occlusion. For example, when two heads overlap, the detector can only detect the frontal head while it will be shared by all filters. *III.* Only observations that are inside the neighbor of the predicted state are kept.

3. HEAD DETECTION ALGORITHM

Any head detection algorithm can be accepted to our system. In this paper, we combine skin&hair color model [4] and template matching to achieve realtime head detection.

5000 skin samples from 50 color images of different ethnicities are clustered and can be fitted by a Gaussian model $N(m, C)$. By using the skin color likelihood, we transform a color image into a gray scale image. Since skin regions are brighter than other parts, it can be segmented to a binary image by appropriate thresholding. By labelling the connected components in the binary image, we find the candidate faces. Then after calculating the center, orientation, width, region and ratio of the region, we resize, rotate and move a template by averaging 200 faces to the position where the candidate face is located. By computing the cross-correlation value, we make a final decision according to an empirical threshold. Likewise the skin color model, we use a hair color model to include the hair area around the faces. Finally, we choose a parametric ellipse to represent the detected head because it can greatly save the computation cost.

4. MOTION ESTIMATION ALGORITHM

Any motion estimation technique can be used in our framework. Here, we use *Adaptive Block Matching (ABM)* in [5].

After dividing each frame into 16×16 blocks, ABM uses two steps to search blocks which belong to the object within a certain window of the current frame. First, blocks that are completely within the desired object are searched. Second, blocks at the object boundary are further divided into 8×8 blocks and then searched to find the object accurately. Finally the motion vector is calculated by the position difference between the current block and the most similar block in the reference frame. To prevent the high degree of freedom, we use a parametric ellipse to fit the output mask of ABM. Then the motion vector which interests us is given by the difference between the center of fitted ellipse in the current frame and the center of the previous ellipse.

5. NOVEL DETECTION-BASED PARTICLE FILTERING FRAMEWORK FOR MHT

Although particle filter has the capacity to keep multi-modality of the posterior density in theory, it usually happens that all particles rapidly migrate to one mode in practical implementation. To consistently maintain the multi-modality arising from both the ambiguity of clutter and the multiple heads, we model the posterior as a mixture of particle filters.

$$p(x_t|Z_{1:t}) = \frac{1}{M}\sum_{m=1}^{M} p_m(x_t^m|Z_{1:t}) = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{N}\pi_t^{m,i}\delta \tag{4}$$

where m is the index of objects. Any parameter model can be used to present the heads while we use ellipse $x_t^m = [cx_t^m, cy_t^m, a_t^m, b_t^m, \theta_t^m]$ where xc and yc are the center, a and b are the major and minor axes, and θ is the orientation. Note that instead of using one component for one mode as [6] and the mixture of Kalman filter [7], each subposterior can be multimodal and associates with a head. By doing so, we avoid the ambiguity that the modes can arise from both heads and clutter.

For each subposterior, we use N labelled particles to approximate it. Thus we have $M \times N$ particles totally.

$$\omega_t^{m,i} = \frac{g(x_t^{m,i})}{q(x_t^{m,i})} \tag{5}$$

$$= \frac{\sum_{j=1}^{N} p(x_t^{m,i}|x_{t-1}^{m,j})\omega_{t-1}^{m,j}}{q(x_t^{m,i}|x_{t-1}^{m,i}, z\_m_t^{m,i}, z\_d_t^{m,i})} p(z_t^{m,i}|x_t^{m,i}) \tag{6}$$

where $p(x_t^{m,i}|x_{t-1}^{m,j})$ is the state transition, $p(z_t^{m,i}|x_t^{m,i})$ is the likelihood, and q is the proposed importance function. After normalization, the system weights are

$$\pi_t^{m,i} = \frac{\omega_t^{m,i}}{\sum_{l=1}^{M}\sum_{j=1}^{N}\omega_t^{l,j}} \tag{7}$$

This procedure makes different particle filters dependent.

To output each head, we choose the mean criteria but do renormalization for corresponding weights first.

$$\gamma_t^{m,i} = \frac{\pi_t^{m,i}}{\sum_{j=1}^{N}\pi_t^{m,j}} \tag{8}$$

Then the output ellipse for each head is

$$\hat{x}_t^m = \sum_{i=1}^{N}\gamma_t^{m,i}x_t^{m,i} \tag{9}$$

5.1. Detection and Motion Based Importance Function

We propose a novel importance function based on detection and motion estimation for MHT.

$$q(x_t^m|x_{t-1}^m, Z\_m_t, Z\_d_t) = \tag{10}$$

$$\begin{cases} \sum\limits_{i=1}^{N} N[(1-\alpha)(x_{t-1}^{m,i} + \Delta x_t^m) + \alpha\, z\_d_t^m, \Sigma_p^m] & 0 < \alpha < 1; \\ \sum\limits_{i=1}^{N} N[x_{t-1}^{m,i} + \Delta x_t^m, \Sigma_p^m] & \alpha = 0. \end{cases}$$

where Z_m_t is the observation of motion estimation, Z_d_t is the observation of detection, and α is a parameter. When the detector fails, α is equal to zero. For each particle, we model the translated density as Gaussian distribution. So the importance function for the filter is a Gaussian mixture.

Motion observation is intermittent and region-based, but discriminant. It can effectively draw the tracker back to objects when the clutter is strong and distracts the tracker from objects. Detection observation is particularly helpful when the background moves with the objects or occlusion occurs.

5.2. Prior for Initialization Based on Head Detection

Since there's no motion information at the initial frame, we model the prior as a Gaussian mixture based on detection.

$$p(x_1|z\_d_1) = MoG = \frac{1}{M}\sum_{m=1}^{M} N(z\_d_1^m, \Sigma_{prior}^m) \quad (11)$$

5.3. State Transition

We use a first order temporal model for each particle filter

$$x_t^m = x_{t-1}^m + N(0, \Sigma_{pred}^m) \quad (12)$$

Any temporal model can be used instead. But even with this simple model, our approach achieves accurate results.

5.4. Likelihood Based on Multi-Information Fusion

Particle filter allows the fusion of different information into the likelihood to make the tracker more stable. Although this fact has been acknowledged before [4], [8], it has not been fully exploited in MHT context. Here, we proposed a general likelihood model fusing different information in a principle manner for MHT. Although we use only edge, color, and detection information, more information such as the motion, sound, appearance can be fused in a direct way.

$$\mu_{total}^{m,i} = \alpha_1\, \mu_{edge}^{m,i} + \alpha_2\, \mu_{color}^{m,i} + \alpha_3\, \mu_{detection}^{m,i} + \cdots \quad (13)$$

where α_1, α_2, and α_3 are the weights.

5.4.1. Edge Information

Any method giving the edge likelihood can be used in our framework. Here, we choose the original model in [8].

Along the object contour, choose Φ points uniformly and find the contour normal line for each point. Then on each normal line, use edge detector to find the candidate edge information in which only one edge is the true contour point Z_ϕ and others are clutter. With the assumption that clutter is a Possion process with spatial density γ and the true observation is normally distributed with standard deviation σ_e, we have the edge likelihood model.

$$p(Z_\phi|\lambda_\phi) \propto \frac{1}{\sqrt{2\pi}\sigma_e p_0 \gamma} exp(-\frac{(min(z_h - \lambda_\phi))^2}{2\sigma_e^2}) \quad (14)$$

where λ_ϕ represents the pixel along normal line ϕ belonging to the particle ellipse, p_0 is a prior probability. By assuming independence between normal lines, we have

$$\mu_{t,edge}^i = \prod_{\phi=1}^{\Phi} p(Z_{t,\phi}|\lambda_\phi) \quad (15)$$

5.4.2. Color Information

Color information is remarkably persistent and robust to changes in pose and illumination. We propose a revised adaptive color model based on detection for MHT.

We generate new color histograms by counting the pixels inside the ellipse of each head after initialization. Instead of using RGB space, we choose YCbCr space and use 8 bins for CbCr and only 4 bins coarsely for luminance. Then, we use Bhattacharyya distance to measure the similarity between the reference histogram h_r and each particle histogram h_t respectively similar to [9].

$$d_B = \sqrt{1 - \prod_{k=1}^{K} \sqrt{h_r(k)h_t(k)}dk} \quad (16)$$

where k is the index of bins. Finally, the likelihood weights are specified by Gaussian distribution with variance σ_c.

$$\mu_{color}^i = \frac{1}{\sqrt{2\pi}\sigma_c} exp\{-\frac{d_B^2}{2\sigma_c^2}\} \quad (17)$$

To handle the lighting changes and full rotation, we also make the model adaptive by updating as [9].

$$h_{r,t}^k = (1-\alpha_h)h_{r,t-1}^k + \alpha_h h_{t,E}^k \quad (18)$$

where h_E^k is the histogram of the mean state vector.

5.4.3. Detection Information

We use Euler distance to measure the similarity.

$$d_E = \sqrt{(cx_d - cx_p)^2 + (cy_d - cy_p)^2} \quad (19)$$

where (cx_d, cy_d) is the center of the detected ellipse, and (cx_p, cy_p) is the center of the ellipse corresponding to each particle. Smaller distance gives larger likelihood weight. Therefore, we can model the likelihood as Gaussian.

$$\mu_{detection}^i = \frac{1}{\sqrt{2\pi}\sigma_d} exp\{-\frac{d_E^2}{2\sigma_d^2}\} \quad (20)$$

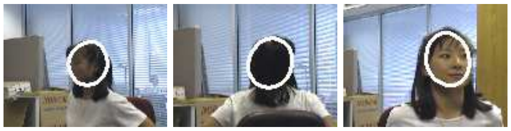

Fig. 2. Zooming, pose change, lighting change, and out of plane rotation. The video is from [10].

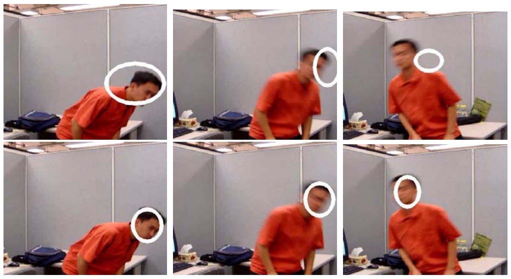

Fig. 3. Comparison of Condensation filter with our approach to fast motion. The first row is using Condensation filter. The second row is using our approach.

6. EXPERIMENTAL RESULTS

We have demonstrated our system on different video sequences. In Fig. 2, we show that our approach is able to handle zooming, head pose change, lighting change, and full rotation. It works well for different poses and full rotation because we exploit motion information and update the color histogram model in the likelihood. In Fig. 3 we show the results of applying the standard Condensation filter [8] (1st row) and our approach (2nd row) respectively to the same sequence with fast motion. It can be seen that Condensation filter is not able to follow the head in time when the person moves rapidly. And the tracker can not recover the desired head again when the strong clutters distract the ellipse away. However, by using detection and motion estimation, our approach can follow the head in fast motion. Even when the ellipse deviates from the object, the detector can guide the tracker to capture it again. In Fig. 4, we show an example of MHT. The tracker is reinitialized with a two-Gaussian mixture according to the detection result when the boy appears. Then the tracker keeps tracking each object individually even during hard occlusion.

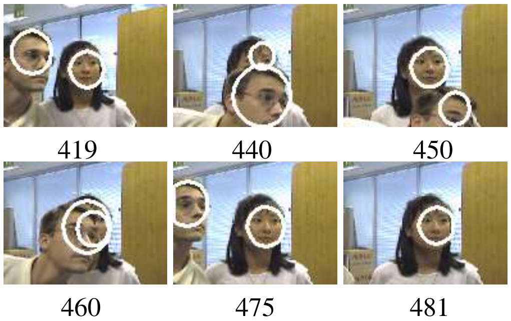

Fig. 4. MHT with hard occlusion. The video is from [10].

7. CONCLUSION

We have presented a novel framework for online multi-head detection and tracking. By integrating detection and motion estimation with particle filter, fusing multi-information into likelihood and solving data association problem, our approach is much more effective and stable than the available methods. By using general object detection method, our system can be easily extended to multi-object tracking.

8. REFERENCES

[1] A.F.M. Smith and A.E.Gelfand, "Bayesian statistics without tears: A sampling-resampling perspective," *American Statistican*, vol. 46, no. 2, pp. 84–88, 1992.

[2] J. S. Liu and R. Chen, "Sequential monte carlo methods for dynamic systems," *J. Amer. Statist. Assoc.*, vol. 93, no. 443, pp. 1032–1044, 1998.

[3] V. S. Zaritskii and L. Shimelevich, "Monte carlo technique in problems of optimal data processing," *Auto. Remo. Cont.*, vol. 12, pp. 95–103, 1975.

[4] J. Yang and A. Waibel, "A real-time face tracker," in *Proc. IEEE Workshop on Applications of Computer Vision (WACV '96)*, 1996, p. 142.

[5] P. R. K. Hariharakrishnan, D. Schonfeld, and F. Yassa, "Fast object tracking using adaptive block matching," *IEEE Trans. Multimedia*, submitted for publication.

[6] J. Vermaak, A. Doucet, and P. Pérez, "Maitaining multi-modality through mixture tracking," in *Proc. IEEE Int. Conf. on Computer Vision*, Oct. 2003, pp. 13–16.

[7] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, 1979.

[8] M. Isard and A. Blake, "Condensation – conditional density propagation for visual tracking," *Int. J. Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[9] K. Nummiaro, E. B. Koller-Meier, and L. Van Gool, "Object tracking with an adaptive color-based particle filter," in *Symposium for Pattern Recognition of the DAGM*, 2002, pp. 355–360.

[10] http://vision.stanford.edu/∼birch/headtracker/.