Adaptive Object Tracking Based on an Effective Appearance Filter

Hanzi Wang, *Member*, *IEEE*,
David Suter, *Senior Member*, *IEEE*,
Konrad Schindler, *Member*, *IEEE*, and
Chunhua Shen

Abstract—We propose a similarity measure based on a Spatial-color Mixture of Gaussians (SMOG) appearance model for particle filters. This improves on the popular similarity measure based on color histograms because it considers not only the colors in a region but also the spatial layout of the colors. Hence, the SMOG-based similarity measure is more discriminative. To efficiently compute the parameters for SMOG, we propose a new technique with which the computational time is greatly reduced. We also extend our method by integrating multiple cues to increase the reliability and robustness. Experiments show that our method can successfully track objects in many difficult situations.

Index Terms—Particle filters, mixture of Gaussians, appearance model, similarity measure, color histogram, visual tracking, occlusion.

1 Introduction

A tracked object can be located by maximizing the similarity measure between a reference window and a candidate window. The maximum can be realized through either a deterministic way or a stochastic way.

Deterministic methods localize the tracked object in each frame by iteratively searching for a region which maximizes the similarity measure between this region and the target window. For example, Comaniciu et al. [3] employed the Mean Shift algorithm for object tracking. These methods are computationally efficient. However, the methods may converge to a local maximum: They are sensitive to background distractors, clutter, occlusions, and quick moving objects. These problems can be mitigated by stochastic methods which maintain multiple hypotheses in the state space and, in this way, achieve more robustness to the local maximum.

Among various stochastic methods, Particle Filters (PF) [4], [6] are very successful. Particle filters are simple, robust, and effective and have obtained success in many challenging tasks. Particle filters simultaneously track multiple hypotheses and recursively approximate the posterior probability density function (pdf) in the state space with a set of random sampled particles. Both the appearance model and the similarity measure are very important to the performance of particle filters. The particles are weighted according to a similarity measure (i.e., the observation likelihood function) and the sampling of particles is also dependent on the similarity measure.

Indeed, the effectiveness and the robustness of the similarity measure greatly affect the performance of both deterministic and stochastic methods. This paper focuses on developing an effective observation model and similarity measure in the context of improving the performance of particle filters. The key contributions of this paper can be summarized as follows:

1. an effective SMOG appearance model and a SMOG-based similarity measure,
2. a new technique for efficiently computing the parameters of SMOG,
3. a new shape similarity measure, and
4. a complete SMOG tracking algorithm.

Experiments and comparisons with several other popular methods show that our method can achieve very promising performance in tracking objects in the presence of challenging situations including fast movements, clutter, scaling, changing appearance, color distractors, and occlusions.

2 The Proposed Method

The histogram-based pdf modeling methods [9], [10] are often preferred because of their simplicity and robustness to scaling and rotation. However, the similarity measure based on color histograms (e.g., Bhattacharyya coefficient) is often not discriminative enough—see Fig. 1. In this section, we propose a SMOG appearance model and a SMOG-based similarity measure.

2.1 Object Representation with SMOG

We represent an object O (which may be distinguished as the target object O_τ or the target candidate O_v) by modeling the spatial-color joint probability distribution of the corresponding region with a mixture of Gaussians. We define $\boldsymbol{S}_i \equiv (\mathrm{x}_i, \mathrm{y}_i)$ to be the spatial feature (i.e., the 2D coordinates) and $\boldsymbol{C}_i \equiv \{\mathrm{C}_i^{\mathrm{j}}\}_{\mathrm{j}=1,\dots,d}$ to be the color feature with d color channels at pixel x_i. We employ the normalized (r, g, I) color space as the color feature in our case, where $r = \mathrm{R}/(\mathrm{R}+\mathrm{G}+\mathrm{B})$; $g = \mathrm{G}/(\mathrm{R}+\mathrm{G}+\mathrm{B})$; $\mathrm{I} = (\mathrm{R}+\mathrm{G}+\mathrm{B})/3$. We can write the features of $\boldsymbol{x}_i$ as the Cartesian product of its spatial position and the color: $\boldsymbol{x}_i \equiv (\boldsymbol{S}_i, \boldsymbol{C}_i) = (\mathrm{x}_i, \mathrm{y}_i, r_i, g_i, \mathrm{I}_i)$. We assume that the spatial feature (S) and the color feature (C) are independent of each other. For the mean and the covariance of the $l\mathrm{th}$ mode of the Gaussian Mixtures at time t, we have $\mu_{l,t} = (\mu_{l,t}^{\mathrm{S}}, \mu_{l,t}^{\mathrm{C}})$ and

$$\Sigma_{l,t} = \begin{pmatrix} \Sigma_{l,t}^{\mathrm{S}}, & 0 \\ 0, & \Sigma_{l,t}^{\mathrm{C}} \end{pmatrix}.$$

We represent both the target object O_τ and the target candidate O_v by a mixture of Gaussians with k modes in a joint spatial-color space

$$\begin{aligned} \hat{\boldsymbol{O}}_{\tau,t} &= \{\hat{N}_{l,t}^{O_\tau}\}_{l=1,\dots,k}, \hat{N}_{l,t}^{O_\tau} \equiv N\Big(\omega_{l,t}^{O_\tau}, \mu_{l,t}^{\mathrm{S},O_\tau}, \mu_{l,t}^{\mathrm{C},O_\tau}, \Sigma_{l,t}^{\mathrm{S},O_\tau}, \Sigma_{l,t}^{\mathrm{C},O_\tau}\Big); \\ \hat{\boldsymbol{O}}_{v,t} &= \{\hat{N}_{l,t}^{O_v}\}_{l=1,\dots,k}, \hat{N}_{l,t}^{O_v} \equiv N\Big(\omega_{l,t}^{O_v}, \mu_{l,t}^{\mathrm{S},O_v}, \mu_{l,t}^{\mathrm{C},O_v}, \Sigma_{l,t}^{\mathrm{S},O_v}, \Sigma_{l,t}^{\mathrm{C},O_v}\Big), \end{aligned} \tag{1}$$

where $\omega_{l,t}$ is the weight of the $l\mathrm{th}$ mode at time t and $\sum_{l=1}^{k} \omega_{l,t} = 1$.

The joint spatial-color density estimate at the point $\boldsymbol{x}_i$ is formulated as

$$\hat{p}_t(\boldsymbol{x}_i) \equiv \hat{p}_t(\boldsymbol{S}_i, \boldsymbol{C}_i) = \sum_{l=1}^{k} \omega_{l,t} p(\boldsymbol{S}_i | l^{\mathrm{S}}) p(\boldsymbol{C}_i | l^{\mathrm{C}}), \tag{2}$$

$$\begin{aligned} \text{where} : p(\boldsymbol{S}_i | l^{\mathrm{S}}) &= \frac{\exp\Big\{-\frac{1}{2}(\boldsymbol{S}_i - \mu_{l,t}^{\mathrm{S}})^T (\Sigma_{l,t}^{\mathrm{S}})^{-1} (\boldsymbol{S}_i - \mu_{l,t}^{\mathrm{S}})\Big\}}{2\pi|\Sigma_{l,t}^{\mathrm{S}}|^{1/2}}; \\ p(\boldsymbol{C}_i | l^{\mathrm{C}}) &= \frac{\exp\Big\{-\frac{1}{2}(\boldsymbol{C}_i - \mu_{l,t}^{\mathrm{C}})^T (\Sigma_{l,t}^{\mathrm{C}})^{-1} (\boldsymbol{C}_i - \mu_{l,t}^{\mathrm{C}})\Big\}}{(2\pi)^{d/2}|\Sigma_{l,t}^{\mathrm{C}}|^{1/2}}. \end{aligned} \tag{3}$$

- *H. Wang is with the Department of Computer Science, The Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218. E-mail: hwang@cs.jhu.edu.*
- *D. Suter and K. Schindler are with the Department of Electrical and Computer Systems Engineering, Monash University, Clayton, Victoria 3800, Australia. E-mail: {d.suter, konrad.schindler}@eng.monash.edu.au.*
- *C. Shen is with National ICT Australia, Locked Bag 8001, Canberra ACT 2601, Australia. E-mail: chunhua.shen@nicta.com.au.*

Manuscript received 9 May 2006; revised 18 Oct. 2006; accepted 16 Jan. 2007; published online 12 Feb. 2007.
Recommended for acceptance by S. Baker.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0362-0506.
Digital Object Identifier no. 10.1109/TPAMI.2007.1112.

0162-8828/07/$25.00 © 2007 IEEE Published by the IEEE Computer Society

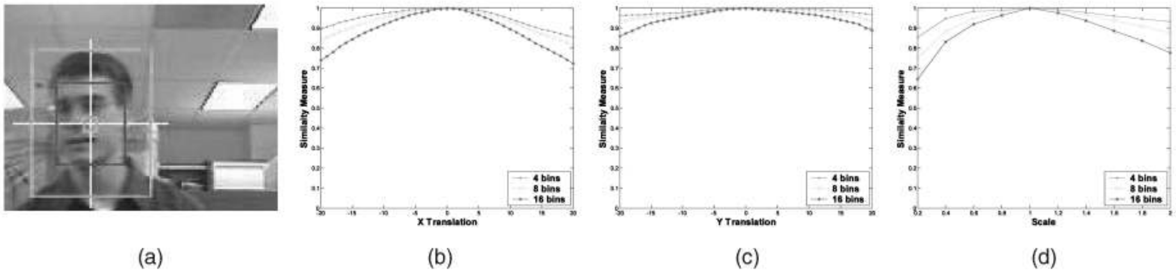


Fig. 1. Color-histogram-based similarity measure. (a) We choose the face (within the middle rectangle with a size of 33×39 pixels) as the target model. The score of the similarity measure over (b) x-translation (in pixels), (c) y-translation (in pixels), and (d) scaling (the ratio to the size of target). When we increase the number of histogram bins, the discriminative ability of the similarity measure slightly improves. However, it still obtains similar scores for different candidate regions even when the number of bins is set to $16 \times 16 \times 16$. (See text and compare with Fig. 2.)

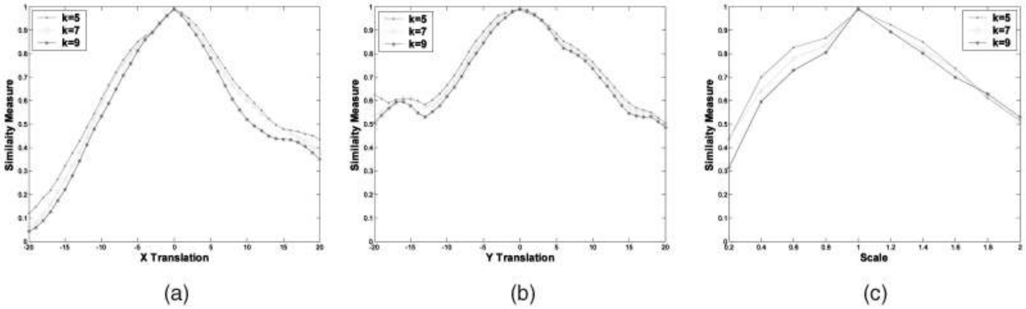


Fig. 2. The scores by the SMOG-based similarity measure with different k values over (a) x-translation (in pixels), (b) y-translation (in pixels), and (c) scale (the ratio to the size of target).

2.2 Similarity Measure Based on SMOG

Our similarity measure is constructed by comparing the *ordered* modes of the Mixture of Gausssian representations of target and candidates—the ordering of modes is discussed in Section 2.4. We denote by

$$\Phi^{\mathrm{S}}_{l,t}\left(\hat{N}^{O_\tau}_{l,t}, \hat{N}^{O_v}_{l,t}\right)$$

and

$$\Phi^{\mathrm{C}}_{l,t}\left(\hat{N}^{O_\tau}_{l,t}, \hat{N}^{O_v}_{l,t}\right),$$

respectively, the spatial match measure and the color match measure between the lth modes of the target object O_τ and the target candidate O_v at time t. We have:

$$\Phi^{\mathrm{S}}_{l,t}\left(\hat{N}^{O_\tau}_{l,t}, \hat{N}^{O_v}_{l,t}\right) = \exp\left\{-\frac{1}{2}(\mu^{\mathrm{S},O_v}_{l,t} - \mu^{\mathrm{S},O_\tau}_{l,t})^T(\hat{\Sigma}^{\mathrm{S}^*}_{l,t})^{-1}(\mu^{\mathrm{S},O_v}_{l,t} - \mu^{\mathrm{S},O_\tau}_{l,t})\right\}, \tag{4}$$

where $\left(\hat{\Sigma}^{\mathrm{S}^*}_{l,t}\right)^{-1} = \left(\Sigma^{\mathrm{S},O_v}_{l,t}\right)^{-1} + \left(\Sigma^{\mathrm{S},O_\tau}_{l,t}\right)^{-1}$.

$$\Phi^{\mathrm{C}}_{l,t}\left(\hat{N}^{O_\tau}_{l,t}, \hat{N}^{O_v}_{l,t}\right) = \min\left(\omega^{O_v}_{l,t}, \omega^{O_\tau}_{l,t}\right). \tag{5}$$

In terms of this, we define the SMOG-based similarity function between the target object O_τ and the target candidate O_v in the joint spatial-color space as

$$\begin{aligned}\Phi_{\mathrm{SMOG}}\left(\hat{O}_{\tau,t}, \hat{O}_{v,t}\right) &= \sum_{l=1}^{k} \Phi^{\mathrm{S}}_{l,t}\Phi^{\mathrm{C}}_{l,t} \\ &= \sum_{l=1}^{k} \exp\left\{-\frac{1}{2}\left(\mu^{\mathrm{S},O_v}_{l,t} - \mu^{\mathrm{S},O_\tau}_{l,t}\right)^T\left(\hat{\Sigma}^{\mathrm{S}^*}_{l,t}\right)^{-1}\right. \\ &\quad \left.\left(\mu^{\mathrm{S},O_v}_{l,t} - \mu^{\mathrm{S},O_\tau}_{l,t}\right)\right\}\min\left(\omega^{O_v}_{l,t}, \omega^{O_\tau}_{l,t}\right).\end{aligned} \tag{6}$$

The SMOG-based likelihood function in our method is then written as

$$\mathcal{L}_{\mathrm{SMOG}}(\mathrm{Y}_{color,t}|\mathrm{X}_t) \propto \exp\left\{-\frac{1}{2\sigma_b^2}(1 - \Phi_{\mathrm{SMOG}}(\hat{O}_{\tau,t}, \hat{O}_{v,t}))\right\}, \tag{7}$$

where X_t and $\mathrm{Y}_{color,t}$ is, respectively, the hidden target state and the image observation at time t. σ_{b} is the observation variance and we experimentally set it to 0.2 in our case. For the influence of the σ_{b} value on the results of particle filters, we refer to [8] for more details.

Spatial information has been utilized in previous work [5], [7], [13], [18]. The proposed method is different from these methods. Compared to [5], [13], we use a modified MOG (i.e., SMOG), instead of the kernel density technique, to estimate the spatial-color joint density distribution. Our method is different from [18] and [7] in that: 1) Neither background modeling nor background subtraction procedures are used in our method. 2) In our method, tracking is driven by both the region appearance of target candidates and the region structure. In [18] and [7], tracking is driven at pixel level. 3) We apply our method in the framework of particle filters, where multiple hypotheses are simultaneously tracked.

Our method is also different from the Spatiograms method in [2]. Instead of computing the spatial mean and covariance of *each bin*, we compute the spatial layout and color distributions of *each mode* of SMOG. Thus, our method is more efficient in computation and needs less storage space.

2.3 Demonstration of Discriminability

In Fig. 2, we test two properties of the proposed similarity measure: 1) its discriminative ability and 2) the influence of the k value (i.e., the number of modes or Gaussians) in SMOG on the results. We repeat the experiment in Fig. 1. From Fig. 2, we can see

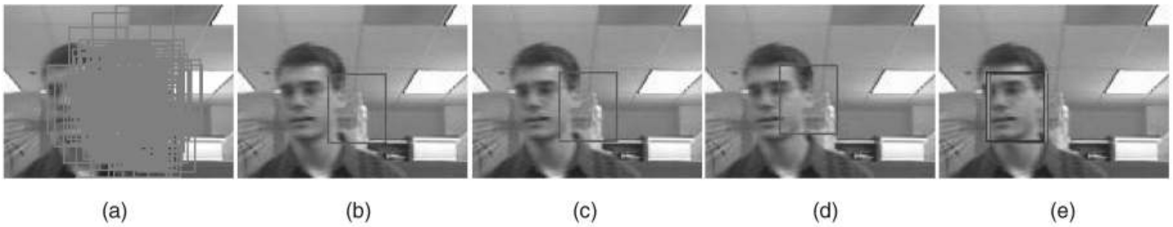


Fig. 3. An example of tracking a face with a weak dynamic model: (a) 100 particles configurations out of the 5,000 particles. (b), (c), and (d) The results using the color histogram-based similarity measure with (b) $4 \times 4 \times 4$ bins, (c) $8 \times 8 \times 8$ bins, and (d) $16 \times 16 \times 16$ bins. (e) The results for the proposed method with $k = 5$, $k = 7$, and $k = 9$.

that: 1) The proposed similarity measure is more discriminative than the color-histogram-based similarity measure and 2) SMOG is more effective in representing the target object than the general color histogram. Even if we use only five Gaussians (i.e., $k = 5$) in SMOG, the similarity measure is more discriminative than that based on the color histogram with a total 4,096 bins (Fig. 1), and 3) when we increase the k value of SMOG from 5 to 9, the proposed method becomes relatively more discriminative (however, the difference is slight).

It is desirable to develop a robust tracking method that can work effectively with weak dynamic models. In Fig. 3, we track the face in Fig. 1a using a weak dynamic model (i.e., we employ a first order AR model [9], [20] despite the fast movement of the face). Fig. 3 shows that the proposed method can work well with a weak dynamic model and outperform the color histogram-based method.

2.4 Calculating and Updating the Parameters of SMOG

Given a region corresponding to the target object O_τ, we initialize the parameters of SMOG *for the target object* by a standard K-means algorithm (with k_0 modes—$k_0 = 7$ in our experiments in Section 6) followed by an EM algorithm. The k_0 modes are sorted by the ratio of the weight of each mode to its standard variance. We choose the first k modes so that

$$k = \arg\min_{k'} \left(\sum_{l=1}^{k'} \omega_{l,1}^{O_\tau} > T_\tau \right),$$

where T_τ is a constant value (e.g., 0.8). The reason that we use k modes out of the k_0 modes ($k \leq k_0$) is that we are more interested in the modes having the most supporting evidence and the least variance. Once we initialize the k modes for the target object, we maintain and update the k modes of the target object in each frame. Similarly to the target, each target *candidate* is represented by k modes and we maintain the ordering of these modes by assigning candidate pixels to the target modes and using the assigned pixels (only) to calculate the parameters for the corresponding candidate mode. Specifically, the parameters of a target candidate can be calculated in the following way:

1. Calculate the Mahalanobis distances of each pixel $\boldsymbol{x}_i$ in the target candidate region O_v to the k modes of SMOG of the target object O_τ in the color space:

$$\hat{\mathcal{D}}_l^2\left(\boldsymbol{C}_i, \hat{N}_{l,t}^{C,O_\tau}\right) = \left(\boldsymbol{C}_i - \mu_{l,t}^{C,O_\tau}\right)^T \left(\Sigma_{l,t}^{C,O_\tau}\right)^{-1} \left(\boldsymbol{C}_i - \mu_{l,t}^{C,O_\tau}\right). \quad (8)$$

2. Assign every pixel $\boldsymbol{x}_i$ to one of the k modes of SMOG:

$$\mathrm{L_B}^*(\mathrm{x_i}) = arg\min_l \left|\hat{\mathcal{D}}_l\right|. \quad (9)$$

The function $\mathrm{L_B}^* : R^2 \rightarrow \{1, \ldots, k\}$ associates to the pixel $\boldsymbol{x}_i$ the index $\mathrm{L}^*(\mathrm{x_i})$ of the k modes.

3. Label all pixels $\{\boldsymbol{x}_i\}_{i=1,\ldots,N}$ as follows:

$$\mathrm{L_B}(\mathrm{x_i}) = \begin{cases} \mathrm{L_B}^*(\mathrm{x_i}) & if\left|\hat{\mathcal{D}}_{\mathrm{L}^*}(\mathrm{x_i})\right| \leq 2.5 \\ 0 & Otherwise. \end{cases} \quad (10)$$

The value 2.5 is used so that 98 percent of a Gaussian distribution (i.e., a mode) is identified as inliers.

4. Calculate the parameters of the target candidate O_v:

$$\begin{aligned} \omega_{l,t}^{O_v} &= \left(\sum_{i=1}^{N} \delta(\mathrm{L_B}(\boldsymbol{x}_i) - l) \right) \Big/ \left(\sum_{l=1}^{k} \sum_{i=1}^{N} \delta(\mathrm{L_B}(\boldsymbol{x}_i) - l) \right) \\ \mu_{l,t}^{O_v} &= \left(\mu_{l,t}^{\mathrm{S},O_v}, \mu_{l,t}^{\mathrm{C},O_v}\right) = \left(\sum_{i=1}^{N} \boldsymbol{x}_i \delta(\mathrm{L_B}(\boldsymbol{x}_i) - l) \right) \Big/ \left(\sum_{i=1}^{N} \delta(\mathrm{L_B}(\boldsymbol{x}_i) - l) \right) \\ \Sigma_{l,t}^{O_v} &= \left(\Sigma_{l,t}^{\mathrm{S},O_v}, \Sigma_{l,t}^{\mathrm{C},O_v}\right) \\ &= \left(\sum_{i=1}^{N} \left(\boldsymbol{x}_i - \mu_{l,t}^{O_v}\right)^T \left(\boldsymbol{x}_i - \mu_{l,t}^{O_v}\right) \delta(\mathrm{L_B}(\boldsymbol{x}_i) - l) \right) \Big/ \left(\sum_{i=1}^{N} \delta(\mathrm{L_B}(\boldsymbol{x}_i) - l) \right), \end{aligned} \quad (11)$$

where δ is the Kronecker delta function. We normalize the coordinate space so that the coordinates of every pixel in the target candidate are within the range of [0, 1].

Similarly to [21] and [14], we assume that the target appearance $\hat{\boldsymbol{O}}_{\tau,t}$ is exponentially forgotten and new information $\hat{\boldsymbol{O}}_t^*$ is gradually added to the appearance model. Here, we use an adaptive learning rate γ_t, the value of which is proportional to the similarity measure $\Phi_{\mathrm{SMOG}}(\hat{\boldsymbol{O}}_{\tau,t}, \hat{\boldsymbol{O}}_t^*)$. We handle occlusion in a heuristic way: We update the appearance only if the value of $\Phi\left(\hat{\boldsymbol{O}}_{\tau,t}, \hat{\boldsymbol{O}}_t^*\right)$ is larger than a threshold value $\mathrm{T_u}$ (e.g., 0.7); otherwise, we stop updating the appearance model.

3 Improving the Computational Efficiency

When we work in the framework of particle filters, the most expensive part in implementing our method is to evaluate the similarity measure in (6). This is because the regions corresponding to the set of particles may have many overlapped areas. The same calculation steps in Section 2.4 may be repeated many times over these areas. Recently, some solutions (e.g., [12], [15]) to overcome the computational inefficiency of such situations have been proposed. In our case, we construct a new feature space using the rectangle features.

Given a target model $\hat{\boldsymbol{O}}_\tau$, we need to calculate the parameters $\{\omega_l^{O_v}, \mu_l^{\mathrm{S},O_v}, \Sigma_l^{\mathrm{S},O_v}\}_{l=1,\ldots,k}$ of a target candidate to evaluate the similarity measure in (6). We can write $(\omega_l^{O_v}, \mu_l^{\mathrm{S},O_v}, \Sigma_l^{\mathrm{S},O_v})$ as follows:

$$\omega_l^{O_v} = n_l \Big/ \sum_{l=1}^{k} n_l; \mu_l^{\mathrm{S},O_v} = (\mu_l^x, \mu_l^y); \Sigma_l^{\mathrm{S},O_v} = \begin{pmatrix} (\sigma_l^{\mathrm{x}})^2 & 0 \\ 0 & (\sigma_l^{\mathrm{y}})^2 \end{pmatrix}, \quad (12)$$

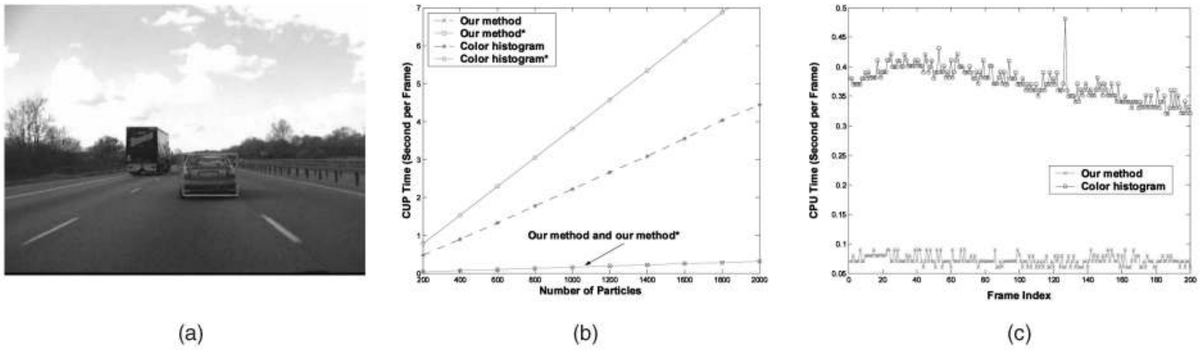


Fig. 4. Evaluation of the computational time per frame (in MATLAB code): (a) A frame. (b) Computational time versus the number of particles and the region size, (c) versus the frame index.

where

$$
\begin{aligned}
n_l &= \sum_{i=1}^{N} \delta(\mathrm{L_B}(\boldsymbol{x}_i) - l);\ \mu_l^x = \left(\sum_{i=1}^{N} \mathrm{x}_i \delta(\mathrm{L_B}(\boldsymbol{x}_i) - l) \right) \Big/ n_l;\\
\mu_l^y &= \left(\sum_{i=1}^{N} \mathrm{y}_i \delta(\mathrm{L_B}(\boldsymbol{x}_i) - l) \right) \Big/ n_l;\\
\sigma_l^{\mathrm{x}} &= \sqrt{\left(\sum_{i=1}^{N} \mathrm{x}_i^2 \delta(\mathrm{L_B}(\boldsymbol{x}_i) - l) \right) \Big/ n_l - (\mu_l^x)^2};\\
\sigma_l^{\mathrm{y}} &= \sqrt{\left(\sum_{i=1}^{N} \mathrm{y}_i^2 \delta(\mathrm{L_B}(\boldsymbol{x}_i) - l) \right) \Big/ n_l - (\mu_l^y)^2}.
\end{aligned}
\tag{13}
$$

In order to avoid repeating the same operators over the overlapped regions of the set of particles, we construct an image $\Theta = \{\chi_i\}_{i=1,\dots,N^*}$ for each mode (corresponding to each label). We write the features of the pixel χ_i in the Θ image as:

$$
\boldsymbol{\chi}_i = \{\mathrm{L_B}(\boldsymbol{x}_i^*), \mathrm{x}_i^*, \mathrm{y}_i^*, (\mathrm{x}_i^*)^2, (\mathrm{y}_i^*)^2\}, \tag{14}
$$

where $\mathrm{L_B}(\boldsymbol{x}_i^*)$ is the label at the pixel x_i^*; x_i^* and y_i^* are, respectively, the x and y coordinate values of that pixel x_i^* in the region R^*.

The detailed procedure of the algorithm is given as follows:

- **Predicting** a region R^*, which covers all regions of the target candidates in the 2D image.
- **Labeling** all pixels $\{x_i^*\}_{i=1,\dots,N^*}$ in the region R^* by Steps 1 to 3 in Section 2.4.
- **Building** an integral Θ image for each label. For each pixel χ_i' in the integral Θ image $\{\boldsymbol{\chi}_i'\}_{i=1,\dots,N^*}$ we have $\chi_i' = \sum_{\mathrm{x}_i^* \le \mathrm{x}_i' \wedge \mathrm{y}_i^* \le \mathrm{y}_i'} \chi_i$, where x_i' and y_i' are respectively the x and y coordinate values of the pixel $\boldsymbol{\chi}_i'$.
- **Calculating** the parameters of target candidates by four table lookup operations, similar to [15].

Fig. 4 shows a rough estimation of the computational time of the proposed method. We implement our method with $k = 9$ and the color histogram-based method using $16 \times 16 \times 16$ bins. In Fig. 4b, we increased the number of particles from 200 to 2,000. We also double the size of the tracked region (those results are marked with a "*"). As we can see, a major advantage of our method compared with the color histogram-based method is that its computational complexity is not significantly influenced by the number of particles or the size of the target candidate regions. Fig. 4c shows that our method can process around 15-20 frames per second (we use 200 particles for the video sequence with image size 384×288 pixels).

4 Fusing Multiple Cues

The SMOG tracking method proposed in the previous section works effectively in most situations. However, it may work poorly when the target appearance experiences greatly changes. Edges (or contours) are robust to the changes of illuminations. However, one problem with the edge features is that images with heavily cluttered backgrounds can lead to a high rate of false alarm. This problem can be significantly reduced by fusing the color appearance with the edge features [1], [11], [19], [20]. Next, we propose a new shape similarity measure considering three features of edge points: their spatial distribution, their gradient intensity, and the size of edge points.

In order to represent the spatial distributions of the object shape, we use a spatial histogram $\hat{\mathbf{H}} = \{\hat{\hbar}_u\}_{u=1,\dots,h}$ with h bins ($h = 16$ in our case) for the edge points along the hypothesized contour (the rectangle in our case). Let $\varsigma : \mathrm{R}^2 \to \{1, \dots, h\}$ be the function which associates to the edge pixel at location $x_i^\dagger$ in the edge image a number $\varsigma(x_i^\dagger)$ corresponding to the index of the histogram bin. The probability of the edge points falling into the uth bin of the spatial histogram can be written as:

$$
\hat{\hbar}_u = \sum_{i=1}^{m} \delta\Big(\varsigma(x_i^\dagger) - u\Big) / \sum_{u=1}^{h} \sum_{i=1}^{m} \delta\Big(\varsigma(x_i^\dagger) - u\Big), \tag{15}
$$

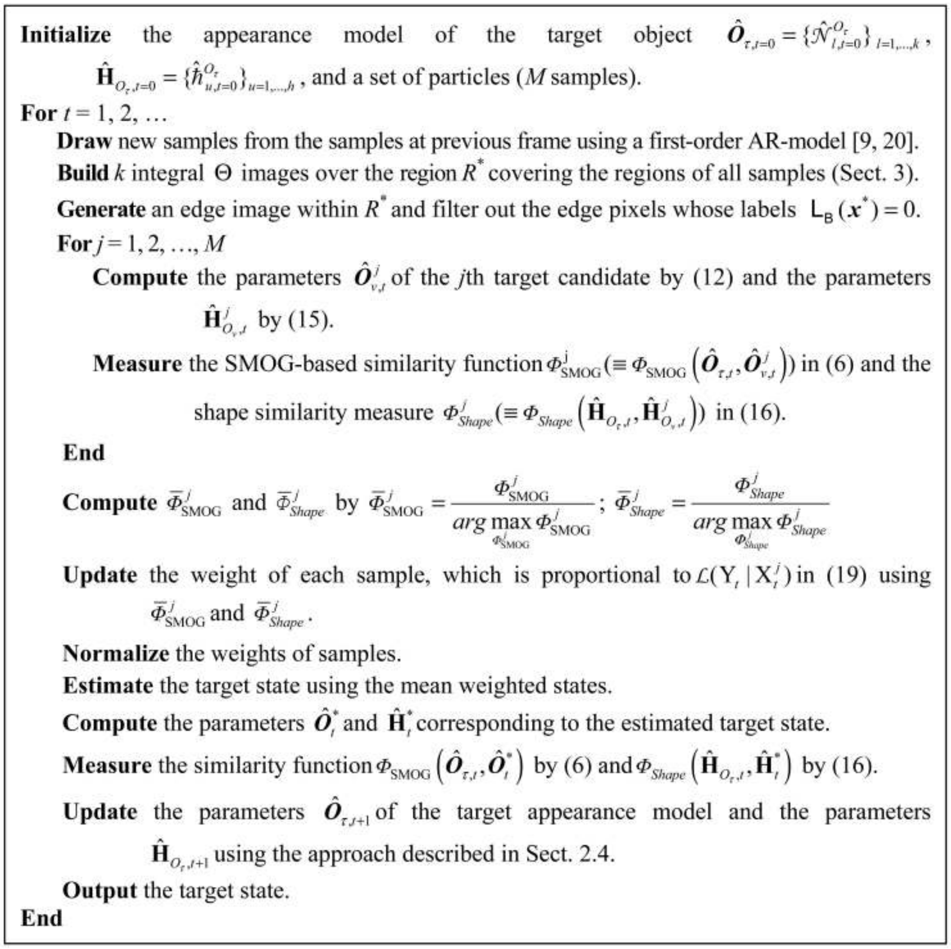
where m is the number of edge points around the object contour and $\sum_{u=1}^{h} \hat{\hbar}_u = 1$.

Let Ω_τ and Ω_v be two sets of the edge points along the contours of the target object O_τ and a target candidate O_v and let $G(x^\dagger)$ be the gradient intensity at the edge point $x^\dagger$. The shape similarity measure between O_τ and O_v is formulated as:

$$
\Phi_{\scriptscriptstyle Shape}\left(\hat{\mathbf{H}}_{O_\tau}, \hat{\mathbf{H}}_{O_v}\right) = \sum_{u=1}^{h} \Big[\min\Big(\hat{\hbar}_u^{O_\tau}, \hat{\hbar}_u^{O_v}\Big) \psi_u^N(O_\tau, O_v) \psi_u^G(O_\tau, O_v) \Big], \tag{16}
$$

where $\psi_u^N(O_\tau, O_v)$ and $\psi_u^G(O_\tau, O_v)$ are, respectively, the match function between the uth bin of O_t and O_v in the data size and the gradient intensity. We can write them as:

$$
\psi_u^N(O_\tau, O_v) = \frac{\min\left(\sum\limits_{x^\dagger \in \Omega_\tau} \delta\Big(\varsigma(x^\dagger) - u\Big), \sum\limits_{x^\dagger \in \Omega_v} \delta\Big(\varsigma(x^\dagger) - u\Big) \right)}{\max\left(\sum\limits_{x^\dagger \in \Omega_\tau} \delta\Big(\varsigma(x^\dagger) - u\Big), \sum\limits_{x^\dagger \in \Omega_v} \delta\Big(\varsigma(x^\dagger) - u\Big) \right)}, \tag{17}
$$



```
Initialize the appearance model of the target object $\hat{\boldsymbol{O}}_{\tau,t=0} = \{\hat{\mathcal{N}}^{O_\tau}_{l,t=0}\}_{l=1,\dots,k}$,
    $\hat{\mathbf{H}}_{O_\tau,t=0} = \{\hat{h}^{O_\tau}_{u,t=0}\}_{u=1,\dots,h}$, and a set of particles ($M$ samples).
For $t = 1, 2, \dots$
  Draw new samples from the samples at previous frame using a first-order AR-model [9, 20].
  Build $k$ integral $\Theta$ images over the region $R^*$ covering the regions of all samples (Sect. 3).
  Generate an edge image within $R^*$ and filter out the edge pixels whose labels $\mathsf{L}_\mathsf{B}(\boldsymbol{x}^*) = 0$.
  For $j = 1, 2, \dots, M$
    Compute the parameters $\hat{\boldsymbol{O}}^j_{v,t}$ of the $j$th target candidate by (12) and the parameters
        $\hat{\mathbf{H}}^j_{O_v,t}$ by (15).
    Measure the SMOG-based similarity function $\Phi^j_{\mathrm{SMOG}} (\equiv \Phi_{\mathrm{SMOG}}(\hat{\boldsymbol{O}}_{\tau,t}, \hat{\boldsymbol{O}}^j_{v,t}))$ in (6) and the
        shape similarity measure $\Phi^j_{Shape} (\equiv \Phi_{Shape}(\hat{\mathbf{H}}_{O_\tau,t}, \hat{\mathbf{H}}^j_{O_v,t}))$ in (16).
  End
  Compute $\bar{\Phi}^j_{\mathrm{SMOG}}$ and $\bar{\Phi}^j_{Shape}$ by $\bar{\Phi}^j_{\mathrm{SMOG}} = \frac{\Phi^j_{\mathrm{SMOG}}}{arg \max_{\Phi^j_{\mathrm{SMOG}}} \Phi^j_{\mathrm{SMOG}}}$; $\bar{\Phi}^j_{Shape} = \frac{\Phi^j_{Shape}}{arg \max_{\Phi^j_{Shape}} \Phi^j_{Shape}}$
  Update the weight of each sample, which is proportional to $\mathcal{L}(\mathrm{Y}_t \mid \mathrm{X}^j_t)$ in (19) using
      $\bar{\Phi}^j_{\mathrm{SMOG}}$ and $\bar{\Phi}^j_{Shape}$.
  Normalize the weights of samples.
  Estimate the target state using the mean weighted states.
  Compute the parameters $\hat{\boldsymbol{O}}^*_t$ and $\hat{\mathbf{H}}^*_t$ corresponding to the estimated target state.
  Measure the similarity function $\Phi_{\mathrm{SMOG}}(\hat{\boldsymbol{O}}_{\tau,t}, \hat{\boldsymbol{O}}^*_t)$ by (6) and $\Phi_{Shape}(\hat{\mathbf{H}}_{O_\tau,t}, \hat{\mathbf{H}}^*_t)$ by (16).
  Update the parameters $\hat{\boldsymbol{O}}_{\tau,t+1}$ of the target appearance model and the parameters
      $\hat{\mathbf{H}}_{O_\tau,t+1}$ using the approach described in Sect. 2.4.
  Output the target state.
End
```

Fig. 5. The complete SMOG tracking algorithm.

$$\psi_u^G(O_\tau, O_v) = \frac{\min\left(\sum_{x^\dagger \in \Omega_\tau} G(x^\dagger)\delta\left(\varsigma(x^\dagger)-u\right) \Big/ \sum_{x^\dagger \in \Omega_\tau} \delta\left(\varsigma(x^\dagger)-u\right), \sum_{x^\dagger \in \Omega_v} G(x^\dagger)\delta\left(\varsigma(x^\dagger)-u\right) \Big/ \sum_{x^\dagger \in \Omega_v} \delta\left(\varsigma(x^\dagger)-u\right)\right)}{\max\left(\sum_{x^\dagger \in \Omega_\tau} G(x^\dagger)\delta\left(\varsigma(x^\dagger)-u\right) \Big/ \sum_{x^\dagger \in \Omega_\tau} \delta\left(\varsigma(x^\dagger)-u\right), \sum_{x^\dagger \in \Omega_v} G(x^\dagger)\delta\left(\varsigma(x^\dagger)-u\right) \Big/ \sum_{x^\dagger \in \Omega_v} \delta\left(\varsigma(x^\dagger)-u\right)\right)}. \quad (18)$$

The joint observation likelihood function is formulated as

$$\mathcal{L}(\mathrm{Y}_t|\mathrm{X}_t) = \mathcal{L}_{SMOG}(\mathrm{Y}_{color,t}|\mathrm{X}_t)\mathcal{L}_{Shape}(\mathrm{Y}_{edge,t}|\mathrm{X}_t), \quad (19)$$

where the likelihood function of shape $L_{Shape}(\mathrm{Y}_{edge,t}|\mathrm{X}_t)$ can be written as

$$\mathcal{L}_{Shape}(\mathrm{Y}_{edge,t}|\mathrm{X}_t) \propto \exp\left\{-\frac{1}{2\sigma_b^2}\left(1 - \Phi_{Shape}\left(\hat{\mathbf{H}}_{O_\tau}, \hat{\mathbf{H}}_{O_v}\right)\right)\right\}. \quad (20)$$

Although we consider the size and the gradient intensity of the edge points in the shape similarity measure (16), the values of the spatial histogram $\{\hat{h}_u\}_{u=1,\dots,h}$ in the target template are adaptive and are updated at each frame. As shown in the experiments in Section 6, our method is robust to both scaling and changes of gradient intensity.

5 The Complete Algorithm

In the previous sections, we have developed all of the ingredients for a robust SMOG tracker. Now, we put them together to yield a complete tracking algorithm (as detailed in Fig. 5).

6 Experiments

In Section 6.1, we compare the proposed method with two popular color histogram-based methods: the Mean Shift (MS) tracker and the Condensation (C) tracker. For the purpose of comparisons between the techniques, we evaluate the proposed method using the SMOG-based color cue only (which we refer to as SMOG1). In Section 6.2, we show the performance of the proposed SMOG tracker using multiple cues (called SMOG2). Quantitative experimental comparisons of the four methods (MS, C, SMOG1, and SMOG2) are given in Section 6.3.

6.1 Using the SMOG-Based Color Cue

We test the behavior of the three trackers in the face of a weak dynamic model, appearance changes, scaling, color distractor, and occlusions. In Fig. 6a, the human face is moving to the left and right very quickly. The first-order AR dynamic model is weak and not well suited to this. Both the MS tracker and the C tracker do not achieve a good result: The MS tracker fails to track the face very soon; the results of Condensation are not accurate. In comparison, our method never loses the target and achieves the most accurate results. The relative ability of the three methods to handle appearance changes and scaling is demonstrated in Fig. 6b. A man rotates his face from one side to another. Both the appearance and the scale of the head keep changing throughout the sequence. We use variable scale for the three methods. Our method achieves the most accurate tracking results and scale estimation. Both the MS tracker and the C tracker perform less accurate in tracking and scale estimation. In Fig. 6d (an 80 frame subsequence of the *girl* sequence), two human faces with very similar colors get close to each other and

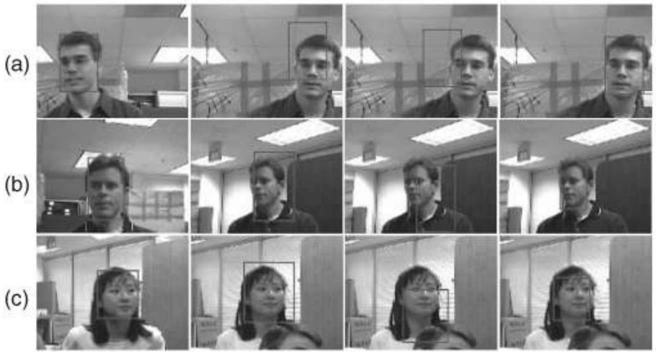


Fig. 6. Tracking results with the MS tracker (second column), the C tracker (third column), and SMOG1 (fourth column). The first column shows the initial target state. For the MS and C trackers, we use $16 \times 16 \times 16$ color histogram bins. For the C and SMOG1 trackers, we use the first-order AR dynamic model [9], [20].

one occludes the other. We use variable scale for all methods. We can see that, when the man's face gets close to and occludes the girl's face, the tracking results with the MS tracker and the C tracker are greatly influenced by the man's face. Our method achieves the best accuracy in both tracking and scale estimation.

6.2 Using Multiple Cues

In this subsection, we will examine the performance of SMOG using both the color and edge cues. Fig. 7 demonstrates the capability of SMOG2. In Fig. 7a, SMOG2 shows the robustness to the changing appearance and resists the influence of the floor (similar color to the person's body) and successfully tracks the person despite the variable illumination, scaling, and occlusion. Fig. 7b shows another advantage of SMOG2. We use 500 frames of the *girl* sequence, which contains rotation, scaling, variable illumination, color distractor, and occlusions. The MS, the C, and the SMOG1 trackers cannot track the girl's face throughout the whole sequence. Only SMOG2 succeeds. In Fig. 7c, we test our method using a challenging video sequence (a *football* sequence). This sequence includes high clutter and color distractors. The fast moving head of a player is tracked (the first-order AR model is weak for such motion) and the appearance of the head changes frequently (including rotation, occlusion, blurring, and changes in the spatial and color distributions of the appearance). SMOG2 succeeds in tracking the head of the sportsman throughout the video sequence.

6.3 Quantitative Performance

Table 1 shows the results of the four methods on 30 video clips, covering challenging difficulties including changing appearance (e.g., rotation, illumination changes, etc.), scaling, weak dynamic model, and occlusions (including both partial and complete occlusions). The length varies from 29 to 500 frames (the average is about 120 frames). A lost track is declared when the centroid of the estimated state does not fall into the target region. We use a variable scale for all methods on the clips with scaling and a fixed scale on the rest of the clips.

Overall, SMOG2 achieves the best results (Table 1). SMOG(1/2) outperform the MS and the C trackers in terms of the number of successful tracks: SMOG2 in 27 out of 30 clips, SMOG1 24 clips, while C succeeds in 13, and MS does in 9. MS is not good at handling occlusion and scaling. The C tracker is slightly better in handling occlusion and scaling, but worse than SMOG(1/2). MS and C are more easily affected by color distractors. When the video includes fast movements, neither MS nor C achieves good results: C succeeds in one out of three clips; MS never succeeds. In contrast, SMOG(1/2) succeed in tracking objects in all the three clips.

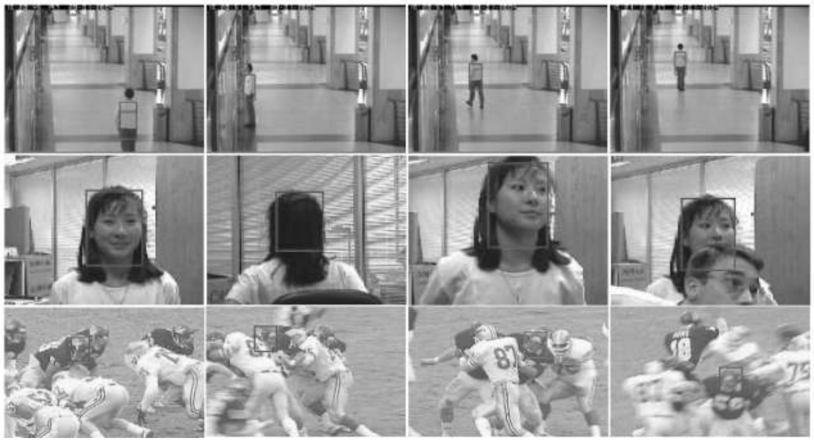


Fig. 7. The tracking results with SMOG2 (a) First row: frames 1, 160, 315, and 400; (b) second row: frames 1, 215, 387, and 453; (c) third row: frames 1, 38, 59, and 73.

TABLE 1
Testing Results of the Four Methods on a Data Set Consisting of 30 Clips

Clip types	O	AS	AW	AO	SO	ASO	ASWO	Total
Number of Clips	5	6	2	5	2	9	1	30
Average frames	58	89	51	116	100	200	73	119
MS	2	3	0	1	2	1	0	9
C	4	3	1	1	1	3	0	13
SMOG1	5	4	2	4	2	6	1	24
SMOG2	5	4	2	4	2	9	1	27

The fourth to the seventh rows show the number of the clips correctly tracked by the corresponding methods. (A: appearance changes. S: scaling. W: weak dynamic model. O: occlusion.).

There are some cases where SMOG(1/2) fail. All failure modes involve several difficult conditions simultaneously. For example, SMOG2 fails in three clips out of 30 testing clips: (first failure) rotation and camera zooming at the same time; (second failure) a man rotates his head with large scaling and with clothes of similar color to his face; (third failure) occlusion and a color distractor occur simultaneously. However, all tested trackers fail on these clips.

7 Conclusion

We have described an effective appearance model (SMOG) in a joint spatial-color space, and an associated similarity measure. The SMOG appearance model and similarity measure consider both the spatial distributions and the color information of objects. Hence, SMOG more effectively represents objects and the SMOG-based similarity measure is more discriminative than the color histogram-based appearance model and similarity measure. We also propose a new technique which greatly improves the computational efficiency of our method with which the number of particles and the size of target candidate region can be greatly increased without significant change in the processing time of the proposed method. We also propose a new *shape* similarity measure which considers the spatial distribution of the size of, and the corresponding gradient intensities of the edge points and we integrate this shape similarity measure with the SMOG-based color similarity measure.

Experiments on a variety of video sequences show SMOG (using color cue only or multiple cues) achieves very promising results in handling a variety of challenges.

Acknowledgments

Parts of the work described in this paper appeared previously in [16], [17]. The authors thank the anonymous reviewers of those papers, as well as the *TPAMI* reviewers, for the valuable comments and suggestions. Hanzi Wang was supported by Australia Research Council grant DP0452416. This work was done while Hanzi Wang was with the Department of Electrical and Computer Systems Engineering, Monash University, Clayton, Victoria 3800, Australia.

References

[1] S. Birchfield, "Elliptical Head Tracking Using Intensity Gradients and Color Histograms," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 232-237, 1998.

[2] S. Birchfield and S. Rangarajan, "Spatiograms versus Histograms for Region-Based Tracking," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1152-1157, 2005.

[3] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-Based Object Tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 5, pp. 564-577, May 2003.

[4] A. Doucet, S. Godsill, and C. Andrieu, "On Sequential Monte Carlo Sampling Methods for Bayesian Filtering," *Statistics and Computing,* vol. 10, no. 3, pp. 197-208, 2000.

[5] A. Elgammal, R. Duraiswami, and L.S. Davis, "Probabilistic Tracking in Joint Feature-Spatial Spaces," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 781-788, 2003.

[6] M. Isard and A. Blake, "Condensation-Conditional Density Propagation for Visual Tracking," *Int'l J. Computer Vision,* vol. 29, no. 1, pp. 5-28, 1998.

[7] S. Khan and M. Shah, "Tracking People in Presence of Occlusion," *Proc. Asian Conf. Computer Vision,* pp. 263-266, 2000.

[8] J. Lichtenauer, M. Reinders, and E. Hendriks, "Influence of the Observation Likelihood Function on Particle Filtering Performance in Tracking Applications," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 767-772, 2004.

[9] K. Nummiaro, E. Koller-Meier, and L.V. Gool, "An Adaptive Color-Based Particle Filter," *Image and Vision Computing,* vol. 21, pp. 99-110, 2003.

[10] P. Perez et al., "Color-Based Probabilistic Tracking," *Proc. European Conf. Computer Vision,* pp. 661-675, 2002.

[11] P. Pérez, J. Vermaak, and A. Blake, "Data Fusion for Visual Tracking with Particles," *Proc. IEEE,* vol. 92, no. 3, pp. 495-513, 2004.

[12] F. Porikli, "Integral Histogram: A Fast Way to Extract Histograms in Cartesian Spaces," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 829-836, 2005.

[13] Y. Sheikh and M. Shah, "Bayesian Modeling of Dynamic Scenes for Object Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 11, pp. 1778-1792, Nov. 2005.

[14] C. Stauffer and W.E.L. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 246-252, 1999.

[15] P. Viola and M. Jones, "Robust Real-Time Face Detection," *Int'l J. Computer Vision,* vol. 52, no. 2, pp. 137-154, 2004.

[16] H. Wang and D. Suter, "Efficient Visual Tracking by Probabilistic Fusion of Multiple Cues," *Proc. Int'l Conf. Pattern Recognition,* pp. 892-895, 2006.

[17] H. Wang, D. Suter, and K. Schindler, "Effective Appearance Model and Similarity Measure for Particle Filtering and Visual Tracking," *Proc. European Conf. Computer Vision,* pp. 606-618, 2006.

[18] C.R. Wren et al., "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 780-785, July 1997.

[19] Y. Wu and T.S. Huang, "Robust Visual Tracking by Integrating Multiple Cues Based on Co-Inference Learning," *Int'l J. Computer Vision,* vol. 58, no. 1, pp. 55-71, 2004.

[20] C. Yang, R. Duraiswami, and L. Davis, "Fast Multiple Object Tracking via a Hierarchical Particle Filter," *Proc. Int'l Conf. Computer Vision,* pp. 212-219, 2005.

[21] S. Zhou, R. Chellappa, and B. Moghaddam, "Visual Tracking and Recognition Using Appearance-Adaptive Models in Particle Filters," *IEEE Trans. Image Processing,* vol. 11, pp. 1434-1456, 2004.