Two-stage real-time head tracking by an active camera based on multimodal information

Dong-Gil Jeong
Dong-Goo Kang
Yu-Kyoung Yang
Jong Beom Ra, MEMBER SPIE
Korea Advanced Institute of Science and Technology
Department of Electrical Engineering and Computer Science
371-1 Guseongdong, Yuseonggu
Daejeon, Korea
E-mail: jbra@ee.kaist.ac.kr

Abstract. We propose a two-stage real-time tracking algorithm for an active camera system having pan-tilt-zoom functions. The algorithm is based on the assumption that a human head has an elliptical shape and that its model color histogram has been acquired in advance. The algorithm consists of two stages, a color-based convergence stage for fast and reliable tracking and a refinement stage for accurate tracking based on multimodal information. In the first color convergence stage, we roughly estimate the target position by using the mean-shift method based on the histogram similarity between the model and a candidate ellipse. To better predict the initial position for the mean shift, the global motion is compensated; to enhance reliability of the mean shift, the model histogram is appropriately updated by referring to the target histogram in the previous frame. In the subsequent refinement stage, we refine the position and size of the ellipse obtained at the first stage by using multimodal information such as color, shape, and quasi-spatial information. In particular, to quantify the quasi-spatial information, we use a spatial color histogram obtained by properly dividing the ellipse into two regions. Extensive experiments verify that the proposed algorithm robustly tracks the head, even when the subject moves quickly, the head size changes drastically, or the background has many clusters and/or distracting colors. Also, the proposed algorithm can perform real-time tracking with a processing speed of about 10 fps on a standard PC.
© *2006 Society of Photo-Optical Instrumentation Engineers.* [DOI: 10.1117/1.2354452]

Subject terms: real-time head tracking; multimodal information; spatial color histogram; mean-shift-based tracking; active camera.

Paper 050688RR received Aug. 24, 2005; revised manuscript received Feb. 19, 2006; accepted for publication Feb. 28, 2006; published online Oct. 5, 2006.

1 Introduction

Real-time person tracking has many application areas such as security systems, video conferences, human-computer interaction (HCI), virtual reality, etc. Head tracking is useful for tracking a person, because head color does not vary much among people and the shape is relatively rigid relative to other parts of the body. Also, head tracking can be used as the initial phase for the identification of a person, facial expression recognition, etc. This work deals with head tracking as an application for a real-time active camera system. The adopted active camera is assumed to have pan-tilt-zoom functions to cover a large viewing area. Hence, to control pan-tilt-zoom operation reliably, we need robust and accurate tracking of the head target.

In previous works, three types of information, namely, motion, shape, and/or color, have mainly been used for target tracking. In the methods using motion information,[1–5] a target is extracted by discriminating its motion from the background. These methods assume that the target always moves and background objects are stationary. However, in an active camera system, since the camera's pan-tilt-zoom operations introduce global motion of translation and scaling, accurate estimation of the global motion vectors is required for reliable tracking. However, accurate motion estimation is a difficult and time-consuming task. Meanwhile, a method using shape information[6] first constructs several representative shape models from many sample datasets of the subject's outline. It then predicts a current outline from the previous ones by using a Kalman filter and attempts to refine the predicted outline based on the models by using the predicted outline and the edge information at the current frame. However, this method is not reliable for complicated backgrounds, because the background region may include distracting strong edges. Birchfield has combined both color and shape information to improve the accuracy of head tracking.[7] The shape of the head is defined as an ellipse. The target similarity is examined by using a weighted sum of the color histogram similarity inside the ellipse and the shape similarity based on gradients on the ellipse boundary. The algorithm provides better performance than those using either shape or color information only. However, it may not be adequate for high performance real-time tracking, since a full search is performed within a wide area for finding the position and scale that maximizes the combined similarity function.

Comaniciu has proposed fast tracking algorithms using color information only.[8,9] The algorithms use the Bhattacharyya coefficient[10] as a similarity measure between two color distributions, and adopt the mean shift[11] as a fast optimization method to maximize the similarity. To improve the accuracy of mean-shift-based tracking, Zhang

0091-3286/2006/$22.00 © 2006 SPIE

Downloaded from SPIE Digital Library on 17 Apr 2011 to 143.248.224.33. Terms of Use: http://spiedl.org/terms

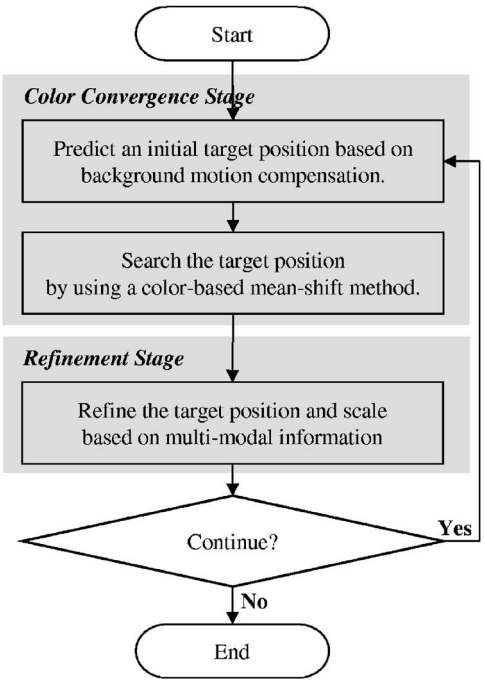


Fig. 1 Flowchart of the proposed real-time tracking algorithm.

and Freedman utilized background mismatching as well as foreground matching with a model color histogram.[12] However, the color histogram adopted in the algorithm represents the global distribution of colors rather than the local (or spatial) distribution. Hence, tracking can be unreliable when distracting colors are included in the nearby background region. Also, the scale change of the target cannot be estimated accurately, because it uses color information only. Note that the existing color-based algorithms use only the initially obtained color histogram of the model, which may not faithfully reflect the color distribution change due to the view-point transition.[7–9,12] However, a direct temporal update of the model or the use of information in previous frames can cause error propagation, if the previous frame includes unwanted outliers.

In this work, we propose a real-time head tracking algorithm, which consists of two stages, a color-based convergence stage for fast and reliable tracking and a refinement stage for accurate tracking based on multimodal information. In Sec. 2, we describe the proposed algorithm in detail. Experimental results are shown in Sec. 3, and we conclude this work in Sec. 4.

2 Proposed Algorithm

The proposed algorithm assumes that a head can be modeled as an ellipse and the ratio of its major and minor axes is constant. At the first frame, a user determines the initial position and scale of the ellipse manually or semiautomatically. The optimal ellipse is then automatically tracked in the following frames by using the results of the previous frame. The proposed two-stage algorithm is presented in Fig. 1. The first color-based convergence stage consists of two steps. Namely, we first predict an initial position for the mean shift by compensating the global motion, and then roughly search the target position by examining the best

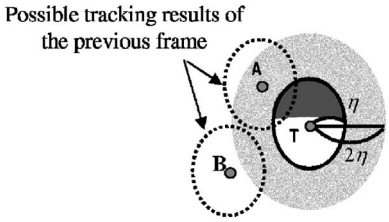


Fig. 2 The region of convergence (ROC), which is represented as a gray ellipse. Convergence to the target is ensured if the tracking result of the previous frame is inside the ROC. Otherwise, the convergence is not guaranteed.

match in terms of color. To ensure fast convergence, the mean shift is adopted as an optimization method. To enhance reliability of the mean shift, we combine the model histogram and the histogram of the previous target ellipse in a robust manner. In the refinement stage, we refine the position and find the scale of the target. To enhance the refinement accuracy, we consider multiple modalities such as color, shape, and quasi-spatial information, and suggest a similarity measure combining them. In particular, we define a spatial color histogram characterizing the quasi-spatial color information, and we devise a reliable measure to quantify shape similarity. This tracking procedure is finished if a user interrupts it or the target is out of screen. The two steps in the color-based convergence stage and the refinement stage are described in detail in the following.

2.1 *Prediction Step in the Color-Based Convergence Stage*

In the color-based convergence stage, we use the mean-shift method to search the optimal position maximizing the color similarity measure. Even though it has been demonstrated that the mean-shift converges,[13] the convergence to the true target is not guaranteed. Hence, we define the "region of convergence" (ROC) such that an initial position inside the region may guarantee the convergence to the true position for the mean shift. Since the ROC consists of the centers of all candidate ellipses that overlap with the target, it becomes an elliptical shape whose radius is twice larger than that of the target ellipse, and the center position is the same as that of the target. Figure 2 illustrates an example of ROC. In contrast with the candidate ellipse (or initial position) A, ellipse B cannot be ensured to converge to the target, because its center is outside the ROC.

When an active camera is used for tracking, the center of the estimated ellipse is frequently outside the ROC. This is due to the large global motion on the image domain that is induced by the change of camera movement. Hence, we model the global motion and compensate it so that the initial position may locate inside the ROC. Since a pan-tilt camera with a fixed axis produces rotational motion only, the corresponding motion field on the consecutive frames can be described using the 2-D parametric model,[14] i.e.,

$$u = -f\Omega_Y + \beta_x x + \frac{\Omega_X}{f}xy - \frac{\Omega_Y}{f}x^2, \tag{1}$$

Downloaded from SPIE Digital Library on 17 Apr 2011 to 143.248.224.33. Terms of Use: http://spiedl.org/terms

$$v = f\Omega_X + \beta_y y + \frac{\Omega_X}{f} y^2 - \frac{\Omega_Y}{f} xy, \tag{2}$$

where f, Ω_X, Ω_Y, and β are the focal length, tilting angle difference, panning angle difference, and zooming factor, respectively. Since this global motion model is used for the rough estimation of the center of the ellipse, we may simplify the model to the following four parameter model by assuming that points (x,y) locate near the image center.

$$u = \alpha_x + \beta_x x, \tag{3}$$

$$v = \alpha_y + \beta_y y, \tag{4}$$

where α_x and α_y denote horizontal and vertical translations, respectively. To estimate the parameters in real time, we propose a projection-based motion estimation scheme. In this scheme, we first project the intensity values of pixels along the vertical and horizontal directions and obtain two sets of projection data for a frame, i.e.,

$$P_i^X(m) = \sum_{n=1}^{N_h} I_i(m,n), \tag{5}$$

$$P_i^Y(n) = \sum_{m=1}^{N_w} I_i(m,n), \tag{6}$$

where $I_i(m,n)$ is the i'th frame of $N_w \times N_h$ pixels. We subsequently extract two segments of a fixed length from each projection data of the previous frame, and then find their best matches in the corresponding projection data of the current frame. For reliable matching, two nonoverlapped segments having the largest variances are selected in the projection data, excluding a projected region of the target ellipse. If X_1 and X_2 (Y_1 and Y_2) denote the centers of two segments in dataset P_i^X (P_i^Y), the corresponding motion vectors MVx_1 and MVx_2 (MVy_1 and MVy_2) can be searched through 1-D matching of the segments to the projection data P_{i-1}^X (P_{i-1}^Y). Then, the motion parameters can be described as

$$\alpha_x = \frac{X_1 MVx_2 - X_2 MVx_1}{X_1 - X_2}, \tag{7}$$

$$\alpha_y = \frac{Y_1 MVy_2 - Y_2 MVy_1}{Y_1 - Y_2}, \tag{8}$$

$$\beta_x = 1 + \frac{MVx_1 - MVx_2}{X_1 - X_2}, \tag{9}$$

$$\beta_y = 1 + \frac{MVy_1 - MVy_2}{Y_1 - Y_2}. \tag{10}$$

Using these motion parameters, the initial position for the mean shift can be compensated as

$$t_x' = \alpha_x + \beta_x t_x, \tag{11}$$

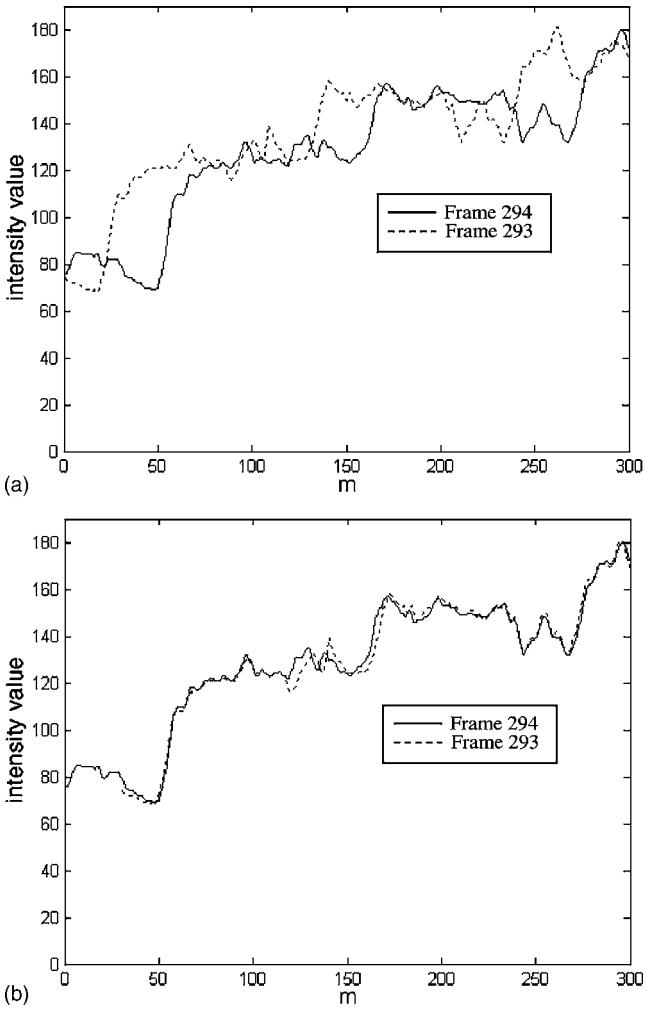


Fig. 3 Two projection data along the horizontal direction, (a) before and (b) after global motion compensation for two consecutive frames of the *Walking* sequence. Dotted and solid lines in the graphs denote the projection data of frames 293 and 294, respectively. The values of α and β are 30.71 and 1.02, respectively.

$$t_y' = \alpha_y + \beta_y t_y, \tag{12}$$

where (t_x, t_y) is the center of the target estimated in the previous frame, and its initial value is determined by a user at the beginning of tracking. Figures 3(a) and 3(b) show the projection data along the horizontal direction in frames 293 and 294 of the *Walking* sequence before and after background motion compensation, respectively. Figure 4 shows a compensation result in a frame.

2.2 *Searching Step in the Color-Based Convergence Stage*

2.2.1 *Color-based mean shift*

Starting from the initial position obtained earlier, the mean shift is performed to maximize the color similarity measure. We adopt the color histogram as the color feature of the target. Generally, chrominance components have been used for the color histogram because they are less sensitive to light changes. However, it is known that the luminance component also has much information regarding an object.

Downloaded from SPIE Digital Library on 17 Apr 2011 to 143.248.224.33. Terms of Use: http://spiedl.org/terms

Fig. 4 Initial position for mean shift that is adjusted by the background motion compensation in frame 294 of the *Walking* sequence. Black and white ellipses represent the initial position obtained before and after background motion compensation, respectively. The black ellipse resides outside the ROC, described by a dashed line.

Given this, Birchifield introduced new color coordinates, B−G, G−R, for chrominance and R+B+G for luminance,[7] and generated a color histogram by allocating 8, 8, and 4 bins to them, respectively. We adopt this color histogram, as it has proved to provide better performance than others.[15]

For color-based searching, we obtain the model color histogram in advance. We then search the parameters (position and scale) of the ellipse that maximize the similarity between its color histogram and the model color histogram. Mean shift is an iterative optimization method based on nonparametric kernel density estimation.[16] The objective function in the mean-shift method has a specific form:

$$\hat{f}_{h,k}(\mathbf{y}) = \frac{c_{k,d}}{Nh^d}\sum_{i=1}^{N} \omega_i k\left(\left\|\frac{\mathbf{y}-\mathbf{x}_i}{h}\right\|^2\right), \tag{13}$$

where $k(\cdot)$ denotes a convex and monotonic decreasing kernel profile, $c_{k,d}$ is the normalization constant that makes $\hat{f}_{h,k}(\mathbf{y})$ integrate to one, h is the bandwidth of the kernel, d is the dimension of the sample, N is the number of pixels that reside inside the confined kernel, $\mathbf{y}$ is a candidate position, $\mathbf{x}_i$ is the position of the i'th sample, and ω_i is the weighting factor at $\mathbf{x}_i$, respectively.

To adapt the color histogram similarity measure to the function given in Eq. (13), a color histogram is first represented as a weighted kernel profile. Namely, the probability of the u'th bin in the weighted color histogram of $\mathbf{y}$ can be given as

$$\hat{p}_u(\mathbf{y}) = C_h\sum_{i=1}^{N} k\left(\left\|\frac{\mathbf{y}-\mathbf{x}_i}{h}\right\|^2\right)\delta[b(\mathbf{x}_i)-u], \tag{14}$$

where $\delta(\cdot)$ is the Kronecker delta function, $b(\mathbf{x}_i)$ is the bin index of the color at $\mathbf{x}_i$, and C_h is a normalization constant, respectively. Note here that we use the Epanechnikov kernel as $k(\cdot)$ in this work.[9] We then adopt the Bhattacharyya coefficient[10] as a similarity measure between the model histogram $\hat{\mathbf{q}}$ and color histogram $\hat{\mathbf{p}}$, i.e.,

$$\rho[\hat{\mathbf{p}}(\mathbf{y}),\hat{\mathbf{q}}] = \sum_{u=1}^{M}[\hat{p}_u(\mathbf{y})\hat{q}_u]^{1/2}, \tag{15}$$

where $\hat{q}_u$ represents the probability of the u'th bin in the kernel-weighted model histogram $\hat{\mathbf{q}}$, and M is the number of bins in the histogram. Note that since the Bhattacharyya coefficient given in Eq. (15) has a value in the range of [0, 1], it makes the comparison easy. Substituting Eq. (14) into Eq. (15) and applying the Taylor expansion, Eq. (15) can be rewritten as[8]

$$\rho[\hat{\mathbf{p}}(\mathbf{y}),\hat{\mathbf{q}}] \approx \frac{1}{2}\sum_{u=1}^{M}[\hat{p}_u(\mathbf{y}_0)\hat{q}_u]^{1/2} + \frac{1}{2}\sum_{u=1}^{M}\hat{p}_u(\mathbf{y})\left[\frac{\hat{q}_u}{\hat{p}_u(\mathbf{y}_0)}\right]^{1/2}. \tag{16}$$

Since the first term in the right side of Eq. (16) is independent of $\mathbf{y}$, we can define the color histogram similarity (CHS) as follows.

$$\mathrm{CHS} \equiv \sum_{i=1}^{N}\omega_i k\left(\left\|\frac{\mathbf{y}-\mathbf{x}_i}{h}\right\|^2\right), \tag{17}$$

where

$$\omega_i = \sum_{u=1}^{M}\left[\frac{\hat{q}_u}{\hat{p}_u(\mathbf{y}_0)}\right]^{1/2}\delta[b(\mathbf{x}_i)-u]. \tag{18}$$

Note here that the CHS given in Eq. (17) is the same as that in Eq. (13), excluding the normalization constants. To determine the maximum value of CHS, Eq. (17) is applied to the mean-shift method. The kernel then moves recursively from the current location $\mathbf{y}_j$ to the next location $\mathbf{y}_{j+1}$ according to the following equation.

$$\mathbf{y}_{j+1} = \frac{\sum_{i=1}^{N}\omega_i\mathbf{x}_i g\left(\left\|\frac{\mathbf{y}_j-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{N}\omega_i g\left(\left\|\frac{\mathbf{y}_j-\mathbf{x}_i}{h}\right\|^2\right)} \quad j=1,2,\cdots, \tag{19}$$

where $g(\cdot)=-k'(\cdot)$.[17] Here, it is assumed that the derivative of $k(x)$ exists for all $x\in[0,\infty)$, except for a finite set of points.

2.2.2 *Color histogram update using adaptively shrunken ellipse*

Since the color distribution of a human head significantly varies according to the change of view position during tracking, the model histogram should be updated to reflect the temporal change. For this purpose, we may use the color information of the target ellipse in the previous frame. However, in conventional studies, the previous target histogram is rarely used, because it may introduce error accumulation due to the unwanted outlier in the previous target ellipse. Therefore, we consider how to effectively incorporate the histogram of the previous target region into the

Downloaded from SPIE Digital Library on 17 Apr 2011 to 143.248.224.33. Terms of Use: http://spiedl.org/terms

mean-shift frame work. If the size of the target ellipse is overestimated, the corresponding histogram may include unwanted colors that do not appear in the model histogram. Hence, to eliminate the unwanted color near the boundary, we shrink the ellipse appropriately. If we assume that the estimated and target ellipses are concentric, then

$$\eta' = r \times \eta, \tag{20}$$

where η and η' are the length of the minor axes of the estimated and shrunken ellipses of the previous frame, respectively, and r is a shrinking ratio.

The ratio r can be adaptively obtained by comparing the color histogram of the estimated ellipse with the model color histogram. Let us first assume that the bins corresponding to the colors in the band between the estimated and shrunken ellipses are zero valued in the model color histogram. If the color histogram of the shrunken (or target) ellipse is the same as the model histogram, the similarity (or the Bhattacharyya coefficient) between the two histograms is maximized. In this case, a portion, which belongs to the shrunken ellipse, of the histogram of the estimated ellipse is a scaled version of the model histogram $\hat{\mathbf{q}}$. Hence, the scale factor can be regarded as the sum of the probabilities that belong to the histogram portion. Meanwhile, since the histogram of the estimated ellipse is obtained after weighting a convex kernel to the ellipse, the scale factor or the sum of the probabilities can be also interpreted as the integrated value of the kernel over the shrunken ellipse. If we adopt the Epanechnikov kernel having a bandwidth of η, namely,

$$K_E(\xi) = \begin{cases} \dfrac{2}{\pi\eta^2}\left(1 - \dfrac{\xi^2}{\eta^2}\right) & \xi \leqslant \eta \\ 0 & \text{otherwise} \end{cases}, \tag{21}$$

as the kernel, its integration within a shrunken ellipse of η' becomes

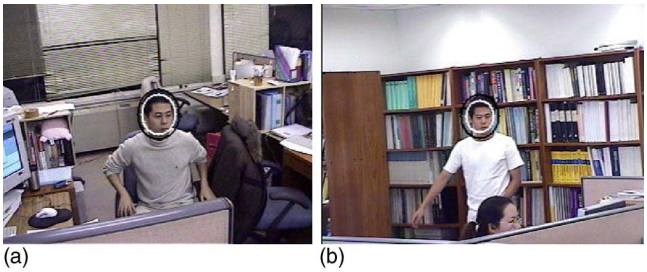
$$\int_{\rho=0}^{\eta'} \int_{\phi=0}^{2\pi} K_E(\xi) d\phi d\xi = \left(\frac{\eta'}{\eta}\right)^2 \left[2 - \left(\frac{\eta'}{\eta}\right)^2\right] = r^2(2 - r^2). \tag{22}$$

Then, by using Eq. (15), we can find that Bhattacharyya coefficient B between the color histogram of the estimated ellipse and the model color histogram has the following condition.

$$B = \sum_{u=1}^{M} \sqrt{\hat{p}_u \hat{q}_u} \leqslant \sum_{u=1}^{M} \sqrt{\hat{p}_u \hat{q}_u}\Big|_{\hat{p}_u = r^2(2-r^2)\hat{q}_u}$$
$$= \sum_{u=1}^{M} [r^2(2 - r^2)\hat{q}_u \hat{q}_u]^{1/2} = r\sqrt{2 - r^2}\rho[\hat{\mathbf{q}}, \hat{\mathbf{q}}] = r\sqrt{2 - r^2}. \tag{23}$$

Therefore,

$$r \geqslant [1 - \sqrt{1 - B^2}]^{1/2}. \tag{24}$$



(a) (b)

Fig. 5 Ellipse size adjustment for proper color histogram extraction. The white ellipse for the histogram extraction is determined by shrinking the black ellipse of the tracking result. (a) Frame 51 of the *Walking* sequence and (b) frame 14 of the *Clutter* sequence.

Based on Eq. (24), we may select the lower bound $[1 - \sqrt{1-B^2}]^{1/2}$ as the value of r for conservative estimation. Figure 5 shows an example of the shrunken ellipse obtained by using this value.

If we regard the color histogram obtained from the shrunken ellipse as the previous histogram, by using the previous and model histograms, the color histogram similarity defined in Eqs. (17) and (18) can be rewritten as

$$\mathrm{CHS} \equiv \sum_{i=1}^{N} \omega_i^{\mathrm{t}} k\left(\left\|\frac{\mathbf{y} - \mathbf{x}_i}{h}\right\|^2\right), \tag{25}$$

where

$$\omega_i^{\mathrm{t}} = \sum_{u=1}^{M} \left\{ \alpha\left[\frac{\hat{q}_u}{\hat{p}_u(\mathbf{y}_0)}\right]^{1/2} + (1-\alpha)\left[\frac{\hat{v}_u}{\hat{p}_u(\mathbf{y}_0)}\right]^{1/2} \right\} \delta[b(\mathbf{x}_i) - u]. \tag{26}$$

Here, $\hat{v}_u$ denotes the probability in the u'th bin of the previous histogram and α denotes a weighting factor, respectively. Based on the CHS given in Eqs. (25) and (26), the mean-shift method moves the kernel recursively from the current location $\mathbf{y}_j$ to the next location $\mathbf{y}_{j+1}$ according to the equation

$$\mathbf{y}_{j+1} = \frac{\sum_{i=1}^{N} \omega_i^{\mathrm{t}} \mathbf{x}_i}{\sum_{i=1}^{N} \omega_i^{\mathrm{t}}} \quad j = 0, 1, 2, \ldots . \tag{27}$$

2.3 *Refinement Stage*

In the previous color-based convergence stage, we roughly estimate the position of the current target ellipse. Now, in the refinement stage, we attempt to determine its accurate position and scale. To enhance the accuracy in the refinement, we introduce a reliable similarity measure by properly combining similarities of multimodal information such as color, spatial domain, and shape.

2.3.1 *Spatial color histogram similarity*

While the color histogram represents the global distribution of colors, it does not contain the information of the local distribution in the target. Thus, even though the color dis-

Downloaded from SPIE Digital Library on 17 Apr 2011 to 143.248.224.33. Terms of Use: http://spiedl.org/terms

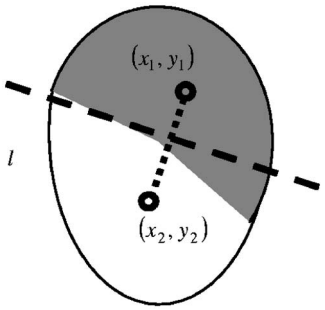


Fig. 6 Partitioning a candidate ellipse into two regions to obtain spatial color histograms.

tribution is clearly distinguishable between the target and the background in the spatial domain, the ellipse parameter values estimated on the basis of the color histogram may have errors if the background includes some colors similar to those in the model color histogram. To alleviate this problem, we introduce a simple but very effective spatial color histogram instead of the global color histogram by using the spatial dimension as well as the color dimension. The spatial color histogram was originally proposed with the goal of achieving more robust image retrieval.[18] In that work, the spatial color histogram was represented as the color distribution density along the annular or angular sectors (or bins). However, increasing the number of bins to obtain better spatial information may reduce the reliability of the similarity measurement, because this significantly decreases the number of samples in each bin, and the measurement becomes sensitive to even a small temporal change of the target appearance. Hence, to obtain a reliable spatial color histogram in our tracking problem, we divide the candidate ellipse into only two sectors, the hair and skin regions, by assuming that the head can be represented with the two dominant colors of the hair and skin. For simple and fast implementation, we may approximate the boundary between the two regions as a straight line, namely,

$$l(x,y) = (\sqrt{m} + \sqrt{n})[(x_2 - x_1)x + (y_2 - y_1)y] - (\sqrt{m}y_2 + \sqrt{n}y_1) \times (y_2 - y_1) + (\sqrt{m}x_2 + \sqrt{n}x_1)(x_2 - x_1) = 0, \quad (28)$$

where (x_1, y_1) and (x_2, y_2) are the gravity centers of pixels located in the hair and skin regions, respectively. m and n denote the number of pixels corresponding to the hair color and skin color, respectively. Note that in Fig. 6, the partitioning line is perpendicular to the line $\overline{(x_1,y_1)(x_2,y_2)}$, and the intersecting point of the two lines divides the line $\overline{(x_1,y_1)(x_2,y_2)}$ with a ratio of $\sqrt{m}/\sqrt{n}$.

The spatial color histogram can be represented in a form similar to the color histogram defined in Eq. (14). Namely, the probability in the u'th bin of the spatial color histogram is represented as

$$\hat{p}_u^s(\mathbf{y}) = C_h \sum_{i=1}^{N} k\left(\left\| \frac{\mathbf{y} - \mathbf{x}_i}{h} \right\|^2\right) \delta[b_s(\mathbf{y} - \mathbf{x}_i) - u], \quad (29)$$

where

$$b_s(\mathbf{y} - \mathbf{x}_i) = \begin{cases} M + b(\mathbf{x}_i) & l(x_{\mathbf{y}-\mathbf{x}_i}, y_{\mathbf{y}-\mathbf{x}_i}) \geq 0 \\ b(\mathbf{x}_i) & \text{otherwise} \end{cases}, \quad (30)$$

and $x_{\mathbf{y}-\mathbf{x}_i}$ and $y_{\mathbf{y}-\mathbf{x}_i}$ denote the x and y components of $\mathbf{y} - \mathbf{x}_i$, respectively. Note in Eq. (30) that the spatial color histogram is the concatenation of two color histograms extracted from different regions. Hence, the number of bins is twice that of the ordinary color histogram, or $2M$. We can now define the spatial-color-histogram similarity, SCHS$(\mathbf{y}, \eta)$, as a Bhattacharyya coefficient, i.e.,

$$\rho[\hat{\mathbf{p}}^s(\mathbf{y}), \hat{\mathbf{v}}^s] = \sum_{u=1}^{2M} [\hat{p}_u^s(\mathbf{y}) v_u^s]^{1/2}, \quad (31)$$

where $\hat{p}_u^s(\mathbf{y})$ and v_u^s denote the probability of the u'th bin in the spatial color histogram $(\hat{\mathbf{p}}^s)$ of the candidate ellipse at position $\mathbf{y}$, and the probability of the u'th bin in the spatial color histogram $(\hat{\mathbf{v}}^s)$ of the previous ellipse, respectively.

Figure 7 demonstrates that the use of the spatial color histogram can improve the tracking performance. In the figure, we compare the two best matching ellipses obtained by using the color histogram and the spatial color histogram, respectively. For fair comparison, we manually extract the accurate ellipse from the previous frame, and use it for both cases. A full search is commonly performed for the same white square range, as depicted in Figs. 7(a) and 7(b), so as to obtain the best matching position where the Bhattacharyya coefficient between the candidate histogram and the previous histogram is maximized. Contrary to the result in Fig. 7(b), Fig. 7(a) demonstrates that the ellipse can be distracted to the background region whose colors occupy a considerable portion of the model color histogram. Figures 7(c) and 7(d) depict the distribution of Bhattacharyya coefficients for both cases. When the spatial color histogram is used, the distribution curve in Fig. 7(d) becomes sharper near the optimal point; the target region is thereby fairly separated from the background, even though a significant portion of color components is common in both regions.

2.3.2 *Color histogram dissimilarity outside ellipse*

If both the histogram of the outerellipse region and the model histogram are similar, the current scale may not be optimal, and it should be enlarged to avoid underestimating the ellipse size. Therefore, the dissimilarity outside the candidate ellipse is an important measure to find a more reliable scale. We define this measure, which is called color histogram dissimilarity outside of the ellipse (CHDO), such that it may provide a negative value of the color similarity between the model histogram and the histogram of the outerellipse region. Here, the outerellipse region represents a band between the candidate ellipse and its concentric ellipse having 1.4 times larger size. CHDO$(\mathbf{y}, \eta)$ can be expressed as

Downloaded from SPIE Digital Library on 17 Apr 2011 to 143.248.224.33. Terms of Use: http://spiedl.org/terms

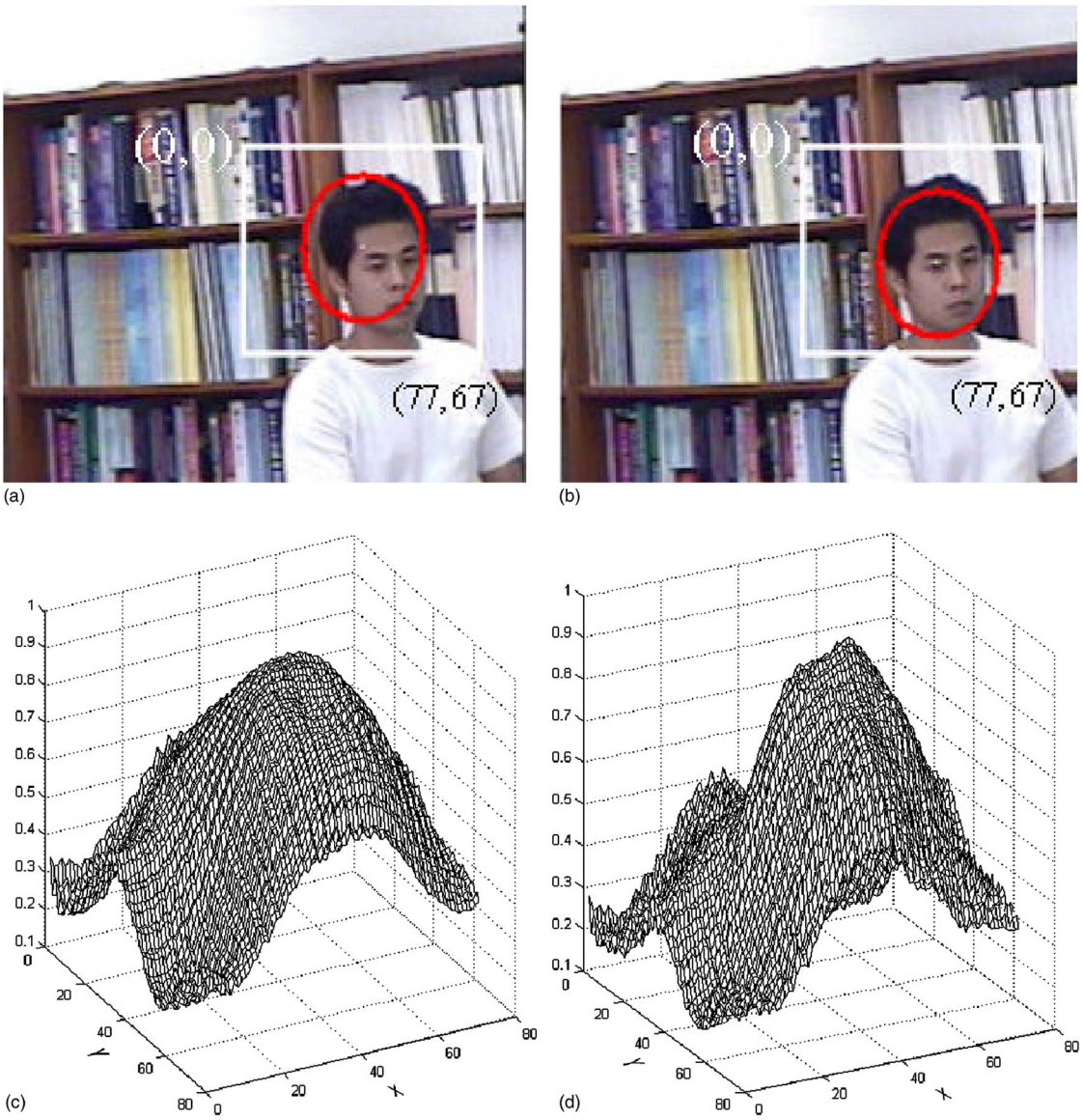


Fig. 7 (a) Tracking result by comparing color histograms of a candidate in the current frame and a user-selected tracking blob in the previous frame. (b) Tracking result by comparing spatial color histograms. (c) and (d) Bhattacharyya coefficient values on the search rage in the cases of (a) and (b), respectively. Frame 38 of the *Clutter* sequence is used to obtain the data.

$$\mathrm{CHDO}(\mathbf{y},\eta) \equiv 1-\rho[\mathbf{p}^{o}(\mathbf{y},\eta),\hat{\mathbf{q}}] = 1-\sum_{u=1}^{M}[p_u^{o}(\mathbf{y},\eta)\hat{q}_u]^{1/2}, \tag{32}$$

where $p_u^o(\mathbf{y},\eta)$ denotes the probability of the u'th bin in the outerellipse color histogram corresponding to a candidate $\mathbf{p}^o(\mathbf{y},\eta)$, and is defined as

$$p_u^{o}(\mathbf{y},\eta) = C_o\sum_{i=1}^{N_o}\delta[b(\mathbf{x}_i)-u]d(\|\mathbf{y}-\mathbf{x}_i\|,\eta), \tag{33}$$

where

$$d(\|\mathbf{y}-\mathbf{x}_i\|,\eta) = \begin{cases} 1 & \eta \leqslant \|\mathbf{y}-\mathbf{x}_i\| \leqslant 1.4\eta \\ 0 & \text{otherwise} \end{cases}. \tag{34}$$

N_o denotes the number of pixels in the range satisfying $d(\|\mathbf{y}-\mathbf{x}_i\|,\eta)=1$, and C_o is a normalizing constant.

2.3.3 *Shape similarity*

In addition to the spatial and color information, we use a shape feature in defining the similarity measure. A gradient is adopted to quantify the shape similarity between the candidate ellipse and the actual head boundary in the current frame. Birchfield quantified the shape similarity by summing gradient values on the ellipse boundary. However, this may not be desirable because the true boundary of a target does not have a precisely elliptical shape. In this work, we

Downloaded from SPIE Digital Library on 17 Apr 2011 to 143.248.224.33. Terms of Use: http://spiedl.org/terms

select a pixel that is close to the boundary and has a significantly large gradient value among those on the line segment perpendicular to the ellipse boundary. We then consider this pixel as an edge point, even though it does not locate exactly on the boundary. In this process, to avoid selecting a pixel having a strong gradient in the background as an edge point and to select a pixel even with a weak gradient on the target boundary, we attempt to minimize the effect of gradient magnitudes by binarizing them, and mainly consider the gradient orientation. Based on this consideration, the shape similarity (SS) measure of the candidate ellipse is defined at position $\mathbf{y}$ and scale η as

$$\mathrm{SS}(\mathbf{y},\eta)=\left[\frac{N_v(\mathbf{y},\eta)}{N_\sigma}\right]^2\cdot\frac{1}{N_\sigma}\times\sum_{i=1}^{N_\sigma}\{|\mathbf{g}[\mathbf{p}_i(\mathbf{y},\eta)]\cdot\mathbf{n}[\mathbf{e}_i(\mathbf{y},\eta)]|\}\left[1-\frac{d_i(\mathbf{y},\eta)}{d_{\mathrm{range}}+1}\right]. \tag{35}$$

Here, N_σ is the number of samples on the ellipse boundary and $\mathbf{e}_i$ denotes the i'th sample. $\mathbf{p}_i$ is the pixel closest to $\mathbf{e}_i$ among those located on the line perpendicular to the boundary at $\mathbf{e}_i$ and whose gradient value is larger than the predefined threshold TH_σ. d_i is the Euclidean distance between $\mathbf{p}_i$ and $\mathbf{e}_i$, $\mathbf{n}[\mathbf{e}_i(\mathbf{y},\eta)]$ is the unit normal vector at $\mathbf{e}_i$, and $\mathbf{g}[\mathbf{p}_i(\mathbf{y},\eta)]$ is the unit gradient vector at $\mathbf{p}_i$. Note here that $\mathbf{g}[\mathbf{p}_i(\mathbf{y},\eta)]$ is set to zero if $\mathbf{p}_i$ is not located within the range of d_{range}. Finally, $N_v(\mathbf{y},\eta)$ is the number of samples, where the value of $|\mathbf{g}[\mathbf{p}_i(\mathbf{y},\eta)]\cdot\mathbf{n}[\mathbf{e}_i(\mathbf{y},\eta)]|$ is larger than TH_σ.

2.3.4 *Parameter optimization*

By combining the three measures described before, namely, the spatial color histogram similarity, color histogram dissimilarity in outside of the ellipse, and shape similarity, we introduce a reliable similarity measure $\mathrm{RSM}(\mathbf{y},\eta)$ as follows.

$$\mathrm{RSM}(\mathbf{y},\eta)=w_1\cdot\mathrm{SCHS}(\mathbf{y},\eta)+w_2\cdot\mathrm{CHDO}(\mathbf{y},\eta)+w_3\cdot\mathrm{SS}(\mathbf{y},\eta), \tag{36}$$

where w_1, w_2, and w_3 denote the fixed weights and $w_1+w_2+w_3=1$. In this work, w_1, w_2, and w_3 are empirically determined as 0.375, 0.25, and 0.375 (or the ratio of 1.5:1.0:1.5), respectively, and are applied to all the experiments. Then, the parameters for the best matching ellipse $(\mathbf{y}_t,\eta_t)$ can be determined as

$$(\mathbf{y}_t,\eta_t)=\arg\max_{(\mathbf{y},\eta)\in R_s}\{\mathrm{RSM}(\mathbf{y},\eta)\}, \tag{37}$$

where R_s denotes the search range. Since the combined similarity measure usually produces a number of local maxima due to its complexity, the exhaustive search is needed for determining the optimal parameters.

3 Experimental Results and Discussions

To examine the performance of the proposed algorithm, we built an active camera system consisting of a pan-tilt unit (SDP-1600, Samsung) and a speed dome camera. The camera has an autofocusing function and provides an analog signal output of National Television System Committee (NTSC) format. The analog video output is converted to a digital signal by a frame grabber (DT3132, Data Translation) with a speed of 30 fps. We prepared three video sequences, *Jumping, Walking*, and *Clutter* sequences, having a frame resolution of 320×240 pixels. We use these sequences to verify the accuracy and robustness of the proposed algorithm relative to the conventional mean-shift-based algorithm, which is a simple hybrid form of Comaniciu's and Birchfield's methods. For all test sequences, we use the same parameter values. In the proposed algorithm, the parameter α in Eq. (26) is set to 0.2. And we use the same search ranges of 7×7 for position and 5 for scale in the refinement stages of both the conventional and proposed algorithms. Specifying a target ellipse at the beginning of tracking, we use the color histogram of that ellipse as a model histogram. Note here that the same ellipse is used at the beginning for the conventional and proposed algorithms.

Figure 8 shows the tracking result for the *Jumping* sequence, in which the target moves quickly along the vertical direction and also turns around. Since its target color histogram changes considerably, the model histogram obtained at the beginning may not be sufficient to provide accurate tracking. In this sense, the figure demonstrates that the proposed algorithm outperforms the conventional algorithm by additionally using the previous color histogram in the mean-shift procedure. We also notice that the conventional algorithm cannot follow the change of the head color histogram when the head rotates, while the proposed one follows the change faithfully, as clearly shown in the fourth row of Fig. 8(b). Figure 9 shows the tracking results for the *Walking* sequence, in which the head scale changes drastically. In this case, exact scale estimation is required during tracking for the zoom control of an active camera. Due to the advantage of fine scale refinement, the proposed algorithm provides a better scale estimation than the conventional one. In particular, unlike in the conventional algorithm, the proposed one covers the whole region of a head even in the frames with noticeable variations of head size. This is mainly because the ellipse searched in the color-based convergence stage is adjusted so as to increase the CHDO used in the refinement stage. In Fig. 10, we test the *Clutter* sequence, whose background has several colors similar to those in the model color histogram and includes many strong edges. Since the proposed algorithm effectively utilizes the quasi-spatial information and adopts the shape-and-color-based refinement step, it provides better performance for this sequence.

Figure 11 compares the tracking errors of the conventional method with those of the proposed method throughout each sequence. The graphs given in Figs. 11(a)–11(c) are obtained from the three sequences, *Jumping, Walking*, and *Clutter*, respectively. Here, scale and position errors constitute the differences between the tracking results and the manually defined results. And they are normalized by a ratio of the true scale to the reference scale to allow more error tolerance for a larger scale of the head. To obtain the data in the graphs, we heuristically set the reference scale to 10 pixels. Note that Figs. 11(a)–11(c) correspond to the results shown in Figs. 8–10, respectively. In the graphs of

Downloaded from SPIE Digital Library on 17 Apr 2011 to 143.248.224.33. Terms of Use: http://spiedl.org/terms

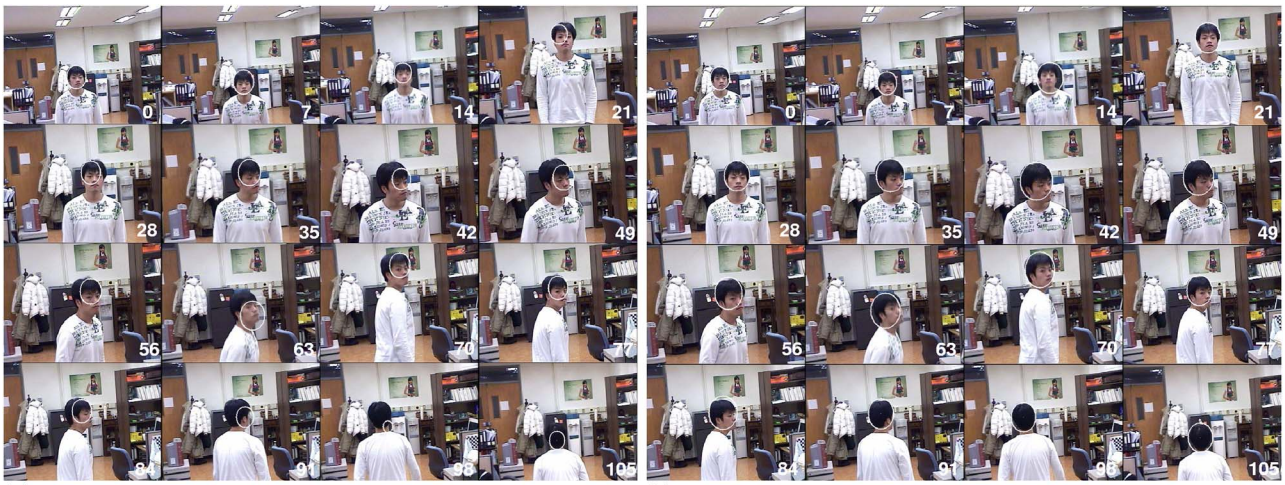


Fig. 8 Tracking results of the *Jumping* sequence obtained by using (a) a conventional method and (b) the proposed method.

the *Jumping* sequence in Fig. 11(a), we notice that the proposed method can consistently track the target even with fast motion of the head, but the conventional method produces large errors in position and scale, because the method is not sufficient to perceive the color histogram change of the target. Figure 11(b) shows that in the *Walking* sequence, the conventional method is unable to recover the face region outside the ellipse, and thereby produces noticeable errors around the 80th frame. Also, the errors of the conventional method become much larger than those of the proposed method after approximately the 100th frame. This is because the head becomes smaller starting from that frame, but the conventional method fails to adapt to the change of head scale. Figure 11(c) shows the tracking errors for the *Clutter* sequence. We note in the graphs that the proposed method can discriminate the target from the cluttered background more precisely than the conventional method. In summary, the graphs in Fig. 11 demonstrate that the proposed algorithm can provide a more accurate and stable tracking performance by reducing tracking errors and their variances.

All experiments are conducted on a PC with a Pentium 4 CPU of 2.8 GHz, and it is verified that the proposed algorithm performs real-time tracking with a processing speed of 10 fps.

The proposed algorithm can be used in a semiautomatic target tracking system. Let us consider the case where an operator monitors multiple screens of different active cameras. If the operator finds a suspicious person in a screen, the operator may specify a target ellipse by pausing on the screen. Then, the ellipse will be automatically tracked by a camera in the following frames and the corresponding video can be recorded. In this way, a single operator can handle several screens simultaneously.

In the case that multiple human heads coexist in the same image, the algorithm consistently tracks the initially

Fig. 9 Tracking results of the *Walking* sequence obtained by using (a) a conventional method and (b) the proposed method.

Downloaded from SPIE Digital Library on 17 Apr 2011 to 143.248.224.33. Terms of Use: http://spiedl.org/terms

Fig. 10 Tracking results of the *Clutter* sequence obtained by using (a) a conventional method and (b) the proposed method.

specified ellipse unless the ellipse overlaps with the other head(s), because the mean-shift-based algorithm makes the ellipse converge to the nearest local maximum in the similarity curve. Meanwhile, in the case of partial occlusion of a target by a background object, the proposed algorithm may also consistently track the target due to the following reason. In the color convergence stage of the proposed algorithm, the ellipse converges to the center of the nonoccluded region of the target if the target color can be discriminated from the color of the occluding background object. Then, in the refinement stage, the converged ellipse is refined to fit to the target boundary. Note here that since the algorithm tries to match the boundary of the fitted ellipse to the original target boundary of elliptical shape without considering the occluded boundary of the nonelliptical shape (see Sec. 2.3.3 for details.), the refined ellipse tends to represent the real target shape with no occlusion.

4 Conclusions

For reliable camera control in an active camera system, accurate position and scale estimation of a target are necessary. In this work, we propose a robust and accurate head tracking algorithm suitable for a real-time active camera system having pan, tilt, and zoom functions. To increase the likelihood of converging to the true target position, the algorithm compensates the global motion and reflects the temporal change of the target color histogram. In addition, the proposed algorithm utilizes various information such as color, shape, and quasi-spatial information so that it may successfully distinguish the target from the background. The proposed algorithm also focuses on real-time tracking. It first rapidly estimates the initial target position by using a color-based mean shift, and subsequently refines the position and scale. To reduce the processing time further, it uses 1-D projection datasets in background motion estimation, which is verified to be appropriate for a pan-tilt-zoom camera system. Experimental results show that the proposed algorithm outperforms existing head tracking algorithms for various sequences and provides real-time tracking in a PC platform.

Downloaded from SPIE Digital Library on 17 Apr 2011 to 143.248.224.33. Terms of Use: http://spiedl.org/terms

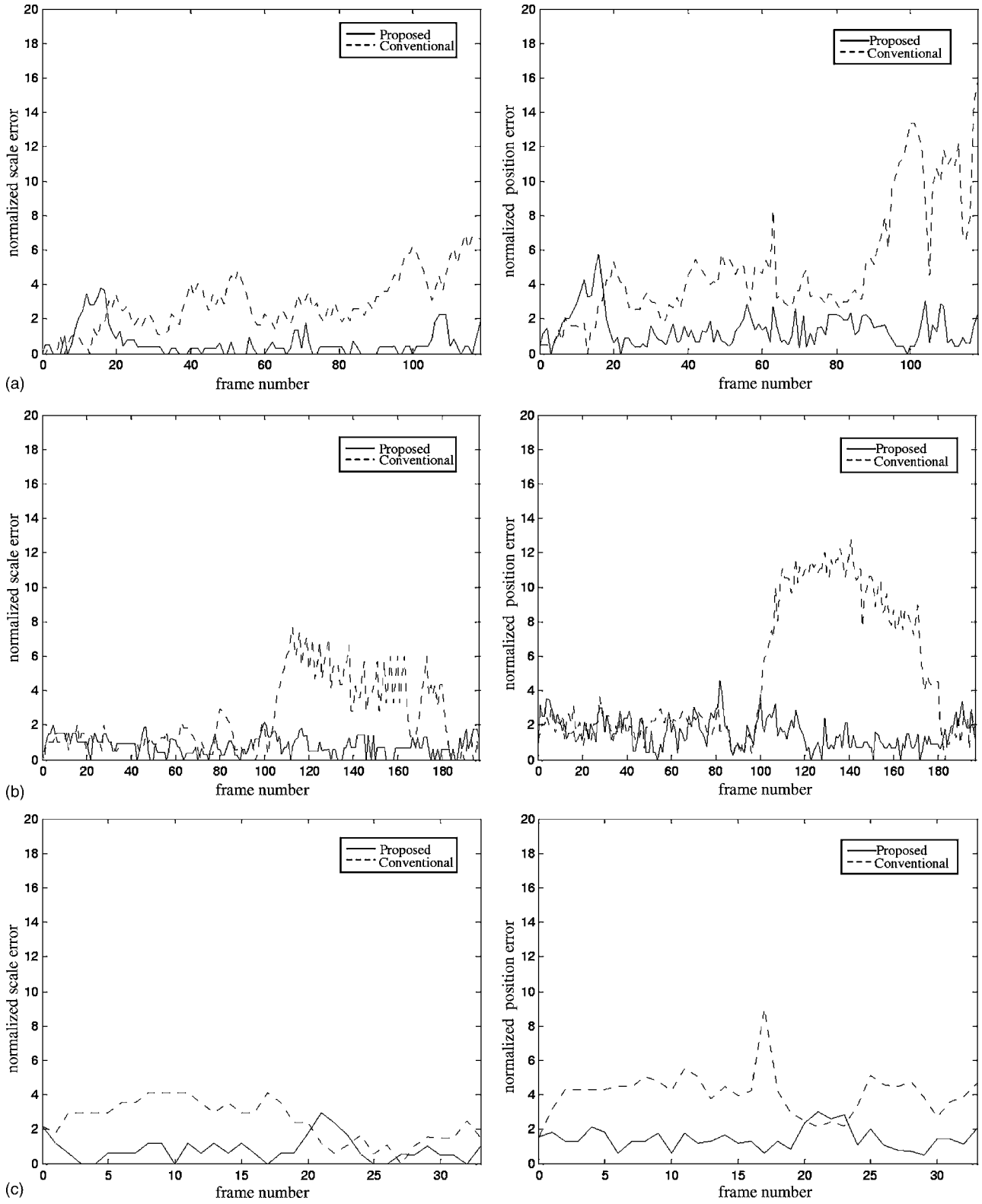


Fig. 11 Normalized scale and position errors between true (manually determined) and estimated tracking results for (a) the *Jumping*, (b) *Walking*, and (c) *Clutter* sequences. Dotted and solid lines in the graphs denote the errors of the conventional and proposed methods, respectively. The unit is pixel.

Downloaded from SPIE Digital Library on 17 Apr 2011 to 143.248.224.33. Terms of Use: http://spiedl.org/terms

Acknowledgments

This research was supported by the Agency for Defense Development, Korea, through the Image Information Research Center at the Korea Advanced Institute of Science and Technology.

References

1. S. M. Smith and J. M. Brady, "ASSET-2: Real-time motion segmentation and shape tracking," *IEEE Trans. Pattern Anal. Mach. Intell.* **17**(8), 814–820 (1995).
2. A. Lipton, "Local application of optic flow to analyze rigid versus non-rigid motion," Technical Report CMU-RI-TR-99-13, Robotics Institute, Carnegie Mellon University (1999).
3. C. Wang and M. Brandstein, "A hybrid real-time face tracking system," *Proc ICASSP'98* **6**, 3737–3741 (1998).
4. D. Murray and A. Basu, "Motion tracking with and active camera," *IEEE Trans. Pattern Anal. Mach. Intell.* **16**(5), 449–459 (1994).
5. R. Gupta, M. D. Theys, and H. J. Siegel, "Background compensation and an active-camera motion tracking algorithm," *Proc. Intl. Conf. Parallel Process.*, pp. 431–440 (1997).
6. A. Baumberg and D. Hogg, "An efficient method for contour tracking using active shape models," *Proc. Workshop Motion Nonrigid Articulated Obj.* IEEE CS Press, Los Alamitos, CA pp. 194–199, (1994).
7. S. Birchfield, "Elliptical head tracking using intensity gradient and color histograms," *Proc. IEEE Conf. Computer Vision Patt. Recog.* pp. 232–237, Santa Barbara, CA (1998).
8. D. Comaniciu, "Robust detection and tracking of human faces with an active camera," *Proc. IEEE Intl. Workshop Visual Surveillance*, pp. 11–18 (2000).
9. D. Comaniciu, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(5), 564–575 (2003).
10. T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.* **15**(1), 52–60 (1967).
11. K. Fukunaga and L. D. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. Inf. Theory* **21**(1), 32–40 (1975).
12. T. Zhang and D. Freedman, "Improving performance of distribution tracking through background mismatch," *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(2), 282–287 (2005).
13. D. Comaniciu and P. Meer, "Mean shift analysis and applications," *Proc. IEEE Conf. Computer Vision* **2**, 1197–1203 (1999).
14. I. Grinias and G. Tziritas, "Robust pan, tilt and zoom estimation," *Proc. IEEE Intl. Conf. Digital Signal Process.* **2**, 679–682 (2002).
15. G. M. Kim, S. H. Yoon, J. H. Kim, and G. T. Hur, "Robust head tracking using hybrid color and 3-D under natural and unspecified environment," *Proc. SPIE* **5150**, 327–333 (2003).
16. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York (1973).
17. D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(5), 603–619 (2002).
18. A. Rao, R. K. Srihari, and Z. Zhang, "Spatial color histograms for content-based image retrieval," *Proc. IEEE Intl. Conf. Tools Artif. Intell.* pp. 183–186 (1999).

Dong-Gil Jeong received his BS degree in electronic, electrical, and communication engineering from Pusan National University, Busan, Korea, in 2003, and his MS degree in electrical engineering and computer science from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2005. He is currently working for the Agency for Defense Development (ADD). His research interests include object tracking and pattern recognition.

Dong-Goo Kang received his BS degree in electronic engineering from Sogang University, Seoul, Korea, in 2000, and his MS degree in electrical engineering and computer science from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2002. He is currently working toward his PhD degree in KAIST. His research interests include medical and video processing.

Yu-Kyung Yang received her BS degree in electronics and information engineering from Chonbuk National University, Jeonju, Korea, in 2002, and her MS degree in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2004. She has been researching mobile applications as a junior engineer of KTFTechnologies, Incorporated, since 2004. Her research interests include image/video processing and its applications.

Jong Beom Ra received his BS degree in electronic engineering in 1975 from Seoul National University, Korea, and his MS and PhD degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, in 1977 and 1983, respectively. From 1983 to 1987, he was an associate research scientist with Columbia University, New York. Since 1987, he has been a professor with the Department of Electrical Engineering and Computer Science, KAIST, where he currently directs the Image Information Research Center. His research interests are digital image processing, video signal processing, 3-D visualization, 3-D display systems and medical imaging, including magnetic resonance imaging (MRI). He is a member of SPIE and a senior member of the IEEE.

Downloaded from SPIE Digital Library on 17 Apr 2011 to 143.248.224.33. Terms of Use: http://spiedl.org/terms