Visual Tracking of Human Head and Arms Using Adaptive Multiple Importance Sampling on a Single Camera in Cluttered Environments

Cheng-Ming Huang, *Member, IEEE*, Yi-Ru Chen, and Li-Chen Fu, *Fellow, IEEE*

***Abstract*— This paper presents a 2D upper body tracking algorithm using a single monocular camera. The proposed method can be applied on a stationary or moving camera platform, and is able to achieve real-time performance in cluttered environments, making it ideal for human-machine interaction. The algorithm extracts body parts even when the target person approaches other objects. This is a common problem for depth-based camera systems. Real-time visual extraction of a subject's head and arms is performed during preprocessing in order to determine their current action and presents two key innovations. First, multiple visual clues are integrated dynamically by an adaptive multiple importance sampling particle filter to generate hypotheses. These hypotheses can efficiently estimate various gestures of arms on images captured from cluttered environments. Second, multiple visual cues of a human face and arms are devised, which quickly and effectively verifies various hypotheses from the multiple importance sampling schemes. To validate the effectiveness of the proposed tracking approach, several experiments are performed whose results appear to be quite promising.**

***Index Terms*— Human pose estimation, visual tracking.**

I. Introduction

FOR human-machine interactions in an intelligent house, the vision system is concentrated on analyzing human behavior after detecting a human figure. As a camera set up on a table or a robot usually focuses on the upper body of a human, due to a limited field of view, this research work contributes the task to effectively tracking a human face and the upper limbs for understanding the human intention during the process of human-machine interaction. So far, there has been much research proposed for human posture estimation. The conventional works estimate the pose by marking the human body parts or by asking the human to wear special clothing [4]. However, this is inconvenient and can only be used in special situations with a specific equipment setup.

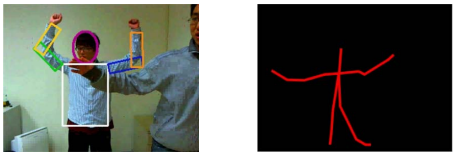
Manuscript received November 29, 2013; revised February 11, 2014; accepted February 13, 2014. Date of publication March 4, 2014; date of current version May 29, 2014. This work was supported by the National Science Council of Taiwan under Grants NSC 102-2221-E-027-084 and NSC 102-2218-E-002-009-MY2. The associate editor coordinating the review of this paper and approving it for publication was Dr. Anna G. Mignani.

C.-M. Huang is with the Department of Electrical Engineering, National Taipei University of Technology, Taipei 106, Taiwan (e-mail: cmhuang@mail.ntut.edu.tw).

Y.-R. Chen and L.-C. Fu are with the Department of Electrical Engineering and Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: b9207058@mail.ntust.edu.tw; lichen@ntu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSEN.2014.2309256



(a) (b)

Fig. 1. Human body pose tracking under other person's distraction. (a) Tracking result labeled by our method. (b) The skeleton obtained by Windows SDK [9], where the estimation result of lower body is confused as only the upper body is observed, and the arm estimations are disturbed due to the distraction from other person.

To improve these methods, there are two kinds of popular visual approaches used for tracking human body parts or estimating posture, namely, background subtraction [6], [7] and depth-based segmentation [8], [9]. A powerful background subtraction method with background initialization, updating, and classification, is presently unlikely to result in real-time tracking when operating on a moving camera platform. Also, using the foreground human segment of the depth image acquired by an infrared or stereo camera module [11] is easily disturbed by distracters around the target as shown in Fig. 1.

Apparently, the depth-based segmentation [12] cannot handle disturbances caused by another object at a similar depth or caused by occlusion. In addition, background subtraction may not be used when the camera moves or when other people move around the target subject. In all cases, these two methods [7], [12] fail to separate an image of a human body from a complex scene. In order to overcome the challenges due to cluttered environments, such as complicated background texture, a moving camera, and even disturbance or occlusion around the target, we avoided using either the background subtraction or the depth-based segmentation approach in our work.

In this work, we focused on the real-time extraction of 2D upper body parts from image data. The sensing results may also be used to recognize a human's intentions [14] or 3D posture [15], [16] and command mechanisms with proper reactions. Thus, the observation processes should be completed quickly since the environment is always varying. The combination of appropriate tracking algorithms can actually reduce the computational load. The particle filter, combined with partitioned sampling [17] or the multiple importance sampling [18], can alleviate the computational cost of estimation in a high dimensional state space. However, the tracking of human beings is recognized to be one of the most challenging tasks. Therefore, the fusion of multiple cues [1], [10] from

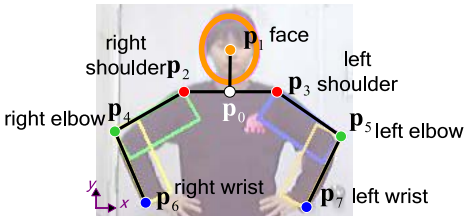


Fig. 2. The 2D human upper body model.

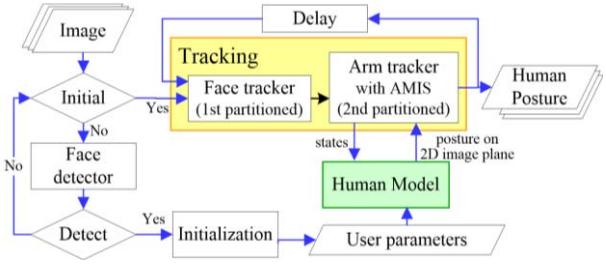


Fig. 3. System flowchart.

various features or sensors is typically applied to evaluate the likelihood of given samples more accurately. The weighting of likelihood fusion over time can also be adaptively renewed [13] to enhance the influence of distinctive cues. If the hypotheses are not effectively sampled, the likelihood fusion cannot achieve maximum efficiency. Hence, drawing and predicting the sample's distribution from multiple clues [25], [26] is a fundamental way for leading the hypotheses generation of the particle filter.

In this paper, the adaptive multiple importance sampling (AMIS) particle filter is proposed to efficiently combine all of the clues for tracking a human's upper body. The number of samples generated from each clue will be dynamically adjusted by evaluating the discriminability functions of these clues. Parallel to that, multiple visual cues are designed as the visual likelihood functions to validate the tracking hypotheses accurately and quickly. When concentrating on 2D human posture estimation with a single monocular camera in a complex scene, as compared with the surveyed works [1], [3], [5], and [10], our proposed algorithm can provide real-time performance while achieving reliable tracking results.

The rest of this paper is organized as follows. In section II, we first introduce the human upper body model. Then, in section III, the tracking algorithm is described. The particle filters with the partitioned sampling method and the AMIS algorithm are presented to efficiently track the posture of the head and arms. The design of likelihood functions in utilizing visual cues is explained in section IV. In section V, we demonstrate several experimental results to validate the effectiveness of the proposed tracking approach. Finally, we conclude this paper in section VI.

II. Human Upper Body Model

In general, the scale of every body part is proportional to the face size, as in the case with the Virtruvian Man [2] created by Leonardo da Vinci. Using the general assumptions on the upper body proportions, a stick model as Fig. 2 can be constructed. The shoulder width, neck length, upper arm width, forearm width, and arm length, belonging to one specified user can be extracted from the initial posture as Fig. 2. In cluttered environments, the automatic initialization would be incorporated with the skin color detector and the extracted human silhouettes using a camera motion estimation method [19]. The user can then present the intention through this initial posture to start the system.

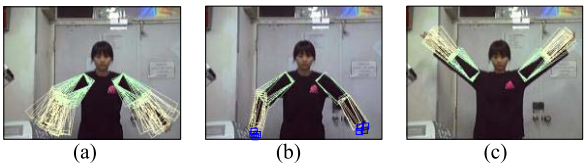
The face and arms of a subject are most likely to reveal one's intentions. For this reason, as long as the head and arms can be seen, the vision system is not concerned with whether the head, torso or shoulders, are moderately slanting or facing directly towards the camera. The joints $\mathbf{p}_0$, $\mathbf{p}_2$ and $\mathbf{p}_3$ on the torso are assumed to be fixed with the state of face. The state can then be simplified and separated into three parts. The first part describes the state of the head $\mathbf{x}_P = [\mathbf{p}_1, r]$, including the center of the head position $\mathbf{p}_1$, and the head scale r, to abstract the simple gross motion of the whole human body. The second and third parts describe the states of the right arm $\mathbf{\Theta}_R = [\mathbf{p}_4, \mathbf{p}_6]$ and the left arm $\mathbf{\Theta}_L = [\mathbf{p}_5, \mathbf{p}_7]$ which involve the positions of elbow joints and wrist joints on the image plane. For obtaining observations from an image, the face is modeled by an ellipse, and the upper arms and forearms are individually modeled by two rectangles, as shown in Fig. 2.

III. Adaptive Multiple Importance Sampling

Given the image observations $\mathbf{z}_{1:t} = \{\mathbf{z}_1, \ldots, \mathbf{z}_t\}$ obtained from one monocular camera up to time t, the tracking problem of a subject's head and arms can be formulated as the conditional probability $p(\mathbf{x}_{P,t}, \mathbf{\Theta}_{R,t}, \mathbf{\Theta}_{L,t} | \mathbf{z}_{1:t})$, where $\mathbf{x}_{P,t}$, $\mathbf{\Theta}_{R,t}$, $\mathbf{\Theta}_{L,t}$ are the states of the face, right arm, and left arm, at time instant t, respectively. According to the human model and the partitioned sampling concept [10], [17], the posterior distribution $p(\mathbf{x}_{P,t}, \mathbf{\Theta}_{R,t}, \mathbf{\Theta}_{L,t} | \mathbf{z}_{1:t})$ can be decomposed into three parts $p(\mathbf{x}_{P,t} | \mathbf{z}_{1:t})$, $p(\mathbf{\Theta}_{R,t} | \mathbf{x}_{P,t}, \mathbf{z}_{1:t})$ and $p(\mathbf{\Theta}_{L,t} | \mathbf{x}_{P,t}, \mathbf{z}_{1:t})$. The face part $p(\mathbf{x}_{P,t} | \mathbf{z}_{1:t})$, which is generally more reliable than other limb partitions, because the movement of face is much more stable than a limb and the features of face are more obvious than a limb, is estimated by the sampling importance resampling (SIR) particle filter. Based on the hypothesis samples of face, the distribution $p(\mathbf{\Theta}_t | \mathbf{x}_{P,t}, \mathbf{z}_{1:t})$ of arms partition, $\mathbf{\Theta}_t$ represents $\mathbf{\Theta}_{R,t}$ or $\mathbf{\Theta}_{L,t}$, are then tracked by the AMIS particle filter. Furthermore, we denote $\mathbf{\Theta}_t = [\mathbf{p}_{e,t}, \mathbf{p}_{w,t}]$, where $\mathbf{p}_{e,t}$ represents the elbow estimation of $\mathbf{p}_4$ or $\mathbf{p}_5$, and $\mathbf{p}_{w,t}$ represents the wrist estimation of $\mathbf{p}_6$ or $\mathbf{p}_7$. Notice that we do not further decompose the arm posture into the upper arm and forearm because the two are tightly connected.

The flowchart of the overall system is summarized in Fig. 3. After detecting the face and completing the initialization of a user, the specific parameters, such as the width or length of one body part, are stored in his/her reference human model. Then, the hypotheses on the body posture are generated and verified through the visual tracking process mentioned above.

By the concept of importance sampling [18], once we have some apparent visual information of a target, the samples can be efficiently drawn from a proposal function. Suppose that there are M different proposal functions



(a) (b) (c)

Fig. 4. The particles generated from multiple proposal functions. (a) Resampling. (b) Inverse kinematic. (c) Line detector.

$q_{1,t}(\mathbf{\Theta}_t), \ldots, q_{M,t}(\mathbf{\Theta}_t)$ for inferring the arm, and each proposal function $q_{i,t}(\mathbf{\Theta}_t)$ can generate $m_{i,t}$ samples $\{\mathbf{\Theta}_{i,t}^{(1)}, \ldots, \mathbf{\Theta}_{i,t}^{(m_{i,t})}\}$ for the model parameters $\mathbf{\Theta}_t$, *i.e.*, $\mathbf{\Theta}_{i,t}^{(j)} \sim q_{i,t}(\mathbf{\Theta}_t)$. We collect all the samples $\mathbf{\Theta}_{i,t}^{(j)}$, $j = 1, \ldots, m_{i,t}$, $i = 1, \ldots, M$, from which the posterior of each arm can be yielded by the approximation with a set of weighted samples $\{\mathbf{\Theta}_{i,t}^{(j)}, \omega_{i,t}^{(j)}\}$:

$$p\left(\mathbf{\Theta}_t \left| \mathbf{x}_{P,t}, \mathbf{z}_{1:t} \right.\right) \approx \alpha \sum_{i=1}^{M} v_{i,t}(\mathbf{\Theta}_t) \sum_{j=1}^{m_{i,t}} \omega_{i,t}^{(j)} \delta\left(\mathbf{\Theta}_t - \mathbf{\Theta}_{i,t}^{(j)}\right), \quad (1)$$

where α is a normalization constant, $v_{i,t}(\Theta_t)$ stands for the discriminability function of proposal function $q_{i,t}(\mathbf{\Theta}_t)$, $\delta(\cdot)$ is the Dirac delta function, and $\omega_{i,t}^{(j)}$ is the corresponding weight of each particle j which belongs to the ith proposal function $q_{i,t}(\Theta_t)$. In order to cover the unpredictable arm movement, three kinds of proposal functions as illustrated in Fig. 4 are applied: the first proposal function $q_{1,t}(\mathbf{\Theta}_t)$ uses the estimates of the latest posterior, the second one $q_{2,t}(\mathbf{\Theta}_t)$ uses the inverse kinematics from the hand position, and the third one $q_{3,t}(\mathbf{\Theta}_t)$ uses a line detector on the edge of arm.

The number of particles $m_{i,t}$ is based on the completeness of arm information provided by the proposal function $q_{i,t}(\mathbf{\Theta}_t)$ and is defined as

$$m_{i,t} = \left(\beta_{i,t}\, v_{i,t}(\mathbf{\Theta}_t) \Big/ \sum_{i=1}^{M} \beta_{i,t}\, v_{i,t}(\mathbf{\Theta}_t)\right) N_s, \quad (2)$$

where N_s is the total number of particles, and $\beta_{i,t}$ is the corresponding weighting of $v_{i,t}(\mathbf{\Theta}_t)$. The discriminability function $v_{i,t}(\mathbf{\Theta}_t)$, which is detailed in the following subsections, is determined by the distinctiveness or the variance of the corresponding proposal function $q_{i,t}(\mathbf{\Theta}_t)$ based on the visual information extracted from the image frame at every time instant. Notice that the total number of particles N_s is constant; however, the particle number $m_{i,t}$ of each proposal function is dynamically determined by the 2D image information at time instant t. When one proposal function has higher level of distinctiveness, the number of particles sampled from that proposal function will increase.

A. Proposal Function from Latest Posterior

Based on the Markov assumption, the SIR scheme is applied to maintain good tracking results around the posterior at previous time. The proposal function here is chosen as the state transition model of an arm:

$$q_{1,t}(\mathbf{\Theta}_t) = p(\mathbf{\Theta}_t | \Theta_{t-1}^{\cdot} \mathbf{x}_{P,t}). \quad (3)$$

We assume that the predicted model parameters of the first proposal function is a normal distribution around the original. The discriminability function $v_{1,t}(\mathbf{\Theta}_t)$ of this proposal $q_{1,t}(\mathbf{\Theta}_t)$ is evaluated by the variance of elements of the state vector samples at previous time instant as follows

$$v_{1,t}(\mathbf{\Theta}_t) = \exp\left[-\sum_{k} \text{var}\left(\mathbf{\Theta}_{t-1}(k)\right)\right], \quad (4)$$

where $\mathbf{\Theta}_{t-1}(k)$ is the kth element of state vector $\mathbf{\Theta}_{t-1}$, and

$$\text{var}\left(\mathbf{\Theta}_{t-1}(k)\right) = \sum_{i,j} \omega_{i,t}^{(j)} \left(\mathbf{\Theta}_{i,t-1}^{(j)}(k)\right)^2 - \left(\sum_{i,j} \omega_{i,t}^{(j)} \mathbf{\Theta}_{i,t-1}^{(j)}(k)\right)^2 \quad (5)$$

is the variance of each element of the samples with respect to the distribution of normalized weighting $\omega_{i,t}^{(j)}$. The estimated state would not converge to the true one when the samples are not concentrated enough. Taking the distribution of weighting into account can suppress the influence from the hypothesis with lower image likelihood.

B. Proposal Function from Inverse Kinematics of Hand

When observing human activity, the hand, which is the end point of an arm, is a useful part of a body for representing the position of an arm. According to the inverse kinematics [20], the variation of joint angles of an arm can be estimated from the displacement of the wrist position $\mathbf{p}_{w,t}$ on the 2D image plane. Since the exact hand position is hard to be determined, we generate many candidates around the previous wrist estimation $\mathbf{p}_{w,t-1}$ according to the color distribution of the hand. The wrist positions $\mathbf{p}_{w,t}^{(j)}$ included in the particles $\mathbf{\Theta}_{2,t^{(j)}}, j = 1, \ldots, m_{2,t}$, are sampled from the following Gaussian distribution of the wrist proposal:

$$\mathbf{p}_{w,t}^{(j)} \sim q\left(\mathbf{p}_{w,t}\right) = P_h \mathcal{N}\left(\mathbf{p}_{w,t-1}, \sigma_w^2\right) \quad (6)$$

where $\mathbf{p}_{w,t-1}$ and σ_w^2 are assigned to be the Gaussian mean and variance of the possible hand position on the image respectively, and P_h is the probability that the hand appeared. The probability P_h is determined by the ratio between the area of detected skin color blob around the previously estimated wrist location and the area of feasible wrist moving region. The variance σ_w^2 is determined by the wrist movement at the previous time instant. Define the joint angle vector $\mathbf{\Psi}_t = [\psi_{s,t}, \psi_{e,t}]$, where $\psi_{s,t}$ and $\psi_{e,t}$ are the angles of shoulder joint and elbow joint, respectively. Then, the joint angle hypothesis $\mathbf{\Psi}_t^{(j)}$ with respect to the wrist hypothesis $\mathbf{p}_{w,t}^{(j)}$ could be derived from the inverse kinematics [20]:

$$\mathbf{\Psi}_t^{(j)} = \mathbf{\Psi}_{t-1} + J^{-1}\left(\mathbf{\Psi}_{t-1}\right)\left[\mathbf{p}_{w,t}^{(j)} - \mathbf{p}_{w,t-1}\right] \quad (7)$$

where $J^{-1}\left(\mathbf{\Psi}_{t-1}\right)$ is the inverse of the Jacobian matrix. Besides, the elbow position $\mathbf{p}_{e,t}^{(j)}$ can also be yielded by giving the joint angle vector $\mathbf{\Psi}_t^{(j)}$ in (7) and two end points, the shoulder estimation $\mathbf{p}_{s,t}$ and the wrist hypothesis $\mathbf{p}_{w,t}^{(j)}$. As the upper body model illustrated in Section 2, the shoulder estimation $\mathbf{p}_{s,t}$, which stands for the joints $\mathbf{p}_2$ or $\mathbf{p}_3$ on the torso, depends on the face estimation $p(\mathbf{x}_{P,t}|\mathbf{z}_{1:t})$.

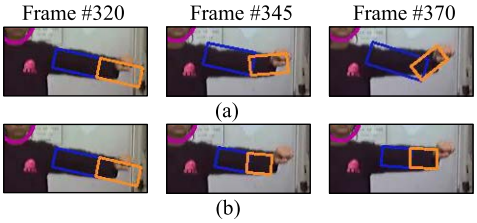


Fig. 5. The influence of the proposal functions from inverse kinematics of the hand and the obvious line of the arm. (a) Least square solution generated from the pseudo inverse. (b) Corrected solution of the cooperation of all proposed proposal functions.

Fig. 6. Detected lines (bold and white) of human posture on the edge image. (a) Putting hands on the waist. (b) Stretching out the full arm.

The proposal function of the inverse kinematics of the hand is therefore defined as $q_{2,t}(\mathbf{\Theta}_t) \propto q(\mathbf{p}_{w,t})$, and the drawn particles are $\mathbf{\Theta}_{2,t}^{(j)} = [\mathbf{p}_{e,t}^{(j)}, \mathbf{p}_{w,t}^{(j)}], j = 1, \ldots, m_{2,t}$. Furthermore, the discriminability function $v_{2,t}(\Theta_t)$ based on the hand appearing probability can also be defined as

$$v_{2,t}(\mathbf{\Theta}_t) = \exp(-P_h) \tag{8}$$

Once there is a lot of skin noise around the hand tracking, the samples generated from this proposal function will be few.

C. Proposal Function from Obvious Edge of Arm

In Fig. 5(a), the arm tracker loses a track of the forearm at frame #320, and then the particles drawn by the inverse kinematics from the hand position updates the model parameters of arm. However, the solution obtained through the inverse kinematics only constrains the forearm to fit the hand position at frame #345. Since the 2D elbow estimation is easily confused by the displacement of the elbow perpendicular to the image plane, the proposal function with the obvious edge of the arm is employed here to fix it.

The Hough line detector [21] could detect the straight and thick lines around the previous arm estimation. If there exists lines whose included angles approach the estimated arm and have length that are longer than a predefined threshold, the tracker will conclude that the user is stretching out his/her full arms as illustrated in Fig. 6(b). Define the detected obvious line with the angle φ_{line}, and thus the joint angle vector $\mathbf{\Psi}_t$ when stretching out the full arm are corrected to be

$$\mathbf{\Psi}_t = \left[\varphi_{\text{line}}, \pi\right]. \tag{9}$$

Based on the shoulder estimation $\mathbf{p}_{s,t}$, the positions of arm joints $\mathbf{\Theta}_{3,t}$ (*i.e.*, the elbow and wrist positions considered in the third proposal function) can be approximated by using the forward kinematics with the arm length l_a and the joint angles in (9). The proposal function of the obvious arm edge can be represented as being proportional to the Gaussian distribution $q_{3,t}(\mathbf{\Theta}_t) \propto \mathcal{N}\left(\mathbf{\Theta}_{3,t}, \sigma_{line}^2\right)$ with the mean $\mathbf{\Theta}_{3,t}$ and the variance σ_{line}^2. Fig. 5(b) presents the corrected elbow estimation that can be compared with Fig. 5(a). Although some particles from this proposal function may be originated from the lines detected in the background, these false hypotheses will be filtered out by the likelihood evaluation.

Since the line obtained through Hough transform only considers the magnitude response of edge points, the information of edge orientation is employed here to determine the orientation consistency of arm edges within multiple lines in textured image region. The orientations of edges or lines are quantized into eight directions [22], analyzed, and then represented as the orientation histograms of edges and lines. The discriminability function $v_{3,t}(\mathbf{\Theta}_t)$ is defined by evaluating the entropy of orientation histograms as

$$v_{3,t}(\mathbf{\Theta}_t) = \exp\left[-c_1 \sum_{b=1}^{8} P_{e,b} \log P_{e,b} - c_2 \sum_{b=1}^{8} P_{l,b} \log P_{l,b}\right] \tag{10}$$

where c_1 and c_2 are the relative weights of two entropy terms, $P_{e,b}$ and $P_{l,b}$ are the probability of bth bin in the normalized orientation histograms of edges and lines, respectively. The arm edge is distinctive when both of its edge and line directions are consistent, and each entropy should be small under such situation. In cluttered environments, the value of this discriminability function $v_{3,t}(\mathbf{\Theta}_t)$ is small since the detected edges or lines would have various orientations.

IV. Likelihood Evaluation

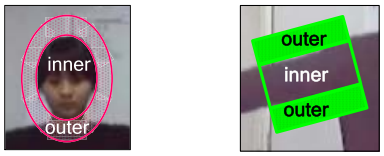
After generating each hypothetical state **x,** the weight of each particle can be evaluated according to multiple visual likelihood functions. In order to operate with a single monocular camera with real-time performance, the designed likelihood functions are simply evaluated with image information through the computation of color and edge contours with geometric constraints. When interaction proceeds in cluttered environments, despite that one visual cue may be undistinguished, other cues can still assist in identifying the human body parts.

A. Color Histogram

The color likelihood [23] uses the Bhattacharyya distance to evaluate the similarity between the reference color histogram defined in the initialization process and the color histogram corresponding to the hypothesized state of each particle, which may stand for the sub-state of head or arm. As shown in Fig. 7, the histogram of the inner part in each particle should be similar to the reference color histogram; however, the histogram of outer part around each particle should be dissimilar to the reference. Note that when the arm is closest to the torso, the joint likelihood [23] will be considered for the overlapping region of different body parts.

B. Enhanced Edge Contour

The motion detection is used here to enhance the edge around the human who is interacting with the vision system



(a) (b)

Fig. 7. The definition of inner/outer part for color likelihood. The area of outer part (within dotted region) is equal to the area of inner part. (a) Head. (b) Upper arm/forearm.

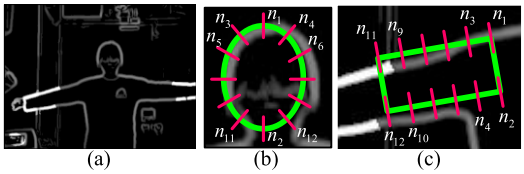

(a) (b) (c)

Fig. 8. The motion enhanced edge and the shape of the contour template. The red line segments indicate the matching direction of each control points. (a) Motion enhanced edge. (b) Head. (c) Upper arm/forearm.

by moving his/her arms. By attaching the edge of this different part between two consecutive frames to the current edge image, we can obtain a motion enhanced edge image as in Fig. 8(a). When considering the motion camera, the edge contour is equivalent to operating on the common edge image and the feature point tracking detailed in the next subsection will assist the tracking of each body part.

The continuous contour of head or arm, with scale r, is represented with N_c discrete control points n_i. The similarity $f_{mat}(\mathbf{x})$ between the reference shape model with the state variable and its neighborhood edge image [24] is to accumulate the distance between the control point and the pixel with a significant edge along the normal direction of the contour. In addition, since the enhanced edge image emphasizes the edge points with higher intensity in the motion area, the contour intensity $f_{int}(\mathbf{x})$ is utilized to differentiate motion areas of humans from others. The smaller value of $f_{int}(\mathbf{x})$ shows that the edge is more significant or that this candidate belongs to the motion area originated from the interacting human, when the camera platform is static.

As illustrated in Fig. 9, all of the three drawn hypotheses may have the same value of their contour matching functions. However, it is only the case in Fig. 9(a) where it is matched with the good length. The contour length function is designed to clarify this:

$$f_{len}(\mathbf{x}) = \left[1 - l/(rl_a)\right] + N_{null}/(N_c/2), \tag{11}$$

where l is the length of the rectangle belonging to one hypothesis, l_a is the initially defined arm length, and N_{null} is the number of control point pairs without matching edge pixel. The contour length function with a smaller value will encourage the hypothesis of a longer rectangle and penalize the segment of a rectangle without edge evidence.

Moreover, the other significant characteristic of a human figure is symmetry. The matching distances d_{2i} and d_{2i-1} of each control point pair n_{2i} and n_{2i-1} defined in Fig. 10(a) should be similar. The unsymmetrical edge may be caused by

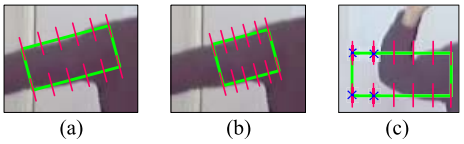

(a) (b) (c)

Fig. 9. Different length of the hypotheses on contour likelihood of upper arm. The blue cross marks denote the invalid matching of control point pairs. (a) Good length. (b) Shorter length. (c) Longer length.

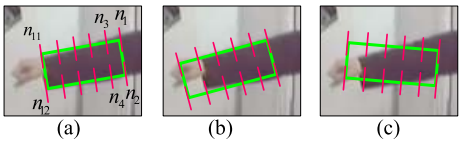

(a) (b) (c)

Fig. 10. Examples of the contour symmetry feature. (a) Symmetrical. (b) Unsymmetrical edge of hand. (c) Unsymmetrical edge by angle.

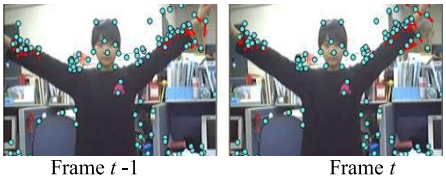

Frame t -1 Frame t

Fig. 11. The displacement of KLT tracked feature points. Blue dots denote feature points, and red arrows denote the displacement of these points.

a hand as illustrated in Fig. 10(b) and by an improper angle as in Fig. 10(c). We define the contour symmetry function as

$$f_{sym}(\mathbf{x}) = \sum_{i=1}^{N_c/2} (d_{2i} - d_{2i-1})/(d_{\max} N_c/2), \tag{12}$$

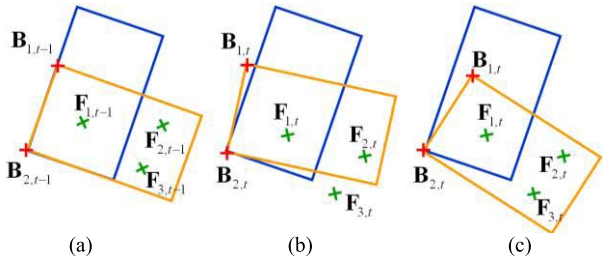
where $d_{\max}$ is the maximum matching distance. The smaller value of $f_{sym}(\mathbf{x})$, which accumulates the difference of matching distances in each control point pair, allows the candidate to fit into a more appropriate position.

C. Geometric Distribution of Feature Points

The Kanade-Lucas-Tomasi (KLT) algorithm [19] is employed to select and track the feature points in an image sequence, and then the displacement across image frames can be derived from the point-tracking results. When the scene and the camera are both moving, the KLT algorithm can still reliably track feature points across frames as shown in Fig. 11. However, the KLT algorithm cannot indicate which feature point belongs to which body part or the background. Since the geometric distribution of feature points belonging to the same body part should be invariant as illustrated in Fig. 12, the desired hypothesis of each body part should contain feature points with the similar distribution as a previous image frame. We formulate the feature point distribution function as

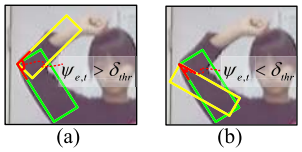
$$f_{\text{point}}(\mathbf{x}) = \frac{\sum_{k=1}^{N_d} \left|\overline{\mathbf{B}_{1,t-1}\mathbf{F}_{k,t-1}} - \overline{\mathbf{B}_{1,t}\mathbf{F}_{k,t}}\right| + \left|\overline{\mathbf{B}_{2,t-1}\mathbf{F}_{k,t-1}} - \overline{\mathbf{B}_{2,t}\mathbf{F}_{k,t}}\right|}{2F_{\max} N_d}, \tag{13}$$

where N_d is the number of feature points $\mathbf{F}_{k,t-1}$ within one estimated body part at time instant $t-1$, $\mathbf{B}_{1,t-1}$ and $\mathbf{B}_{2,t-1}$



(a) (b) (c)

Fig. 12. Feature point distribution in a target region for likelihood evaluation. (a) Reference feature points distribution. (b) One illustrated sample with undesired feature points distribution. (c) One illustrated sample with a similar feature points distribution as (a).



(a) (b)

Fig. 13. The overlapping false alarm of the elbow angle on 2D image. (a) Non-overlapping. (b) Overlapping.

are the reference points in the estimated body part, $\mathbf{F}_{k,t}$ is the kth tracked feature point at time instant t, and $\mathbf{B}_{1,t}$ and $\mathbf{B}_{2,t}$ are reference points in one particle of the body part. Here, the corners of the rectangle for modeling are the arms, which are taken as the reference points. If the feature point $\mathbf{F}_{k,t}$ is out of the region of one particle, the distances of $\overline{\mathbf{B}_{1,t}\mathbf{F}_{k,t}}$ and $\overline{\mathbf{B}_{2,t}\mathbf{F}_{k,t}}$ will be penalized with a larger penalty value $F_{\max}$. Moreover, the selected feature points will be updated every several frames to re-distribute the feature points when they are occluded or when the viewing direction is varying.

D. Elbow Angle on Image Plane

For the arm tracker, since one end of the upper arm is fixed to the face and the torso, the estimation of the upper arm is more stable than that of the forearm. The candidate for the forearm is attached to the upper arm; however, these two connected parts can overlap and be confusing when a user flexes their arm as in Fig. 13. The appearance of each body part on a 2D image is assumed to be free of overlap with one another. In order to avoid the overlapping false alarm, the estimated elbow angle $\psi_{e,t}$ between the upper arm and the forearm hypothesis on 2D image is constrained to be larger than the threshold δ_{thr}.

V. Experimental Results

The experimental video sequences captured by a Logitech webcam are processed by a PC with an Intel Core2 2GHz processor and 1GB RAM. The image resolutions are 320×240 pixels. Each partition of head and arms employs 30 particles. The corresponding weightings $\beta_{1,t}, \beta_{2,t}, \beta_{3,t}$ of three discriminability functions are assumed to be 0.4, 0.4 and 0.3, respectively. The relative weights c_1 and c_2 in (10) are set to 0.4 and 0.6, respectively. During the likelihood evaluation, we let the maximum matching distance $d_{\max} = 20$ in (12) and the penalty value $F_{\max} = 80$ in (13). The threshold δ_{thr} of elbow angle constraint is set to $2\pi/9$. In the results, the green and yellow rectangles stand for the right upper arm and forearm, respectively, whereas blue and orange rectangles stand for the left upper arm and forearm, respectively. In the following experiments, the statistical data, such as the root-mean-square (RMS) error, the standard deviation (STD) of the errors, and the average of computational time, are evaluated by repeating the experiments with ten times over the whole frames.

Fig. 14 shows the snapshots of tracking under a complex environment on a moving camera. Our proposed AMIS algorithm with the likelihood functions mentioned in Section 4 is compared with the tracker using the color likelihood (named *Color*), or using the color likelihood plus edge contour likelihood (named *Color+Edge*). All of the approaches consider the constraint of elbow angle on the image plane. We have manually labeled the upper body parts on the 2D image as the ground truth. The RMS and STD of the errors in the 2D joint position, the size of body part, and the arm angle are listed in Tables I and II.

In the experiment of comparing the likelihoods, it can be seen that the color likelihood [10] without the foreground or depth segmentation is hard to correctly track the motion of arms. The color likelihood combined with the motion enhanced edge likelihood (where the edge of the moving body parts cannot be efficiently enhanced with the moving camera platform) has a minimal contribution to the tracking, and may be confused by the cluttered background. However, by considering multiple visual clues in drawing particles, the AMIS method can assist in recovering the failures in tracking with these poor likelihood functions. The comparisons also imply that our method, with the designed likelihood functions, outperforms the compared existing methods.

Snapshots of 3D motion are shown in Fig. 15. It can be seen that the direction of movement of the target is parallel to the optical axis of the camera, produced by a person wearing a T-shirt. The estimations of particle filters with three different kinds of proposal functions are compared here, including our algorithm *AMIS*, particles drawn from the inverse kinematics of hand by sequential importance sampling (*SIS*), and particles drawn by sampling importance resampling (*SIR*) from the previous estimation. These three methods employ the same likelihood functions in Section 4. The tracking errors of each method are provided in Tables III and IV.

Without the depth information from a 2D image, the estimation of 3D motion is rather difficult. From frame #250 to #310 in the AMIS and SIS results of Fig. 15, the estimations of the right forearm and elbow are slightly misled by the particles from the proposal function with the inverse kinematics of the hand. The particle drawing methodology based on the inverse kinematics in (7) may not be effective for the hand motion in arbitrary 3D directions. The hand should be visible on the image; otherwise, the joint angles cannot be correctly solved through use of the inverse kinematics. In addition, the 3D arm motion, which contains the displacement perpendicular to the image plane, would result in significant ambiguity when it is projected on the image plane. Since it is an ill-conditioned problem, some postures cannot be detected from the solution of inverse kinematics. This error can be

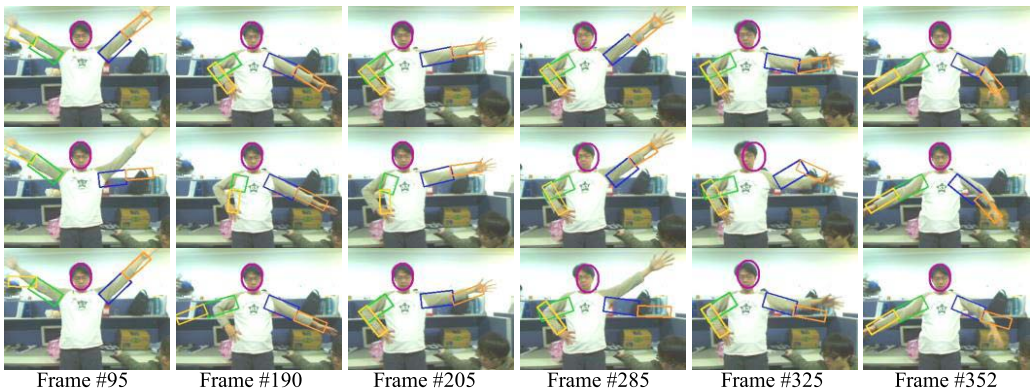


Fig. 14. Comparisons of tracking with different likelihood functions on a moving camera. (Top row: $AMIS$; middle row: $Color$; bottom row: $Color + Edge$.)

TABLE I

RMS ERROR AND STD OF THE ERROR IN 2D JOINT POSITION AND SIZE OF BODY PART.

(L: LEFT ARM, R: RIGHT ARM, W: WRIST, E: ELBOW, S: SHOULDER

(Unit: pixel)		Head position	L_W position	L_E position	L_S position	R_W position	R_E position	R_S position	L_upper width	L_upper height	L_fore width	L_fore height	R_upper width	R_upper height	R_fore width	R_fore height
AMIS	RMS	1.50	17.91	11.30	7.05	14.54	8.51	6.86	3.96	11.71	5.59	17.05	4.11	10.76	7.06	7.70
	STD	1.15	7.93	7.30	3.79	6.35	4.79	3.65	2.43	5.54	3.06	8.11	2.61	5.83	3.81	4.47
Color	RMS	1.62	21.36	17.87	8.08	35.84	19.06	7.67	3.96	18.96	5.67	21.47	4.13	10.58	7.05	8.77
	STD	1.22	13.44	11.49	4.13	24.28	12.51	4.32	2.39	6.00	3.14	8.06	2.67	5.96	3.87	5.45
Color	RMS	1.62	31.98	14.22	8.61	30.57	15.57	8.25	4.05	18.88	5.65	20.22	4.11	11.05	7.15	8.55
+Edge	STD	1.25	21.15	8.03	4.67	18.94	9.49	4.78	2.46	5.60	3.13	8.53	2.59	5.88	3.80	5.00

TABLE II

RMS ERROR AND STD OF THE ERROR IN ARM JOINT ANGLE.

(L: LEFT ARM, R: RIGHT ARM, E: ELBOW, S: SHOULDER)

(Unit: radius)	L_E angle		L_S angle		R_E angle		R_S angle	
	RMS	STD	RMS	STD	RMS	STD	RMS	STD
AMIS	0.27	0.17	0.15	0.10	0.18	0.14	0.11	0.07
Color	0.53	0.45	0.35	0.30	0.19	0.15	0.23	0.15
Color+Edge	0.45	0.34	0.20	0.14	0.21	0.17	0.19	0.13

TABLE III

RMS ERROR AND STD OF THE ERROR IN 2D JOINT POSITION AND SIZE OF BODY PART.

(L: LEFT ARM, R: RIGHT ARM, W: WRIST, E: ELBOW, S: SHOULDER)

(Unit: pixel)		Head position	L_W position	L_E position	L_S position	R_W position	R_E position	R_S position	L_upper width	L_upper height	L_fore width	L_fore height	R_upper width	R_upper height	R_fore width	R_fore height
AMIS	RMS	1.56	21.82	14.71	6.32	20.94	9.98	7.40	2.75	13.13	5.15	13.75	3.47	14.08	4.41	11.55
	STD	0.93	11.31	7.45	3.14	11.07	6.27	3.85	1.67	5.90	2.42	6.52	2.03	6.42	2.68	6.65
SIS	RMS	1.64	30.15	21.79	6.49	35.49	15.88	7.66	2.77	14.49	5.18	13.43	3.47	14.57	4.45	13.19
	STD	1.02	18.48	13.99	3.54	23.67	10.85	4.02	1.70	5.82	2.42	6.20	2.04	6.86	2.69	7.59
SIR	RMS	1.64	45.31	25.01	6.49	34.60	16.69	7.70	2.79	15.82	5.14	14.49	3.49	14.17	4.43	12.69
	STD	1.03	27.28	13.01	3.37	23.09	10.29	4.02	1.71	6.73	2.40	6.90	2.05	6.91	2.69	7.54

TABLE IV

RMS ERROR AND STD OF THE ERROR IN ARM JOINT ANGLE.

(L: LEFT ARM, R: RIGHT ARM, E: ELBOW, S: SHOULDER)

(Unit: radius)	L_E angle		L_S angle		R_E angle		R_S angle	
	RMS	STD	RMS	STD	RMS	STD	RMS	STD
AMIS	0.31	0.23	0.20	0.13	0.19	0.15	0.13	0.10
SIS	0.38	0.28	0.34	0.27	0.32	0.26	0.25	0.20
SIR	0.41	0.27	0.33	0.23	0.45	0.36	0.21	0.16

fixed by the AMIS when the arm is captured more clearly. However, the SIS method presents several mistakes in the elbow estimation and hand extraction, where the estimation of the left arm is extended over the real one at several frames of the SIR. The SIS and SIR methods do not efficiently generate particles at the proper arm locations, the fusion of multiple visual likelihood functions mentioned in Section 4 cannot contribute to the arm tracking. Nevertheless, we can see that

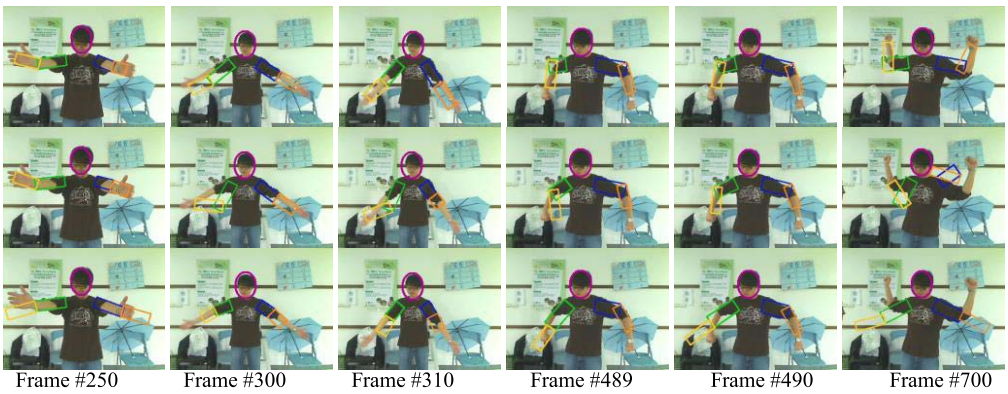


Fig. 15. Comparisons of particle filter with different proposal functions on a static camera. (Top row: $AMIS$; middle row: SIS; bottom row: SIR.)

TABLE V
COMPARISON OF THE COMPUTATIONAL TIME

(Unit: ms.)		*AMIS*	*SIS*	*SIR*
Head	Draw samples / Predict / Update	1.100 / 0.129 / 9.167	1.100 / 0.130 / 9.401	1.100 / 0.130 / 9.267
Arms	Draw samples / Predict / Update	21.501 / 0.130 / 29.009	15.267 / 0.133 / 29.036	8.633 / 0.133 / 27.637
total		61.036 (≈ 16 fps)	55.067 (≈ 18 fps)	46.900 (≈ 21 fps)

Fig. 16. Snapshots of tracking with a moving camera and external disturbances.

the AMIS produces particles that diversify in state and are complementary to each other.

The computational time of each major process is also analyzed in Table V. The proposed algorithm AMIS, which spends a lot of time on drawing particles and verifying them, is a little bit slower than the SIR and SIS methods; however, it still can achieve near real-time ability (about 16 fps) and has better 2D estimation than that from the SIR or SIS method.

The remaining experiments are then performed in an on-line and real-time manner to validate our tracking ability. The video sequence shown in Fig. 16 presents the system's ability to deal with the challenging scenario of a moving camera and under the disturbances of another person. This cluttered environment also contains the complex background and the non-uniform illumination. Initially, the system detects two human faces. However, only the person with the posture in Fig. 2 will be recognized as the subject with whom interaction will take place. The non-tracked person waves their hands in the background, and even in the foreground, of the tracked person, in order to disturb or occlude arm tracking. The pose estimation approaches based on the background subtraction or the depth-based segmentation would easily fail in such cluttered environment, however, the proposed tracking algorithm can successfully recognize the correct posture of target person and even overcome the temporal occlusion.

VI. CONCLUSION

This paper presents a particle filter methodology for tracking the face and arms of the human upper body in cluttered environments with a monocular camera. The adaptive multiple importance sampling particle filter of the arm tracker combines the hand position, the arm edge, and temporal information to efficiently generate hypotheses with various characteristics. The number of particles generated from each importance function is dynamically changing according to the distinctive visual clues of human posture. The likelihood model, which takes the visual cues of the motion, feature point distribution, color appearance, and shape of the human face and arms to verify various hypotheses from the multiple importance sampling schemes, is designed using color and edges observed on the 2D image plane. In order to easily apply the algorithm to human-machine interaction, our approach can handle the situation of a single moving camera platform and achieve near real-time performance. Unlike other human pose estimations, which require the background or the depth information to identify a human silhouette, our method is not constrained by a static camera or a depth sensor. The proposed algorithm demonstrates excellent ability of extracting the arms in cluttered environments, even with disturbance from occasional occlusions.

When the upper body tracking system is implemented on a robot in the future, an extra benefit is the mobility of that robot. A static monocular camera may not obtain sufficient

information from a 2D image since human posture is dynamic and diverse. Some postures can be undistinguishable in a 2D image. The mobility of the robot allows it to move to a better position to acquire enriched image observations. Moreover, the action recognition for human-machine interaction will be performed by utilizing the current 2D visual tracking results, and the 3D posture of a human body will be further estimated.

REFERENCES

[1] L. Sigal and M. J. Black, "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation," in *Proc. IEEE Comput. Soc. Conf. Comput Vis. Pattern Recognit.*, Jun. 2006, pp. 2041–2048.

[2] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon, "The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 103–110.

[3] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman, "Long term arm and hand tracking for continuous sign language TV broadcasts," in *Proc. 19th Brit. Mach. Vis. Conf.*, 2008, pp. 1105–1114.

[4] K. Dong-Wan and O. Jun, "Postures of a human wearing a multiple-colored suit based on color information processing," in *Proc. ICME*, vol. 1. 2003, pp. 261–264.

[5] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2D articulated human pose estimation and retrieval in (almost) unconstrained still images," *Int. J. Comput. Vis.*, vol. 99, no. 2, pp. 190–214, 2012.

[6] L. Mun Wai and R. Nevatia, "Human pose tracking in monocular sequence using multilevel structured models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 27–38, Jan. 2009.

[7] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Understand.*, vol. 104, nos. 2–3, pp. 90–126, 2006.

[8] S. Knoop, S. Vacek, and R. Dillmann, "Fusion of 2D and 3D sensor data for articulated body tracking," *Robot. Autom. Syst.*, vol. 57, no. 3, pp. 321–329, 2009.

[9] L. Jing-Feng, X. Yi-Hua, C. Yang, and J. Yun-De, "A real-time 3D human body tracking and modeling system," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2006, pp. 2809–2812.

[10] C. Liu, P. Liu, J. Liu, J. Huang, and X. Tang, "2D articulated pose tracking using particle filter with partitioned sampling and model constraints," *J. Intell. Robot. Syst.*, vol. 58, no. 2, pp. 109–124, 2010.

[11] L. Calderita, J. Bandera, P. Bustos, and A. Skiadopoulos, "Model-based reinforcement of Kinect depth data for human motion capture applications," *Sensors*, vol. 13, no. 7, pp. 8835–8855, 2013.

[12] Y. Zhu and K. Fujimura, "A Bayesian framework for human body pose tracking from depth image sequences," *Sensors*, vol. 10, no. 5, pp. 5280–5293, 2010.

[13] M. Yin, Y. Bo, G. Zhao, and W. Zou, "Adaptive block-fusion multiple feature tracking in a particle filter framework," in *Proc. IEEE 3rd Annu. Int. Conf. Cyber Tech. Autom., Control Intell. Syst. (CYBER)*, May 2013, pp. 400–404.

[14] S. Calderara, R. Cucchiara, and A. Prati, "Action signature: A novel holistic representation for action recognition," in *Proc. IEEE 5th Int. Conf. AVSS*, Sep. 2008, pp. 121–128.

[15] S. Ishikawa, T. Joo Kooi, K. Hyoungseop, and S. Ishikawa, "3-D recovery of a non-rigid object from a single camera view," in *Proc. SICE Annu. Conf. (SICE)*, Sep. 2011, pp. 447–450.

[16] A. Elgammal and L. Chan-Su, "Inferring 3D body pose from silhouettes using activity manifold learning," in *Proc. IEEE Comput. Soc. Conf. CVPR*, vol. 2. Jul. 2004, pp. 681–688.

[17] A. F. Garcia-Fernandez, J. Grajal, and M. R. Morelande, "Two-layer particle filter for multiple target detection and tracking," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 49, no. 3, pp. 1569–1588, Jul. 2013.

[18] E. Veach and L. J. Guibas, "Optimally combining sampling techniques for Monte Carlo rendering," in *Proc. 22nd Annu. Conf. Comput. Graph. Interact. Tech.*, 1995, pp. 419–428.

[19] C. Panagiotakis, E. Ramasso, G. Tziritas, M. Rombaut, and D. Pellerin, "Shape-motion based athlete tracking for multilevel action recognition," in *Articulated Motion Deformable Objects*, vol. 4069, F. Perales and R. Fisher, Eds. Berlin, Germany: Springer-Verlag, 2006, pp. 385-394.

[20] S. R. Buss and J. S. Kim, "Selectively damped least squares for inverse kinematics," *J. Graph. Tools*, vol. 10, no. 3, pp. 37–49, 2005.

[21] S. Georgiana and C.-D. Caleanu, "Sparse feature for hand gesture recognition: A comparative study," in *Proc. Int. Conf. TSP*, 2013, pp. 858–861.

[22] M.-B. Kaâniche and F. Brémond, "Recognizing gestures by learning local motion signatures of HOG descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2247–2258, Nov. 2012.

[23] C. Rasmussen and G. D. Hager, "Probabilistic data association methods for tracking complex visual objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 560–576, Jun. 2001.

[24] V. Robert, C. Arthur, S. Istvan, and N. Sergiu, "Efficient real-time contour matching," in *Proc. IEEE Int. Conf. Intell. Comput. Commun. Process.*, Sep. 2012, pp. 193–199.

[25] S. Sedai, M. Bennamoun, and D. Q. Huynh, "A Gaussian process guided particle filter for tracking 3D human pose in video," *IEEE Trans. Image Process.*, vol. 22, no. 11, pp. 4286–4300, Nov. 2013.

[26] H. B. Yu, G. H. Wang, Q. Cao, and Y. Sun, "A novel particle filtering algorithm based on state fusion," in *Proc. IET Int. Radar Conf.*, 2013, pp. 1–5.

Cheng-Ming Huang received the B.S. degree from National Chiao Tung University, Hsinchu, Taiwan, the M.S. degree from National Cheng Kung University, Tainan, Taiwan, and the Ph.D. degree from National Taiwan University, Taipei, Taiwan, in 2000, 2002, and 2009, respectively. From 2009 to 2011, he was a Post-Doctoral Researcher with the Department of Electrical Engineering, National Taiwan University. Since 2011, he has been an Assistant Professor with the Department of Electrical Engineering, National Taipei University of Technology. His research interests include visual tracking, visual servoing, multicamera cooperation, and microaerial vehicle.

Yi-Ru Chen received the B.S. and M.S. degrees in electrical engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, and the National Taiwan University, Taipei, in 2007 and 2009, respectively. Since 2013, she has been with Nook Media LLC, Taipei, where she is currently a Software Engineer, developing Android applications. Her research interests include visual tracking and human posture estimation.

Li-Chen Fu (F'04) received the B.S. degree from National Taiwan University, Taiwan, and the M.S. and Ph.D. degrees from the University of California, Berkeley, USA, in 1981, 1985, and 1987, respectively.

Since 1987, he has been a member of the faculty, and is currently a Full Professor with the Department of Electrical Engineering and the Department of Computer Science and Information Engineering, National Taiwan University, where he received the Lifetime Distinguished Professorship Award in 2007. He has received numerous academic recognitions, such as Distinguished Research Awards from the National Science Council, Taiwan, the Irving T. Ho Chair Professorship, and the Macronix Chair Professorship. He serves as the Editor-in-Chief of the *Asian Journal of Control*, the President of the Asian Control Association, and a Distinguished Lecturer of the IEEE Control Systems Society from 2013 to 2015. His research interests include robotics, smart home, visual detection and tracking, intelligent vehicle, production scheduling, virtual reality, and control theory and applications.