The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Visual Tracking of Human Head and Arms with a Single Camera

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

Abstract—**This paper presents an upper body tracking algorithm with a single monocular camera. In order to be suitable for human robot interaction, the designed method should be free to work on the moving camera platform and also can achieve real-time performance. The dimension of human posture model is extremely high, and we hereby focus on the visual extraction of head and arms. A hierarchical structure model is proposed to solve the tracking problem by particle filter with partitioned sampling in the order of head, upper arm and the forearm. The hand position, straight edge of arm and temporal information are combined by the multiple importance sampling particle filter to efficiently estimate the irregular gesture of arms on image frames. The visual clues of the motion, appearance and shape to human face and arms are to verify the various hypotheses from the multiple importance sampling schemes. To validate the effectiveness of the proposed tracking approach, extensive experiments have been performed, of which the results appear to be quite promising.**

I. Introduction

ACCORDING to the image captured by camera, there are many different kinds of information which can be derived in computer vision and robotics fields, such as those about the environment, contents, instructions, etc. For example, to improve the interactive function for human-robot interaction in a scenario with home service robot, trying to understand the meaning of human behavior is a more effective way to let the robot be socialized. Because the camera mounted on a robot, which focuses on the upper body of the human for interaction, has limited field of view, this research work restricts the task to tracking of the face and upper limbs only.

In general, there are two kinds of popular approaches which are used for human tracking, namely, background subtraction [1] and depth-based segmentation [2]. The background subtraction is a problem which has attracted extensive interests in the field of computer vision. However, a powerful background subtraction method with background initialization, updating, and classification usually suffers from heavy computation cost when operating on a moving camera platform. Thus, it is unlikely to achieve real time tracking based on this kind of approach with possibly mobile camera platform. On the other hand, the dense disparity map, which is acquired from the stereo camera platform or auxiliary sensors such as multiple cameras, sonar, and laser, etc., is a critical basis for achieving depth-based segmentation. The particular requirement of auxiliary sensor setup and the excessive computational cost of depth data observation also cause serious concerns.

So far, there have been many researches proposed for human or posture detection in the literature. The traditional works estimate the human by marking the human body parts [3] or by asking the human to wear some special clothes [4]. However, this is inconvenient and can only be used in special situations with complete setup of equipments. To improve these methods, many researches provide alternative approaches without relying on markers worn or attached onto the human body. Among them, the learning-based method [5] is applied to detect the human's position. The work in [6] gave an overview on human tracking via tracking of some body parts, such as face, hands, fingers, etc. Using background subtraction or foreground segmentation techniques, the silhouette-based methods [7] can be utilized to detect humans with a static camera. The dense disparity map is constructed for estimating human posture by using the 3D information from a multi-camera system [2]. In order to integrate the posture detection algorithm with the human-robot interface, the tracking system apparently can not assume that the environment is with static camera and simple background since both the human and the camera mounted on the robot may move while both are interacting.

This work presents a monocular vision-based tracking algorithm that aims to fast and accurately track the upper body posture of a human. Concerning the processing time, applying some detection technique in each frame during the course of tracking a human generally, however, will surely be shown inefficient. Combining some appropriate tracking algorithms can actually reduce the solution complexity. The particle filter, which is a popular tracking algorithm, can successfully solve the non-Gaussian state estimation problems in nonlinear systems. The particle filter is cooperated with the partitioned sampling scheme [10] to alleviate the intense need for a large number of particles in the case with estimation in high dimensional state space. Besides, tracking human beings is recognized to be one of the most difficult tasks, simply because human's motions are fast, nonlinear, unpredictable, and there are so many kinds of possible human posture. The multiple importance sampling (MIS) [11] is employed here to fuse all of the clues, such as the hand, obvious line of arm and temporal association, that we can obtained for dealing with the human's upper body tracking problem. Parallel to that, we also adopt the appearance color and shape edge information of the head and

Y. R. Chen and C. M. Huang are with the Department of Electrical Engineering, National Taiwan University, Taiwan, ROC.

L. C. Fu is with the Department of Electrical Engineering and Computer Science and Information Engineering, National Taiwan University, Taiwan, ROC (e-mail: lichen@ntu.edu.tw).

arms to design the evaluation of likelihood function for verifying the tracked parts of human's upper body.

The rest of this paper is organized as follows. In section 2, we first introduce the human upper body model. And then in section 3, the tracking algorithm is described. The particle filters with partitioned sampling method and multiple importance sampling algorithm are presented to efficiently track the posture of head and arms. The design of likelihood functions utilizing the visual cues is explained in section 4. In section 5, we demonstrate several experimental results to validate the effectiveness of the proposed tracking approach. Finally, we conclude this paper in section 6.

II. Human Upper Body Model

A. Definition

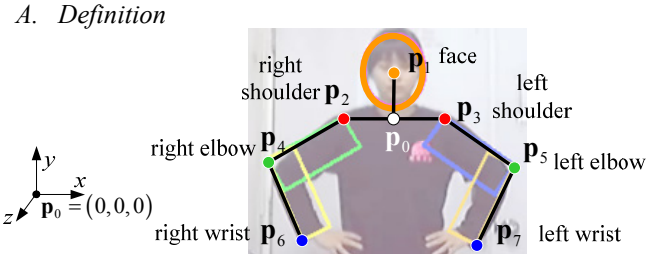


Fig. 1. The 3D human model and its projection on the 2D image plane.

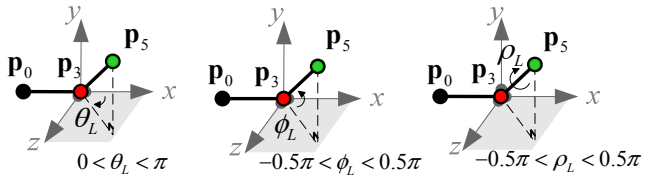


Fig. 2. Definition of joint angles for the left shoulder with limited range.

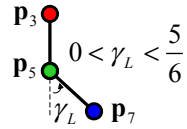


Fig. 3. Definition of the joint angle for the left elbow with limited range. The stick between $\mathbf{p}_5$ and $\mathbf{p}_7$ lies on the plane decided by the angle ρ_L.

As discussed in the previous work [6], there are several kinds of human model which can be categorized into 2D models and 3D models. Although we focus on the 2D tracking in the image plane in this paper, the 3D stick model [6] is still applied here to avoid the unfeasible joint positions simply generalized from 2D model and be able to estimate the 3D posture further [6]. As shown in Fig. 1, $\mathbf{p}_0$ is set as the origin of the coordinate system. In general, the scale of every body part is proportional to the face size. We model the face as an ellipse whose ratio between the length of the minor axis to that of the major axis is 1:1.2. Also, we approximate the scale between the width of the face and the torso as 1:2. The length of the upper arm and the forearm is about 1.5 times of the width of the face, and the length of the upper arm is as same as that of the forearm. Using these general assumptions of upper body proportions, a stick model as Fig. 1 can be constructed after the states of the arm angles are determined.

The joint positions, such as positions of the elbows $\mathbf{p}_4$ and $\mathbf{p}_5$, and positions of the wrists $\mathbf{p}_6$ and $\mathbf{p}_7$, can be determined by the transformation matrices with the states of both arms. The definitions and limitations of the involved angles are mentioned as Fig. 2 and Fig. 3. The state can be separated into three parts. The first part describes the state of face $\mathbf{x}_P$, including the center of face position (u,v), and the face scale r. The second and third parts describe the posture of the right arm $\mathbf{\Theta}_R=[\theta_R,\phi_R,\rho_R,\gamma_R]$ and the left arm $\mathbf{\Theta}_L=[\theta_L,\phi_L,\rho_L,\gamma_L]$, respectively. The full state $\mathbf{X}$ can be defined as

$$\mathbf{X} = [\mathbf{x}_P, \mathbf{\Theta}_R, \mathbf{\Theta}_L]. \quad (1)$$

The hypothesis predicted from the 3D human model is then projected on the *xy* plane. In order to obtain the observations of human model from image, the face is modeled by an ellipse, and the arms are modeled by four rectangles, in which two rectangles for the upper arms and two rectangles for the forearms. Here, we assume that the human's head and back should not be bended purposely during the human-robot interaction, and each body part will not be fully occluded. In addition, the person who would like to do the interaction should face to the camera, and the body orientation is assumed to be a noise during the tracking process.

B. Initialization

In general, the size of every part is proportional to the face size, like the Virtruvian Man created by Leonardo da Vinci. But the accurate shoulder width $w_{shoulder}$, neck length l_{neck}, upper arm width $w_{upperarm}$, forearm width $w_{forearm}$, and arm length l_{arm} are variable from one person to another. An initial posture should be picked to provide the angle and length information of the parameters and initial states. The initial posture should also be easy and general, so that the system does not easily mislead the posture. We define that the user putting hands on the waist as the initial posture like Fig. 1. When the user doing this initial posture, it forms an obvious included angle between the upper arm and forearm and shows the explicit size of each upper body part.

Then, the line detector [8] which finds lines in edge image by using Hough transform below the detected face can be applied to find the straight edges of the torso and arms as illustrated in Fig. 6 (a). These detected lines will be categorized by different angle and position to decide what body part they belong to. Some lines with the disturbance originated from the background or clothes will be filtered out by taking the symmetry of human body parts and averaging the group of lines belonging to the same part. The shoulder width $w_{shoulder}$ is decided by the distance between the vertical lines on the sides of the torso, and the width of the upper arm $w_{upperarm}$ and forearm $w_{forearm}$ can be evaluated by the non-vertical lines belonging to the sides of each arm. The joint position of shoulder, elbow and wrist can also be determined at the ends of the non-vertical lines. Hence, the arm length l_{arm} and neck length l_{neck} of one specified user are obtained now. Except improving the accuracy of the human model, the user can present the intention of the human-robot interaction through doing this initial posture to

start the intelligent system.

III. Tracking Methodology

As mentioned in Section 2, the dimension of the human model described in (1) is very high. If the particle filter estimates the state $\mathbf{X}$ at one time instant t, denoted as $\mathbf{X}_t$, then the system must generate many particles to cover the whole high dimensional state space. The large number of particles will cost a lot of computational time. Furthermore, many particles will generate with negligible weight will cause the system ineffective. Given the image observation $\mathbf{z}_t$ obtained from one monocular camera, the tracking problem can be formulated as the following probabilistic form:

$$p(\mathbf{X}_t|\mathbf{z}_t) = p(\mathbf{x}_{P,t}, \boldsymbol{\Theta}_{R,t}, \boldsymbol{\Theta}_{L,t}|\mathbf{z}_t), \quad (2)$$

where $\mathbf{x}_{P,t}$, $\boldsymbol{\Theta}_{R,t}$, $\boldsymbol{\Theta}_{L,t}$ are the states of face, right arm and left arm, at time instant t, respectively. Since the head and arms are fixed on the torso, the state arms can be estimated easier once the location of head has been obtained. Moreover, assume that the left arm and right arm will not occlude each other during this human-machine interaction, and states of them are independent. The posterior distribution in (2) can then be decomposed and reduced as

$$\begin{aligned} p(\mathbf{x}_{P,t}, \boldsymbol{\Theta}_{R,t}, \boldsymbol{\Theta}_{L,t}|\mathbf{z}_t) &= p(\boldsymbol{\Theta}_{R,t}, \boldsymbol{\Theta}_{L,t}|\mathbf{x}_{P,t}, \mathbf{z}_t)\, p(\mathbf{x}_{P,t}|\mathbf{z}_t), \quad (3) \\ &= p(\boldsymbol{\Theta}_{R,t}|\mathbf{x}_{P,t}, \mathbf{z}_t)\, p(\boldsymbol{\Theta}_{L,t}|\mathbf{x}_{P,t}, \mathbf{z}_t)\, p(\mathbf{x}_{P,t}|\mathbf{z}_t). \end{aligned}$$

From the derivative of (3), a hierarchical algorithm is developed by the particle filter with partitioned sampling concept [10]. First, we focus on the face tracking $p(\mathbf{x}_{P,t}|\mathbf{z}_t)$ by the sampling importance resampling (SIR) particle filter [9]. Second, based on the face tracking result and the human model, the location of shoulder can be estimated. The distribution $p(\boldsymbol{\Theta}_{R,t}|\mathbf{x}_{P,t}, \mathbf{z}_t)$ and $p(\boldsymbol{\Theta}_{L,t}|\mathbf{x}_{P,t}, \mathbf{z}_t)$ of arms are tracked by the multiple importance sampling (MIS) particle filter [11]. The flow chart of the overall system is summarized as Fig. 4.

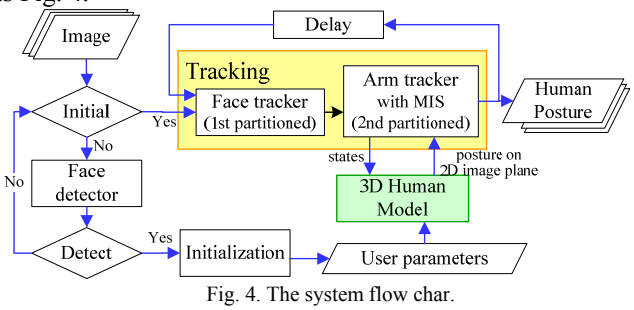
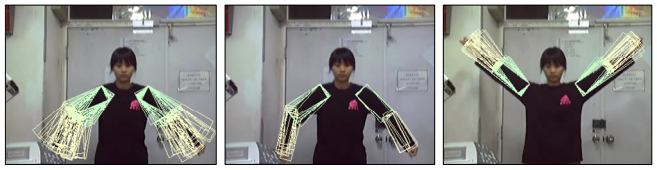


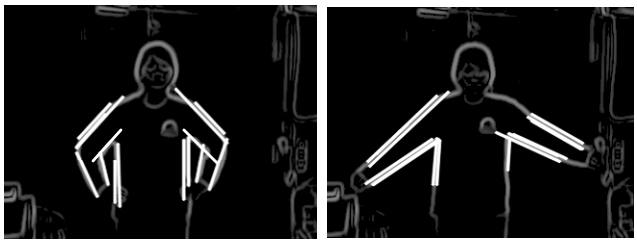
Fig. 4. The system flow char.

In order to avoid the impoverishment phenomenon of high dimensional arm tracking, the number of particles and the prediction variance should be increased to raise the diversity. However, this process would cause the heavy computational complexity and some particles with negligible weight. The MIS particle filter [11] uses several proposal functions to generate particles from current 2D image observation and latest states. The multiple importance sampling scheme is applied here to minimize the estimation variance due to the fusion of various proposal functions. The estimations of right arm and left arm are independent, and the algorithm applied for each of them is the same. We redefine the states of one arm at time instant t as $\boldsymbol{\Theta}_t$, *i.e.*, $\boldsymbol{\Theta}_t$ represents for $\boldsymbol{\Theta}_{R,t}$ or $\boldsymbol{\Theta}_{L,t}$.



(a) resampling; (b) inverse kinematic; (c) line detector;

Fig. 5. The particles generated from multiple proposal functions.



(a) putting hands on the waist; (b) stretching out the full arm;

Fig. 6. Detected lines (bold and white) of human posture on the edge image.

Three kinds of proposal functions for covering the unpredictable movement of arm are applied on previous states and current 2D image observations as illustrated in Fig. 5. The first proposal function $q_{1,t}(\boldsymbol{\Theta}_t)$ uses the estimates of the latest posterior, the second proposal function $q_{2,t}(\boldsymbol{\Theta}_t)$ uses the inverse kinematics from hand position, and the third proposal function $q_{3,t}(\boldsymbol{\Theta}_t)$ uses the line detector for the obvious edge of arm as shown in Fig. 6. The number of particles $m_{i,t}$ is based on the fullness of arm information provided by the proposal function $q_{i,t}(\boldsymbol{\Theta}_t)$ and denoted as

$$\begin{aligned} m_{1,t} &= N_s - (m_{2,t} + m_{3,t}), \\ m_{2,t} &= P_h(N_s - m_{3,t})/2, \quad (4) \\ m_{3,t} &= \begin{cases} N_s/3, & \text{if there is an obvious line detected} \\ 0, & \text{otherwise} \end{cases}, \end{aligned}$$

where N_s is the total number of particles, and P_h is the probability of the hand appeared. The particle number $m_{i,t}$ of each proposal function is dynamically decided by the 2D image information at time instant t. When one proposal function $q_{i,t}(\boldsymbol{\Theta}_t)$ has better observation, the particle number sampled from that proposal function $q_{i,t}(\boldsymbol{\Theta}_t)$ will increase more quantity. Collect all the samples $\boldsymbol{\Theta}_{i,t}^{(j)} \sim q_{i,t}(\boldsymbol{\Theta}_t)$, $j = 1, \ldots, m_{i,t}$, $i = 1, 2, 3$, the posterior of each arm in (3) can then be yielded by the following Monte Carlo approximation with a set of weighted samples $\{\boldsymbol{\Theta}_{i,t}^{(j)}, \omega_{i,t}^{(j)}\}$:

$$p(\boldsymbol{\Theta}_t|\mathbf{x}_{P,t}, \mathbf{z}_t) \approx \alpha \sum_{i=1}^{3} \sum_{j=1}^{m_{i,t}} \omega_{i,t}^{(j)} \delta(\boldsymbol{\Theta}_t - \boldsymbol{\Theta}_{i,t}^{(j)}), \quad (5)$$

where α is a normalization constant, $\delta(\cdot)$ is the Dirac delta function, and $\omega_{i,t}^{(j)} = \omega_{i,t-1}^{(j)} p(\mathbf{z}_t|\boldsymbol{\Theta}_{i,t}^{(j)}) p(\boldsymbol{\Theta}_{i,t}^{(j)}|\boldsymbol{\Theta}_{i,t-1}^{(j)}, \mathbf{x}_{P,t})$

$/q_{i,t}(\mathbf{\Theta}_{i,t}^{(j)})$ is the corresponding weight of each particle.

IV. Likelihood Evaluation

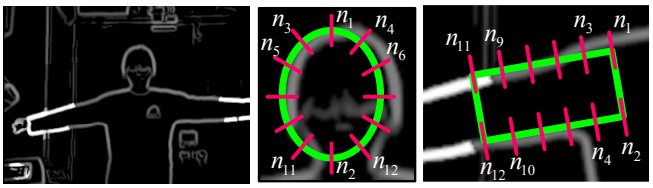
By projecting the human model onto the image plane, we can get the image observation for each hypothesis state **X** and evaluate the weight of each particle. The weight of each particle is measured by the evaluation of likelihood function. For operating on a single monocular camera with real-time performance, the designed likelihood function is simply evaluated by image information through the computation of color and edge contour with the geometry constraints.

A. Color Histogram

The color likelihood [13] uses the Bhattacharyya distance to evaluate the similarity between the reference color histogram defined in the initialization process and the color histogram corresponding to each particle. The histogram of inner part in each particle should be similar to the reference color, however, the histogram of outer part around each particle should be dissimilar to the reference one. Note that, when the arm closes to the torso, the color of each part may be the same due to the clothes. Under this situation, the joint likelihood [15] will be considered to deal with the overlapping region of different body parts.

B. Edge Contour

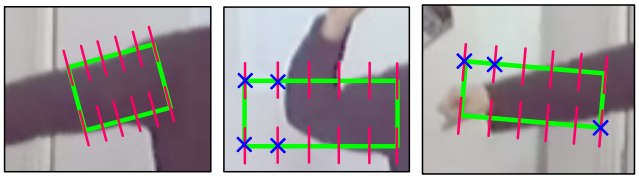
We classify the scenario types into tracking with static camera and with motion camera, which can be determined by the encoder of the camera motion platform. With the static camera, two consecutive images are compared to find the different part. In order to distinguish the human body and environment, the motion detection is used to enhance the edge around the human who is interacting with the robot by moving arms. Attaching the edge of this different part to current edge image, we can obtain the motion enhanced edge image as Fig. 7(a). When considering the motion camera, the edge contour is equivalent to operate on the common edge image and the optical flow detailed in the next subsection will assist the tracking of each body part.



(a) motion enhanced edge; (b) head; (c) upper arm/forearm;

Fig. 7. The motion enhanced edge and the shape of contour template. The red line segments indicate the matching direction of each control points.

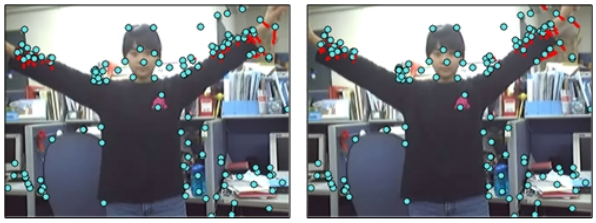
The contour of face is modeled by an ellipse template as Fig. 7(b), and the contour of arms is modeled by the rectangle template as Fig. 7(c). The continuous contour with is represented with N_c discrete control points n_i. Here, we employ the following four likelihood functions to evaluate the visual cue of edge contour at these control points: contour matching likelihood, contour intensity likelihood, contour length likelihood, and contour symmetry likelihood. The contour matching likelihood [14] computes the similarity between the reference shape model with state variable and its neighborhood edge image. Since the enhanced edge image emphasizes the edge in the motion area, *i.e.*, the edge points in motion area with higher intensity indicates that this point is belonging to the motion area which originates from the interacting human body when the camera platform is static. The contour length likelihood encourages the hypothesis with longer rectangle for reducing the case as shown in Fig. 8(a) and penalizes the segment of rectangle without edge evidence like illustrated as Fig. 8(b). Moreover, the other significant characteristic of human body part is symmetry. The contour symmetry likelihood accumulates the difference of matching distance in each control point pair as Fig. 8(c), lets the candidate fit with more appropriate position.



(a) shorter length; (b) longer length; (c) unsymmetrical;

Fig.8. Several false hypotheses on contour likelihood of upper arm. The cross marks denote the invalid matching of control point pairs.

C. Optical Flow



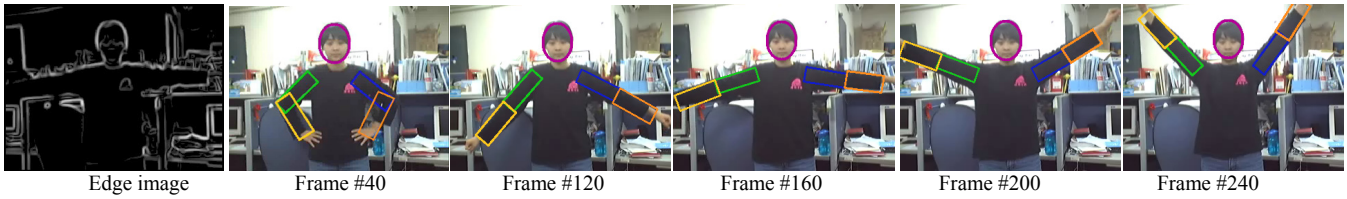
Frame t -1 Frame t

Fig. 9. The displacement of feature points in image sequence. The blue dots denote the feature points, and the red arrows denote the optical flows.

From an image sequence which is captured from a moving camera to the static scene, we can see that there is a displacement for the same scene in the image space at different time instant. This displacement is also called the optical flow. We employ the Kanade-Lucas-Tomasi (KLT) feature tracker [12] to obtain the optical flow in an image sequence. On the other hand, the moving object may result in the displacement whose magnitude and orientation are different from that of the static scene. Fig. 9 presents the optical flows originate from the motion of forearm. Although KLT algorithm can track feature points, it can not identify which feature point is belonging to which body part or the background. Hence, the geometry of feature points belonging to a tracked body part should be invariant between two consecutive image frames.

V. Experimental Results

The video sequence is processed by a PC with Intel Core2 2GHz processor and 1GB RAM, and images are captured by a Logitech webcam. The image resolution is 320×240 pixels. The particle number of each partition as mentioned in Section 3.A is 30 particles. In the results, the green and yellow



Edge image Frame #40 Frame #120 Frame #160 Frame #200 Frame #240

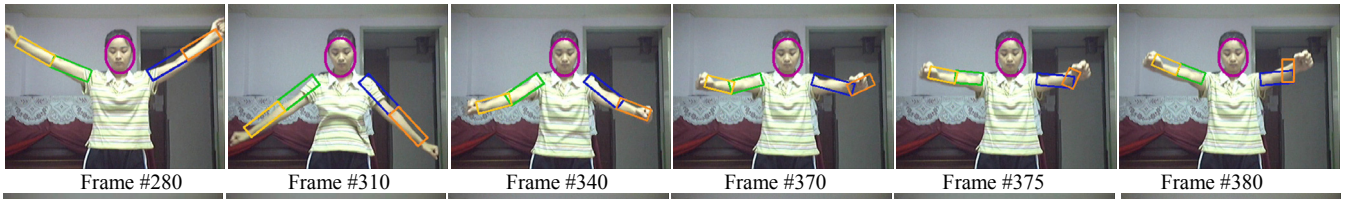
Fig. 10. Snapshots of tracking under complex environment with a static camera.

position (pixel)	face	shoulder		elbow		wrist	
		right	left	right	left	right	left
RMS	0.57	0.70	3.49	2.50	3.94	5.77	4.71
STD	0.42	0.55	2.62	2.89	3.24	4.54	3.56

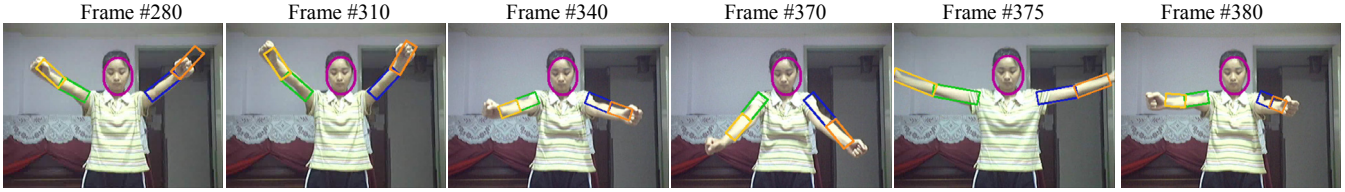
Table. 1. The RMS error and standard deviation of the error in 2D joint position.

angle (degree)	upper arm		forearm	
	right	left	right	left
RMS	4.41	4.12	1.82	2.51
STD	3.82	3.07	1.53	1.87

Table. 2. The RMS error and standard deviation of the error in arm angle.



Frame #280 Frame #310 Frame #340 Frame #370 Frame #375 Frame #380



Frame #390 Frame #400 Frame #430 Frame #460 Frame #490 Frame #520

Fig. 11. Snapshots of 3D motion of human arms with a static camera.

position (pixel)	face	shoulder		elbow		wrist	
		right	left	right	left	right	left
RMS	0.60	0.61	2.91	2.28	5.41	3.93	6.07
STD	0.46	0.47	2.61	1.70	4.47	2.87	4.63

Table. 3. The RMS error and standard deviation of the error in 2D joint position.

angle (degree)	upper arm		forearm	
	right	left	right	left
RMS	3.82	4.15	2.09	5.35
STD	3.18	3.47	1.71	4.31

Table. 4. The RMS error and standard deviation of the error in arm angle.

rectangles stand for the right upper arm and forearm, respectively, whereas blue and orange rectangles stand for the left upper arm and forearm, respectively.

Fig. 10 shows the snapshots in the complex environment with a static camera. The background has many confusing edge noises as shown in the edge image, so the estimated arm width may be disturbed by the noise. Even though the rectangle is thinner than the real arm, the results are still good by using the color feature. We manually label the upper body parts on the 2D image as the ground truth. The error of each 2D joint position between the estimation and ground truth is computed from $\mathbf{p}_1$ to $\mathbf{p}_7$, and the error of each arm angle is computed from the estimated rectangles of upper arm and forearm. The root-mean-square (RMS) error and standard deviation (STD) of the errors in joint position and arm angle are listed in Table. 1 and Table. 2, respectively.

Fig. 11 shows the snapshots of 3D motion, which means that the moving direction of the target would be parallel to the optical axis of the camera, produced by the person wearing the T-shirt and captured on a static camera. Without the depth information from 2D image, the estimation of 3D motion is rather difficult. The MIS particle filter of arm tracker mentioned in Section 3.C uses multiple visual clues as the importance functions to generate the particles. From frame #370 to #380, the estimations of the left forearm and elbow are unfortunately mislead by the particles from the proposal function with the inverse kinematics of hand, however, this mistake is fixed by the proposal function with the obvious edge of arm when the left arm is captured more clearly. On the other hand, the estimate of the left arm is extended over the real one at frame #390, but the particles drawn from the inverse kinematics with the hand position correct the results after several frames. We can see that the mechanism of multiple importance sampling produces the particles which have the diversity of states and are complementary to each other. Table. 3 and 4 list the corresponding RMS and STD errors in joint position and arm angle of Fig. 11. Although the result of Fig. 11 indicates larger estimate error then that of Fig. 16, it still presents the acceptable performance that can be utilized for human-robot interaction.

The video sequence shown in Fig. 12 presents the system ability to deal with the scenario on a motion camera and under the disturbances of another person. The system detects two human faces in the beginning, however, only the person acting the initial posture as Fig. 1 will be recognized as the person who has intentions to do the interaction. The non-tracked person moves on the back of the tracked person and waves his hands to disturb the arm tracking during frame #20 to #160. From frame #200 to #280, the non-tracked even waves his hands in the front of the left arm of the tracked person. The proposed tracking algorithm can successfully overcome the temporal occlusion and filter out the false hypotheses originated from the hand of non-tracked person. As shown in the rest frames of Fig. 12, this result also implies

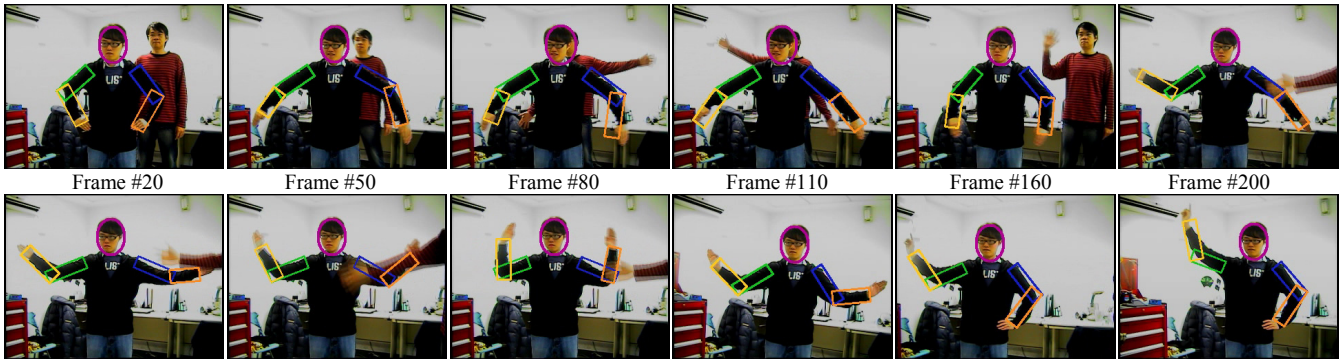


Fig. 12. Snapshots of tracking with a motion camera and under disturbances.

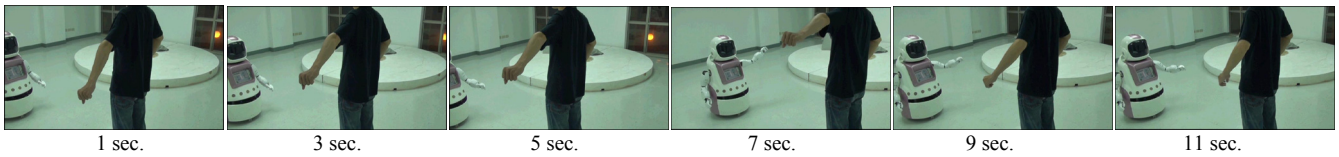


Fig. 13. Snapshots of human robot interaction.

that the proposed algorithm is not constrained to operate on a static camera. Moreover, if the tracking system is conscious that the estimated posture is ambiguous due to the 3D motion, it can automatically control the motion of camera platform to clarify the human intention.

At last, we test the real-time efficiency of the human-robot interaction as shown in Fig. 13. The robot tries to repeat the same posture as human arms. The human posture obtained by the tracking system then acts as the instruction of the human-robot interaction. In the future, the different views of human during camera moving could also be used to control the camera motion and correct the human posture.

VI. Conclusion

This paper presents a particle filter methodology for tracking the face and arms of the human upper body on a monocular camera. Although we focus on the 2D tracking performance in the image plane now, the 3D stick model is still applied here to avoid the unreasonable human postures simply from a 2D model and be able to estimate the 3D posture further. By the partitioned sampling method, we identify the upper body parts in the order of human face, right arm and left arm, so that the tracker can reduce the dimension of estimated states. The MIS particle filter of arm tracker combines the hand position, straight edge of arm and temporal information to efficiently generate the particles with various characteristic. The likelihood model, which takes the visual clues of the motion, appearance and shape to human face and arms to verify the various hypotheses from the multiple importance sampling schemes, is designed by just using color and edge observed on 2D image plane. In order to easily apply the algorithm to human-robot interaction, our approach can handle the situation with a single moving camera platform.

Now, the computational time for the overall system is about 10-15 fps, which achieves near real-time performance. For the future work, we will examine and analysis the system in more complicated scenarios and aim to reduce the overall computational time. The 3D posture of human body will be further estimated by a monocular camera.

References

[1] M. Lee and R. Nevatia, "Human pose tracking in monocular sequence using multilevel structured models," *IEEE Trans. Pattern Anal. Mach. Intell.* , vol. 31, pp. 27-38, 2009.

[2] N. Jojic, M. Turk, and T. S. Huang, "Tracking self-occluding articulated objects in dense disparity maps," in *Proc. of the Seventh IEEE Int. Conf. on Computer Vision*, pp. 123-130 vol.1, 1999.

[3] J. Lee, J. Chai, P. S. A. Reitsma, J. K. Hodgins, and N. S. Pollard, "Interactive control of avatars animated with human motion data," *ACM Trans. on Graphics archive,* vol. 21, pp. 491-500, 2002.

[4] K. Dong-Wan and J. Ohya, "Estimating Postures of a human wearing a multiple-colored suit based on color information processing," in *Proc. of Int. Conf. on Multimedia and Expo*, pp. I-261-I-264 vol.1, 2003.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in. *Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, 2005, pp. 886-893, vol. 1*, 2005*.

[6] J. J. Wang and S. Singh, "Video analysis of human dynamics--a survey," *Real-Time Imaging,* vol. 9, pp. 321-346, 2003.

[7] W. Liang, T. Tieniu, N. Huazhong, and H. Weiming, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 25, pp. 1505-1518, 2003.

[8] R.O. Duda and P.E. Hart, "Use of the Hough transform to detect lines and curves in pictures," *Comm. ACM*, vol. 15, pp. 11-15, 1972.

[9] T. B. Moeslund, A. Hilton and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding,* vol. 40, pp. 90-126, 2006.

[10] J. MacCormick and M. Isard, "Partitioned sampling, articulated objects and interface-quality hand tracking," *Lecture Notes in Computer Science,* vol. 1843, pp. 3-19, 2000.

[11] E. Veach and L. J. Guibas, "Optimally combining sampling techniques for Monte Carlo rendering," in *Proc. of the 22nd annual conference on Computer graphics and interactive techniques*: ACM, 1995.

[12] J. Shi and C. Tomasi, "Good features to track," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 593-600, 1994.

[13] P. Pérez, J. Vermaak, A. Blake, "Data fusion for visual tracking with particles," in *Proc. IEEE*, vol. 92(3), pp. 495-513, 2004.

[14] M. Isard and A. Blake, "Contour tracking by stochastic propagation of conditional density," in *Proc. of European Conf. of Computer Vision*, vol. 1, pp. 343-356, 1996.

[15] C. Rasmussen and G. D. Hager, "Probabilistic data association methods for tracking complex visual objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, pp. 560-576, 2001.