REMARKS ON MARKERLESS HUMAN MOTION CAPTURE USING MULTIPLE IMAGES OF 3D ARTICULATED HUMAN CG MODEL

Kazuhiko Takahashi[1], *Takashi Oida*[2], *Jun-ichiro Hori*[3] *and Masafumi Hashimoto*[1]

[1]Doshisha University
610-0321, Kyoro, Japan
{katakaha, mhashimo}@mail.doshisha.ac.jp

[2]Ricoh Co., Ltd.
ex-affiliation: Graduate School of Doshisha University
Tokyo, Japan

[3]NTT DATA Co.
ex-affiliation: Doshisha University
Tokyo, Japan

ABSTRACT

This paper proposes markerless human motion capture using a 3D articulated computer generated (CG) model of the human body. The method of estimating human body posture is based on 2D matching between human silhouettes extracted from camera images and model silhouettes projected onto the corresponding camera planes in virtual space. Candidates for human model silhouette are generated using a Monte Carlo filter-based algorithm, and the normalised core-weighted XOR distance is introduced to calculate the likelihood rate between silhouettes. Experimental results show the feasibility and effectiveness of the proposed method for achieving markerless human motion capture.

1. INTRODUCTION

The demand for human motion capture is increasing in various applications such as advanced human-machine interface systems, visual communications, virtual reality applications and video game systems. Computer vision technology is increasingly being expected to be used for sensing human information [1], and the use of contact-type sensors or markers to obtain motion parameter information would no longer be required. As a result, there have been many studies of human motion capture using computer vision [2, 3, 4]. The authors have also proposed real-time methods of human motion capture from images captured by CCD cameras [5, 6].

In this paper, markerless human motion capture using 3D articulated human body model is proposed. The proposed method is based on 2D matching between human silhouettes extracted from input images captured by CCD cameras in actual space and silhouettes of a 3D human model projected onto the corresponding camera planes in virtual space. To generate candidates for the human model silhouette, a Monte Carlo filter (MCF) [7], which is a robust filtering technique based on a Bayesian framework, is introduced. In the fitting between the observed human silhouette and the candidate silhouette, XOR (exclusive or) distance [8] is used for a precise comparison of the two silhouettes. To improve the estimation accuracy with respect to body parts such as arms and legs, which are thin, a cost function based on weighted distance functions with equal weight on the shape skeleton obtained from the silhouette is considered. In section 2, an algorithm for 3D human body posture estimation in human motion capture is described. In section 3, computational experiments on estimating 3D human body postures are presented. In section 4, experimental results of markerless human motion capture using an actual multi-camera system are presented.

2. HUMAN MOTION CAPTURE SYSTEM

2.1 Outline of Human Motion Capture

Figure 1 shows an overview of our proposed human motion capture system based on multiple cameras in both actual space and virtual space. The approach to human motion capture proposed here is based on the projection of human silhouettes, in which the human silhouettes correspond to human body areas in the images. The human silhouette is typically extracted by calculating the difference between the background image and the input image at each pixel and then thresholding the difference at that pixel. The thresholded image, in which each pixel has a value indicating whether it is the human silhouette or the background, is called a silhouette image. The proposed human motion capture method consists of three processes: human silhouette extraction using background subtraction, generation of the 3D CG human model postures as candidates and calculation of silhouette matching between the human silhouettes and the CG human model silhouettes. Here, the candidates for human body posture are generated using a CG model in parallel. The posture of the CG model that has the highest matching rate indicates the estimated human body posture.

2.2 Human CG Model

For 3D human body posture estimation, the articulated human body model shown in Fig. 2 is introduced. The human model is designed manually using a 3D graphics software Poser and LightWave 3D (D-Storm, Inc.). The model has 10 joints, indicated by circles in Fig. 2, and consists of 11 segments representing the head, upper half of the body, lower half of the body, upper arms, lower arms, upper legs and lower legs. In the model, limits are introduced on the

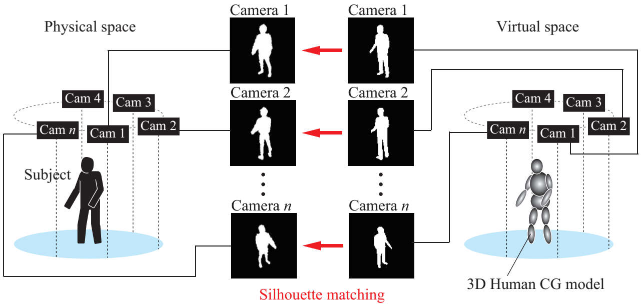

Figure 1: Configuration of actual and virtual multi-camera systems for human motion capture.

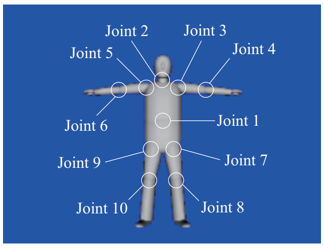

Figure 2: Articulated 3D CG human model.

joint angles to constrain the body posture. The shape of the CG model is determined in advance by measuring the shape of a subject captured by a CCD camera in actual space.

To generate a variety of body postures, a state space model of joint angles is introduced, and an MCF-based control algorithm is applied to the model to control the joint angles. Assuming the second-order difference of the joint angles is 0, the state space model for the j-th joint ($j = 1,2,\cdots,10$) with noise is defined as follows.

$$\boldsymbol{x}_j(k+1) = \boldsymbol{F}\boldsymbol{x}_j(k) + \boldsymbol{G}\boldsymbol{w}_j(k) \qquad (1)$$

$$\boldsymbol{y}_j(k) = \boldsymbol{H}\boldsymbol{x}_j(k) + \boldsymbol{v}_j(k) \qquad (2)$$

Here $\boldsymbol{x}_j^{\mathrm{T}}(k) = [\phi_x(k)\ \phi_x(k-1)\ \phi_y(k)\ \phi_y(k-1)\ \phi_z(k)\ \phi_z(k-1)]$, $\phi_i(k)$ is the joint angle with respect to the i-th axis ($i = x,y,z$), k is the sampling number and $\boldsymbol{w}_j(k)$ and $\boldsymbol{v}_j(k)$ are noise process vectors. The system matrix, disturbance matrix, and observation matrix are defined as follows: $\boldsymbol{F} = [F_{ij}]$ where $F_{11} = F_{33} = F_{55} = 2$, $F_{12} = F_{34} = F_{56} = -1$, $F_{21} = F_{43} = F_{56} = 1$ and otherwise is 0, $\boldsymbol{G} = [G_{ij}]$ where $G_{11} = G_{32} = G_{53} = 1$ and otherwise is 0, $\boldsymbol{H} = [H_{ij}]$ where $H_{11} = H_{23} = H_{35} = 1$ otherwise is 0.

Based on an MCF, the state vector is updated as follows:

Step 1 Generation of particles $\{\boldsymbol{q}_j^{(i)}(0)\}_{i=1}^m$ from an initial distribution of state $p_0(\boldsymbol{x}_j)$.

Step 2 Generation of particles for system noise $\{\boldsymbol{w}^{(i)}(k)\}_{i=1}^m$ from the system noise distribution.

Step 3 Calculation of particles representing the predicted distribution $\{\boldsymbol{p}_j^{(i)}(k)\}_{i=1}^m$ using the state space equation.

Step 4 Calculation of the likelihood rate of the particle $\boldsymbol{p}_j^{(i)}(k)$ using the silhouette images (see section 2C). The particle that has the highest likelihood rate is chosen as the state vector $\boldsymbol{x}_j(k)$.

Step 5 Calculation of the filter distribution's particles $\{\boldsymbol{q}_j^{(i)}(k)\}_{i=1}^m$ by resampling the particles $\{\boldsymbol{p}_j^{(i)}(k)\}_{i=1}^m$ with the probability $\mathtt{Pr}\left[\boldsymbol{q}_j^{(i)}(k) = \boldsymbol{p}_j^{(i)}(k)\right]$. The probability distribution is defined so that the posture candidate with a higher matching rate is more frequently selected.

Step 6 Return to Step2.

2.3 Model Matching with Silhouette Images

The normalised core-weighted XOR distance in the l-th camera ($l = 1,2,\cdots,n$) is introduced as follows in order to calculate the likelihood rate between the input silhouette and the model silhouettes:

$$d_l(S,T) = \frac{1}{w_b}\left\{\sum_{T(n)=1} D(S) + \sum_{T(n)=0}\frac{\beta D(S_c)}{D(S_c)+D(S_s)}\right\} \qquad (3)$$

S is the silhouette image extracted from the input image, T is the silhouette image of the human model, S_c is the silhouette contour, S_s is the skeleton of the silhouette shape, the function D denotes the distance transform, w_b is the fixed bounding window and n is the number of pixels in the window. The likelihood rate of the particle $\boldsymbol{p}_j^{(i)}$ at the l-th camera is then obtained by applying the following Gaussian function to the distance.

$$\mathtt{prob}_j^i(l) = \frac{1}{\sqrt{2\pi\sigma_l^2}}\exp\left(-\frac{d_l^2(S_l,T_l^i)}{2\sigma_l^2}\right) \qquad (4)$$

When multiple cameras are used in the estimation, the total likelihood rate is calculated by the following equation.

$$\overline{\mathtt{prob}}_j^i = \prod_{l=1}^{n}\mathtt{prob}_j^i(l) \qquad (5)$$

2.4 Human Body Posture Estimation

The human body posture estimation consists of three processes. First, the model, which has the highest likelihood rate for the combination of head, chest (upper half of body) and waist (lower half of body), is selected. Next, one body part is added to the model, and the model matching is calculated. This estimation is undertaken for the right arm, left arm, right leg and left leg in parallel. Then, a model of the whole human body is constructed by combining the candidates for each body part, and the likelihood rate is re-evaluated with respect to the whole body model. The candidate, which has the highest likelihood rate, is defined as the estimation result.

3. COMPUTATIONAL EXPERIMENTS OF HUMAN MOTION CAPTURE

To evaluate the feasibility of the proposed method, computational 3D human body posture estimation was examined experimentally. The estimation method was coded in C++ with Direct X graphics system. In this computational experiment, the target human motion was generated in virtual space using a 3D human CG model, and input images were obtained by projecting the human CG model onto virtual camera planes instead of using images captured with CCD cameras. Six cameras were used in the estimation. Three cameras at 120 degree intervals were arranged in the same horizontal plane as the target CG model, and the other three, also arranged at 120 degree intervals, were set to overlook the target CG model from an angle of 45 degrees. In the MCF, the number of particles (human model) m was 4000. Two kinds of motion were used in the experiment: walking in place (motion 1) and turning both hands in front of the body (motion 2). Motions 1 and 2 were expressed for 30 frames and 60 frames, respectively.

Figure 3 shows examples of human body posture estimation in each camera, and Fig. 4 shows the averaged estimation errors for the sequence shown in Fig. 3. Here, the averaged estimation error is the mean of the Euclidean distance between the target and the estimation result in all the

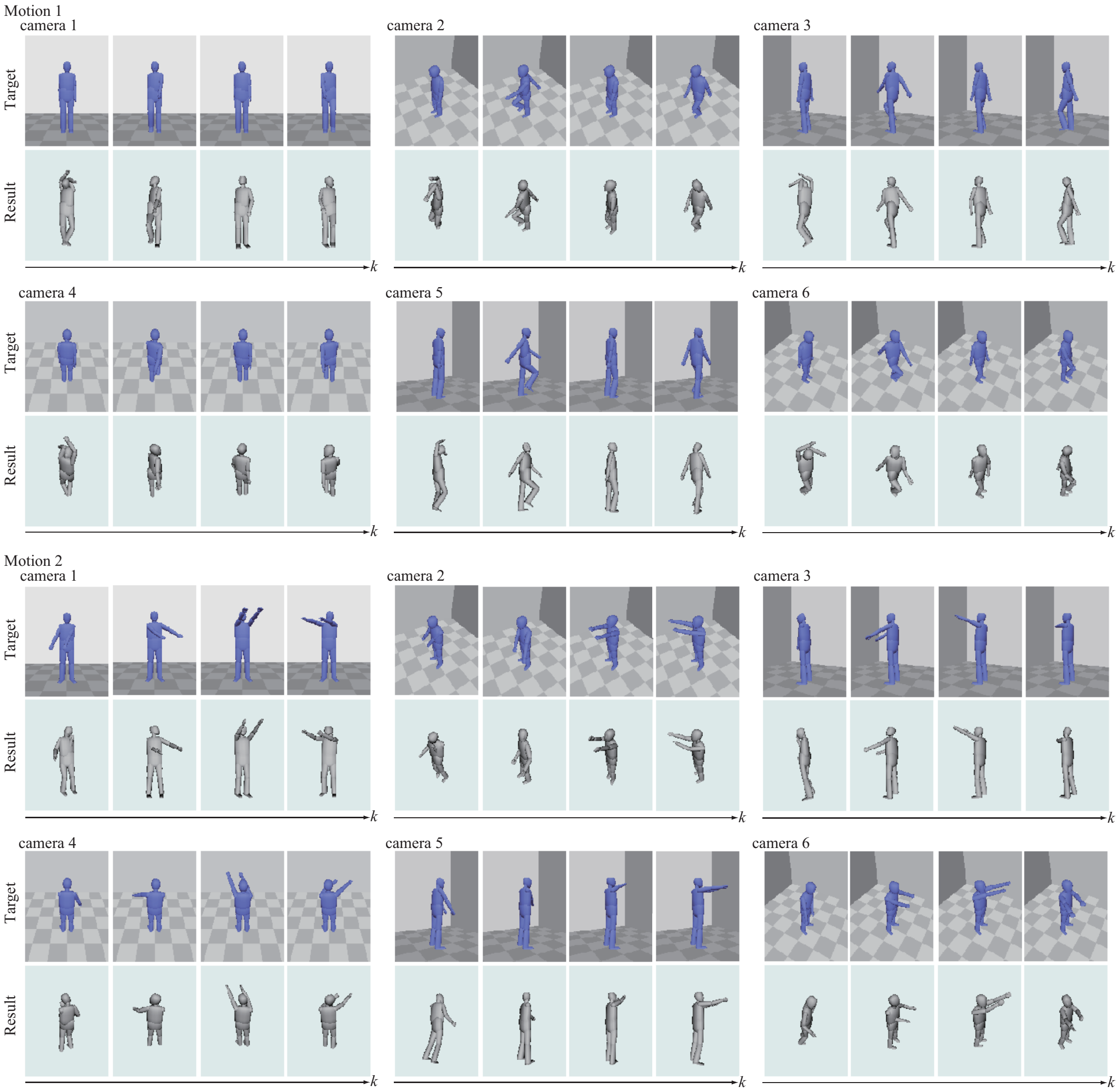

Figure 3: Estimation results (top: target posture, bottom: estimated posture with the highest likelihood, left to right: frame k = 1, 7, 13 and 19).

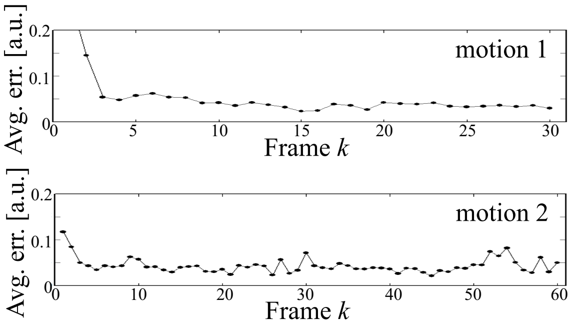

Figure 4: Averaged estimation error.

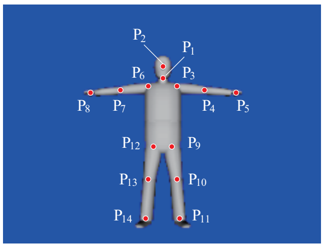

Figure 5: Feature points of the human body.

human body feature points P_q $(q = 1, 2, \cdots, 14)$, which are shown in Fig. 5. In Fig. 3, the top panel contains pictures of the sequence of target posture, and the bottom panel contains

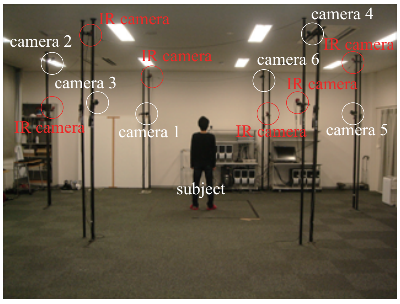

Figure 6: Overview of multi-camera system.

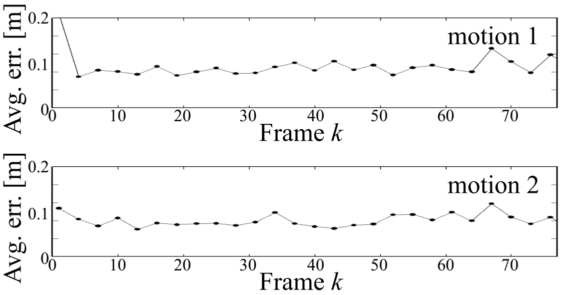

Figure 8: Averaged estimation error.

of the sequences of estimated postures that have the highest likelihood. As these figures show, at the beginning of the estimation (frame 1), posture estimation is unstable, and the likelihood is low. This is because the estimation in the first frame is like a random estimation. However, the likelihood converges after the second frame due to the MCF, and posture estimation is successful. Even though self-occlusion of the limbs in the target motion sequences occurs in some of the silhouette images obtained from each camera's viewpoint, the human body's posture can be estimated successfully in every frame. These results show the feasibility and effectiveness of the proposed estimation method of human motion capture.

4. MOTION CAPTURE EXPERIMENT USING A MULTI-CAMERA SYSTEM

To evaluate the efficiency of the proposed human motion capture, estimation experiments using the multi-camera system [5] shown in Fig. 6 were carried out. The multi-camera system consists of a server-client system. The communication between the server (Dell Precision Workstation 650, Intel(R) Xeon CPU 3.2GHz, 3.50GB RAM, Windows XP SP3) and clients (Dell Precision 390, Intel(R) Core 2 Duo CPU 2.66GHz, 512MB RAM, Windows XP SP3) are achieved by using socket communication with TCP/IP, a local 1000Base-T network and Winsock2 programming. Images from CCD cameras (IEEE 1394 camera Dragonfly2, Point Grey Research, Inc.) were digitised into the client computers with a 640-by-480 pixel resolution via an IEEE 1394 interface in real time (frame rate of 60 Hz). Six hexagonally arranged cameras are used in the system. Three cameras are located near the ceiling, and other three are set at middle height. The measured space in front of the cameras is $2[m] \times 2[m] \times 2[m]$. The cameras were calibrated using Microsoft's easy camera calibration tool [9]. The silhouette image is extracted by using background subtraction in each client computer connected to each camera and is then transferred to the server. As a reference for the proposed motion capture method, the movement of the subject is simultaneously measured by optical (infrared ray) motion capture (OptiTrack$^{\text{TM}}$ FLEX:V1007, ARENA$^{\text{TM}}$ Motion Capture Software, NaturalPoint, Inc.).

To apply the proposed method to actual images, the virtual camera parameters of the DirectX graphics system (i.e. location and posture in virtual space, direction vector of the gaze point, rotation of the camera, angle of view) must be defined so as to correspond to the actual cameras of the multi-camera system. The virtual camera parameters are estimated by searching for the parameters that make a well-known calibration pattern captured by the virtual camera match with the calibration pattern captured by the actual cameras. To search for matching parameters, a genetic algorithm is used.

In the experiments, the motions were the same as in the computational experiments: the motion 1 was walking in place, and the motion 2 was turning both hands in front of the body. In the MCF, the numbers of particles m were 9000 and 16000 for estimating motions 1 and 2, respectively. Figure 7 shows examples of human body posture estimation in each camera, and Fig. 8 show the averaged estimation errors for the sequence shown in Fig. 7. Here, the averaged estimation error is defined by the mean of the Euclidean distance between the 3D coordinates of all the human body's feature points measured by optical motion capture and the 3D coordinates estimated by the proposed method. In Fig. 7, the top figures show the sequence of input images, and the bottom figures are the sequences of estimated postures that have the highest likelihood. As these figures show, at the beginning of the estimation (frame 1), posture estimation is unstable: however, posture estimation is successful after the second frame due to the MCF. The average estimation error is about 0.1 m. These results show the feasibility of the proposed estimation method to achieve markerless human motion capture in an actual system.

5. CONCLUSIONS

This paper proposed a markerless human motion capture system based on 2D matching between human silhouettes extracted from camera images captured by multiple cameras and 3D articulated human CG model silhouettes projected onto the corresponding virtual camera planes. A Monte Carlo filter-based algorithm was proposed to generate candidates for the human model silhouette, and the normalised core-weighted XOR distance was introduced to calculate the likelihood rate between silhouettes. Computational experiments of human body posture estimation and a human motion capture experiment using a multi-camera system confirmed the feasibility and effectiveness of the proposed markerless human motion capture system.

REFERENCES

[1] L. Wang, W. Hu, and T. Tan, Recent Developments in Human Motion Analysis, Pattern Recognition, Vol. 36, No. 3, pp.585-601, 2003.

[2] T. B. Moeslund, A. Hilton, and V. Kruger, A Survey of Advances in Vision-based Human Motion Capture and

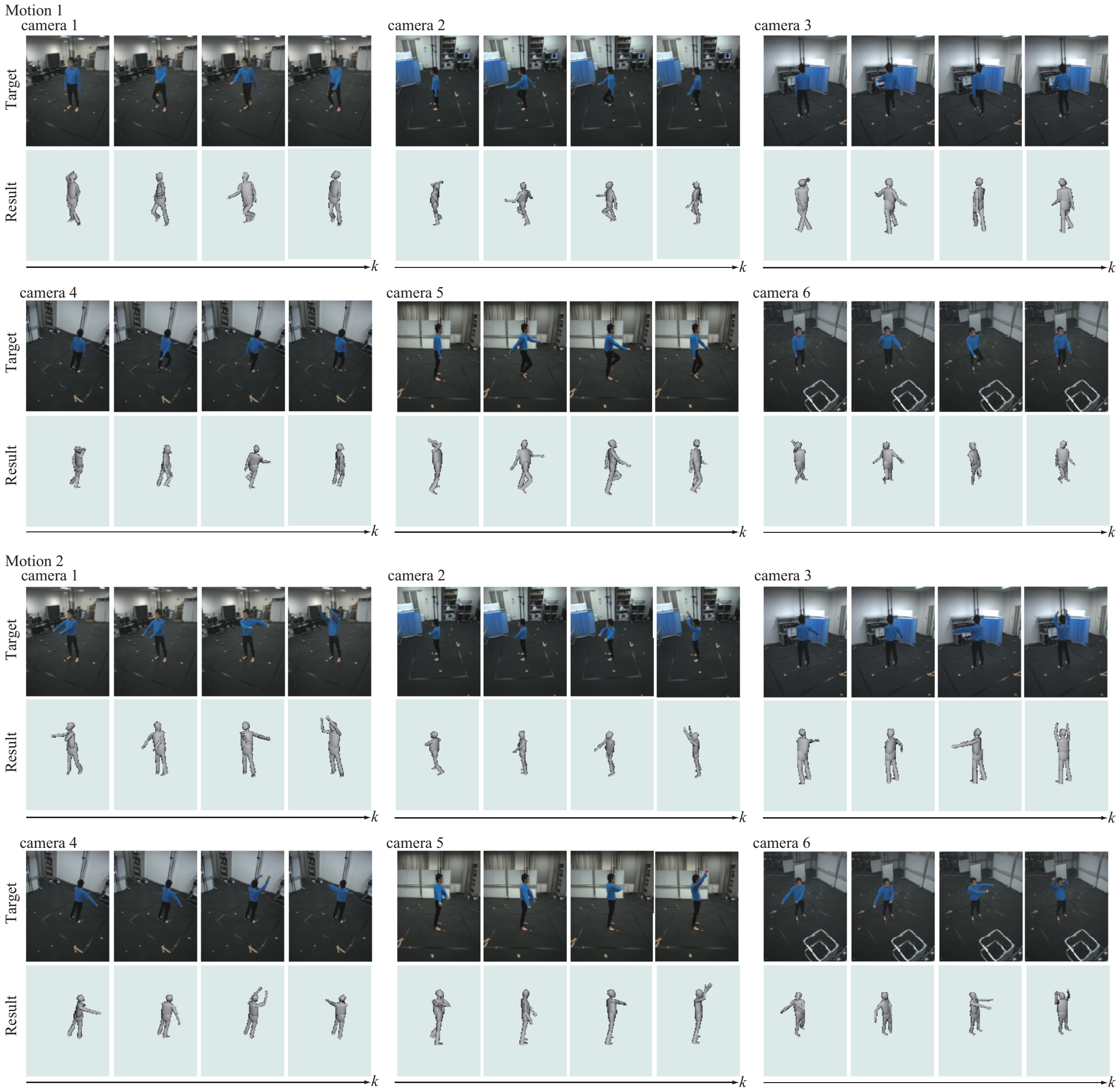

Figure 7: Estimation results (top: target posture, bottom: estimated posture with the highest likelihood, left to right: frame k = 1, 7, 13 and 22 for motion 1, and 1, 9, 17 and 25 for motion 2).

Analysis, Computer Vision and Image Understanding, Vol. 104, No. 2, pp.90-126, 2006.

[3] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan, Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis, Foundations and Trends in Computer Graphics and Vision, Vol. 1, pp.77-254, 2006.

[4] R. Poppe, Vision-Based Human Motion Analysis: An Overview, Computer Vision and Image Understanding, Vol. 108, pp.4-18, 2007.

[5] K. Takahashi, Y. Nagasawa, and M. Hashimoto, Markerless Human Motion Capture from Voxel Reconstruction with Simple Human Model, JSME Journal of Advanced Mechanical Design, Systems, and Manufacturing, Vol. 2, No. 6, pp.985-997, 2008.

[6] T. Kodama and K. Takahashi, Remarks on Markerless Motion Capture Using Heuristic Rules, Proceedings of 5th international Conference of Image and Graphics, pp.808-813, 2009.

[7] B. Ristic, S. Arulampalam, and N. Gordon, Beyond the Kalman Filter - Particle Filters for Tracking Applications, Artech House Publishers, Boston-London, 2004.

[8] R. Okada, R and B. Stenger, A Single Camera Motion Capture System for Human-Computer Interaction, IEICE Transactions, Information and Systens E, Vol. 91, No. 7, pp.1855-1862, 2008.

[9] Z. Zhang, "A Flexible New Technique for Camera Calibration", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.11, pp.1330-1334, 2000.