1st International Conference on Sensing Technology
November 21-23, 2005 Palmerston North, New Zealand

Comparison of Human Body Posture Estimation Method Using Silhouette Image

Kazuhiko Takahashi[†,‡] and Masahide Naemura[‡]
†Faculty of Engineering, Doshisha University, Kyoto, Japan
katakaha@mail.doshisha.ac.jp
‡ATR MIS Laboratories, Kyoto, Japan
naemura@atr.jp

Abstract

This paper presents human body posture estimation methods based on analysis of human silhouette and investigates their characteristics. Two types of estimation method are presented: one method is based on both the heuristically extraction method of estimating the significant points of human body and the contour analysis of the human silhouette, and the other is using artificial neural network. In the former, the 2D coordinates of the significant points are located by applying the heuristically extraction method to the human silhouette and the contour analysis of the human silhouette. In the latter, the input feature vector of the ANN is composed with the result of analyzing a human silhouette image and the output vector of the ANN indicates the 2D coordinates of the significant points. In both methods, the estimated results are optimized and tracked by using Kalman filter. Experimental results show both the feasibility and the characteristics of the presented methods for estimating human body postures.

Keywords: Human posture estimation, image processing, silhouette image, neural networks, kalman filter

1 Introduction

Recently, computer vision technology has been increasingly expected to achieve sensing human information. Using computer vision can lighten the burden and stress of users since people no longer need to utilize contact-type sensors. To recognize non-verbal information, such as gestures and sign language, and to understand human actions and/or motions, the importance of measuring the human body posture or motion parameters is increasing in human-machine interface applications. Specifically, demands of human body posture estimation are increasing in a number of applications such as advanced human-machine interface systems, visual communications, virtual reality applications, and video game systems [11]. Therefore, many studies have been undertaken on estimation methods using computer vision [1, 3, 6, 7, 8, 9, 10, 15] and the authors have also proposed real-time estimation methods [5, 12, 13, 14] that are based on a contour analysis of human silhouette. Their characteristics are: (1) high-speed and robust processing, (2) no markers on the human body, and (3) a small computing power requirement.

In this paper, we present two types of human body posture estimation method from human silhouette image: one is based on both the heuristically extraction method of estimating the significant points of human body called as "WMOSURA"[5] and the contour analysis of the human silhouette [12], and the other is based on artificial neural network (ANN)[14], and compare their characteristics for estimating human body postures. To optimize and track the position of the estimated significant points, Kalman filter is introduced in the both methods. Section 2 describes human body posture estimation methods. Section 3 presents the experiments that tested how the methods could estimate human body postures and the results are summarized in section 4.

2 Posture Estimation Method

Figure 1 shows the outline of the human body posture estimation method based on both the heuristically extraction and the contour analysis of the human silhouette. It is composed of three processes: human silhouette extraction using background subtraction, human body's significant point estimation from the human silhouette image, and the significant point tracking using Kalman filter. Figure 2 shows the outline of the human body posture estimation method based on the ANN. It is composed of four processes: human silhouette extraction using background subtraction, feature extraction from the human silhouette using image processing, human body's significant point estimation using ANN, and the significant point tracking using Kalman filter. To simplify the process, the following assumptions are used in the both method: (1) no other object exists within the field besides the human, and (2) the camera is facing the front of the human.

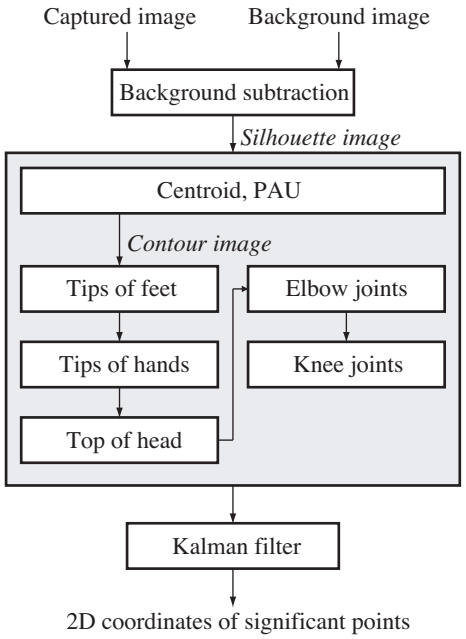

Figure 1: Outline of estimation method based on heuristically extraction and the contour analysis.

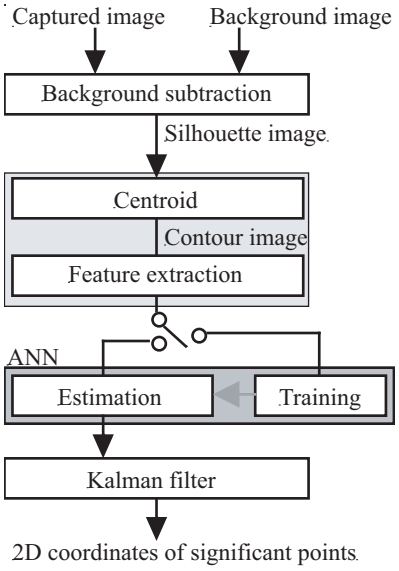

Figure 2: Outline of estimation method based on ANN.

2.1 Background subtraction

The human silhouette is extracted by calculating the difference at each pixel between the background image and the input image and then thresholding the difference at that pixel. The thresholded image, in which each pixel has a value indicating that the human silhouette or the background, is called a "silhouette image". Several very promising silhouette extraction methods that can be applied to backgrounds with complex textures has been developed [4, 16], however, we use a simple background subtraction where the background has an uniform color. RGB color system is used as

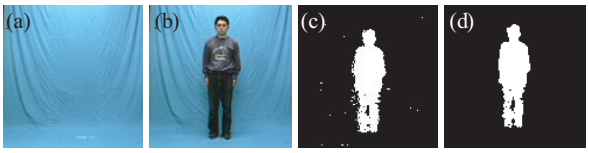

Figure 3: Background subtraction using looking-up table of RGB color system. ((a) background image, (b) input image, (c) silhouette image without noise reduction, (d) silhouette image after noise reduction)

the background model. Figure 3 shows an example of background subtraction result.

2.2 Significant point estimation based on heuristically extraction and the contour analysis

2.2.1 Centroid of and orientation of the body

First, the centroid of the human silhouette, $\left[\begin{array}{cc} x_g & y_g \end{array}\right]$, is located. To accurately obtain the centroid regardless of different poses of the arms and legs, the distance transform is applied to the silhouette image f_{xy} where $f_{xy} = 1$. The coordinates of the centroid are calculated as follows.

$$\left[\begin{array}{cc} x_g & y_g \end{array}\right] = \left[\begin{array}{cc} \frac{M_d(1,0)}{M_d(0,0)} & \frac{M_d(0,1)}{M_d(0,0)} \end{array}\right] \quad (1)$$

$$M_d(m,n) = \sum_x \sum_y x^m y^n d_{xy} \quad (2)$$

Here d_{xy} is the distance-transformed image of the silhouette image f_{xy}. Next, the principal axis of the upper half of the body (PAU) is obtained as the inclination ϕ of the human silhouette's principal axis of inertia as follows.

$$\tan^2\phi + \frac{M_g(2,0) \quad M_g(0,2)}{M_g(0,0)} \tan\phi \quad 1 = 0 \quad (3)$$

$$M_g(m,n) = \sum_x \sum_y \{x \quad x_g\}^m \{y \quad y_g\}^n g_{xy} \quad (4)$$

Here, g_{xy} is the distance-transformed image of the silhouette image in the upper half of the body.

The contour image of the human silhouette is obtained by a border tracking technique in the silhouette image f_{xy} as shown in figure 4(a). A temporary top of the head, which is the intersection of the PAU and the contour, is then chosen as shown in figure 4(a).

2.2.2 Representative points

Tips of feet : To extract the tips of the feet, the local maxima of the distance transformed image d_{xy} is calculated as shown in figure 4(b). This image includes many edges and end points that are candidates of the tips of the feet. The tips are chosen from the foot candidates that satisfy the following conditions: (1) they exist in the lower half of the body or below the centroid,

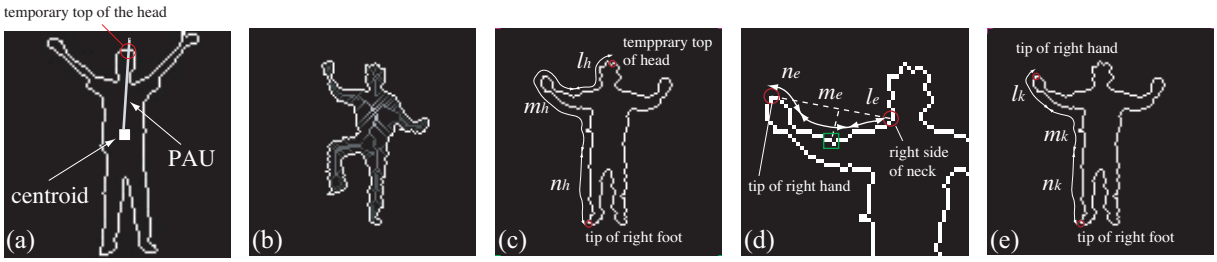

Figure 4: Significant points extraction from contour image ((a) Centroid of the human body, PAU, and temporary top of the head on contour image of human silhouette, (b) Local maxima of distance transformed image, (c) Locating tip of hand, (d) Locating elbow joint, (e) Locating knee joint).

and (2) they are the farthest ones from the centroid on the left and right hand side of the centroid, respectively. Here the distance between the centroid $\begin{bmatrix} x_g & y_g \end{bmatrix}$ and the ith contour pixel $\begin{bmatrix} x_c(i) & y_c(i) \end{bmatrix}$ is defined in the following form.

$$D(i) = \sqrt{\{x_g \quad x_c(i)\}^2 + \{y_g \quad y_c(i)\}^2} \tag{5}$$

Tips of hands : The contour pixels between the temporary top of the head and the tip of right (left) foot are divided into three segments using a predetermined pixel number ratio $l_h : m_h : n_h$ (in our current implementation, $l_h : m_h : n_h = 1 : 2 : 2$); this ratio can be applied to most humans, as shown in figure 4(c). The middle segment, corresponding to the ratio m_h, is the hand candidate contour. To extract the tip of the hand, we define following function [12] by using the centroid $\begin{bmatrix} x_g & y_g \end{bmatrix}$, the temporary top of the head $\begin{bmatrix} x_{th} & y_{th} \end{bmatrix}$, and the ith contour pixel $\begin{bmatrix} x_c(i) & y_c(i) \end{bmatrix}$ that located in this contour.

$$L(i) = \sqrt{D(i)^2 + \{x_{th} \quad x_c(i)\}^2 + \{y_{th} \quad y_c(i)\}^2} \tag{6}$$

The contour pixel that has the maximum value of the function $L(i)$ corresponds to the tip of the hand.

Neck sides and top of head : The contour pixels between the temporary top of the head and the tip of the right (left) hand are divided into three segments using a predetermined pixel number ratio $l_p : m_p : n_p$ (in our current implementation, $l_p : m_p : n_p = 1 : 2 : 2$). The right (left) side of the neck is assumed to be included in the middle segment corresponding to the ratio m_p. The neck side pixel is chosen as the contour pixel that has the shortest distance to the PAU. The central point between both neck side points along the contour is then determined as the top of the head.

2.2.3 Major joint positions

Elbow joints : To estimate the position of the elbow joint, the contour pixels between the right (left) neck side and the tip of the right (left) hand are divided into three segments using a predetermined pixel number ratio $l_e : m_e : n_e$ (in our current implementation, $l_e : m_e : n_e = 1 : 2 : 2$) as shown in figure 4(d). The middle segment, corresponding to the ratio m_e, is the elbow candidate contour. Even though the elbow does not always have salient features in the contour of the silhouette, the coordinates of the right (left) elbow joint is defined in the elbow candidate contour by using the farthest position from the straight line that connects between the right (left) side of the neck and the tip of the right (left) hand. When an arm is straight, the end position of the elbow candidate contour that is near the tip of the hand is defined as the elbow joint.

Knee joints : To estimate the position of the knee joint, the contour pixels between the tip of the right (left) hand and the tip of the right (left) foot are divided into three segments using a predetermined pixel number ratio $l_k : m_k : n_k$ (in our current implementation, $l_k : m_k : n_k = 1 : 2 : 2$) as shown in figure 4(e) and the segment, corresponding to the ratio n_k, is then chosen as the knee candidate contour. Even though the knee does not always have salient features in the contour of the silhouette, the coordinates of the right (left) knee joint is defined in the knee candidate contour by using the farthest position from the straight line that connects between the tip of the right (left) foot and the end position of the knee candidate contour that is near the tip of the hand. When a leg is straight, the end position of the knee candidate contour that is near the tip of the hand is defined as the knee joint.

2.3 Significant point estimation based on ANN

2.3.1 Feature extraction

In the contour image as shown in figure 4(a), the contour pixel is sampled with the distance of L/N (where N is the sampling rate) through the counter-clockwise direction from the point of the temporary head. Then we define the feature vector, V as follows:

$$V = \begin{bmatrix} \bar{x}_{p_0} & \bar{y}_{p_0} & \bar{x}_{p_1} & \bar{y}_{p_2} & \cdots & \bar{x}_{p_N} & \bar{y}_{p_N} \end{bmatrix} \tag{7}$$

where

$$\bar{x}_{p_i} = \frac{x_{p_i} \quad x_g}{D_X}, \quad \bar{y}_{p_i} = \frac{y_{p_i} \quad y_g}{D_Y}$$

x_{p_i} and y_{p_i} are the 2D coordinates of the sampled contour pixel, and D_X and D_Y are the pixel resolutions in horizontal and vertical directions, respectively. Thus the relative positions of the sampled contour with respect to the centroid are composed the feature vector. As a result, the dimension of the feature vector V is $2(N+1)$. This feature vector V is used as an input feature input vector in the stage of estimating significant points of the human body using ANN. This feature vector composition is more simple than that presented in [13].

2.3.2 Significant point estimation using ANN

The ANN is a three-layer PDP model and its relationship between inputs and outputs is given as follows:

$$U_l = f(\sum_{j=1}^{q} W_{2_{lj}} f(\sum_{i=1}^{r} W_{1_{ji}} V_i + \theta_{1_j}) + \theta_{2_l}) \tag{8}$$

where V_i is the input to the ith neuron in the input layer, U_l is the output of the lth neuron in the output layer, $W_{k_{ji}}$ and θ_{k_j} $(k = 1,2)$ are the weight and threshold, f is a sigmoid function, and r and q are the number of neuron unit in the input and hidden layer. In this study, 11 significant points of human body (head, hands, feet, shoulder joints, elbow joints, and knee joints) are considered. Thus the outputs from the ANN U_l are corresponded to the significant points as follows.

$$\begin{aligned} U^T \quad = \quad & [\, x_H \; y_H \; x_{s_L} \; y_{s_L} \; x_{e_L} \; y_{e_L} \; x_{h_L} \; y_{h_L} \; x_{k_L} \; y_{k_L} \; x_{fL} \; y_{f_L} \\ & x_{f_R} \; y_{f_R} \; x_{k_R} \; y_{k_R} \; x_{h_R} \; y_{h_R} \; x_{e_R} \; y_{e_R} \; x_{s_R} \; y_{s_R} \,] \end{aligned}$$

Here, the subscript H is the index of the top of the head, s_L is the index of the left shoulder joint, e_L is the index of the left elbow joint, h_L is the index of the left hand tip, k_L is the index of the left knee joint, f_L is the index of the left foot tip, f_R is the right foot tip, k_R is the index of the right knee joint, h_R is the index of the right hand tip, and e_R is the right elbow joint, and s_R is the right shoulder joint. By using back-propagation algorithm, the learning of the ANN is carried out to minimize the cost function $J = \frac{1}{2}\sum^{\rho}\sum_{l=1}^{s}(U_{d_l} \quad U_l)^2$ where U_{d_l} is the teaching signal, s is the number of neuron unit in the output layer, and ρ is the total number of the training data. In our study, the teaching data sets are gathered by capturing various postures from CCD camera with a D_X-by-D_Y pixel resolution in our laboratory, and the teaching data of the significant points are then extracted manually from images. Thus the teaching vector U_d is defined by normalizing with respect to the pixel resolution as follows.

$$\begin{aligned} U_d^T \quad = \quad & [\, \frac{x_{d_H} \quad x_g}{D_X} \; \frac{y_{d_H} \quad y_g}{D_Y} \; \frac{x_{d_{s_L}} \quad x_g}{D_X} \; \frac{y_{d_{s_L}} \quad y_g}{D_Y} \\ & \frac{x_{d_{e_L}} \quad x_g}{D_X} \; \frac{y_{d_{e_L}} \quad y_g}{D_Y} \; \frac{x_{d_{h_L}} \quad x_g}{D_X} \; \frac{y_{d_{h_L}} \quad y_g}{D_Y} \\ & \frac{x_{d_{k_L}} \quad x_g}{D_X} \; \frac{y_{d_{k_L}} \quad y_g}{D_Y} \; \frac{x_{d_{f_L}} \quad x_g}{D_X} \; \frac{y_{d_{h_L}} \quad y_g}{D_Y} \\ & \frac{x_{d_{f_R}} \quad x_g}{D_X} \; \frac{y_{d_{f_R}} \quad y_g}{D_Y} \; \frac{x_{d_{k_R}} \quad x_g}{D_X} \; \frac{y_{d_{k_R}} \quad y_g}{D_Y} \\ & \frac{x_{d_{h_R}} \quad x_g}{D_X} \; \frac{y_{d_{h_R}} \quad y_g}{D_Y} \; \frac{x_{d_{e_R}} \quad x_g}{D_X} \; \frac{y_{d_{e_R}} \quad y_g}{D_Y} \\ & \frac{x_{d_{s_R}} \quad x_g}{D_X} \; \frac{y_{d_{s_R}} \quad y_g}{D_Y} \,] \end{aligned} \tag{9}$$

In the following estimation experiment, the network topology of the ANN is first optimized using a small number of teaching data composed of 40 images and a 100-100-22 network topology is selected by considering computational costs. Next, the training of the ANN is carried out using the total amount of 140 teaching data.

2.4 Significant point tracking with Kalman filter

To optimize and track the positions of the human body's significant point, we assume the following model for every significant point.

$$X_{t+1} = AX_t + Bw_{t_i} \tag{10}$$

$$Y_t = CX_t + \omega_{t_i} \tag{11}$$

where

$$X_t^T = \left[\; x_{t_H}^T \quad x_{t_{s_L}}^T \quad \cdots \quad x_{t_g}^T \;\right]$$

$$x_{t_i}^T = \left[\; x_{t_i} \quad \dot{x}_{t_i} \quad y_{t_i} \quad \dot{y}_{t_i} \;\right]$$

$$w_t^T = \left[\; w_{x_t} \quad w_{y_t} \;\right]$$

$$A = \mathbf{diag}(a)$$

$$a = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$B^T = \left[\; b \quad b \quad \cdots \quad b \;\right]$$

$$b = \begin{bmatrix} \frac{T^2}{2} & 0 \\ T & 0 \\ 0 & \frac{T^2}{2} \\ 0 & T \end{bmatrix}$$

$$Y_t^T = \left[\; y_{t_H}^T \quad y_{t_{s_L}}^T \quad \cdots \quad y_{t_g}^T \;\right]$$

$$y_{t_i}^T = \left[\; x_{t_i} \quad y_{t_i} \;\right]$$

$$\omega_t^T = \left[\; \omega_{x_{t_H}} \quad \omega_{y_{t_H}} \quad \cdots \quad \omega_{y_{t_g}} \;\right]$$

$$C = \begin{bmatrix} c & \mathbf{0}_{2\ 4} & \cdots & \mathbf{0}_{2\ 4} \\ \mathbf{0}_{2\ 4} & c & \cdots & \mathbf{0}_{2\ 4} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{2\ 4} & \mathbf{0}_{2\ 4} & \cdots & c \end{bmatrix}$$

$$c = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Here, t is the sampling number, T is the sampling rate, w and ω are white noise processes. By using Kalman filter, the state vector X_t is estimated as follows:

$$\hat{X}_{t|t} = \hat{X}_{t|t\ \ 1} + K_t(Y_t \quad C\hat{X}_{t|t\ \ 1}) \tag{12}$$

$$\hat{X}_{t+1|t} = A\hat{X}_{t|t} \tag{13}$$

$$K_t = \hat{P}_{t|t\ \ 1}C^T(I + C\hat{P}_{t|t\ \ 1}C^T)^{\ 1} \tag{14}$$

$$\hat{P}_{t|t} = \hat{P}_{t|t\ \ 1} \quad K_tC\hat{P}_{t|t\ \ 1} \tag{15}$$

$$\hat{P}_{t+1|t} = A\hat{P}_{t|t}A^T + \frac{\sigma_w^2}{\sigma_\omega^2}BB^T \tag{16}$$

where K is the Kalman gain matrix, σ_w and σ_ω are the deviations of observation noise. The initial conditions are $\hat{X}_{0|\ \ 1} = \mathbf{0}$ and $\hat{P}_{0|\ \ 1} = \varepsilon I$ where $\varepsilon > 0$. The measurement vector Y_t is the estimation results of the significant points.

3 Estimation Experiment

To evaluate the feasibility of our estimation method in real-time application, the proposed method was coded in the Visual C and implemented on a personal computer (DELL Dimension 8300 Pentium 4 2.6GHz, Windows XP). The images from the CCD camera (SONY EVI-D30) were digitized into the computer via flame grabber (PHOTRON FDM-PCI IV). The entire process for estimating human postures ran in real time (20 frames/sec).

Figure 5 shows examples of estimation results. The images of the left column of figure 5 are the captured images, and the images of the middle and right column show the estimated significant points that are indicated with small squares on the contour image. As shown in figure 5, all of the postures could be estimated successfully without depending on the posture, position where the person was standing. Figure 6 shows an example of tracking results of the significant point in the image sequence shown in figure 5. Here the significant points of human body are the left and right hands. In figure 6, the real locations of significant points were obtained manually at each frame of the sequences. Although the tracking error increases when the significant point moves widely, the output from Kalman filter can track the real locations of the significant points. By comparing the results, the method based on the heuristically extraction and the contour analysis shows accurate estimation results than those of using the ANN. These experimental results indicate both the feasibility of both methods for estimating human body postures, however, the estimation accuracy of the significant points should be improved for practical applications.

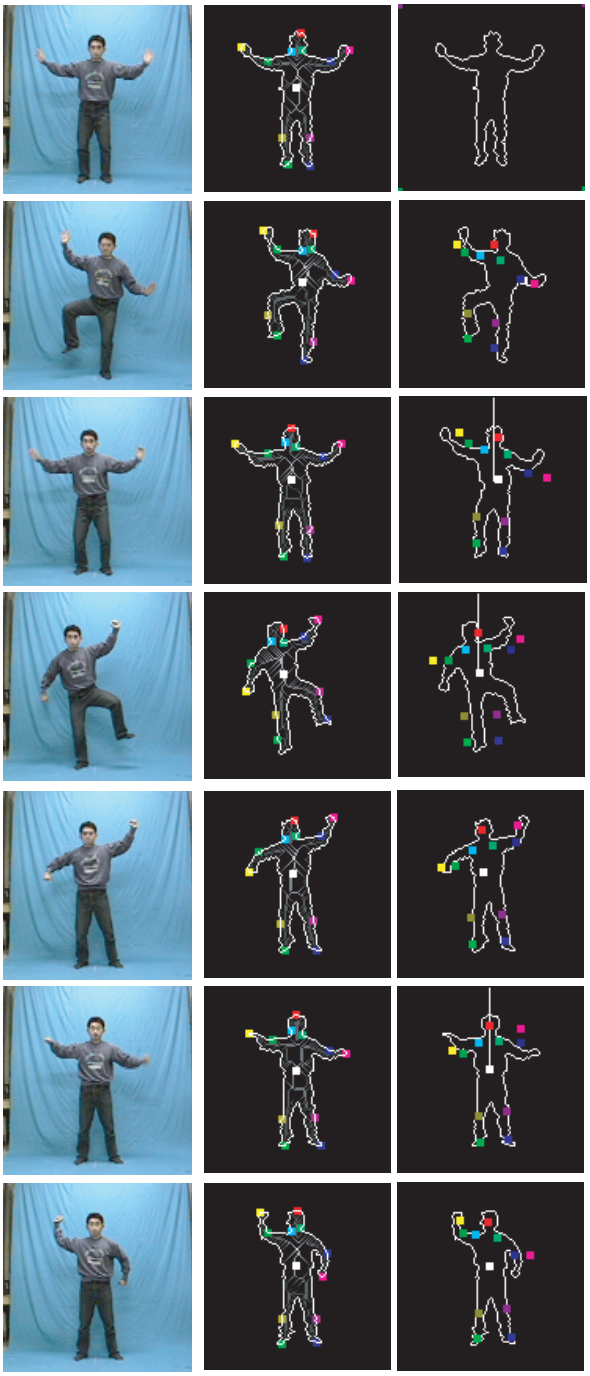

Figure 5: Examples of estimated human body postures (left : original camera image, middle : estimated significant points using heuristically extraction and the contour analysis, right : estimated significant points using ANN).

4 Conclusions

This paper has presented human body posture estimation methods based on analysis of human silhouette

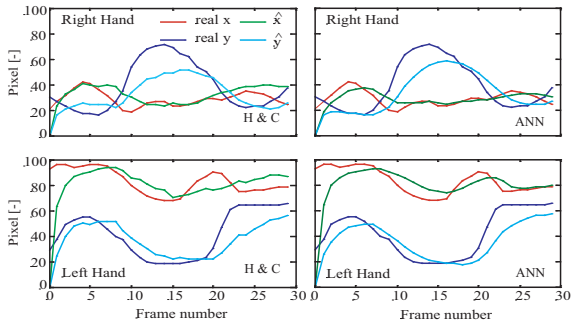

Figure 6: Examples of tracking results of significant points (left : estimated significant points using heuristically extraction and the contour analysis, right : estimated significant points using ANN).

and Kalman filter. Two types of estimation method are investigated their characteristics. Experimental results show the feasibility of the presented methods for estimating human body postures. The method based on both heuristically extraction and contour analysis shows accurate estimation results than those based on the ANN.

5 Acknowledgements

This work was supported by MEXT Grant-in-Aid for Young Scientists (B) 15700160, and a grant for "Research on Interaction Media for High-Speed and Intelligent Networking " from the National Institute of Information and Communications Technology, Japan.

References

[1] D. M. Gavrila, "The Visual Analysis of Human Movement: A Survey", *Computer Vision and Image Understanding*, Vol. 73, No. 1, pp. 82-98 (1999).

[2] E. H-Jaraba, C. O-Urunuela, and J. Senar, "Detected Motion Classification with a Double-Background and a Neighborhood-Based Difference", *Pattern Recognition Letters*, Vol. 24, pp. 2079-2092 (2003).

[3] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Who? When? Where? What? A Real Time System for Detecting and Tracking People", *Proceedings of Third IEEE International Conference on Face and Gesture Recognition*, pp. 222-227 (1998).

[4] I. Haritaoglu, D. Harwood, and L. S. Davis, "A Fast Background Scene Modeling and Maintenance for Outdoor Surveillance", *Proceedings of 15th International Conference on Pattern Recognition*, Vol. 4, pp. 179-183 (2000).

[5] S. Iwasawa, J. Ohya, K. Takahashi, T. Sakaguchi, S. Kawato, K. Ebihara, and S. Morishima, "Real-time, 3D Estimation of Human Body Postures from Trinocular Images", *Proceedings of International Conference on Computer Vision Workshop on Modeling People*, pp. 3-10 (1999).

[6] I. A. Kakadiaris and D. D. Metaxas, "Three-Dimensional Human Body Model Acquisition from Multiple Views", *International Journal of Computer Vision*, Vol. 30, No. 3, pp. 191-218 (1998).

[7] I. A. Karaulova, P. M. Hall, and A. D. Marshall, "Tracking People in Three Dimensions Using a Hierarchical Model of Dynamics", *Image and Vision Computing*, Vol. 20, No. 12, pp. 691-700 (2002).

[8] Y. Li, A. Hilton, and J. Illingworth, "A Relaxation Algorithm for Real-Time Multiple View 3D-Tracking", *Image and Vision Computing*, Vol. 20, No. 12, pp. 841-859 (2002).

[9] R. Marks, R. Deshpande, C. M. Wideman, V. Kokkevis, S. Sargaison, and E. Larsen,"Real-Time Motion Capture for Interactive Entertainment", *Proceedings of ACM SIGGRAPH 2003 Emerging Technologies* (2003).

[10] T. B. Moeslund and E. Granum, "A Survey of Computer Vision-Based Human Motion Capture", *Computer Vision and Image Understanding*, Vol. 81, No. 3, pp. 231-268 (2001).

[11] J. Ohya, J. Kurumisawa, R. Nakatsu, K. Ebihara, S. Iwasawa, D. Harwood, and T. Horprasert, "Virtual Metamorphosis", *IEEE Multimedia*, Vol. 6, No. 2, pp. 29-39 (1999).

[12] K. Takahashi, T. Sakaguchi, and J. Ohya, "Remarks on No-wear, Non-contact, 3D Real-time Human Body Posture Estimation Method", *Systems and Computers in Japan*, Vol. 31, No. 14, pp. 1-10 (2000).

[13] K. Takahashi, T. Uemura, and J. Ohya, "Remarks on Neural-Network-Based Real-Time Human Body Posture Estimation", *2000 IEEE Workshop on Neural Networks for Signal Processing*, pp. 477-486, 2000.

[14] K. Takahashi and T. Tanigawa, "Remarks on Real-Time Human Posture Estimation from Silhouette Image Using Neural Network," *Proceedings of 2004 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 370-375 (2004).

[15] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp. 780-785 (1997).

[16] K. Yasuda, T. Naemura, and H. Harashima, "Thermo-key: Human Region Segmentation from Video Using Thermal Information", *Proceedings of ACM SIGGRAPH 2003 Emerging Technologies* (2003).