Remarks on Human Body Posture Estimation From Silhouette Image Using Neural Network

Kazuhiko Takahashi
Faculty of Engineering
Yamaguchi University
Tokiwadai Ube, Yamaguchi, Japan
kylyn@yanaguchi-u.ac.jp

Abstract

This paper proposes a human body posture estimation method using neural network. The input feature vector of the neural network is composed with the result of analyzing a human silhouette extracted from camera image and the output vector of the neural network indicates the 2D coordinates of the human body's significant points. The proposed method is implemented on a personal computer and runs in real-time. Experimental results show both the feasibility and the effectiveness of the proposed method for estimating human body postures.

1 Introduction

Artificial neural network (ANN) originated from studies on the mechanisms and structures of the brain has excellent capabilities, such as nonlinear mapping and learning. Studies on biological systems have led to the development of new computational models with application to complex problems such as pattern recognition, fast information processing, learning, and adaptation. In the field of computer vision, ANN is often utilized in many applications since the capabilities of ANN provide very powerful tools for solving pattern classification and/or recognition problems [10]. While computer vision technologies have widely been studied in various engineering fields, recent expectations have been for them to find application in sensing human information [2, 9]. To recognize non-verbal information, such as gestures and sign language, and to understand actions or motions, awareness has been growing on the importance of being able to measure the human body posture or motion parameters. More specifically, human body posture estimation is important for a number of applications including advanced human-machine interface systems, visual communications, virtual reality applications, and video game systems.

Human body posture estimation based on computer vision can lighten the burden and stress of users since they no longer need to utilize contact methods such as magnetic sensors. Therefore, several studies have been undertaken on estimation methods using computer vision [1, 3, 5, 7, 8, 14, 15]; they often lack the ability for real-time performance, have limits on the body parts able to be detected, and require large computational costs. The authors have also proposed real-time human body posture estimation methods [4, 12, 13]. The methods are based on both a contour analysis of human silhouette and their characteristics are: (1) high-speed and robust processing, (2) no markers on the human body, and (3) a small computing power requirement. However, the range of acceptable postures is still limited and a priori knowledge about the relationship between the analysis results and the human body parts is required.

This paper proposes a human body posture estimation method using ANN. In the proposed method, a human silhouette image extracted from camera image is analyzed to compose of input feature vector to the ANN and the ANN outputs the 2D coordinates of the human body's significant points in the image. By using the learning ability of the ANN to obtain the relationship between the human silhouette and the significant points, no a priori knowledge is necessary to achieve the human body posture estimation. Section 2 describes our human body posture estimation method in detail. Section 3 presents the experiments that tested how well the proposed method could estimate human body postures and the results are summarized in section 4.

2 Posture Estimation Method

Figure 1 shows the outline of our method for estimating human body postures. It is composed of two processes: feature extraction which uses image processing and significant

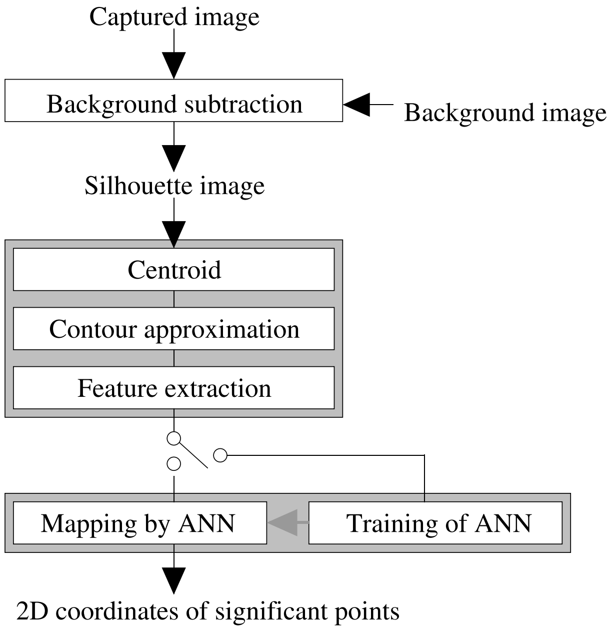

Figure 1. Outline of estimation method.

point estimation which uses ANN. To simplify the process, we use the following assumptions: (1) no other moving object exists within the field besides the human, and (2) the camera is facing the front of the human.

2.1 Feature extraction from silhouette image

Here, we describe how to compose an input feature vector from a human silhouette image obtained from camera images. As a preprocess, 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. First, the centroid of the human silhouette, $[x_g \; y_g]$, is located as follows.

$$x_g = \frac{M_d(1,0)}{M_d(0,0)} \qquad (1)$$

$$y_g = \frac{M_d(0,1)}{M_d(0,0)} \qquad (2)$$

$$M_d(m,n) = \sum_x \sum_y x^m y^n d_{xy} \qquad (3)$$

Here, d_{xy} is the silhouette image, x and y are the vertical and horizontal coordinates of the image, respectively. Next, an intersection of the human silhouette's contour and an axis that is projected outwards from the centroid is obtained [11]. The intersection point is the furthest contour edge along the axis. As shown in the top of Fig. 2, the intersection points are located at every Θ_0 degrees around the centroid with the clockwise direction from point A on the contour pixel. As shown in the bottom of Fig. 2, a figure which connected each intersection point sequentially is a circumscribed polygon of the silhouette image. This approximated contour has similar characteristics to active shape models [6].Then, the distance between the intersection point $[x_{p_i} \; y_{p_i}]$ and the centroid is calculated as follows.

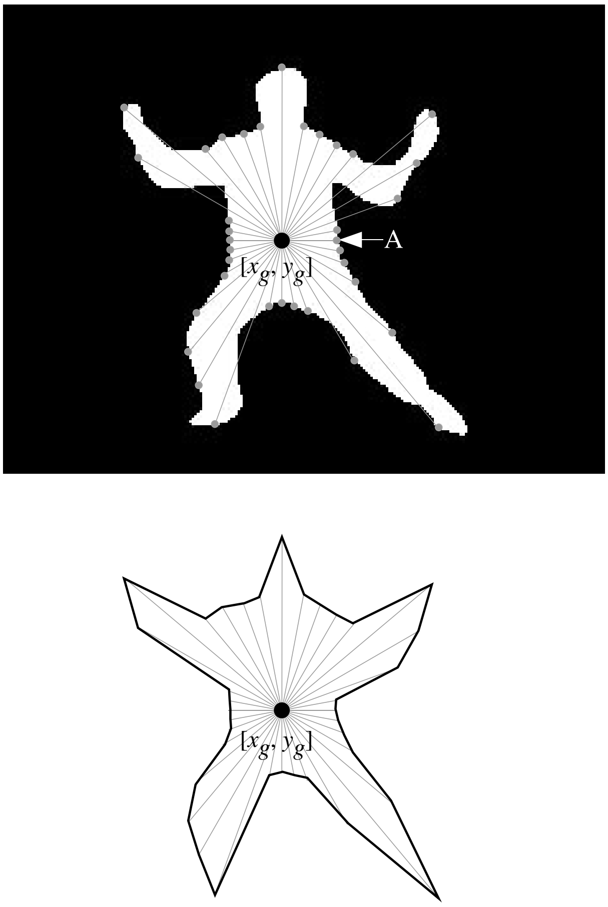

Figure 2. Extracting feature vector from human silhouette image (top: axes projected at specific angles on silhouette image, bottom: contour image approximated by furthest contour edge from the centroid).

$$D_i = \sqrt{(x_{p_i} - x_g)^2 + (y_{p_i} - y_g)^2} \qquad (4)$$

The distance D_i can be considered as a feature from the human silhouette. However, it is hard to find mathematical model that defines the relationship between the feature and the part of the human body. Therefore, we estimate the relationship by using ANN. The feature vector can be defined as follows.

$$\boldsymbol{z}^T = \left[\begin{array}{ccccc} D_1 & D_2 & D_3 & \ldots & D_N \end{array}\right] \tag{5}$$

The feature vector is normalized in order to have zero mean and unit variance and is scale, location and rotation invariant. The normalized feature vector $\boldsymbol{z}^*$ is used as a feature input vector in the stage of estimating significant points of the human body using ANN.

$$z_i^* = \frac{z_i - \mu_z}{\sqrt{\sigma_z}} \tag{6}$$

Here μ_z and σ_z are the mean and the variance of the element of the feature vector $\boldsymbol{z}$, respectively.

2.2 Significant point estimation using ANN

In this section, we describe how to design ANN for estimating significant points of the human body. The ANN is a three-layer PDP model with no inner feedback loops and no direct connections from the input layer to the output layer. The following sigmoid function, f, is used as an activation function of neurons in the hidden and output layers.

$$f(u) = \frac{1 - exp(-au)}{1 + exp(-au)} \tag{7}$$

Here a is the parameter of the sigmoid function. The relationship between inputs and outputs of the ANN is given by the following equation.

$$v_l = f(\sum_{j=1}^{M} w_{2_{lj}} f(\sum_{i=1}^{N} w_{1_{ji}} z_i + \theta_{1_j}) + \theta_{2_l}) \tag{8}$$

where z_i is the input to the ith neuron in the input layer, v_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, and N and M are the number of neuron unit in the input and hidden layer. The learning of the ANN is carried out according to the generalized δ-rule with an adaptive learning rate to minimize the cost function J.

$$\begin{aligned} \left[\begin{array}{c} w_{k_{ji}}(t+1) \\ \theta_{k_j}(t+1) \end{array}\right] &= \left[\begin{array}{c} w_{k_{ji}}(t) \\ \theta_{k_j}(t) \end{array}\right] \\ -\eta(t) \left[\begin{array}{c} \frac{\partial J}{\partial w_{k_{ji}}(t)} \\ \frac{\partial J}{\partial \theta_{k_j}(t)} \end{array}\right] &+ \alpha \left[\begin{array}{c} \Delta w_{k_{ji}}(t) \\ \Delta \theta_{k_j}(t) \end{array}\right] \end{aligned} \tag{9}$$

$$\eta(t) = \begin{cases} (1+\gamma)\eta(t-1) & \text{if } J(t) < J(t-1) \\ (1-\beta\gamma)\eta(t-1) & \text{if } J_(t) > J(t-1) \\ \eta(t-1) & \text{otherwise} \end{cases} \tag{10}$$

$$J(t) = \frac{1}{2} \sum^{P} \sum_{l=1}^{L} (v_{d_l} - v_l(t))^2 \tag{11}$$

where η is the learning factor, α is the momentum factor, $\Delta w_{k_{ji}}(t)$ is the weight increments at the tth iteration, $\Delta\theta_{k_j}(t)$ is the threshold increments at the tth iteration, γ and β are the adaptive factor in the learning, v_{d_l} is the teaching signal, L is the number of neuron unit in the output layer, and P is the total number of the training data.

3 Estimation Experiment

In this study, 9 significant points of human body (head, hands, feet, elbow joints, and knee joints) are considered. Thus we define the relationship between the ANN's outputs and the significant points as follows.

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

Here, the subscript H is the index of the top of the head, 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.

In order to achieve the relationship between the input feature vector and the significant points of human body part by learning of the ANN, various patterns of body posture are necessary as a teaching data of the ANN. However it is hard to collect everything of every body posture even if the posture is limited to only front view. In our experiment, the teaching data is made by extracting postures from the teaching materials video of Chinese shadow boxing (Tai Chi Quan) because the movements of Chinese shadow boxing have various combinations with complexity of hands and feet. The total amount of 1064 teaching data ($P = 1064$) was extracted manually from the Chinese shadow boxing instruction video with a 320-by-240 pixel resolution. Thus the teaching vector $\boldsymbol{v}_d$ is defined by normalizing with respect to the centroid of the human silhouette [x_g y_g] as follows.

$$\begin{aligned} \boldsymbol{v}_d^T \quad = \quad & [\, \frac{x_{d_H} - x_g}{x_g}\ \frac{y_{d_H} - y_g}{y_g}\ \frac{x_{d_{e_L}} - x_g}{x_g}\ \frac{y_{d_{e_L}} - y_g}{y_g} \\ & \frac{x_{d_{h_L}} - x_g}{x_g}\ \frac{y_{d_{h_L}} - y_g}{y_g}\ \frac{x_{d_{k_L}} - x_g}{x_g}\ \frac{y_{d_{k_L}} - y_g}{y_g} \\ & \frac{x_{d_{f_L}} - x_g}{x_g}\ \frac{y_{d_{h_L}} - y_g}{y_g}\ \frac{x_{d_{f_R}} - x_g}{x_g}\ \frac{y_{d_{f_R}} - y_g}{y_g} \\ & \frac{x_{d_{k_R}} - x_g}{x_g}\ \frac{y_{d_{k_R}} - y_g}{y_g}\ \frac{x_{d_{h_R}} - x_g}{x_g}\ \frac{y_{d_{h_R}} - y_g}{y_g} \\ & \frac{x_{d_{e_R}} - x_g}{x_g}\ \frac{y_{d_{e_R}} - y_g}{y_g}\,] \end{aligned} \tag{12}$$

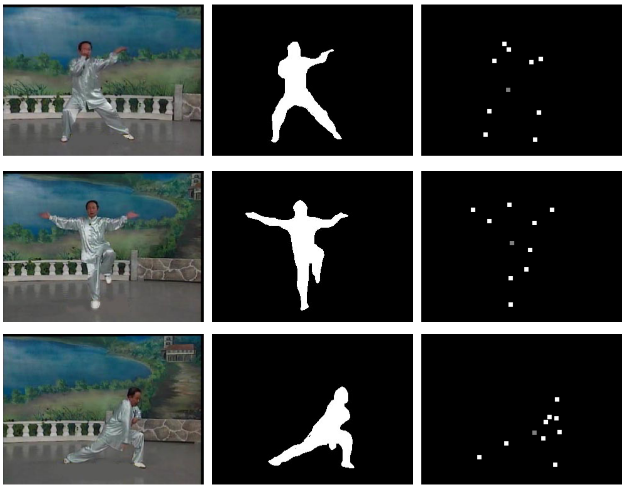

Figure 3. Examples of estimating significant points in close testing (left column: original image, middle column: human silhouette image, right column: estimated human body's significant point image).

The dimension of the feature input vector was 36 since the intersection points are located at every 10 degrees ($\Theta_0 = 10$). The number of neuron in the hidden layer was defined by trial and error in order to converge the ANN learning. In the experiment, the ANN structure was a 36-76-18 network ($N = 36$, $M = 76$, and $L = 18$). The initial weight matrices were randomly selected from the interval $[-1, 1]$. The parameters were $a = 1$, $\eta(0) = 10^{-4}$, $\gamma = 10^{-2}$, $\beta = 10$, and $\alpha = 0.9$. The maximum allowable error in each output unit was 10^{-2} while the learning iteration was limited to 10^5 for each training data. After the learning of the ANN was completed, the weight matrices were fixed and the ANN was used in the significant point estimation process. Figure 3 shows examples of the estimation result in the close testing. Here the silhouette images are manually obtained in the middle of Fig.3 and the estimated significant points of human body are shown with small squares in the right of Fig.3. As shown in the right figures, the significant points and the centroid are located successfully.

To evaluate the feasibility of our estimation method, we carried out experiments using real camera images. The proposed method was coded in the C language and implemented on a personal computer (Gateway GP7 Pentium III 800MHz, Windows 98). The images from the CCD camera (SONY EVI-D30) were digitized into the computer with a 160-by-120 pixel resolution via flame grabber (MATROX METEOR-II). The background subtraction was carried out by using the statistical background modeling, threshold selection, subtraction operation, and pixel classification [3] and an image processing of noise cleaning was also applied. The entire process for estimating human postures ran in real time (approximately 20 frames/sec).

Figure 4 shows examples of estimation results and Table 1 shows the estimation error that is defined as follows.

$$\epsilon_i = \sqrt{(x_{i_m} - x_i)^2 + (y_{i_m} - y_i)^2}$$

Here the real locations of significant points, [x_{i_m} y_{i_m}], were obtained manually from images and the estimated locations of significant points, [x_i y_i], were the outputs from the ANN (i = H, e_L, h_L, k_L, f_L, f_R, k_R, h_R, e_R). Fig. 4(a) is the captured images, Fig. 4(b) is the human image extracted by using the background subtraction, Fig. 4(c) is the significant points that are intersections of the human silhouette's contour and axes that are projected outwards from the centroid, and Fig. 4(d) is the significant points estimated by the ANN. In Fig. 4(d), the estimated significant points of human body are indicated with small squares. In each estimation result, the significant points are indicated with small squares (white: significant point, gray: centroid). As shown in Fig. 4, all of the significant points could be extracted successfully without depending on the posture, position where the person was standing, background condition, and image resolution. Although the ANN was trained using only the limited data extracted from the Chinese shadow boxing video and the human image extraction with the background subtraction is not always completely successful, the proposed method could achieve robust estimation of the significant points with the ANN's generalization ability. These experimental results indicate both the feasibility and effectiveness of our proposed method for estimating human body postures. However the estimation of the significant points is not good enough accurate as shown in Table 4. In order to improve the estimation accuracy, the learning or network topology of the ANN should be redesigned and the human extraction method from image should also be improved.

4 Conclusions

This paper has proposed a real-time human body posture estimation method using ANN. The input feature vector of

Table 1. Estimation error of the posture shown in Fig. 4.

	posture 1	posture 2	posture 3
min ϵ_i [pixel]	1.00	5.83	5.39
max ϵ_i [pixel]	10.03	21.59	15.62
$\sum \epsilon_i$ [pixel]	47.78	112.75	88.23
$\sum \epsilon_i/9$ [pixel]	5.31	12.53	9.80

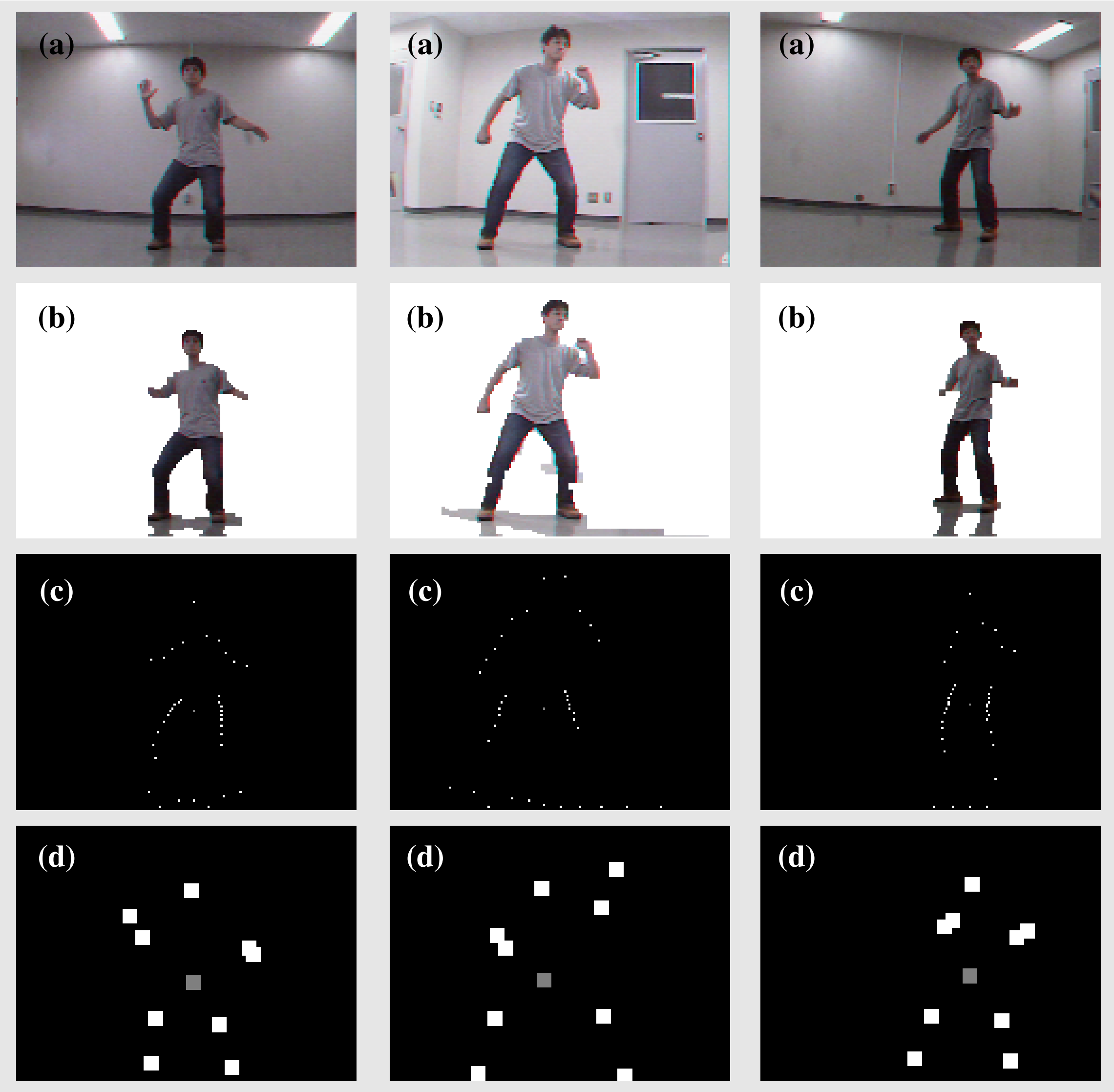

Figure 4. Examples of estimating significant points(left column: posture 1, middle column: posture 2, right column: posture 3). (a) original camera image, (b) human image obtained by background subtraction, (c) extracted significant point image, and (d) estimated human body's significant point image

the ANN is extracted from a human silhouette's contour image, and the output of the ANN indicates the 2D coordinates of the human body's significant points. Using image data extracted from Chinese shadow boxing video, the ANN is trained. The proposed method is implemented on a personal computer and runs in real time. Experimental results confirm both the feasibility and the effectiveness of the proposed method for estimating human body postures.

Acknowledgment

This work was supported by JSPS Grant-in-Aid for Young Scientists (B) 15700160.

References

[1] A. Bottino and A. Laurentini. A silhouette based technique for the reconstruction of human movement. *Computer Vision and Image Understanding*, 83(1):79–95, 2001.

[2] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[3] I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Who? when? where? what? a real time system for detecting and tracking people. In *Proc. of Third IEEE Int. Conf. on FG'98*, pages 222–227, 1998.

[4] 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. In *Proc. of Int. Workshop on mPeople*, pages 3–10, 1999.

[5] I. A. Kakadiaris and D. D. Metaxas. Three-dimentional human body model acquisition from multiple views. *Int. Journal of Computer Vision*, 30(3):191–218, 1998.

[6] A. Koschan, S. Kang, J. Paik, B. Abidi, and M. Abidi. Color active shape models for tracking non-rigid objects. *Pattern Recognition Letters*, 24:1751–1765, 2003.

[7] M. K. Leung and Y. H. Yang. First sight: A human body outline labeling system. *IEEE Trans. on PAMI*, 17(4):359–377, 1995.

[8] Y. Li, A. Hilton, and J. Illingworth. A relaxation algorithm for real-time multiple view 3d-tracking. *Image and Vision Computing*, 20(12):841–859, 2002.

[9] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[10] M. S. Obaidat. Editorial: Artificial neural networks to systems, man, and cybernetics: Characteristics, structures, and applications. *IEEE Trans. on SMC – Part B:Cybernetics*, 28(4):489–495, 1998.

[11] K. Tabb, N. Davey, S. George, and R. Adams. Detecting partial occlusion of humans using snakes and neural networks. In *Proc. of 5th Int. Conf. on EANN'99*, pages 34–39, 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*, 31(14):1–10, 2000.

[13] K. Takahashi and T. Uemura. Real-time human body posture estimation using neural networks. *JSME Int. Journal (C)*, 44(3):618–625, 2001.

[14] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. on PAMI*, 19(7):780–785, 1997.

[15] M. Yamamoto, T. Kondo, T. Yamagiwa, and K. Yamanaka. Skill recognition. In *Proc. of Third IEEE Int. Conf. on FG'98*, pages 604–609, 1998.