Vision-based Human Body Posture Recognition Using Support Vector Machines

Chia-Feng Juang, Chung-Wei Liang, Chiung-Ling Lee
Department of Electrical Engineering
National Chung-Hsing University
Taichung 402, Taiwan, ROC
e-mail: cfjuang@dragon.nchu.edu.tw

I-Fang Chung
Institute of Biomedical Informatics
National Yang Ming University
Taipei 112, Taiwan, ROC
e-mail: ifchung@ym.edu.tw

***Abstract*—This paper proposes a vision-based human posture recognition method using a support vector machine (SVM) classifier. Recognition of four main body postures is considered in this paper, and they are standing, bending, sitting, and lying postures. First of all, two cameras are used to capture two sets of image sequences at the same time. After capturing the image sequences, a RGB-based moving object segmentation algorithm is used to distinguish the human body from background. Two complete and corresponding silhouettes of the human body are obtained. The Discrete Fourier Transform (DFT) coefficients and length-width ratio are calculated from horizontal and vertical projections of each silhouette. Finally, these features are fed to a Gaussian-kernel-based SVM to recognize postures. Experimental results show that the proposed method achieves a high recognition rate.**

Keywords- computer vision, object segmentation, posture recognition, discrete fourier transform, support vector machines.

I. INTRODUCTION

The aim of an intelligent visual surveillance system is to replace the traditional passive video surveillance. Visual surveillance has been implemented on understanding and describing human behaviors [1]. In dynamic scenes, visual surveillance includes the following stages: modeling environments, motion object detection, object tracking, understanding and describing of behaviors. An efficient moving object segmentation algorithm can be employed to distinguish moving objects from background. The background subtraction is a very common segmentation approach, where the background model is registered and subtracted by each current image for foreground object segmentation. In the past, most human body segmentation algorithms used a gray-level model to distinguish moving objects from background [2]-[6]. As color cameras are becoming more and more popular, several object segmentation algorithms have been proposed by using color models. Previous studies have proposed segmentation algorithm using a red, green, and blue (RGB) color model [7, 8] and a normalized RGB color model [9].

After successfully segmenting human body from a sequence of images, the next step is human behavior understanding. Behavior understanding involves the analysis and recognition of body postures or motion patterns. Examples are human gesture recognition [11], sign language recognition [12], and human motion analysis [13]. This paper focuses on recognition of four main body postures, including standing, bending, sitting, and lying postures. Some studies on this topic have been proposed [2, 6, 14-16]. Li et al. [14] used a multiscale morphological method to recognize human postures. Spagnolo et al. [15] proposed an algorithm based on an unsupervised clustering approach for posture recognition. Juang and Chang [6] used discrete Fourier transform (DFT) of vertical and horizontal projections of the silhouette of human body as features. A neural fuzzy network based on training error minimization is used as a classifier. In these studies, recognition is based on monocular image. Different postures may have very similar body silhouette when using monocular image. An example is standing and bending back to the camera. To address this problem, this paper proposes recognition using two cameras.

This paper proposes a new method to recognize standing, bending, sitting, and lying postures using images from two cameras. An RGB-based human body segmentation method is used to segment human bodies from each of the two images from the two cameras. The recognition features are obtained from DFT coefficients and the length-width ratio of human body of each image. The DFT values and length-width ratio of human body are obtained by using vertical and horizontal histogram projections of the silhouette of the human body in each image. The features from the two images are combined and sent to a Gaussian-Kernel-based SVM classifier. The SVM classifier is used because of its good generalization ability.

This paper is organized as follows. Section II introduces the overall system and the RGB-based human body segmentation method. Section III introduces the features used for recognition. Section VI introduces the SVM for multi-class recognition. Experimental results are presented in Section V in order to demonstrate the effectiveness of the proposed method. Finally, conclusions are presented in Selection VI.

II. RGB-BASED HUMAN BODY SEGMENTATION METHOD

The flowchart of the proposed method is shown in Fig. 1. Two cameras are used to capture human body images from two different viewpoints. Moving object segmentation and feature extraction are implemented independently in each of the two images. The segmentation algorithm consists of frame difference, background registration, background difference, and object segmentation. The flowchart of the segmentation algorithm is shown in Fig. 2. This section introduces the RGB-

978-1-4673-2112-9/12/$31.00 ©2012 IEEE

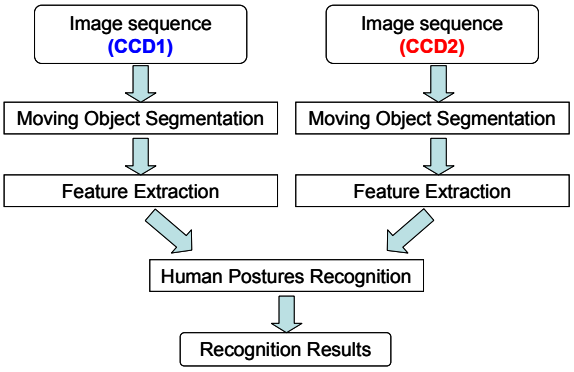

Fig. 1. Flowchart of entire system.

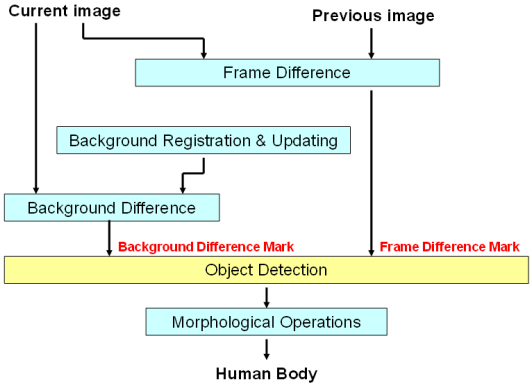

Fig. 2. Flowchart of segmentation algorithm.

based segmentation method as follows.

A. *Frame Difference*

The difference between the current frame and the previous frame is calculated as follows:

$$FD_n(x,y)=\max\left\{\begin{matrix}|R_n(x,y)-R_{n-1}(x,y)|,\\ |G_n(x,y)-G_{n-1}(x,y)|,\\ |B_n(x,y)-B_{n-1}(x,y)|\end{matrix}\right\} \quad (1)$$

$$FDM_n(x,y)=\begin{cases}1 & if\ FD_n \ge Th_{FD}\\ 0 & Otherwise\end{cases} \quad (2)$$

where $R_n(x,y)$, $G_n(x,y)$, and $B_n(x,y)$ represent the red, green, and blue color value of a pixel with position (x, y) in the nth frame. If the frame difference $FD_n(x,y)$ is smaller than a pre-defined threshold value the corresponding pixel is classified as a stationary pixel and marked using $FDM_n(x,y)$.

B. *Background Registeration*

According to $FDM_n(x,y)$, pixels remaining still for a long time (30 frames in this paper) are considered to be reliable background pixels and are registered in the background buffer. A registered background buffer pixel is updated using following equation:

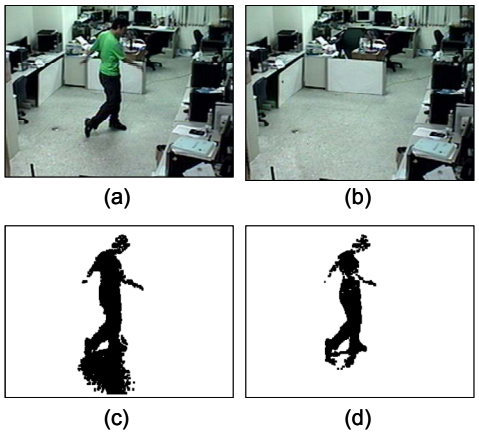

Fig. 3. (a) Original image (b) The registered background(c) Segmentation result (d) Segmentation result after shadow removal.

$$\begin{aligned}&if\ \{\quad |R_n(x,y)-\mu_{Rn}(x,y)|<2\sigma_{Rn}(x,y)\\ &\quad \&\,|G_n(x,y)-\mu_{Rn}(x,y)|<2\sigma_{Gn}(x,y)\\ &\quad \&\,|B_n(x,y)-\mu_{Rn}(x,y)|<2\sigma_{Bn}(x,y)\ \}\\ &then\\ &\begin{cases}\mu_{Rn}(x,y)=\alpha\mu_{Rn-1}(x,y)+(1-\alpha)R_n(x,y)\\ \sigma_{Rn}{}^2(x,y)=\alpha\sigma_{Rn-1}{}^2(x,y)+(1-\alpha)\left(R_n(x,y)-\mu_{Rn}(x,y)\right)^2\end{cases}\\ &\begin{cases}\mu_{Gn}(x,y)=\alpha\mu_{Gn-1}(x,y)+(1-\alpha)G_n(x,y)\\ \sigma_{Gn}{}^2(x,y)=\alpha\sigma_{Gn-1}{}^2(x,y)+(1-\alpha)\left(G_n(x,y)-\mu_{Gn}(x,y)\right)^2\end{cases}\\ &\begin{cases}\mu_{Bn}(x,y)=\alpha\mu_{Bn-1}(x,y)+(1-\alpha)B_n(x,y)\\ \sigma_{Bn}{}^2(x,y)=\alpha\sigma_{Bn-1}{}^2(x,y)+(1-\alpha)\left(B_n(x,y)-\mu_{Bn}(x,y)\right)^2\end{cases}\end{aligned} \quad (3)$$

where $\sigma_{Rn}(x, y)$, $\sigma_{Gn}(x, y)$, and $\sigma_{Bn}(x, y)$ are the standard deviations of each color component of a pixel with position (x, y) in the n-th frame. Fig. 3 (a) and (b) illustrate the original images and the registered background buffer.

C. *Background Difference*

The background difference distinguishes moving objects from the background, and its operations are shown as follows:

$$\begin{aligned}BD_n(x,y)=\exp\Big(&\frac{|R_n(x,y)-\mu_{Rn}(x,y)|}{\sigma_{Rn}(x,y)}\\ &+\frac{|G_n(x,y)-\mu_{Gn}(x,y)|}{\sigma_{Gn}(x,y)}\\ &+\frac{|B_n(x,y)-\mu_{Bn}(x,y)|}{\sigma_{Bn}(x,y)}\Big)\end{aligned} \quad (4)$$

$$BDM_n(x,y)=\begin{cases}1, & if\ BD_n(x,y)\ge Th_{BD}\\ 0, & if\ BD_n(x,y)<Th_{BD}\end{cases} \quad (5)$$

where $BDM_n(x,y)$ is the background difference mark of the pixel with position (x,y) in n-th frame.

D. *Object Segmentation*

A pixel is classified within a moving object when $BDM_n(x,y)$ equals 1. After the segmentation process, we use a morphological operator that performs one erosion operation

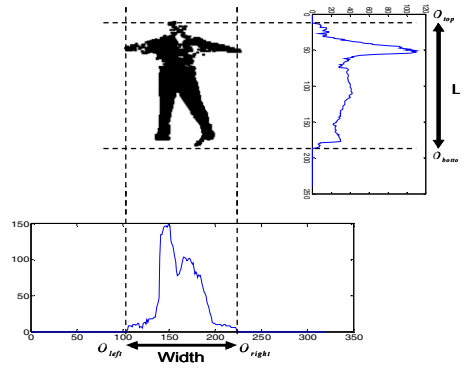

Fig. 4. The length and width calculation of a segmented human body.

followed by one dilation operation, each of which with 3×3 structuring features, to eliminate small noise. The erosion operation is shown as follows:

$$E(x,y) = P \cap (P_1 \cap P_2 \cap P_3 \cap P_4 \cap P_5 \cap P_6 \cap P_7 \cap P_8) \quad (6)$$

And then, the dilation operation is shown as follows

$$D(x,y) = P \cup (P_1 \cup P_2 \cup P_3 \cup P_4 \cup P_5 \cup P_6 \cup P_7 \cup P_8) \quad (7)$$

Fig. 3 (c) shows the segmentation result after the morphological operator, where shadow is included. To remove the shadow, the included angle between the background and a segmented pixel in the RGB space is computed. The angle value measures the difference in lightness. Fig. 3(d) shows the shadow removal result by setting a threshold on the included angle.

III. Feature Extraction

After human body segmentation, the next step is feature extraction from the body silhouette. The length-width ratio is a simple feature of an object. In order to get the length-width ratio we find the vertical and horizontal projections of the body silhouette in each of the two images. First, we use horizontal projection of the body silhouette. Scan each row from top to bottom of the image, count the pixel number of human body in each row, and find the one with the maximum number of pixels. Then we search from the found row to the top of the image, if the row data is less than a threshold, we mark it as the top of human body (O_{top}). Also, we search from the found row to the bottom of the picture, if the row data is less than a threshold, we mark it as the bottom of human body (O_{bottom}). Second, we find vertical projection of the body silhouette. Scan each column from left to right of the picture, and count the pixel number of human body in each column from O_{top} to O_{bottom}. Find the column with the maximum number of pixels. Then search from the maximum column to the left of the picture, if the column data is less than a threshold, we mark it as the left boundary of human body (O_{left}). Also, we search from the maximum column to the right of the picture, if the column data is less than a threshold,

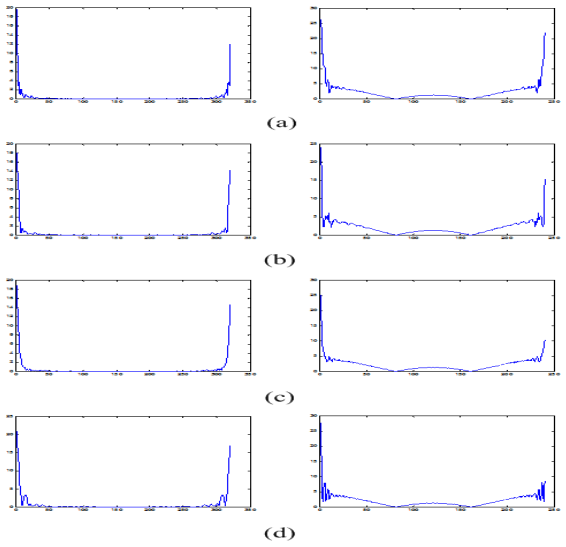

Fig. 5. DFT of histograms of the vertical (left column) and horizontal (right column) projections for postures of (a) standing, (b) bending, (c) sitting, and (d) lying in each of the two captured images.

we mark it as the right boundary of the moving object (O_{right}). The result in a captured image is shown in Fig. 4. Third, we can get the height of a human body by using

$$O_{length} = O_{bottom} - O_{top}. \quad (8)$$

The width of a human body can be derived as follows:

$$O_{width} = O_{right} - O_{left}. \quad (9)$$

Finally, the length-width ratio ($L-W\ ratio$) can be derived as follows:

$$L-W\ ratio = \frac{O_{length}}{O_{width}}. \quad (10)$$

After the horizontal and vertical projections of a human body have been calculated, we take DFT of the projections and find the coefficients as in [6]:

$$F[u] = \sum_{i=0}^{i=N-1} projection(x)_i \exp[-j2\pi ui / N]$$
$$F[v] = \sum_{i=0}^{i=N-1} projection(y)_i \exp[-j2\pi vi / N] \quad (11)$$
$$u, v = 0,1,2,.....,N-1.$$

Illustrative results of DFT of the four classes of body postures are shown in Fig. 5. Then we can compress the feature size by using the first N =20 features and normalize them by $F[1]$. The normalized DFT ($nF[u]$) coefficients are computed as follows:

$$nF[u] = \frac{F[u]}{F[1]} \qquad where\ u = 0,1,2,....,N-1 \quad (12)$$

For each image, the length-width ratio and the 40 normalized DFT coefficients are used the recognition features. The 41

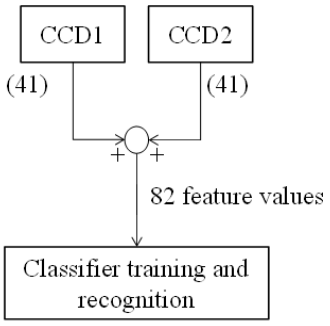

Fig. 6. The feature vector of human body.

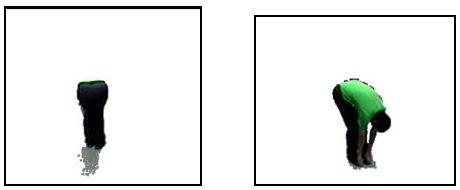

Fig. 7. The compensated image pair by using dual-camera.

features from each of the two images are combined together, resulting in a total of 82 features. These 82 features are fed to a classifier for final posture determination, as shown in Fig. 6.

In [6], recognition using a camera was proposed. Different postures may have similar body silhouettes and, therefore, similar feature information when using only a monocular image. This may lead to a misclassification. An example is shown in Fig. 7. When the man is bending back to the camera, the body silhouette is very similar to that of a standing posture. When a second camera is used, the information from the second body silhouette helps achieve a higher recognition rate.

IV. Multi-class Psoture Classification Using Support Vector Machine Classifiers

This paper uses a Gaussian kernel–based SVM classifier because of its good generalization ability [17]. A Gaussian-kernel-SVM can be trained to classify two classes with desired output being 1 or –1 . For multi-class classification, the approaches of one-against-all and one-against-one [18] are applied. In one-against-all, each posture and the remaining three postures are trained to be separated into two classes. There are four SVMs in the four-class classification problem. For a test pattern, four outputs y_1 , y_2 , y_3 , and y_4 are obtained. If y_i is the largest one, then the unknown posture is classified as the i th posture. Fig. 8 and 9 show the training and test procedures.

In one-against-one, each pair of the four postures is trained at a time, so there are six kinds of training, including postures classes 1 to 2, 1 to 3, 1 to 4, 2 to 3, 2 to 4, and 3 to 4. There are six outputs after test. We use a voting strategy to decide the final result with the maximum number of votes. The training and test procedures are shown in Fig. 10 and 11.

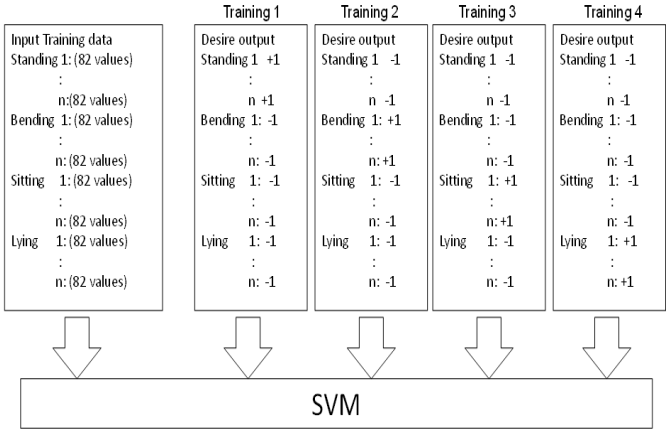

Fig. 8. The training procedure by SVM using the one-against-all classification approach.

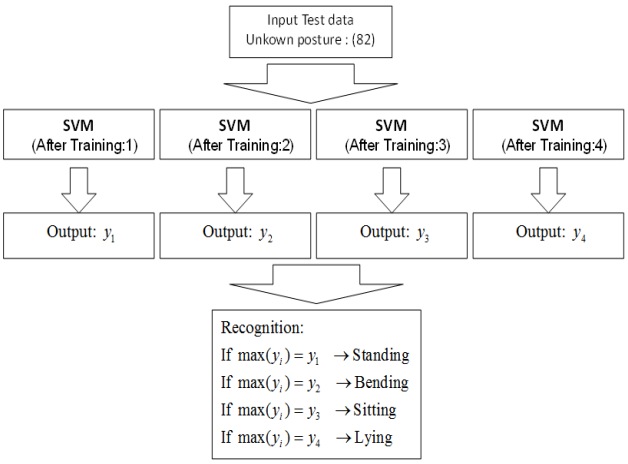

Fig. 9. The test procedure by SVM using the one-against-all classification approach.

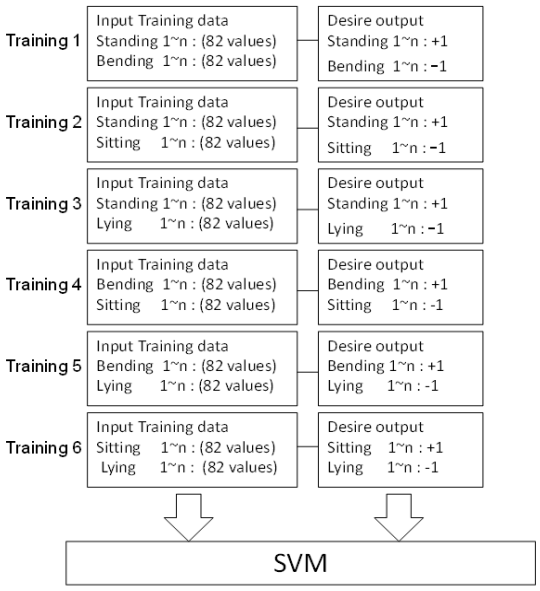

Fig. 10. The SVM training procedure with one-against-one approach.

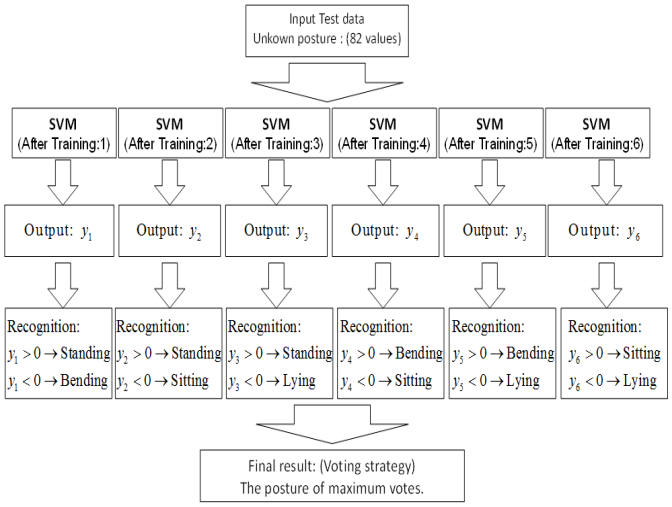

Fig. 11. The SVM test procedure with the one-against-one approach.

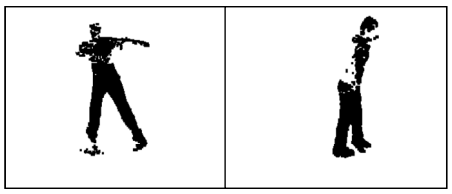

Fig. 12. Illustrative segmented results of the four main postures.

TABLE I. TEST RESULTS BY GAUSSIAN-SVM USING 400 TRAINING DATATABLE TYPE STYLES

Classifiers	SVM One-against-one	SVM One-against-all
# SVs	232	279
# Parameters	19522	23210
Test rate (%)	95.66	95.06

TABLE II. TEST RESULTS BY SONFIN USING 400 TRAINING DATATABLE TYPE STYLES

Methods	Method [6] Camera 1	Method [6] Camera 2	W4 [2]	SVM One-against-one
Test rate (%)	85.53	88.94	84.43	95.66

V. EXPERIMENTS

This section shows the posture recognition results by using Gaussian-kernel-based SVMs. There were a total of 400 training samples, with 100 samples for each of the four main postures. An illustrative human body segmentation result is shown in Fig. 12. For testing, a set of 1843 consecutive images were used, with 334 for standing, 560 for bending, 553 for sitting, and 396 for lying postures. For Gaussian- kernel-based SVM training, the cost parameter C and Gaussian kernel width parameter γ are selected form the sets {1, 10, 100, 100} and {2, 512, 1024}, respectively. There were 12 kinds of combinations for training and we chose the one with the best test performance for comparison. Table I shows the test rates of SVMs with the one-against-one and one-against-all approaches. The results show that the one-again-one approach slightly outperforms the one-against-all approach.

For the purpose of comparison, different recognition approaches [2, 6] were applied to the same problem. The first method used for comparison was the W4 method [2]. In this method, average normalized horizontal and vertical projections of human body silhouettes were used as templates. Template matching method was used for classification. Here, the templates were obtained from the 400 samples from two images. Table II shows the recognition rate of this method. The results show that the proposed method outperforms the W4 method.

In [6], recognition using DFT coefficients from a monocular image and a neural fuzzy network classifier [19] was proposed. To improve the recognition rate of this approach, a larger training data set containing 1200 training samples was used for classifier design, where there were 300 samples for each of the four main postures. Table II shows the recognition rates when using a monocular image from camera one or camera two. The results show that the recognition rate using camera one or two is similar and both rates are lower than the proposed method.

VI. CONCLUSION

This paper proposes a new posture recognition method using two cameras. RGB-based human body segmentation approach is proposed to effectively segment moving objects from the background. The body length-width ratio and the DFT coefficients of horizontal and vertical histograms in the two images are combined to from a new recognition feature for recognition performance improvement. Both of the two kinds of SVM training approaches used for classification show good classification results. In the future, new recognition features will be studied for recognition performance improvement. In addition, the proposed approach will be applied to real-time falling down detection in a homecare system.

REFERENCE

[1] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. System, Man, and Cyber., Part C: Applications and Reviews,* vol.34, no.3, pp. 334-352, Aug. 2004.

[2] I. Haritaoglu, D. Harwood and L. S. Davis, "W^4 real-time surveillance of people and their activities," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809 – 830, Aug. 2000.

[3] S. Y. Chen, S. Y. Ma, and L. G. Chen, "Efficient moving object segmentation algorithm using background registration technique," *IEEE Trans. Circuits and Systems for Video Technique*, vol. 12, no. 7, pp. 577-586, July 2002.

[4] C. Kim, J. Cho, and Y. Lee, "The relational properties among results of background subtraction", *Proc. of Int. Conf. on Advanced Communication Technology*, vol. 3, pp. 1887-1890, 2008.

[5] C. F. Juang, C. M. Chang, J. R. Wu, and D. M. Lee, "Computer vision-based human body segmentation and posture estimation," *IEEE Trans. Syst., Man, and Cyber., Part A: Systems and Humans*, vol. 39, no. 1, pp. 119-133, Jan. 2009.

[6] C. F. Juang and C. M. Chang, "Human body posture classification by a neural fuzzy network and home care system application," *IEEE Trans. Syst., Man, and Cyber., Part A: Systems and Humans,* vol. 37, no. 6, pp. 984-994, Nov. 2007.

[7] Q. Zhou and J.K. Aggarwal, "Tracking and classifying moving objects from video", *Proc. IEEE Int. Workshop Performance Evaluation of Tracking and Surveillance,* pp. 52-59, Dec. 2001.

[8] C. F. Juang, S. H. Chiu, and S. J. Shiu, "Fuzzy system learned through fuzzy clustering and support vector machine for human skin color segmentation," *IEEE Trans. Syst., Man, and Cyber., Part A: Systems and Humans*, vol. 37, no. 6, pp. 1077-1087, Nov. 2007.

[9] M. T. Razali and B. J. Adznan, "Detection and classification of moving object for smart vision sensor," *Proc. Information and Communication Technologies*, vol. 1, pp. 733-737, 2006.

[10] R. C. Gonzalez and R. E. Woods, Digital Image Processing 2/e, Prentice Hall, 2008

[11] C. F. Juang and K. C. Ku, "A recurrent fuzzy network for fuzzy temporal sequence processing and gesture recognition," *IEEE Trans. Syst., Man, and Cyber.,-Part B: Cyber.*, vol. 35, no. 3, pp. 646-658, Aug. 2005.

[12] T. Starmer, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer-based video," *IEEE Transactions Pattern Analysis Machine Intelligence*, vol. 20, pp. 1317-1375, Dec. 1998.

[13] H. Fujiyoshi and A. J. Lipton, "Real-time human motion analysis by image skeletonization," *Proc. IEEE Workshop on Applications of Computer Vision*, pp. 15-21, Oct. 1998.

[14] Y. Li, M. Songde, and L. Hanqing, "A multiscale morphological method for human posture recognition," *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, pp. 56-61, Apr. 1998.

[15] P. Spagnolo, M. Leo, A. Leone, G. Attolico, and A.Distante, "Posture estimation in visual surveillance of archaeological sites," *Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance*, pp. 277-283, July 2003.

[16] M. Singh, A. Basu and M. Kr. Mandal, "Human activity recognition based on silhouette directionality," *IEEE Trans. Circuits and Systems For Video Technology,* vol.18, no. 9, pp. 1280-1292, Sep. 2008.

[17] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[18] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415–525, Mar. 2002.

[19] C.F. Juang and C.T. Lin, "An on-line self-constructing neural fuzzy inference network and its applications," *IEEE Trans. Fuzzy Systems*, vol.6, no.1, pp.12-32, Feb. 1998.