Human Behavior Analysis Using Multiple 2D Features and Multicategory Support Vector Machine

Hao-Cheng Mo, Jin-Jang Leou, and Cheng-Shian Lin

Department of Computer Science and Information Engineering
National Chung Cheng University,
Chiayi, Taiwan 621, Republic of China
{mhc95m, jjleou, lchh95p}@cs.ccu.edu.tw

Abstract

In this study, a human behavior analysis system using multiple 2D (two-dimensional) features and a multicategory support vector machine is proposed. In the proposed system, three kinds of features, namely, human star skeleton, angles of six sticks in the star skeleton, and object motion vectors, are employed to train the human posture classifier and recognize human postures. Based on the recognized human postures, a backward search strategy is proposed to recognize human actions. Based on the experimental results obtained in this study, in terms of recall and precision rates, the proposed system has good performance and is superior to the comparison system.

1. Introduction

Most traditional surveillance systems may have the only function of recording video events. Recently, a surveillance system can analyze video contents and recognize human postures and actions. Because most elders stay in their homes for a long time, there is a high probability that many accidents for elders happen in their homes. If dangerous events of elders can be detected as soon as possible, elders' injuries will decrease. Many human behavior analysis systems for homecare are proposed to monitor human actions and recognize elders' behaviors, such as falling down, sitting, and bending down. Once some dangerous events are detected, the systems will automatically give some necessary responses.

Based on the data type, existing human action recognition approaches can be classified into two major types [1]: 2D based or 3D based. A 3D based approach may have higher accuracy with higher computational complexity, whereas a 2D based approach may have lower accuracy with lower computational complexity. To recognize human actions, some human features should be extracted from video frames by background/foreground separation [2]. Yoo, Nixon, and Harris [3] proposed a new human feature extraction method and used a 2D stick figure to build the human body model. For the 2D contour approach, Haritaoglu, Harwood, and Davis [4] used human silhouette boundaries to extract extreme points of silhouettes as human features.

Human behavior analysis involves recognizing human actions in video sequences and high-level description of human actions. Existing human behavior analysis approaches can be generally classified into four categories, namely, dynamic time warping (DTW), finite state machine (FSM), hidden Markov model (HMM), and support vector machine (SVM). Cuntoor, Kale, and Chellappa [5] analyzed different human features for human identification by DTW. FSM is a state-transition function containing a finite number of states, which are used to judge which reference sequence matches the test sequence. HMM is a kind of stochastic state model containing many hidden and observable parameters. Guo and Miao [6] proposed a homecare surveillance system, in which HMMs are used to build the motion model and recognize human postures. In [7], multidimensional discrete HMMs are used to model and recognize human actions. Schuldt, Laptev, and Caputo [8] used a local SVM approach to recognize complex human actions. In this study, a human behavior analysis system using multiple 2D features and a multicategory SVM is proposed.

The paper is organized as follows. The proposed system is addressed in Section 2. Simulation results are addressed in Section 3, followed by concluding remarks.

2. Proposed Human Behavior Analysis System

As shown in Fig. 1., the proposed system contains three parts, namely, foreground/background segmentation, feature extraction, and human action recognition.

2.1. Foreground/Background Segmentation

Initially, the proposed system uses the first N frames of a video sequence to build the background pixel model. The intensity for a stationary background pixel can be modeled as a Gaussian distribution and the statistical background subtraction method [9] is used. That is, for a stationary background pixel, its intensity at ith frame, x_i, can be modeled as:

$$p(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i - m)^2}{2\sigma^2}\right], \qquad (1)$$

where m and σ^2 are the mean and variance values, respectively, which will be updated using new video frames. Then, $p(x_i)$ is used to determine a pixel by:

This work was supported in part by National Science Council under Grants NSC 95-2221-E-194-020-MY3 and NSC 96-2221-E-194-033-MY3.

$$x_i = \begin{cases} foreground, & if\ p(x_i) < \gamma, \\ background, & otherwise, \end{cases} \quad (2)$$

where γ is a threshold. Due to noise and illumination variation, a foreground human object may contain some fragmented regions. Here, two fundamental morphological operators, namely, dilation and erosion, are used to remove fragmented regions and fill in small holes in each human object.

2.2. Feature Extraction

Using the whole human contour to represent a human posture is inefficient. Here, star skeleton [10] is employed, in which terminal points on the human object contour are detected and each terminal point is connected to the human object centroid. Traditional star skeleton has to detect local maxima of the human object contour, which is computationally expensive and sensitive to noise. Here, a modified star skeleton representation for human object contour is proposed, which is described as follows.

1. Determine the centroid (x_c,y_c) of the human object contour.
2. Divide the human object contour into two (left and right) parts via its centroid. Find the highest, leftest, and lowest points in the left bounding box and find the highest, rightest, and lowest points in the right bounding box.
3. As shown in Fig. 2, calculate six Euclidean distances d_i, i = 1, 2, …, 6, from the six terminal points to the object centroid. Then, the six stick lengths, d_1, d_2, …, d_6, are six extracted features.

For a human star skeleton containing six sticks, if the angle between the ith stick and the horizontal line is denoted by θ_i, the six angles $\theta_1, \theta_2, ..., \theta_6$ in the human star skeleton are other useful features. Note that the angle for the line segment connecting a terminal point (x_t, y_t) and the object centroid (x_c, y_c) in the star skeleton can be determined by

$$\theta = \begin{cases} \tan^{-1}\dfrac{y_t - y_c}{x_t - x_c}, & if\ x_t > x_c\ and\ y_t \geq y_c, \\ 90^\circ, & if\ x_t = x_c\ and\ y_t > y_c, \\ 180^\circ - \tan^{-1}\dfrac{y_t - y_c}{x_t - x_c}, & if\ x_t < x_c\ and\ y_t > y_c, \\ 180^\circ, & if\ x_t < x_c\ and\ y_t = y_c, \\ 180^\circ + \tan^{-1}\dfrac{y_t - y_c}{x_t - x_c}, & if\ x_t < x_c\ and\ y_t < y_c, \\ 270^\circ, & if\ x_t = x_c\ and\ y_t < y_c, \\ 360^\circ - \tan^{-1}\dfrac{y_t - y_c}{x_t - x_c}, & otherwise. \end{cases} \quad (3)$$

Additionally, object motion can be extracted as another feature. Here, each video frame is divided into equal-sized blocks with block size being 8×8. A simple match measure for motion estimation, namely, the sum of absolute differences (SAD) is employed, which is defined as

$$SAD(m,n) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left|I_c(i,j) - I_r(i+m, j+n)\right|, \quad p \geq m, n \geq -p, \quad (4)$$

where I_c and I_r are the pixel intensity values of a block in the current frame and the corresponding block in the reference frame, respectively, (i,j) denotes the pixel coordinate, and p is the search range in x and y directions. To reduce motion estimation computations, pixel-by-pixel differences between a block in the current frame and the corresponding frame in the reference frame can be computed first. If all pixel-by-pixel differences are less than a threshold, the block is treated as a static background block, whose motion estimation will be avoided. Here, only the 25 largest motion vectors, $[v_x^1, v_x^2, ..., v_x^{25}]$ and $[v_y^1, v_y^2, ..., v_y^{25}]$ in each frame are extracted as features.

As a summary, the combined feature vector for the tth frame of a video sequence is defined as:

$$F_t = [d_1, ..., d_6, \theta_1, ..., \theta_6, v_x^1, ..., v_x^{25}, v_y^1, ..., v_y^{25}]. \quad (5)$$

2.3. Human Behavior Analysis

In this study, human behavior analysis contains two parts, namely, human posture recognition and human action recognition.

2.3.1 Human Posture Recognition

In this study, the posture of a person is represented by a combined feature vector in Eq. (5). To recognize human postures, a multicategory support vector machine (MSVM) based classifier is employed to label a sequence of frames as one of k categories, such as walking, bending down, and sitting. Here, a multicategory support vector machine from library for support vector machines (LIBSVM) [11] is employed. Because extracted features are not linear, extracted features are mapped from the original feature space into a high-dimensional feature space by a kernel function. A popular radial basis function (RBF) is selected as the kernel function in LIBSVM, which is defined as:

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2), \quad \text{for } \gamma > 0, \quad (6)$$

where x_i and x_j are the feature vectors, and γ is the kernel parameter to be determined. To train the proposed human posture recognition classifier, the cross-validation approach is employed. In cross-validation, all training samples are partitioned into N subsets of equal size. Then, one subset tests the classifier trained by the remaining N-1 subsets sequentially. Each unknown human posture of each frame will be recognized by the proposed human posture classifier as one of the main human postures.

2.3.2 Human Action Recognition

For human action recognition, a human action is recognized by recognizing successive human postures. When a person is executing some action, many different postures will appear over a time period. As he changes his original action to another action, he will execute a transition action, which does not belong to any meaningful action. The transition action might consist of several transition postures. While recognizing which human action in the current frame, a backward search strategy is employed. If previous 4 frames (frames n-4, n-3, n-2, n-1) and the current frame (frame n) are recognized as the same human posture, these frames will be recognized. Otherwise, the current frame will be tentatively skipped and the next frame will be recognized by the same backward search strategy. Other "undetermined" frames will be determined as "transition" frames.

3. Simulation Results

In the proposed system, recall and precision for human action recognition are used as the metrics for performance assessment, which are defined as:

$$recall = \frac{TP}{TP + FN}, \tag{7}$$

$$precision = \frac{TP}{TP + FP}, \tag{8}$$

where TP means true positive, FN means false negative, and FP means false positive.

Six video sequences are employed to evaluate the performance of the proposed system. Because the proposed system is applied on homecare surveillance, it is assumed that each video sequence is taken by a fixed camera and each video sequence has a "stationary" background over some time period. Here, the five main human actions include "walking," "swinging," "bending down," "sitting," and "falling down." The first and second sequences having 800 and 1446 frames, respectively, contain the five main human actions. The third, fourth, and fifth sequences having 417, 598, and 654 frames, respectively, contain four main human actions. The sixth sequence having 515 frames contains three main human actions. To deal with the transition time period, a special action, "transition action," is defined, which is not any of the five main human actions. Each video frame is 352×240 in size with frame rate = 29.97. To evaluate the performance of our proposed system, the human action recognition system proposed by Chen et al. [12] is implemented in this study for comparison. Some recognized human actions by the proposed system for the first video sequence are shown in Fig. 3. The recall and precision rates of the comparison system [12] and the proposed system for the first video sequence are shown in Table 1. Performance comparison between the comparison system [12] and the proposed system for the six video sequences is listed in Table 2.

4. Concluding Remarks

In this study, a human behavior analysis system using multiple 2-D features and a mutlicategory support vector machine is proposed. Three kinds of features, namely, human star skeleton, angles of six sticks in the star skeleton, and object motion vectors are employed to train the human posture recognition classifier and recognize human postures. Based on the recognized human postures, a backward search strategy is proposed to recognize human actions. Based on the experimental results obtained in this study, in terms of recall and precision rates, the proposed system has good performance and is superior to the comparison system.

References

[1] D. Y. Chen, S. W. Shih, and H. Y. Mark Liao, "Human action recognition using 2-D spatio-temporal templates," in Proc. of IEEE Int. Conf. on Multimedia and Expo, July 2007, pp. 667-670.

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

[3] J. H. Yoo, M. S. Nixon, and C. J. Harris, "Extracting human gait signatures by body segment properties," in Proc. of IEEE Southwest Symposium on Image Analysis and Interpretation, Apr. 2002, pp. 35-39.

[4] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: real-time surveillance of people and their activities," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 22, no. 3, pp. 809-830, Aug. 2000.

[5] N. Cuntoor, A. Kale, and R. Chellappa, "Combining multiple evidences for gait recognition," in Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Apr. 2003, pp. (III)113-116.

[6] P. Guo and Z. Miao, "A home environment posture and behavior recognition system," in Proc. of IEEE Int. Conf. on Convergence Information Technology, Nov. 2007, pp. 175-180.

[7] M. Ahmad and S. W. Lee, "Human action recognition using multi-view image sequences features," in Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, Apr. 2006, pp. 523-528.

[8] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in Proc. of IEEE Int. Conf. on Pattern Recognition, Aug. 2004, vol. 3 pp. 32-36.

[9] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: real-time tracking of the human body," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 780-785, July 1997.

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

[11] C. C. Chang and C. J. Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[12] H. S. Chen, H. T. Chen, Y. W. Chen, and S. Y. Lee, "Human action recognition using star skeleton," in Proc. of the 4th ACM Int. Workshop on Video Surveillance and Sensor Network, Oct. 2006, pp. 171-178.

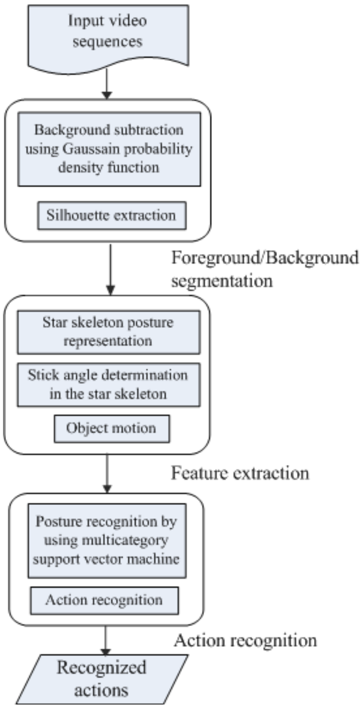
Input video sequences
Background subtraction using Gaussain probability density function
Silhouette extraction
Foreground/Background segmentation
Star skeleton posture representation
Stick angle determination in the star skeleton
Object motion
Feature extraction
Posture recognition by using multicategory support vector machine
Action recognition
Action recognition
Recognized actions

Figure 1. The proposed system architecture.

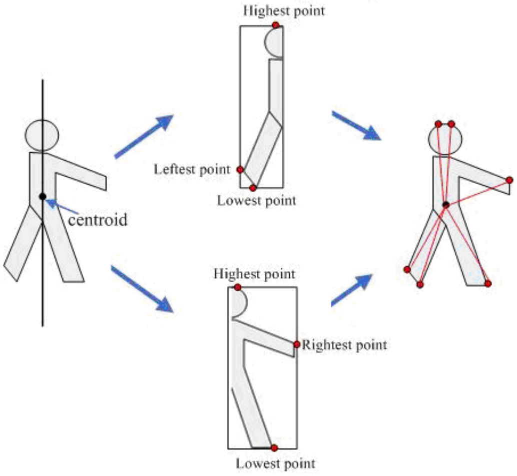

Figure 2. The star skeletonization procedure.

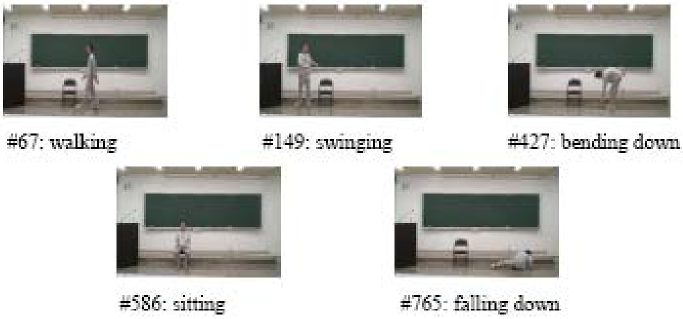

Figure 3. Some recognized human actions for the first video sequence.

Table 1. The recall and precision rates of the comparison system and the proposed system for the first video sequence.

		recall (%)	*precision* (%)
Walking	Comparison	80.45	64.5
	Proposed	97.74	89.52
Swinging	Comparison	87.05	69.16
	Proposed	92.94	84.94
Bending down	Comparison	86.66	88.35
	Proposed	95.23	91.59
Sitting	Comparison	90.51	83.33
	Proposed	98.27	95.79
Falling down	Comparison	86.56	100
	Proposed	95.22	95.52
Average	Comparison	84.82	73.81
	Proposed	96.55	90.92

Table 2. Performance comparison between the comparison system and the proposed system for the six video sequences.

		recall (%)	*precision* (%)
Sequence 1	Comparison	84.82	73.81
	Proposed	96.55	90.92
Sequence 2	Comparison	76.39	65.93
	Proposed	89.56	82.26
Sequence 3	Comparison	75.7	62.79
	Proposed	89.09	83.24
Sequence 4	Comparison	85.83	76.28
	Proposed	86.51	81.21
Sequence 5	Comparison	82.43	63.84
	Proposed	93.5	77.1
Sequence 6	Comparison	85.43	79.63
	Proposed	91.39	87.6
Average	Comparison	81.77	70.38
	Proposed	91.1	83.72