2009 IEEE International Symposium on Intelligent Control
Part of 2009 IEEE Multi-conference on Systems and Control
Saint Petersburg, Russia, July 8-10, 2009

Fuzzy Rule Inference Based Human Activity Recognition

Jyh-Yeong Chang, *Member, IEEE* Jia-Jye Shyu, and Chien-Wen Cho

Abstract—**Human activity recognition plays an essential role in e-health applications, such as automatic nursing home systems, human-machine interface, home care system, and smart home applications. Many of human activity recognition systems only used the posture of an image frame to classify an activity. But transitional relationships of postures embedded in the temporal sequence are important information for human activity recognition.**

In this paper, we combine temple posture matching and fuzzy rule reasoning to recognize an action. Firstly, a fore-ground subject is extracted and converted to a binary image by a statistical background model based on frame ratio, which is robust to illumination changes. For better efficiency and separability, the binary image is then trans-formed to a new space by eigenspace and canonical space transformation, and recognition is done in canonical space. A three image frame sequence, 5:1 down sampling from the video, is converted to a posture sequence by template matching. The posture sequence is classified to an action by fuzzy rules inference. Fuzzy rule approach can not only combine temporal sequence information for recognition but also be tolerant to variation of action done by different people. In our experiment, the proposed activity recognition method has demonstrated higher recognition accuracy of 91.8% than the HMM approach by about 5.4 %.

I. Introduction

Human activity recognition plays an important role in applications such as automatic surveillance systems, human-machine interface, home care system and smart home applications. For example, an automatic system will trigger an alarm condition when the automated surveillance system detect and recognize suspicious human activities. Human activity recognition can also be used in extracting semantic descriptions from video clips to automate the process of video indexing. However, there is no rigid syntax and well-defined structure as that of the gesture and sign language which can be used for activity recognition. Therefore, this makes human activity recognition become a more challenging task**.**

Several human activity recognition methods have been proposed in the past few years. A detailed survey is introduced in [1]. Most of human activity recognition methods can be classified into two categories depending on the features being used. The first one makes use of motion-based features [2], [3]. In [2], Bobick and Davis recognized the human activities by comparing motion-energy and motion-history of template images with temporal images. In [3], Hamid *et al*. extracted spatio-temporal features such as the relative distance between two hands and their velocities; furthermore thy used dynamic Bayesian networks to recognize human activities such as writing, drawing and erasing on a white board. On the other hand, 2-D and 3-D shape features were used to recognize activities [4], [5]. In [4], shape was represented by edge data obtained from canny edge detector, and key frames were defined for each activity. In [5], the authors presented a view-independent 3-D shape description for classifying and identifying human activity using SVM.

If we only adopt the motion-based and shape-based features to recognize an activity, many activities remain unidentified since the temporal information is discarded. Hence, this motivates us to design a robust method that uses temporal information, which is implicitly inherent in the human activity recognition. People have the same postures and posture sequences when they perform a specific activity. Therefore, we use shape features to classify each image frame into postures we defined. Then, we use the frame sequences of key postures to recognize which activity one does. Besides, a human body has almost constant natural frequency when one performs an action. It is the congenital restrictions of people. There are few differences between two image frames if they are captured in a short period. Hence, we can down sample the video frame instead of using all the thirty frames per second. Down sampling can also ease the intensive computational and memory loads encountered in a video signal processing.

II. Video Frame Preprocessing for Activity Recognition

A. Object Extraction

The first step of human activity recognition system is object extraction. We have to construct a background model for object extraction. There are many well-known background models. The most common one is that applies frame difference with a threshold. W^4 is such a typical example with some modifications [6]. It records the maximum and minimum grayscale and the maximum inter-frame difference of each pixel in a background video. Then each image frame subtracts the maximum and minimum grayscale of each pixel. If the pixel's absolute value of the

Manuscript received January 22, 2009. This work was supported in part by the National Science Council under grants NSC 94-2213-E-009-097, NSC95-2752-E-009-011-PAE, and 95-EC-17-A-02- S1-032 Taiwan, R.O.C.

Jyh-Yeong Chang and Jia-Jye Shyu are with the Department of Electrical and Control Engineering, National Chiao Tung University, Hsinchu, Taiwan, R. O. C. (e-mail: jychang@mail.nctu.edu.tw and sjj.zyca@msa.hinet.net).

Chien-Wen Cho was with the Department of Electrical and Control Engineering, National Chiao Tung University, Hsinchu, Taiwan, R. O. C. and is now working in Hsinchu, Taiwan, R. O. C.. (e-mail: cwcho.ece90g@nctu.edu.tw).

subtraction operation is larger than the maximum inter-frame difference, the pixel is classified to a foreground one. W^4 admits some rules make the background model be adaptive to varying environment. Following W^4 approach, we describe the background scene as a statistical model. We obtain a background model from pure background video by calculating the maximum, minimum gray level and frame ratio of each pixel in the images.

Although extraction of foreground based on frame difference approach is the most famous method in image processing, the drawback involves the robustness of illumination changes. If we film an environment at a standstill, background modeling based on frame difference may still invoke errors due to the illumination changes. As a result, noises will be detected and the quality of object extraction will be affected. We have proposed a method utilizing frame ratio, instead of frame difference, which has been proved robust to the illumination changes.

B. Posture Representation

In video and image processing, the dimensions of image data are often extremely large. Because there are great deals of redundancies in the images, it is common to transform image from one space to another space to reduce redundancy. Many methods like Fourier Transformation, wavelet, Principal Component Analysis (PCA) and eigenspace transformation (EST) has actually been demonstrated to be a potent scheme to this end. However, PCA based on the global covariance matrix of the full set of image data is not sensitive to class structure in the data. In order to increase the discriminatory power of various activity features, Etemad and Chellappa [11] use Linear Discriminant Analysis (LDA), also called Canonical Analysis (CA) [7], which can be used to optimize the class separability of different activity classes and improve the classification performance. The features are obtained by maximizing between-class and minimizing within-class variations. Unfortunately, this approach has high computation cost when applying to large images. It was only tested with small images. Here we call this approach canonical space transformation (CST). Combining EST based on PCA with CST based on CA, our approach reduces the data dimensionality and optimizes the class separability of different action sequences simultaneously.

Images in high-dimensional image space are converted to low-dimensional eigenspace using PCA. The obtained vector thus is further projected to a smaller canonical space using CST. Recognition is accomplished in the canonical space.

III. Video Frame Activity Recognition Procedure

A. Activity Template Selection

There are few differences between two postural image frames if they are captured in a short interval. Besides, a human body is a rigid body, thus has its natural frequency; namely, it has restriction on action speed when doing some specific actions. Therefore, we select some key frames from a sequence to represent an activity. Cameras usually capture image frames in a high frequency. In our approach, we select one image frame, as called the essential template image, with a fixed interval instead of each image.

These essential templates are transformed to a new space by eigenspace transformation (EST) and canonical space transformation (CST). As described in Sec. 2, we only utilize k largest eigenvalues and their associated eigenvectors to approximate the image. The approximation can decrease data dimension, but it would also lose slight information of image with few differences. However, two similar image frames will converge to two near points after eigenspace and canonical space transformation. The images of similar postures done by difference people also barely converge to one point. Consequently, we select only essential templates rather than use all sequences for human activity recognition.

B. Construction of Fuzzy Rules from Video Streams

Transitional relationships of postures in temporal sequence are important information for human activity classification. If we only utilize one image frame to classify the action, classification result may be failed easily because human's actions may have similar postures in two different activity sequences. For example, the action of "jumping" and "crouching" both have the same postures called common states of "Standing Right."

Human activities have lots of ambiguity, so we propose a fuzzy rule base approach which not only can combine temporal sequence information for recognition but also can be tolerant to variation of actions done by different people. Fuzzy rule base classification has known with the ability to absorb data difference by learning and has been successfully used in many applications. In our system we view each transformed vector of the temporal image as a different feature.

In our approach, EST and CST methods are used to extract features. Assume that there are c clusters in the system. As described above, each image frame is transformed to a $(c-1)$-dimensional vector by EST and CST methods.

We make use of the membership functions to represent the features' possibility to each cluster. Many types of membership functions, e.g., bell-shaped, triangular, and trapezoid ones, are frequently used in a fuzzy system. We choose the Gaussian type membership function to represent the features because the Gaussian type membership function can reflect the similarity via the first order and second order statistics of clusters and is differentiable.

Firstly, when the k-th training image frame $\mathbf{x}_k$ is inputted, the feature vector $\mathbf{a}_k$ is extracted by

$$\mathbf{a}_k = \mathbf{H}\,\mathbf{x}_k. \tag{1}$$

At the same time, $\mathbf{a}_k$ can be rewritten as

$$\mathbf{a}_k = \left[a_k^1, a_k^2, \cdots, a_{i,j}^{c-1}\right]^{\mathrm{T}} \tag{2}$$

If we assume the dimensions of the feature vectors are independent, a local measure of similarity between the training vector and each template vectors can be computed. Let $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ denote respectively the covariance matrix and mean vector of all essential template vectors and C_i denote the i-th class of essential templates. The membership function is given by

$$\begin{aligned} P_k &= \mathrm{M}(\mathbf{a}_k | C_i) \\ &= \arg\max_i \max_j \left\{ \frac{1}{(2\pi)^{\frac{c-1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} (\mathbf{a}_k - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{a}_k - \boldsymbol{\mu}) \right] \right\} \\ &= \arg\max_i \max_j \left\{ \prod_{m=1}^{c-1} \frac{1}{\sqrt{2\pi}\sigma_m} \exp\left[-\frac{1}{2} \frac{(a_k^m - \mu_{i,j}^m)^2}{\sigma_m^2} \right] \right\} \end{aligned} \quad (3)$$

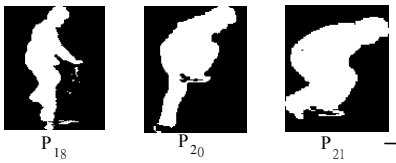
where m is the number of dimension and j is the training model, i. e., action person, index. P_k denotes the grade of membership function of maximal category of the k-th image frame.

As developed by Wang and Mendel [10], fuzzy rules can be generated by learning from examples. Three contiguous images are combined as a group (I_1, I_2, I_3) in our approach. We view the transformation of the three images as three features, and form a feature vector set [$\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$]. An image sequence with feature vector set [$\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$] is associated with its output of corresponding activity, to lead to an input-output pair being learned in the fuzzy rule base. For example, an image sequence, its corresponding CST transformations, $\mathbf{a}_1^1, \mathbf{a}_2^1$, and $\mathbf{a}_3^1$, of image 1, image 2, image 3, where images 1, 2, and 3 are the three consecutive 5:1 down-sampled images captured by the camera, is given by $\left[\mathbf{a}_1^1, \mathbf{a}_2^1, \mathbf{a}_3^1; D_1\right]$ and D_1 is the corresponding belonging action category. The class of each image is obtained by Eq. (3) above for key posture classification. Let images 1, 2, 3 belong to postures P_i, P_j, P_k, respectively. Finally, a rule is supported by these three images as given by

Rule q. IF the activity's I_1 is P_i AND its I_2 is P_j AND its I_3 is P_k, THEN the activity is D_l. (4)

where I_i is Image i and P_j is Posture j.

Due to a large number of training activities, some conflicting rules may be generated. The conflicting rules have the same antecedent conditions but lead to different consequent conditions. For a set of antecedent conditions, we can have only one rule to reflect it. Therefore, we have to choose one from the two or more conflicting rules from each qualified cluster. To this end, we choose the rule that is supported by a maximum number of examples. Furthermore, to prune redundant or inefficient fuzzy rules, if the supporting actions of a rule are less than a threshold, the rule is excluded from defining an IF-THEN rule. Fig. 1 demonstrates a fuzzy rule learned to classify action "climbing up."



IF the activity's I_1 is P_{18} AND its I_2 is P_{20} AND its I_3 is P_{21}, THEN the activity is C_{UP}

Fig. 1. A fuzzy rule learned to classify action C_{UP}.

C. *Activity Classification*

When a video stream is inputted for recognition, we extract image frames from the video first. Then we utilize background model of Section 2.1 to extract foreground subject from the scene. The foreground object is a binary image. Suppose that we have given three consecutive down-sampled images to recognize their action type. Theses images need all the pre-processes including object extraction and normalization.

The set of these three images is inferred to all fuzzy rules we constructed and compute the similarity between the image sequence set and the postural sequence of all rules in the training data base. For example, there is a rule "IF the activity's I_1 is P_{n_1} AND its I_2 is P_{n_2} AND its I_3 is P_{n_3}, THEN the activity is D_n" in the rule base. In order to calculate the similarity, we take out the membership functions r_{k-2,n_1}, r_{k-1,n_2} and r_{k,n_3} which are corresponding to the three category of linguistic labels, P_{n_1}, P_{n_2} and P_{n_3}, in the rule and have been calculated by Eq. (3). The summation of r_{k-2,n_1}, r_{k-1,n_2} and r_{k,n_3} is the similarity between current image sequence and the postural sequence of this rule. We can obtain the similarity related to all fuzzy rules of training data base in the same manner. The rule, which has the highest value of similarity, is selected and the activity is classified to the activity of this rule.

IV. Experimental Results

In our experiment, we test our system on real temporal images. There are six action datasets, which are done by six persons. The camera is set up at a fixed location and kept stationary. The camera has a frame rate of thirty frames per second and the image resolution is 640×480 pixels.

Each person performed all the six actions: "walking from left to right," "walking from right to left," "jumping," "crouching," "climbing up" and "climb down."; they are abbreviated as W_{LR}, W_{RL}, $\mathrm{J}_{\mathrm{UMP}}$, $\mathrm{C}_{\mathrm{ROUCH}}$, C_{UP}, and C_{DN}, respectively. The action "climbing up" is to climb up on the table from ground. The action "climbing down" is to climb down on ground from the table. Hence we have six model actions, each contains six actions above as their typical snapshots are shown in Fig. 2. Six lab members did there six actions at their pleasure. Besides, a video of pure background with no subject in the scene is adopted in our experiment and this is used as a background model. One video chosen randomly from the six action datasets is used for recognition

Fig. 2. Typical snapshots of six actions, one for each action, to be recognized.

and the other five are used for training, and this procedure is repeated in turn for six times.

In order to decrease the numbers of fuzzy set, we select templates to represent a video sequence. The postures of certain activities vary slightly between two image frames if their interval is fewer than five frames in video stream. Therefore, we selected one frame every fifth frame as the template image of posture, and on the other hand the interval is equal to one-sixth second in our experiment.

We chose six kinds of essential postures for "walking from right to left," "walking from left to right" and "climbing down," respectively; five for "climbing down," three for "crouching" and two for "jumping." There are totally 28 kinds of essential postures, and called 28 classes, as shown in Fig. 3. Each essential template is a cluster with five template images which are selected from five (training) persons. Images are resized until its height equals to 128 pixels or width equals to 96 pixels. Images of stand posture usually resize according to its height.

Training is accomplished in off-line setting. Therefore, we collected three consecutive 5:1 down-sample images from different start points to train fuzzy rules. A threshold of minimal support of rules should be set after all training patterns have been learned. The threshold is used to abandon the rules whose occurrence is too few. If some conflicting rules are generated, we choose the rule that is supported by a maximum number of training instances. For example, for the threshold chosen to be three, we have obtained 131 rules via training from action video sequences except for the first one person. As a result, these 131 rules will be used to test the action video sequence by the first person.

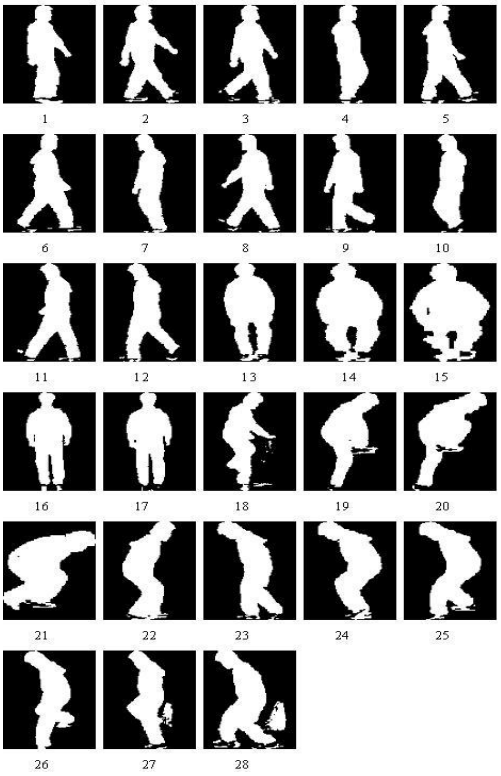


Fig. 3. 28 essential postures chosen for six action recognition.

The templates and the test patterns of fuzzy rules are both sampled with a rate of five image frames. An activity should appear in proper order directly perceived through our sense. For example, P_1 through P_6 are the six linguistic labels of the activity "walking from left to right." The activity of "walking from left to right" should contain the six rules with the posture sequence directly perceived through the following senses: (P_1, P_2, P_3), (P_2, P_3, P_4), (P_3, P_4, P_5), (P_4, P_5, P_6), (P_5, P_6, P_1), (P_6, P_1, P_2). We called these rules essential rules. There would be totally 24 essential rules for the six activities. But there are only 18 essential rules found in our experiment for threshold at three. The appeared essential rules are less than 24 because fuzzy rule base combines some similar rules to one rule. It is evident that the number of fuzzy rules is many more than the essential rules. This is because essential rules are based on the view of spatiotemporal space but our fuzzy rule base is generated from the view of canonical space. The constructed fuzzy rule base is able to learn the hidden and/or replaceable modes existent in these actions. We integrated the results of the same activity starting from different beginning image frame, and Table I shows the recognition rate of our system.

TABLE I
THE RECOGNITION RATE OF EACH ACTIVITY

Test data	Recognition rate (%)					
	W_{LR}	W_{RL}	C_{ROUCH}	J_{UMP}	C_{UP}	C_{DN}
Psn 1	100.0	92.3	71.0	78.4	78.1	94.6
Psn 2	100 .0	82.5	97.1	61.8	100.0	94.3
Psn 3	100.0	100.0	74.4	94.1	100.0	45.3
Psn 4	100.0	93.7	100.0	91.3	93.6	76.7
Psn 5	100.0	100.0	100.0	100.0	90.7	100.0
Psn 6	100.0	100.0	97.6	100.0	100.0	100.0
Avg	91.8					

A. Comparison Between Fuzzy Rule Base Approach and Hidden Markov Model Approach

Yamato and Ohya have proposed a human activity recognition system based on Hidden Markov Model [8]. Hidden Markov Model (HMM) is a kind of stochastic state sequential transit model and is possible to deal with time-sequential data. HMM approach is used to recognize human activity of our video datasets. In the experiment we adopted three for the number of states, and the number of key postures still was set to 28. The length of the observation sequence was set to three. The recognition rate comparison between HMM approach and fuzzy rule base approach is shown in Table II. The fuzzy rule base approach leads to a higher recognition rate. The fuzzy rule base approach improved recognition rate by 5.4%. Consequently, the fuzzy rule base approach has shown a better performance on human activity recognition in our experiment.

TABLE II
THE COMPARISON OF RECOGNITION RATE BETWEEN HMM APPROACH AND FUZZY RULE BASE APPROACH

	HMM	Fuzzy Rule Base
Psn 1	79.2	84.6
Psn 2	90.0	91.0
Psn 3	80.3	87.2
Psn 4	90.0	93.3
Psn 5	91.8	97.7
Psn 6	88.9	99.5
Avg	86.4	91.8

V. CONCUSION

In this paper, we present a fuzzy rule base approach in human activity recognition. In our approach, the illumination variation is decreased by adopting frame ratio method. CST and EST are used to reduce data dimensionality and optimize the class separability simultaneously. The frame sequences of video are then converted to one of 28 key frame postures. At last, fuzzy rule base for activity recognition is obtained by learning from three temporal postures. In the testing phase, a three posture sequences is processed by fuzzy rule base, and the recognition result is determined as the action which best matches the posture sequence in the fuzzy rules. Furthermore, fuzzy rule base is able to learn the hidden mode of the training data and is tolerant to variation of activities done by different people.

Experiment results have shown that the recognition rate for six activity classification is 91.8% without referring any geographic information such as location, path and velocity of the moving object. In comparison with HMM approach, our approach can provide a better recognition rate by about 5.4%.

To investigate further, we will try a large scale experiment and further refine feature extraction. In addition, recognition from a different viewing direction, extension of test environment and more complicated activities are our future work.

ACKNOWLEDGMENT

This research was supported in part by the National Science Council under grants NSC 94-2213-E-009-097, NSC95-2752-E-009-011-PAE, and 95-EC-17-A-02-S1-032 Taiwan, R.O.C.

REFERENCES

[1] R. Poppe, "Vision-based human motion analysis: an overview," *Comput. Vision and Image Understan*., vol. 108, pp. 4–18, 2007.

[2] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Machine Intell*., vol. 23, no. 3, 2001.

[3] R. Hamid, Y. Huang, and I. Essa, "ARGMode–Activity recognition using graphical models", in *Proc. Conf. Comput. Vision Pattern Recog*., vol. 4, pp. 38–45, Madison, Wisconsin, 2003.

[4] S. Carlsson and J. Sullivan, "Action recognition by shape matching to key frames,"in *Proc. IEEE Comput. Soc. Workshop Models versus Exemplars in Comput. Vision*, pp. 263–270, Miami, Florida, 2002.

[5] R. Cucchiara, C. Grana, A. Prati, and R. Vezzani, "Probabilistic posture classification for human-behavior analysis," *IEEE Trans. Syst. Man, and Cybern. A*, vol. 35, no. 1, pp. 42–54, 2005.

[6] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: Real-time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Machine Intell*., vol. 22, no. 8, pp. 809–830, August 2000.

[7] P. S. Huang, C. J. Harris, and M. S. Nixon, "Canonical space representation for recognizing humans by gait or face," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation*, pp. 180–185, 1998.

[8] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *Proc. IEEE CVPR*, pp. 379–385, 1992.

[9] F. Niu and M. Abdel-Mottaleb, "View-invariant human activity recognition based on shape and motion features," in *Proc. IEEE Sixth Int. Symposium Multimedia Softw. Eng*., pp. 546–556, 2004.

[10] L. X. Wang and J. M. Mendel, "Generating fuzzy rules by learning from examples," *IEEE Trans. Syst., Man Cybern*., vol. 22, no. 6, pp. 1414–1427, 1992.

[11] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *J. Opt. Soc. Am. A*, Vol. 14, pp. 1724–1733, 1997.