

國立交通大學

電機與控制工程研究所

碩士論文

根據法則於人類正常與異常動作辨識

Rule Based Human Normal and Abnormal Activity Recognition



研究生：陳冠廷

指導教授：張志永

中華民國九十七年六月

根據法則於人類正常與異常動作辨識

Rule Based Human Normal and Abnormal Activity Recognition

學 生：陳冠廷

Student : Kuan-Ting Chen

指導教授：張志永

Advisor : Jyh-Yeong Chang

國立交通大學

電機與控制工程學系



Submitted to Department of Electrical and Control Engineering

College of Electrical Engineering and Computer Science

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of Master in

Electrical and Control Engineering

June 2008

Hsinchu, Taiwan, Republic of China

中華民國九十七年六月

Rule Based Human Normal and Abnormal Activity Recognition

STUDENT: Kuan-Ting Chen

ADVISOR: Dr. Jyh-Yeong Chang

Institute of Electrical and Control Engineering
National Chiao-Tung University

ABSTRACT

Human activity recognition plays an essential role in applications such as automatic surveillance systems, human-machine interface, home care system and smart home applications. It is insufficient that a human activity recognition system uses only the posture of an image frame to classify an activity. On the other hand, transitional relationships of postures embedded in the temporal sequence are important information for human activity recognition.

In the thesis, we combine template posture matching and fuzzy rule reasoning to recognize an action. Firstly, a foreground subject is extracted and converted to a binary image by a statistical background model based on frame ratio. The binary image is then transformed 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 the variation of action done by different people. During the training of image sequences, we can compute the mean and standard deviation of each pre-defined activity. These numbers can be employed to determine whether an input image belongs to one of the pre-defined actions or an unknown

action. Lastly, we will also use Unsupervised Clustering Algorithm to generate some key postures for unknown activities. Our action recognition system not only can recognize an pre-defined action but also can signal an unknown action, which enhance the capability and recognition accuracy of activity recognition.



ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. Jyh-Yeong Chang for valuable suggestions, guidance, support and inspiration he provided. Without his advice, it is impossible to complete this research. Thanks are also given to all of my lab members for their suggestion and discussion. Finally, I would like to express my deepest gratitude to my family for their concern, supports and encouragements.



Content

摘要	i
ABSTRACT	ii
ACKNOWLEDGEMENTS	iv
Content	v
List of Figures	vii
List of Tables	ix
Chapter 1 Introduction	1
1.1 Motivation of this research	1
1.2 Foreground Subject Extraction	4
1.3 Eigenspace and Canonical Space Transformation	4
1.4 Image Frame Classification and Activity Recognition	5
1.5 Thesis Outline	7
Chapter 2 Basic Concept	8
2.1 Fundamentals of Eigenspace and Canonical Space Transform	8
2.1.1 Eigenspace Transformation (EST)	9
2.1.2 Canonical Space Transformation (CST)	11
2.2 Unsupervised Clustering Algorithm.....	13

Chapter 3 Human Activity Recognition System	16
3.1 Object Extraction	16
3.1.1 Background Modeling by Frame Ratio	16
3.1.2 Extraction of Foreground Object	18
3.2 Activity Template Selection	21
3.3 Construction of Fuzzy Rules from Video Streams	23
3.4 Classification Algorithm	27
Chapter 4 Experimental Results	30
4.1 Background Model and Object Extraction	31
4.2 Fuzzy Rule Construction for Action Recognition	34
4.3 The Recognition Rate of Activities Using Fuzzy Rule Base Approach	42
4.4 Extract the New Key Postures	44
Chapter 5 Conclusion	48
References.....	49



List of Figures

Fig. 1.1.	The block diagram of human activity recognition system	3
Fig. 3.1.	Histogram of binary image projection in X and Y direction	20
Fig. 3.2.	The binary image of extracted foreground region	20
Fig. 3.3.	One image frame is selected as template with an interval	21
Fig. 3.4.	Common states of two different activities	23
Fig. 3.5.	The structure of the human activity recognition algorithm	29
Fig. 4.1.	The environment of the classroom	30
Fig. 4.2.	An example of foreground region extraction at different threshold, k , values. (a) An image frame, (b) $k=1.0$, (c) $k=1.1$, (d) $k=1.2$, (e) $k=1.3$, (f) $k=1.4$	32
Fig. 4.3.	An example of foreground region extraction. (a) An image frame, (b) binary image after background analysis, (c) project of (b) onto X direction, (d) projection of (b) onto Y direction, (e) foreground region extracted....	33
Fig. 4.4.	Some “essential templates of posture” of person 1.....	35
Fig. 4.5.	Some “essential templates of posture” of person 5.....	36
Fig. 4.6.	Two examples of fuzzy rules. (a) Walking from left to right, (b) Climbing down.....	41
Fig. 4.7.	“The new key postures,” of person 1.....	45

Fig. 4.8. “The new key postures,” of person 5.....46

Fig. 4.9. “The final new key postures,” of person 1.....46

Fig. 4.10. “The final new key postures,” of person 5.....47



List of Tables

TABLE I.	THE RULE NUMBERS AT DIFFERENT THRESHOLD.....	38
TABLE II.	THE OBTAINED FUZZY RULE BASE GENERATED FROM THE TRAINING DATA EXCEPT PERSON 5.....	39
TABLE III.	THE RECOGNITION RATES OF KEY POSTURES SELECTED MANUALLY....	40
TABLE IV.	THE RECOGNITION RATES OF KEY POSTURES SELECTED BY UNSUPERV- ISED CLUSTERING ALGORITHM.....	40
TABLE V.	THE MEANS AND STANDARD DEVIATIONS OF SIX ACTIVITIES' MATCHING DEGREE OF TRAINING MODEL EXCEPT PERSON 5.....	41
TABLE VI.	THE RECOGNITION RATE OF PERSON 5 WITH DIFFERENT STARTING FRAME.....	43
TABLE VII.	THE RECOGNITION RATES OF EACH ACTIVITY.....	44
TABLE VIII.	THE DETECTION ACCURACY OF REAL-TIME NEW KEY POSTURES BELONGING TO THE UNKNOWN ACTION VIDEO.....	47
TABLE IX.	THE DETECTION ACCURACY OF FINAL NEW KEY POSTURES BELONGING TO THE UNKNOWN ACTION VIDEO.....	47

Chapter 1 Introduction

1.1 Motivation of this research

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. 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 [1], [2]. In [1], Bobick and Davis recognized the human activities by comparing motion-energy and motion-history of template images with temporal images. In [2], R. Hamid *et al.* extracted spatio-temporal features such as the relative distance between two hands and their velocities; furthermore they 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 [3], [4]. In [3], shape was represented by edge data obtained from canny edge detector, and key frames were defined for each activity. In [4], 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.

The system flowchart is illustrated in Fig 1.1. Our system can be separated into three components. The first component is foreground subject extraction. The second component is transformation of image data into a space that is smaller and easier for posture recognition. The third component is automatic key posture frame selection of the video frames for activity classification. Using this key frame selection technique, we can discriminate a new, namely unknown, activity from the previous defined activities. A certain amount of unknown activity detected will signal us to request update learning of the activity recognition rule base. Note that the proposed key posture selection scheme can also generate the new key postures for unknown activities, which accelerates the update learning of our activity recognition rule base.

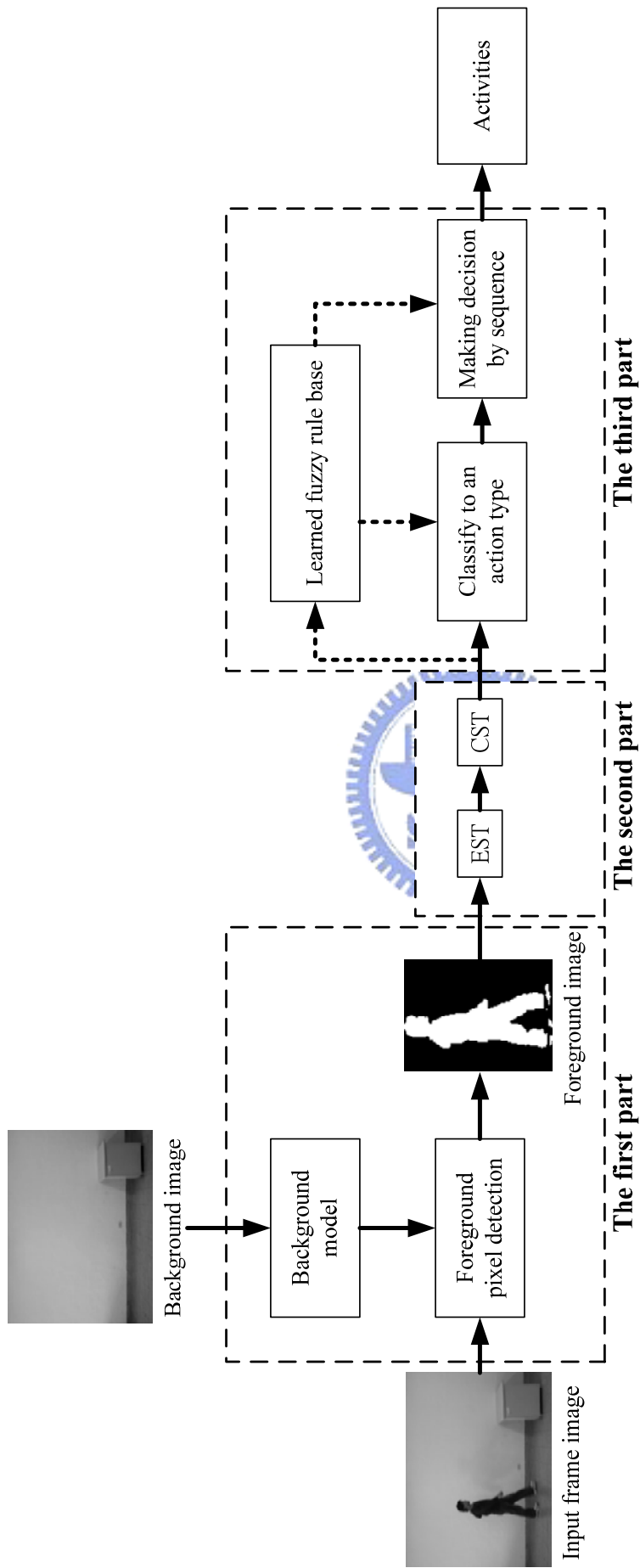


Fig. 1.1 The block diagram of human activity recognition system.

1.2 Foreground Subject Extraction

Background subtraction is widely used for detecting moving objects from image frames of static cameras. The rationale of this approach is to detect the moving objects by the difference between the current frame and a reference frame, often called the “background image,” or “background model.” A review is given in [5] where many different approaches were proposed in recent years. These approaches are all based on background subtraction. Basically, the background image is a representation of the scene with no moving objects; besides, background images are usually kept regularly update so as to adapt to the varying luminance conditions.

In order to solve the effect of varying luminance conditions, we develop a method which is robust to the illumination changes. The method use frame ratio rather than frame difference. After building a background model, we can extract foreground subject from video frames by subtracting each pixel value of background model from that of current image frame. The resulting image is converted to a binary one by setting a threshold. The binary image mainly contains foreground subject with only little noise. Therefore, we can set a threshold in the histogram of the binary image to extract a rectangle image, which is the most resemble shape of a person, of the target subject. The rectangle image is resized to the same measurements.

1.3 Eigenspace and Canonical Space Transformation

In most of video and image processing, the size of frame is usually very large and it usually exists some redundancy. The redundancy possesses little information of an image. Hence, some space transformations are introduced to reduce redundancy of

an image by reducing the data size of the image. The first step of redundancy reduction often transforms an image from spatiotemporal space to another data space. The transformation can use fewer dimensions to approximate the original image. There are many well-known transformation methods such as Fourier transformation, wavelet transformation, Principal Component Analysis and so on. Our transformation method combines eigenspace transformation and canonical space transformation which are described as follows.

Eigenspace transformation (EST), based on Principal Component Analysis, has been demonstrated to be a potent scheme used below: automatic face recognition proposed in [7], [8]; gait analysis proposed in [9]; and action recognition proposed in [10]. The subsequent transformation, Canonical space transformation (CST) based on Canonical Analysis, is used to reduce data dimensionality and to optimize the class separability and improve the classification performance. Unfortunately, CST approach needs high computation efforts when the image is large. Therefore, we combine EST and CST in order to improve the classification performance while reducing the dimension, and hence each image can be projected from a high-dimensional spatiotemporal space to a single point in a low-dimensional canonical space. In this new space the recognition of human activities becomes much simpler and easier.

1.4 Image Frame Classification and Activity Recognition

In this thesis, images are transformed into an image feature vector by extracting features from images. We utilize eigenspace and canonical space transformation method which is used to extract image features. We group three feature consecutive vectors from three contiguous images. Consequently, the time-sequential images are

converted to a posture sequence by using these three feature vectors. The posture sequence is dignified by the number of the templates. In the learning phase, we build a transition model in terms of three consecutive posture sequences which is the category symbol of the posture template. For human action recognition, the model which best matches the observed posture sequence is chosen as the recognized action category.

After transforming image frames to eigenspace and canonical space domain, some data information have been omitted. By using fuzzy rule-base techniques, the activity analysis task is tolerant to uncertainty, ambiguity and irregularity. Relevant articles using the fuzzy theory are described as follows. Wang and Mendel proposed that fuzzy rules to be generated by learning from examples in [11]. Su [12] presented a fuzzy rule-based approach to spatio-temporal hand gesture recognition. This approach employs a powerful method based on hyperrectangular composite neural networks (HRCNNs) for selecting templates. Ushida and Imura [13] introduced a real-time human-motion recognition method by means of Fuzzy Associative Memories Organizing Units System.

In our system, we propose a fuzzy rule-base approach for human activity recognition. Training data of each activity is represented in the form of crisp IF-THEN rules that is extracted from the posture sequences of the training data. Each crisp IF-THEN rule is then fuzzified by employing an innovative membership functions in order to represent the degree indicating the similarity between a pattern and the corresponding antecedent part in the training data. When an unknown activity is to be classified, each sample of the unknown activity is tested by each fuzzy rule. The accumulated similarity measure associated with three consecutive samples of the input image frame is to match the posture sequence representing an activity model of the training database, and the unknown activity is classified to the activity yielding the

highest accumulative similarity.

In classifying an action in the video frames, we compute the smallest dissimilarity membership degree among the IF-THEN rules. In the training phase, we compute the rule's mean and standard deviation of each pre-defined activity. If the dissimilarity of the testing video frames, which is greater than the pre-defined activity to be performing a new action. Then, these three images frames will be input to the unsupervised clustering algorithm for possible new key posture selection.

1.5 Thesis Outline

The thesis is organized as follows. The theory of eigenspace transform, canonical space transform and unsupervised clustering algorithm is firstly introduced in Chapter 2. In Chapter 3, we describe our human activity recognition system in detail. In this chapter, we also use 5 : 1 down sampled video frames to build a fuzzy rule database for activity recognition. We collect three consecutive images as a feature vector. Then by training from the known data, we can extract transitional rules of templates for activity recognition. The fuzzy rules play an important role in our activity recognition system. Likewise, we also use action's dissimilarity to the learned rule bases to find new activity. Last, we use unsupervised clustering algorithm to obtain the new key postures. In Chapter 4, the experiment results of our recognition system are shown. At last, we conclude this thesis with a discussion in Chapter 5.

Chapter 2 Basic Concept

In this chapter, we briefly explain the basic concepts of eigenspace and canonical space transform. Then unsupervised clustering algorithm concept is introduced.

2.1 Fundamentals of Eigenspace and Canonical Space Transform

In video and image processing, the dimensions of image data are often extremely large. There are many well-known transformation methods to reduce the size of data such as Fourier transformation, wavelet, principal component analysis (PCA), eigenspace transformation (EST) and so on. However, PCA based on the global covariance matrix of the full set of image data is not sensitive to the class structure existent in the data. In order to increase the discriminatory power of various activity features, Etemad and Chellappa [14] used linear discriminant analysis (LDA), also called canonical analysis (CA), 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. Here we call this approach canonical space transformation (CST). Combining EST with CST, our approach reduces the data dimensionality and optimizes the class separability among classes.

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

Assume that there are c classes to be learned. Each class represents a specific posture, which assumes of testers various forms existing in the training image data. $\mathbf{x}'_{i,j}$ is the j -th image in class i , and N_i is the number of images in the i -th class. The total number of images in training set is $N_T = N_1 + N_2 + \dots + N_c$. This training set can be written as

$$[\mathbf{x}'_{1,1}, \dots, \mathbf{x}'_{1,N_1}, \dots, \mathbf{x}'_{2,1}, \dots, \mathbf{x}'_{c,N_c}], \quad (1)$$

where each $\mathbf{x}'_{i,j}$ is an image with n pixels.

At first, the intensity of each sample image is normalized by

$$\mathbf{x}_{i,j} = \frac{\mathbf{x}'_{i,j}}{\|\mathbf{x}'_{i,j}\|}. \quad (2)$$

Then we can get the mean pixel value for training image as

$$\mathbf{m}_x = \frac{1}{N_T} \sum_{i=1}^c \sum_{j=1}^{N_i} \mathbf{x}_{i,j}. \quad (3)$$

The training set can be rewritten as a $n \times N_T$ matrix \mathbf{X} . And each image $\mathbf{x}_{i,j}$ forms a column of \mathbf{X} , that is

$$\mathbf{X} = [\mathbf{x}_{1,1} - \mathbf{m}_x, \dots, \mathbf{x}_{1,N_1} - \mathbf{m}_x, \dots, \mathbf{x}_{c,N_c} - \mathbf{m}_x]. \quad (4)$$

2.1.1 Eigenspace Transformation (EST)

Basically EST is widely used to reduce the dimensionality of an input space by mapping the data from a correlated high-dimensional space to an uncorrelated low-dimensional space while maintaining the minimum mean-square error to avoid

information loss. EST uses the eigenvalues and eigenvectors generated by the data covariance matrix to rotate the original data coordinates along the direction of maximum variance.

If the rank of the matrix \mathbf{XX}^T is K , then K nonzero eigenvalues of \mathbf{XX}^T , $\lambda_1, \lambda_2, \dots, \lambda_K$, and their associated eigenvectors, $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_K$, satisfy the fundamental relationship

$$\lambda_i \mathbf{e}_i = \mathbf{R} \mathbf{e}_i, \quad i = 1, 2, \dots, K \quad (5)$$

where $\mathbf{R} = \mathbf{XX}^T$ and \mathbf{R} is a square, symmetric matrix. In order to solve Eq. (5), we need to calculate the eigenvalues and eigenvectors of the $n \times n$ matrix \mathbf{XX}^T . But the dimensionality of \mathbf{XX}^T is the image size, it is too large to be computed easily. Based on singular value decomposition theory, we can get the eigenvalues and eigenvectors by computing the matrix $\tilde{\mathbf{R}}$ instead, that is

$$\tilde{\mathbf{R}} = \mathbf{X}^T \mathbf{X}, \quad \mathbf{X} : \text{data matrix} \quad (6)$$

in which the matrix size of $\tilde{\mathbf{R}}$ is $N_T \times N_T$ which is much smaller than $n \times n$ of \mathbf{R} . Assume that the matrix $\tilde{\mathbf{R}}$ has K nonzero eigenvalues $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_K$ and K associated eigenvectors $\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2, \dots, \tilde{\mathbf{e}}_K$ which are related to those in \mathbf{R} by

$$\begin{cases} \lambda_i = \tilde{\lambda}_i \\ \mathbf{e}_i = (\tilde{\lambda}_i)^{-\frac{1}{2}} \mathbf{X} \tilde{\mathbf{e}}_i \end{cases} \quad i = 1, 2, \dots, K \quad (7)$$

These K eigenvectors are used as an orthogonal basis to span a new vector space. Each image can be projected to a point in this K -dimensional space. Based on the theory of PCA, each image can be approximated by taking only the $k \leq K$ largest

eigenvalues $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_k|$ and their associated eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k$.

This partial set of k eigenvectors spans an eigenspace in which $\mathbf{y}_{i,j}$ are the points that are the projections of the original images $\mathbf{x}_{i,j}$ by the equation

$$\mathbf{y}_{i,j} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k]^T \mathbf{x}_{i,j}, \quad i=1, 2, \dots, c; j=1, 2, \dots, N_c \quad (8)$$

We called this matrix $[\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k]^T$ the eigenspace transformation matrix. After this transformation, each original image $\mathbf{x}_{i,j}$ can be approximated by the linear combination of these k eigenvectors and $\mathbf{y}_{i,j}$ is a one-dimensional vector with k elements which are their associated coefficients.

2.1.2 Canonical Space Transformation (CST)

Based on canonical analysis in [15], we suppose that $\{\phi_1, \phi_2, \dots, \phi_c\}$ represents the classes of transformed vectors by eigenspace transformation and $\mathbf{y}_{i,j}$ is the j -th vector in class i . The mean vector of entire set can be written as

$$\mathbf{m}_y = \frac{1}{N_T} \sum_i \sum_j \mathbf{y}_{i,j}, \quad i=1, 2, \dots, c; j=1, 2, \dots, N_i \quad (9)$$

The mean vector of the i -th class can be presented by

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{\mathbf{y}_{i,j} \in \Phi_i} \mathbf{y}_{i,j}. \quad (10)$$

Let \mathbf{S}_w denote the within-class matrix and \mathbf{S}_b denote the between-class matrix, then

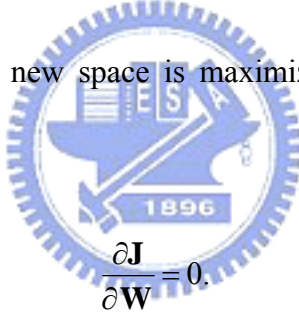
$$\mathbf{S}_w = \frac{1}{N_T} \sum_{i=1}^c \sum_{\mathbf{y}_{i,j} \in \phi_i} (\mathbf{y}_{i,j} - \mathbf{m}_i)(\mathbf{y}_{i,j} - \mathbf{m}_i)^T$$

$$\mathbf{S}_b = \frac{1}{N_T} \sum_{i=1}^c N_i (\mathbf{m}_i - \mathbf{m}_y)(\mathbf{m}_i - \mathbf{m}_y)^T$$

where \mathbf{S}_w represents the mean of within-class vectors distance and \mathbf{S}_b represents the mean of between-class distance vectors distance. The objective is to minimize \mathbf{S}_w and maximize \mathbf{S}_b simultaneously, which is known as the generalized Fisher linear discriminant function and is given by

$$\mathbf{J}(\mathbf{W}) = \frac{\mathbf{W}^T \mathbf{S}_b \mathbf{W}}{\mathbf{W}^T \mathbf{S}_w \mathbf{W}}. \quad (11)$$

The ratio of variances in the new space is maximized by the selection of feature transformation \mathbf{W} if



$$\frac{\partial \mathbf{J}}{\partial \mathbf{W}} = 0. \quad (12)$$

Suppose that \mathbf{W}^* is the optimal solution where the column vector \mathbf{w}_i^* is a generated eigenvector corresponding to the i -th largest eigenvalues λ_i . According to the theory presented in [15], we can solve Eq. (12) as follows

$$\mathbf{S}_b \mathbf{w}_i^* = \lambda_i \mathbf{S}_w \mathbf{w}_i^*. \quad (13)$$

After solving Eq. (11), we will obtain $c-1$ nonzero eigenvalues and their corresponding eigenvectors $[\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_c]$ that create another orthogonal basis and span a $(c-1)$ -dimensional canonical space. By using these bases, each point in eigenspace can be projected to another point in canonical space by

$$\mathbf{z}_{i,j} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{c-1}]^T \mathbf{y}_{i,j}, \quad (14)$$

where $\mathbf{z}_{i,j}$ represents the new point and the orthogonal basis $[\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{c-1}]^T$ is called the canonical space transformation matrix. By merging equation (8) and (14), each image can be projected into a point in the new $(c-1)$ -dimensional space by

$$\mathbf{z}_{i,j} = \mathbf{H} \cdot \mathbf{x}_{i,j}, \quad (15)$$

in which $\mathbf{H} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{c-1}]^T [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k]^T$.

2.2 Unsupervised Clustering Algorithm

To further automatize in key frame selection of action recognition system, we will propose an automatic clustering scheme for this purpose. Clustering is a powerful technique used in various disciplines such as pattern recognition [16], speech analysis [17], and information retrieval [18], etc. In [19], an unsupervised clustering based approach was introduced to determine key frames within a shot boundary. In this section, we introduce a new clustering approach to key frames extraction in video sequences [20].

Given a video shot $s = \{f_1, f_2, \dots, f_N\}$ obtained from a shot foreground extraction algorithm [6], we cluster the N frames into M clusters of center similarity, say $\sigma_1, \sigma_2, \dots, \sigma_M$. The similarity of two frames is defined as the similarity of their visual content, where the visual content could be color, texture, shape of the salient object of the frame, or the combination of the above. In this thesis, we select the binary values of our foreground detected frames as our visual content, although

other visual contents are readily integratable into the algorithm. The size of each frame we used is 128×96, raster-scanned to become 1×12288 vector. The similarity between frames i and j is thus defined as :

$$\left\| \sum_{y=1}^{12288} B_i(1, y) - B_j(1, y) \right\|_2 \quad (16)$$

Any clustering algorithm has a threshold parameter δ to determine whether a new frame will be classified into a certain cluster which in turns control the density of clustering. The similarities between the new frame node and the centroid of the existing cluster must be computed first. If the smallest similarity is less than δ , it means this node is close enough to be a member representing the cluster; otherwise, this node is not close enough to be a member representing the cluster. So we need to create a new cluster to represent the current image frame. The unsupervised clustering algorithm can be summarized as follows :



1. Initialization : $\sigma_1 \leftarrow f_1$, $f_1 \rightarrow$ the centroid of σ_1 (denoted as c_{σ_1}), $1 \rightarrow NumClu$;
2. Get the next frame f_i . If the frame pool is empty, Goto 6;
3. Calculate the similarities between f_i and existing clusters $\sigma_k (k = 1, 2, \dots, NumClu)$: $Sim(f_i, \sigma_k)$, based on Eq. 16;
4. Determine which cluster is the closest to f_i by clustering $MinSim$. Let

$$MinSim = \min_{k=0}^{NumClu} Sim(f_i, \sigma_k).$$

If $MinSim > \delta$, it means that f_i is not close enough to be put in any of the clusters, goto 5; otherwise, put f_i into the cluster which has $MinSim$, and Goto 6.

5. $NumClu = NumClu + 1$. A new cluster is formed : $f_i \rightarrow \sigma_{NumClu}$.

6. End the clustering algorithm.

It is natural that a big enough cluster would be qualified to be a representative of a dataset. In this thesis we say a clustering is big enough if its size is bigger than N/M (N : training data numbers; M : clustering numbers), the average size of clusters. With above constraint in mind, we use Eq. (17) to compute the representative key frames which can represent its cluster.

$$\text{Representative keyframes} = \min_i \left(\sum_{j \in \text{CluMem}} \|M_i - M_j\|_2 \right), \quad (17)$$

where $i=1, 2, \dots, n$; $j=1, 2, \dots, n$. Number n is the number of cluster members.



Chapter 3 Human Activity Recognition System

The first step of human activity recognition system is foreground subject extraction. We have to construct a background model for subject 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 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 to be adaptive to varying environment. In our approach, we describe the background scene as a statistical model. We obtain a background model from background only video by calculating the maximum, minimum gray level and frame ratio of each pixel in the images.

3.1 Object Extraction

3.1.1 Background Modeling by Frame Ratio

We assume the image captured by a camera can be described as

$$I_i(x, y) = S_i(x, y)r_i(x, y), \quad (18)$$

where I_i is the intensity of the scene, S_i is the spatial distribution of source

illumination, r_i is the distribution of scene reflectance, (x, y) is the location of a pixel in the image and i is the image sequence index. If the camera is fixed stationary and moving objects are not permitted to show up in the scene, the reflectance of the background may remain the same at any time. That is,

$$r_i(x, y) = r(x, y). \quad (19)$$

Although the reflectance is not changed, the effect of illumination is still going on. The frame ratio between two consecutive frames can respectively be written as

$$\begin{aligned} \log\left(\frac{I_i(x, y)}{I_{i-1}(x, y)}\right) &= \log\left(\frac{S_i(x, y)r(x, y)}{S_{i-1}(x, y)r(x, y)}\right) \\ &= \log\left(\frac{S_i(x, y)}{S_{i-1}(x, y)}\right) \\ &= \log(S_i(x, y)) - \log(S_{i-1}(x, y)), \end{aligned} \quad (20)$$

where I is the intensity of captured images, S is the spatial distribution of source illumination.

We propose to utilize the frame ratio to build the background model. Each pixel of background scene is characterized by three statistics: minimum intensity value $n(x, y)$, maximum intensity value $m(x, y)$ and maximum inter-frame ratio $d(x, y)$ of a background video. Because these three values are statistical, we need a background video, without any moving objects, for background model training. Let I be an image frame sequence and contains N consecutive images. $I_i(x, y)$ be the intensity of a pixel which is located at (x, y) in the i -th frame of I . The background model, $[m(x, y), n(x, y), d(x, y)]$, of a pixel is obtained by

$$\begin{bmatrix} m(x, y) \\ n(x, y) \\ d(x, y) \end{bmatrix} = \begin{cases} \begin{bmatrix} \max_i \{I_i(x, y)\} \\ \min_i \{I_i(x, y)\} \\ \max_i \{I_i(x, y)/I_{i-1}(x, y)\} \end{bmatrix} & \text{if } I_i(x, y)/I_{i-1}(x, y) \geq 1 \\ \begin{bmatrix} \max_i \{I_i(x, y)\} \\ \min_i \{I_i(x, y)\} \\ \max_i \{I_{i-1}(x, y)/I_i(x, y)\} \end{bmatrix} & \text{otherwise} \end{cases} \quad (21)$$

$$i = 1, 2, \dots, N.$$

3.1.2 Extraction of Foreground Object

Foreground objects can be segmented from every frame of the video stream. Each pixel of the video frame is classified to either a background or a foreground pixel by the difference between the background model and a captured image frame. We utilize the maximum intensity $m(x, y)$, minimum intensity $n(x, y)$ and maximum inter-frame ratio $d(x, y)$ of the training background model to segment a foreground by

$$B(x, y) = \begin{cases} 0, & \text{a background pixel} & \text{if } \begin{cases} I_i(x, y)/m(x, y) < kd(x, y) \\ \text{or} \\ I_i(x, y)/n(x, y) < kd(x, y) \end{cases} \\ 255, & \text{a foreground pixel} & \text{otherwise.} \end{cases} \quad (22)$$

where $I_i(x, y)$ be the intensity of a pixel which is located at (x, y) , $B(x, y)$ is the gray level of a pixel in a binary image and k is a threshold. Threshold k is determined by experiments according to difference environments. The value of k affects the

mount of information retained in binary image B .

According to binary image B , we extract the region of foreground object to minimize the image size. Foreground region extraction can be accomplished by simply introducing a threshold on the histograms in X and Y direction. Fig. 3.1 shows an example of foreground region extraction. We utilize the binary image and project it to X and Y directions. The interested section has higher counts in the histogram. We obtain the boundary coordinates x_1, x_2 of X axis and y_1, y_2 of Y axis from the projection histogram. We can use these boundary coordinates as the corner of a rectangle to extract foreground region. Fig. 3.2 is the extracted foreground region.



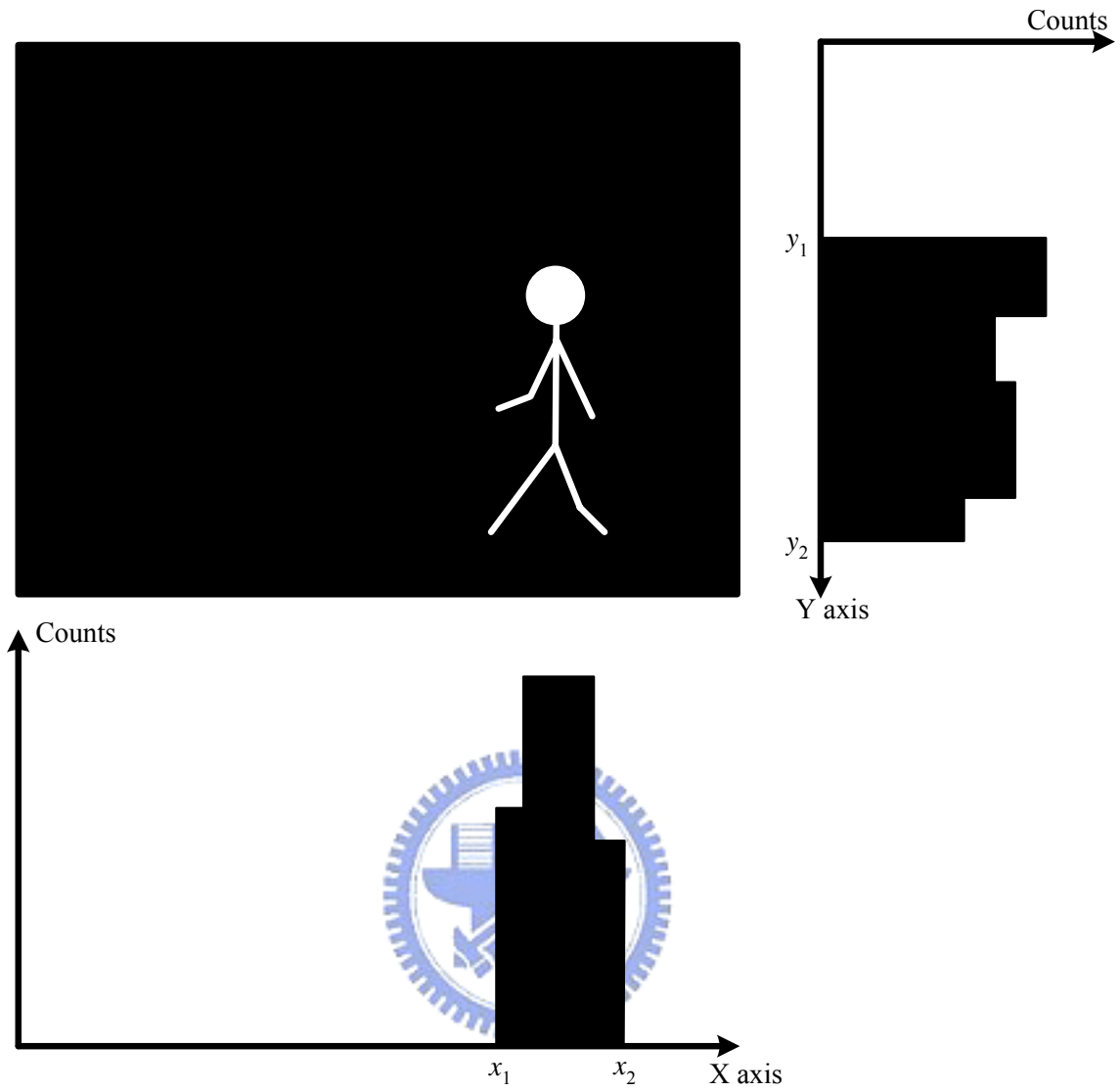


Fig. 3.1 Histogram of binary image projection in X and Y direction.



Fig. 3.2 The binary image of extracted foreground region.

3.2 Activity template selection

A human body is a rigid body, thus has its natural frequency. It has restriction on action speed when doing some specific actions. Because cameras usually capture image frames in a high frequency, i.e., 30 frames /sec, there is little difference between two consecutive postural image frames in such a short interval. In the following, we develop an automatic key posture frame selection of the video frames for activity classification. Firstly, we need to determine how many key posture frames we want in this video sequence. These images frames will be inputted to the unsupervised clustering algorithm for possible key posture selection and the key frames usually should pass the mean number of constituent of member support for enough representative. Namely, we will delete the clusters whose constituent member number is too small. After the above clustering procedure, if the clustering number is not larger than the predefined threshold (threshold=28 the suitable key postures for six actions), we will delete clusters that are not representative enough; otherwise, we will find the first 28 cluster centers whose mutual distances are farthest in distance each other. After the clustering steps, we call them as the essential template image, or equivalently, key posture frame. An example is shown in Fig. 3.3. After the template selection, each activity is represented by several essential templates.

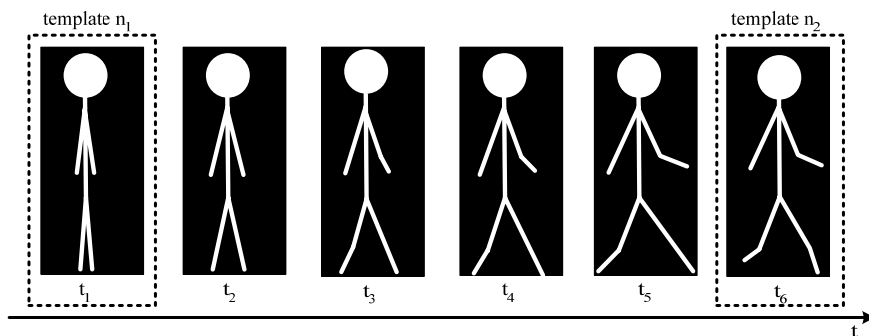
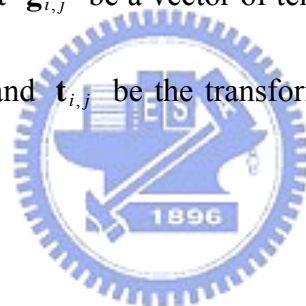


Fig. 3.3 One image frame is selected as template with an interval.

These essential templates are transformed to a new space by eigenspace transformation (EST) and canonical space transformation (CST). 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.

As described in Chapter 2, each image frame is transformed to a $(c-1)$ -dimensional vector by EST and CST methods. Assume that there are n training models and c clusters in the system. Therefore, we have N_t templates, where N_t is equal to n multiplied by c . Let $\mathbf{g}_{i,j}$ be a vector of template image of the j -th training model and the i -th category and $\mathbf{t}_{i,j}$ be the transformed vector of $\mathbf{g}_{i,j}$. Vector $\mathbf{t}_{i,j}$ is computed by



$$\mathbf{t}_{i,j} = \mathbf{H} \cdot \mathbf{g}_{i,j}, \quad i = 1, 2, \dots, c; \quad j = 1, 2, \dots, n \quad (23)$$

where \mathbf{H} denote the transformation matrix combing EST and CST and n is the total number of posture images in the i -th cluster. $\mathbf{t}_{i,j}$ is a $(c-1)$ -dimensional vector and each dimension is supposed to be independent. Hence, vector $\mathbf{t}_{i,j}$ is rewritten as

$$\mathbf{t}_{i,j} = [t_{i,j}^1, t_{i,j}^2, \dots, t_{i,j}^{c-1}]^T. \quad (24)$$

3.3 Construction of Fuzzy Rules from Video Streams

Transitional relationships of postures in a 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 as shown in Fig. 3.4. Besides, the posture sequence of each activity is dissimilar in different people.

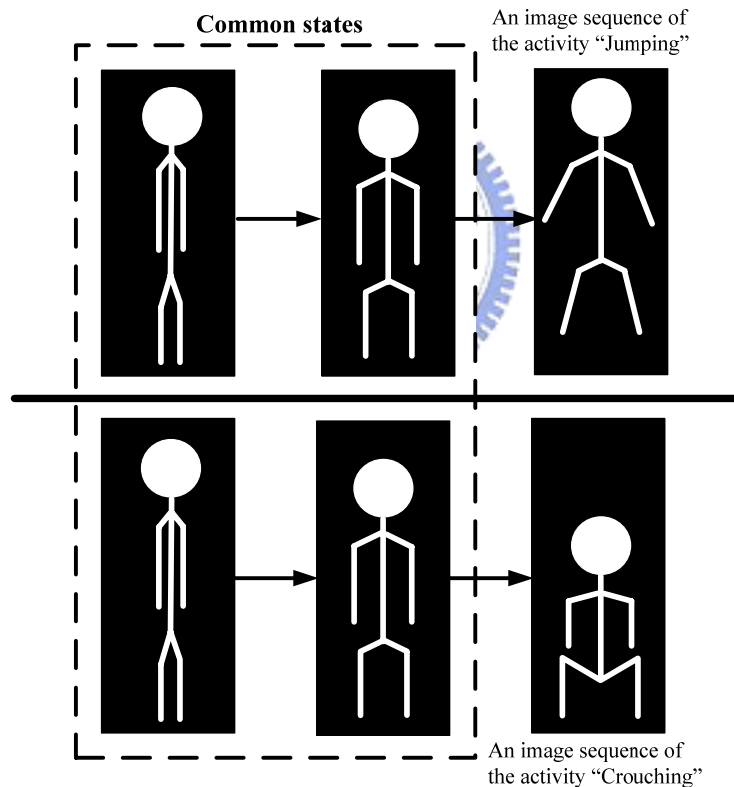


Fig. 3.4 Common states of two different activities.

Hence, we propose a method which not only combines temporal sequence information for recognition but also is tolerant to variations of different people. We

use the fuzzy rule-base approach to design our system. The fuzzy rule-base approach also has been proposed in gesture recognition in [12]; it is capable of absorbing the data difference by learning.

We use the Euclidean distance between the image frames of CST transformed feature vector to indicate the chance to be belonging to the modeling clusters of key posture. Firstly, when the k -th training image frame \mathbf{x}_k is inputted, the feature vector \mathbf{a}_k is extracted by

$$\mathbf{a}_k = \mathbf{H} \cdot \mathbf{x}_k. \quad (25)$$

As same as $\mathbf{t}_{i,j}$ in Eq. (24), \mathbf{a}_k can be rewritten as

$$\mathbf{a}_k = [a_k^1, a_k^2, \dots, a_k^{c-1}]^T \quad (26)$$

If we assume the dimensions of the feature vectors are independent, a local measure of similarity between the input image frame and each modeling key postures can be computed. The class index as well as the distance between the input image frame and the best matched key posture frames can be computed as

$$\hat{i}_k = \arg \min_i \left(\min_j \left(\|a_k - t_{i,j}\|_2 \right) \right), \quad (27)$$

$$r_{\hat{i},k} = \min_i \left(\min_j \left(\|a_k - t_{i,j}\|_2 \right) \right), \quad (28)$$

where j is the training model number. Number $r_{\hat{i},k}$ denotes the distance between the k -th image frame and most resembled key frame posture \hat{i} .

The minimal distance computed above is the most likely posture of frame k .

Moreover, we collect three consecutive image frames to integrate the temporal information. If we use too many images to form a basis, we could contain too many images to be processed and reduce the throughput rate. If we use too few images, it may not have enough timing information to represent an activity.

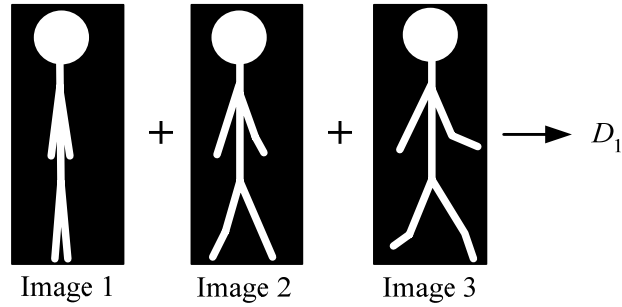
Using minimal distance criterion, we can determine the belonging key posture of the input image. Namely, each image frame can be represented by one of 28 key postures. To employ the temporal relation existing among the video frames, we combine three contiguous 5 : 1 down-sampled images to a group (I_1, I_2, I_3) . The transformation through CST of the image group will result in a feature vector $[a_1, a_2, a_3]$. Using Eqs. (27) and (28), we can classify a_i to P_i , $i = 1, 2$, and 3. Thus classify $[a_1, a_2, a_3]$ to $[P_1, P_2, P_3]$. There are c^3 combinations of the feature vector to form a clue for action identification. Each combination represents the possible transition states of the three images that compile part of an action. Naturally, an image sequence with $[P_1, P_2, P_3]$ sequence will be associated with its output of corresponding activity.

As developed by Wang and Mendel [11], fuzzy rules can be generated by learning from examples. Such image sequence constitutes an input-output pair to be learned in the fuzzy rule base. In this setting, the generated rules are a series of associations of the form

“IF antecedent conditions hold, **THEN** consequent conditions hold.”

The number of antecedent conditions equals the number of features. Note that antecedent conditions are connected by **“AND.”**

$$[P_1^1, P_2^1, P_3^1; D_1] \quad (29)$$

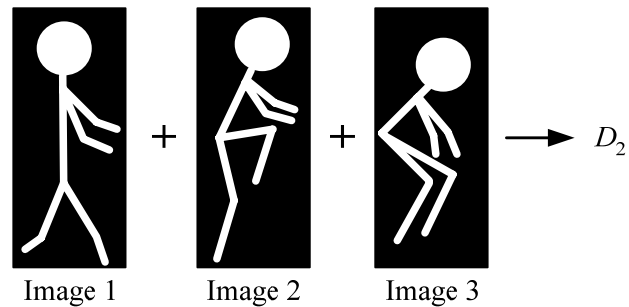


For example, suppose that Image 1, Image 2 and Image 3 belong to key posture 1, key posture 2 and key posture 3 respectively. Therefore, we assign the image sequences, whose CST transformed feature vector is $[a_1^1, a_2^1, a_3^1]$, to the key posture sequence Posture 1, Posture 2 and Posture 3 respectively; i.e. $[P_1, P_2, P_3]$. Finally, according to the feature-target association implies this image sequence to support the rule of

Rule 1. IF the activity's I_1 is P_1^1 AND its I_2 is P_2^1 AND its I_3 is P_3^1 ,
THEN the activity is D_1 . (30)

Similarly, for the second

$$[P_1^2, P_2^2, P_3^2; D_2] \quad (31)$$



where \mathbf{a}_1^i , \mathbf{a}_2^i , and \mathbf{a}_3^i denote the identified key postures of Image 1, Image 2, and Image 3 of the activity, respectively, and D_1 is the corresponding belonging object category of the activity.

$[P_1^2, P_2^2, P_3^2; D_2]$ can imply the rule of

Rule 2. IF the activity's I_1 is P_1^2 AND its I_2 is P_2^2 AND its I_3 is P_3^2 ,
THEN the activity is D_2 . (32)

After the learning steps of action video, some rules that obtained enough member of supporting fire strength may be representative to describe an action in video. In this thesis, a rule with at least four supporting input image frames is selected and compiled to constitute the knowledge rule base of our action recognition system. During the training of image sequences, we can compute the mean and standard deviation of each pre-defined activity. In this thesis, we have six pre-defined activities, thus we can compute six activity group's means and standard deviations and use them to determine whether the input image belongs to one of the pre-defined actions or an unknown action. Lastly, we will also generate some extra key postures for unknown actions.

3.4 Classification Algorithm

After constructing the rule base, we can match the input image sequence to each fuzzy rule by calculating their matching degrees. The distance between the input image sequence and a rule base is the sum of distance between three corresponding key postures in Eq. (28). The consequence of the rule having smallest key posture

difference is identified to be the action of the current image sequence. However, if this posture difference sum falls outside 3.5 times standard deviations of this action type, then the current image frame is not similar enough to this action, and we will identify it to be an unknown action instead. Moreover, we can also generate the extra key postures for unknown action. Fig. 3.5 shows the structure of human activity recognition algorithm.



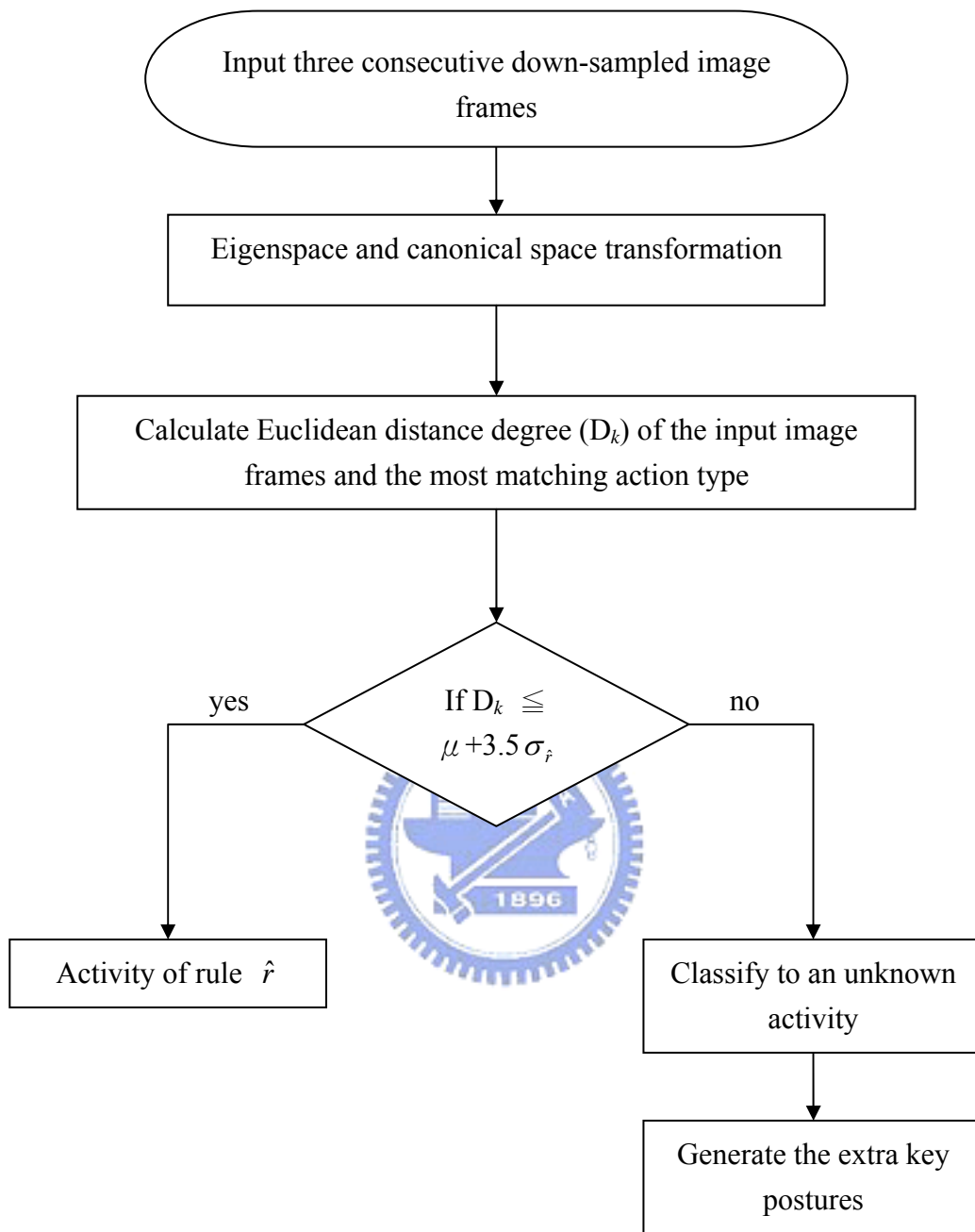


Fig. 3.5 The structure of the human activity recognition algorithm.

Chapter 4 Experimental Results

In our experiment, we test our system on video sequence containing a subject action. There are eight model actions, which are demonstrated by the research members of our Intelligence Technology Lab of the Department of Electrical and Control Engineering at National Chiao Tung University (NCTU). The video is taken in a classroom at the 5th Engineering Building in NCTU. The light source is fluorescent lamps and is stable. The background is not complex and we equip a table in the scene. 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 320×240 pixels. The environment of the classroom is shown in Fig. 4.1.

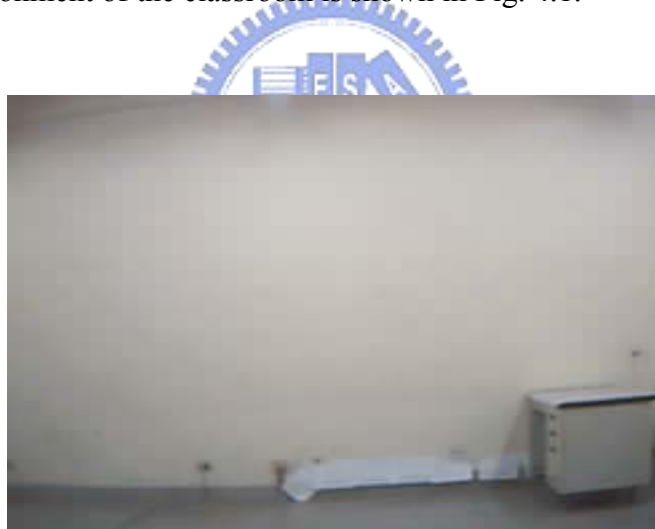


Fig. 4.1 The scene environment of our system.

Each member performed eight actions: “walking from left to right,” “walking from right to left,” “jumping,” “crouching,” “climbing up,” “climbing down,” “cavorting ” and “sitting.” The action “climbing up” is to climb up on the table from the ground. The action “climbing down” is to climb down to the ground from the table. Hence we have eight model actions as described above. First six actions are the

pre-defined, or called known, actions and the last two actions, “cavorting” and “sitting,” are not defined, or called unknown, actions. Five lab members did these eight 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 five model actions is used for testing, i.e., recognition and the rest four are used for training. The video of each subject model is used for testing in turn.

4.1 Background Model and Object Extraction

A background model is used for segmenting the foreground subject or object. If the background model is affected by the illumination change, there will be some noise or wrong segmented region left in the extracted image. From our experience, the frame ratio method can eliminate the influence of illumination variations.

A threshold k is applied in frame ratio approach to obtain binary image $B(x, y)$ in Eq. (22) described in Section 3.1.2. The value of k is chosen by experiment and varies with different environment. Hence, we ran a series of experiments to determine the optimal threshold k and the corresponded binary images are shown in Fig. 4.2. The threshold value $k = 1.3$ was adopted in our experiment.

Foreground object region is extracted from binary image $B(x, y)$ in order to minimize the size of images. Foreground region extraction is accomplished by simply taking a threshold along X and Y directions. Fig. 4.3 shows an example of foreground region extraction. Fig. 4.3(a) is a image frame of the video stream. Fig. 4.3(b) is the binary image after performing background model analysis. Figs. 4.3(c) and 4.3(d) show the projection of Fig. 4.3(b) onto the X and Y directions, respectively. We can

find the boundary coordinates of X and Y directions by observing the projection histogram. We used these boundary coordinates to define a rectangle to extract foreground region from Fig. 4.3(b). Fig. 4.3(e) is the extracted foreground region.

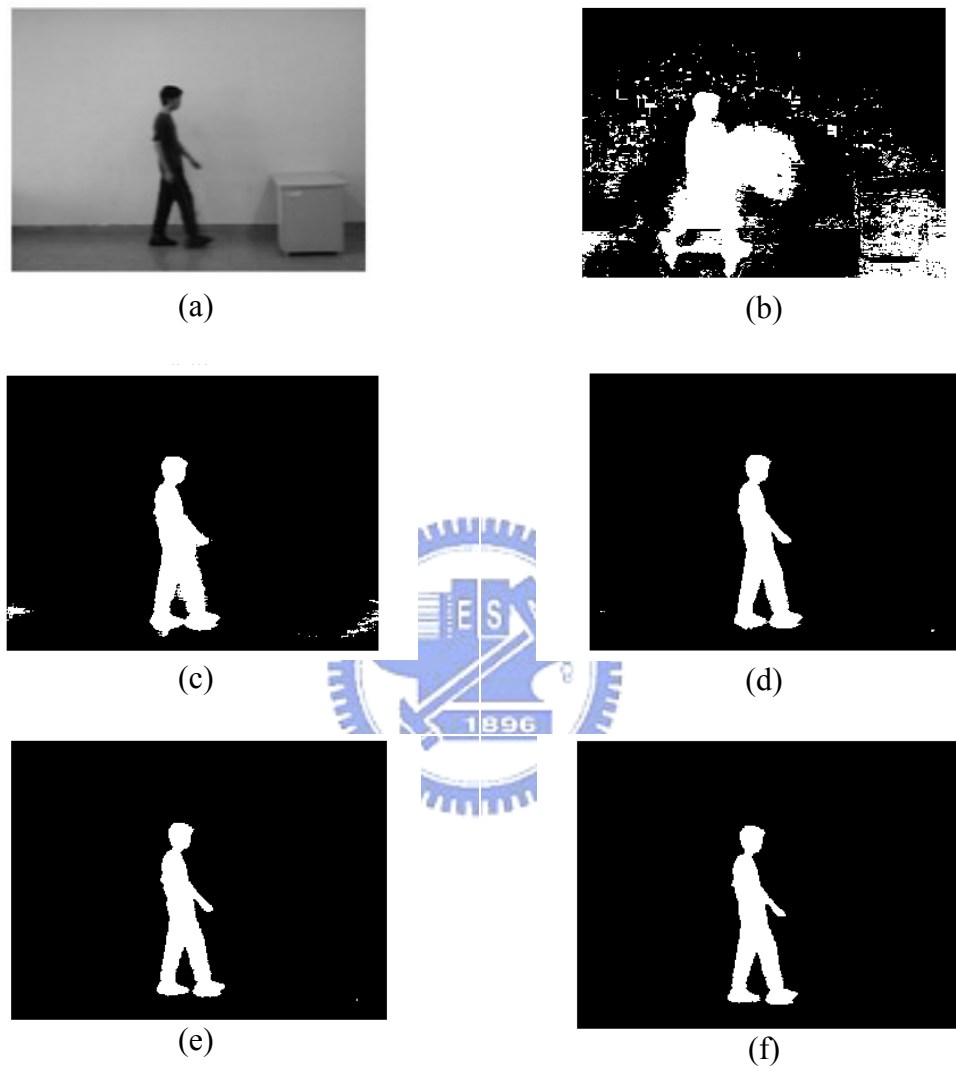


Fig. 4.2 An example of foreground region extraction at different threshold, k , values. (a) An image frame, (b) $k = 1.0$, (c) $k = 1.1$, (d) $k = 1.2$, (e) $k = 1.3$, (f) $k = 1.4$.

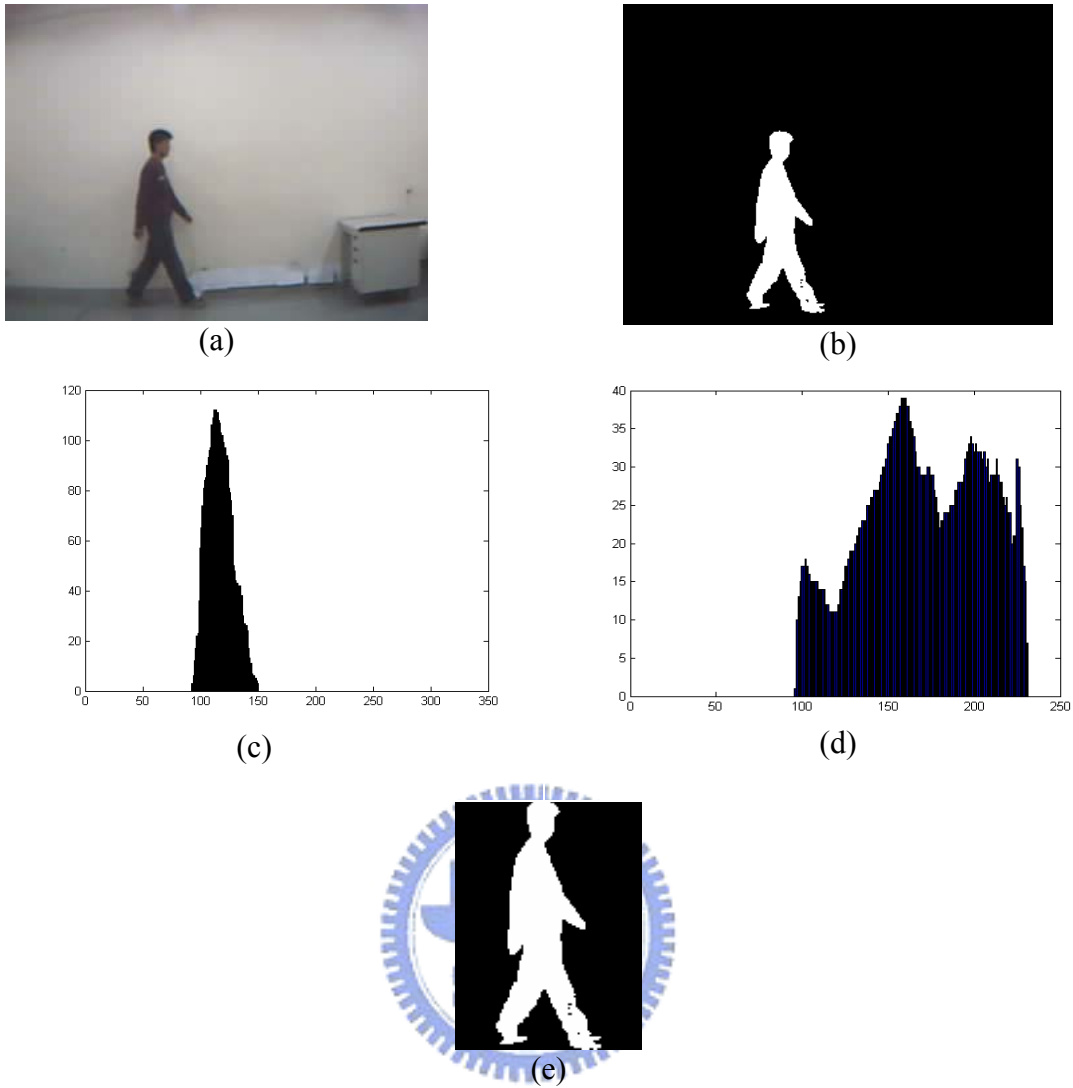


Fig. 4.3 An example of foreground region extraction. (a) An image frame, (b) binary image after background analysis, (c) projection of (b) onto X direction, (d) projection of (b) onto Y direction, (e) foreground region extracted.

4.2 Fuzzy Rule Construction for Action Recognition

For activity recognition, we use the key frame selection technique to automatically select essential templates of the video frames for activity clustering. We chose six essential templates 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 essential templates, and comprising 28 classes. The essential template numbers of each activity depend on how long the activity takes for a complete cycle. Each essential template is a representative cluster center around every five images, which are extracted from five different training persons and have similar postures. Figs. 4.4 and 4.5 are two examples of some templates of two training models.

As shown in Figs. 4.4 and 4.5, if a model bend down or squat down, the bodies in template images are wider than others. For normalization purpose, every segmented image is resized until its height equals to 128 pixels or width equals to 96 pixels. Images of stand posture usually resize to its height to be the ratio of 96 to 128 pixels. On the contrary, when the height of body shape is small, the magnifying factor of the image becomes large.



Class 2



Class 6



Class 7



Class 9



Class 15



Class 19



Class 21



Class 25



Class 27

Fig. 4.4 Some “essential templates of posture” of person 1.

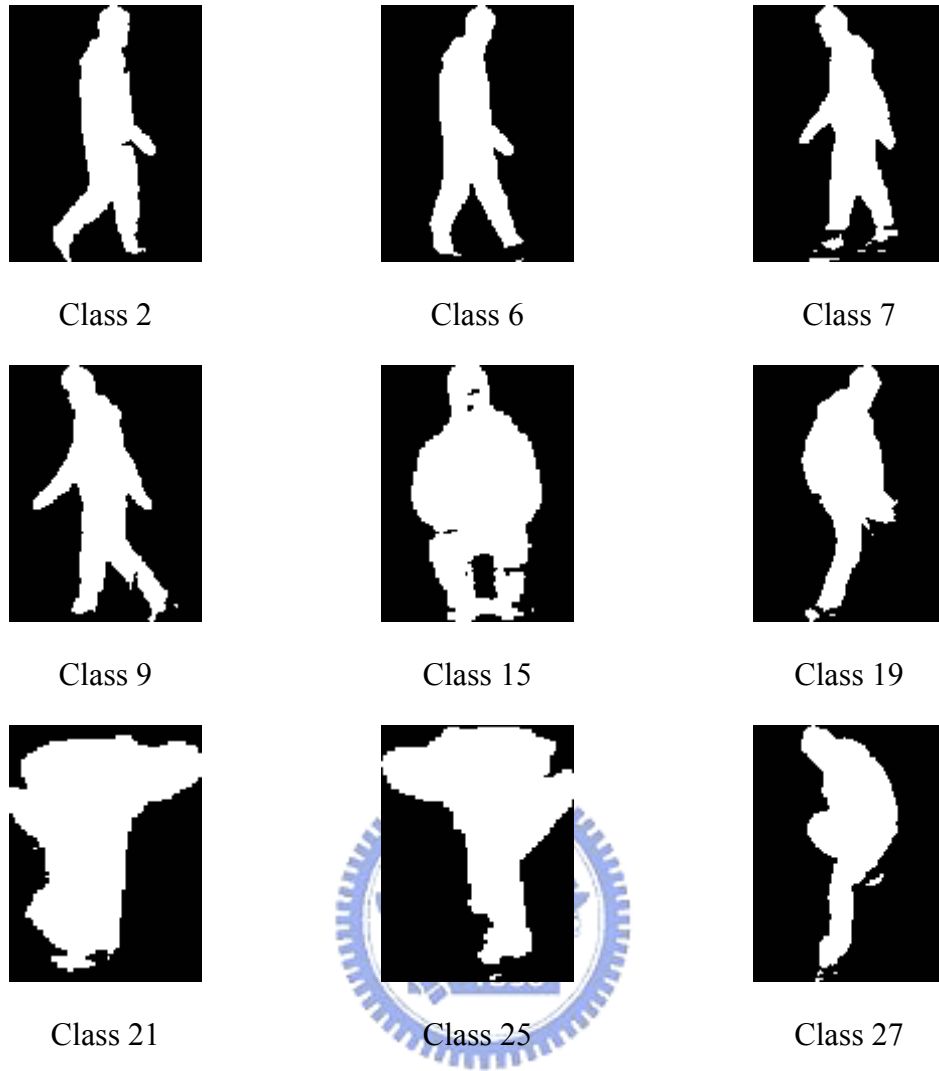


Fig. 4.5 Some “essential templates of posture” of person 5.

The template images are transformed to canonical space by the methods described in Chapter 2. Each essential template image of a training model was treated as a center. Hence, there were 112 center vectors because of four subject models and 28 class centers of essential templates of each subject models. Using leave-one-out strategy, there were five test subject models to be tested.

In the testing phase, the training video frames are inputted for activity recognition. The smallest essential template to each image frame is calculated by using Eq. (28) in Section 3.4. We gathered three consecutive 5 : 1 sub-sampled images

as a group in order to include temporal information. Training is accomplished in off-line manner. Therefore, we gathered three images from different start points to train for constructing fuzzy rules. For examples: the first frame, the 6-th frame and the 11-th frame are gathered together as an input training data; the second frame, the 7-th frame and 12-th frame are gathered together as another input training data; and the third frame, the 8-th frame and the 13-th frame are gathered together as an other input training data, etc. Different start points of image frames, as described above, are used for training fuzzy rules in our experiment, in corresponding to the starting testing video frame may not be the same, either. By utilizing different starting images, the system will be robust to be insignificant to the starting position of the video frames.

The group of the threes images is converted to the posture sequence which has the summation of the three minimal Euclidean distances to the essential templates by Eq. (28). Each posture sequence will support the corresponding rule once. If the corresponding rule is not existent, a new rule is generated in the **IF-THEN** form as represented in Section 3.4.

A threshold has to be set after all training patterns have been learned. The threshold is used to abandon the **IF-THEN** rule whose cumulative occurrence times are relative few. The numbers of rules being selected varies with different threshold selection. Table I shows the rule numbers versus different threshold values. Five model excluding one subject are chosen from the training data be utilized for rule construction. Obviously, the higher the threshold we choose, the fewer rules we will obtain. Although higher threshold can reduce rules, fewer rules will lose the tolerance for small difference observed even for the same activity. If some conflicting rules are generated, we choose the rule that is supported by a maximum number of training instances.

TABLE I

THE RULE NUMBERS AT DIFFERENT THRESHOLD

Training models except	Threshold = 3	Threshold = 4	Threshold = 5
Person 1	200	137	92
Person 2	190	138	104
Person 3	184	126	94
Person 4	185	128	91
Person 5	210	150	107

The test images are similarly 5 : 1 down-sampled of video frames. An activity should appear in a 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 have the rules with the posture sequence directly perceived through the 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) . With the threshold was set at four, a set of fuzzy rules generated from the training data except person 5 are listed in Table II. Two of the learned fuzzy rules of above are represented together with template images are shown in Fig. 4.6. After training all of image sequences, we can compute each the mean and standard deviation of pre-defined activity’s matching degree. In this thesis, we have six pre-defined activities, this we can compute these six activities means and standard deviations and use them to determine whether a input image is belonging to one of pre-defined activity or an unknown activity. In Section 3.2, we discuss the key postures selected by manually and unsupervised clustering algorithm. Table III is the

recognition rates of key postures selected manually. Table IV is the recognition rates of key postures selected by unsupervised clustering algorithm. Table V is the means and standard deviations of person 5.

TABLE II

THE OBTAINED FUZZY RULE BASE GENERATED FROM THE TRAINING DATA EXCEPT
PERSON 5

Number	Image 1	Image 2	Image 3	Class
1	P ₁	P ₁	P ₁	W _{LR}
2	P ₁	P ₁	P ₂	W _{LR}
3	P ₁	P ₁	P ₄	W _{LR}
⋮	⋮	⋮	⋮	⋮
50	P ₇	P ₁₁	P ₈	W _{RL}
⋮	⋮	⋮	⋮	⋮
90	P ₁₄	P ₁₃	P ₁₃	C _{ROUCH}
⋮	⋮	⋮	⋮	⋮
110	P ₁₇	P ₁₇	P ₁₇	J _{UMP}
⋮	⋮	⋮	⋮	⋮
120	P ₁₉	P ₂₀	P ₂₁	C _{UP}
⋮	⋮	⋮	⋮	⋮
148	P ₂₇	P ₂₇	P ₂₈	C _{DOWN}
149	P ₂₇	P ₂₈	P ₂₇	C _{DOWN}
150	P ₂₇	P ₂₈	P ₂₈	C _{DOWN}

TABLE III

THE RECOGNITION RATES OF KEY POSTURES SELECTED MANUALLY

Testing data	Recognition rate (%)					
	W_{LR}	W_{RL}	C_{ROUCH}	J_{UMP}	C_{UP}	C_{DOWN}
Person 1	100	100	100	100	100	62.26
Person 2	100	94.62	100	100	100	93.33
Person 3	99.11	100	100	100	100	72.61
Person 4	97.06	100	71.70	100	83.54	100
Person 5	100	100	100	100	82.61	100
Average	96.21					



TABLE IV

THE RECOGNITION RATES OF KEY POSTURES SELECTED BY UNSUPERVISED CLUSTERING ALGORITHM

Testing data	Recognition rate (%)					
	W_{LR}	W_{RL}	C_{ROUCH}	J_{UMP}	C_{UP}	C_{DOWN}
Person 1	100	95.74	100	100	100	83.02
Person 2	100	89.25	100	100	100	95.56
Person 3	99.11	100	100	100	87.16	32.88
Person 4	97.06	98.78	85.85	97.17	81.01	91.38
Person 5	100	89.66	100	100	89.86	100
Average	94.79					

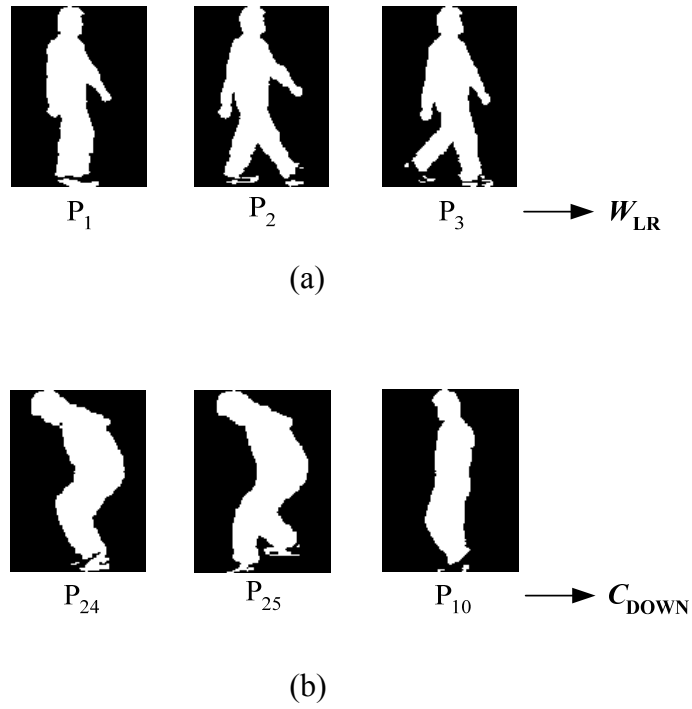


Fig. 4.6 Two examples of fuzzy rules. (a) Walking from left to right, (b) Climbing down.



THE MEANS AND STANDARD DEVIATIONS OF SIX ACTIVITIES' MATCHING DEGREE OF TRAINING MODEL EXCEPT PERSON 5

Activity	Mean	Standard deviation (σ)
W_{LR}	7060.04	1819.92
W_{RL}	7043.74	1870.61
C_{ROUCH}	5227.48	1332.18
J_{UMP}	3630.96	1168.53
C_{UP}	6282.87	2544.87
C_{DOWN}	7131.24	2804.25

4.3 The Activity Recognition Using Fuzzy Rule Base

Approach

The activity recognition system in our experiment is off-line presented and tested; therefore, the testing video is not done in real time phase. We input the testing video from different starting frames which is similar to the way for the training phase. Namely, we recognize the video from the first frame, the second frame, the third frame and the fourth frame, etc. with the down sampling intervals of five frames. The testing video was not used for constructing templates and fuzzy rules. Hence, there are five video databases for training and testing.

An example of recognition rate of a testing video start from different frames is shown in Table VI. In this table, W_{LR} is the activity “walking form left to right,” W_{RL} is the activity “walking from right to left,” J_{UMP} is the activity “jumping,” C_{ROUCH} is the activity “crouching,” C_{UP} is the activity “climbing up,” C_{DOWN} is the activity “climbing down,” C_{AVORT} is the activity “cavorting” and S_{IT} is the activity “sitting.” The threshold selected for this model is four and we also employ mean plus 3.5 standard deviations as matching degree a boundary to differentiate between the predefined and unknown activities.

TABLE VI

THE RECOGNITION RATE OF PERSON 5 WITH DIFFERENT STARTING FRAME

Starting frame	Recognition rate (%)							
	W_{LR}	W_{RL}	C_{ROUCH}	J_{UMP}	C_{UP}	C_{DOWN}	C_{AVORT}	S_{IT}
From the 1 st , 6 th , ... frame	100	94.44	100	100	85.71	100	64.71	89.66
From the 2 nd , 7 th , ... frame	100	83.33	100	100	78.57	100	52.94	87.93
From the 3 rd , 8 th , ... frame	100	88.24	100	100	78.57	100	47.06	91.23
From the 4 th , 9 th , ... frame	100	88.24	100	100	64.29	100	68.75	89.47
From the 5 th , 10 th , ... frame	100	94.12	100	100	61.54	100	87.50	85.96

Table VII shows the recognition rate of our system on the five testing subject models. The threshold used to construct fuzzy rules is four. For each activity, the recognition are obtained from different frame and then averaged.

TABLE VII
THE RECOGNITION RATES OF EACH ACTIVITY

Testing data	Recognition rate (%)							
	W_{LR}	W_{RL}	C_{ROUCH}	J_{UMP}	C_{UP}	C_{DOWN}	C_{AVORT}	S_{IT}
Person 1	100	95.74	74.34	100	100	83.02	17.54	96.65
Person 2	100	89.25	94.03	87.04	95.96	68.89	14.42	94.74
Person 3	78.57	97.25	100	100	83.49	32.88	84.56	100
Person 4	97.06	98.78	59.43	70.75	49.37	91.38	4.85	100
Person 5	100	89.66	100	100	73.91	100	63.86	88.85
Average	84.63							

4.4 Extract the New Key Postures



After the recognition scheme is complete, we can generate some unknown image frames from unknown activities of cavorting and sitting. We input these image frames not belonging to the essential templates to unsupervised clustering algorithm. Moreover, we can then generate the extra key postures for unknown action. After generate the key postures, we will compute the distances between the newly found key postures 112 pre-defined key postures. If this distance is greater than the threshold, $Th=8500$, then the key posture is not similar enough to the pre-defined key postures, and we will thus identify it to be the extra key posture instead. Figs. 4.7 and 4.8 are two obtained extra key postures of two testing models. Then, imposing at least having $Th=8500$ deviation from a pre-defined key postures on these key postures. Figs. 4.9 and 4.10 are obtained the final key postures of these two testing model. The

detection accuracies of real-time and final new key postures belonging to the unknown action video are summarized in Tables VIII and IX.

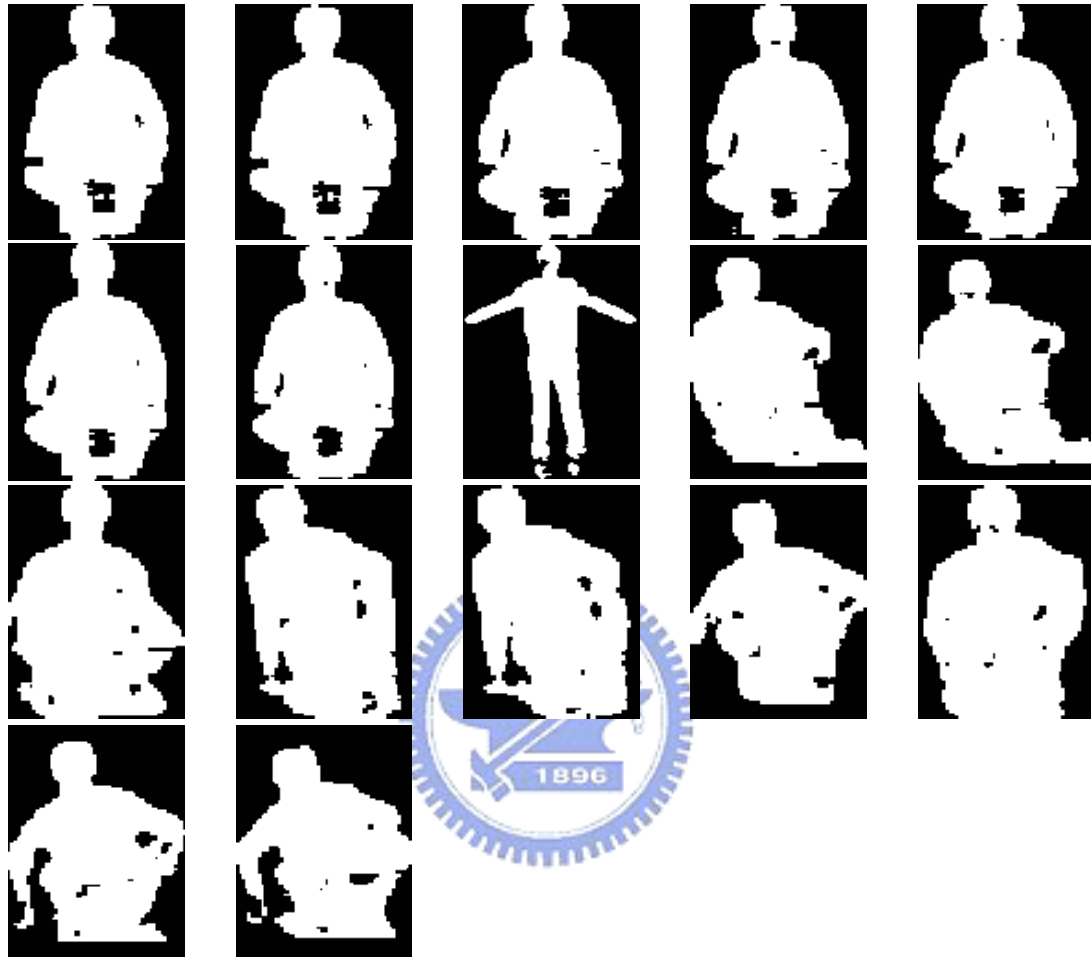


Fig. 4.7 “The new key postures,” of person 1.

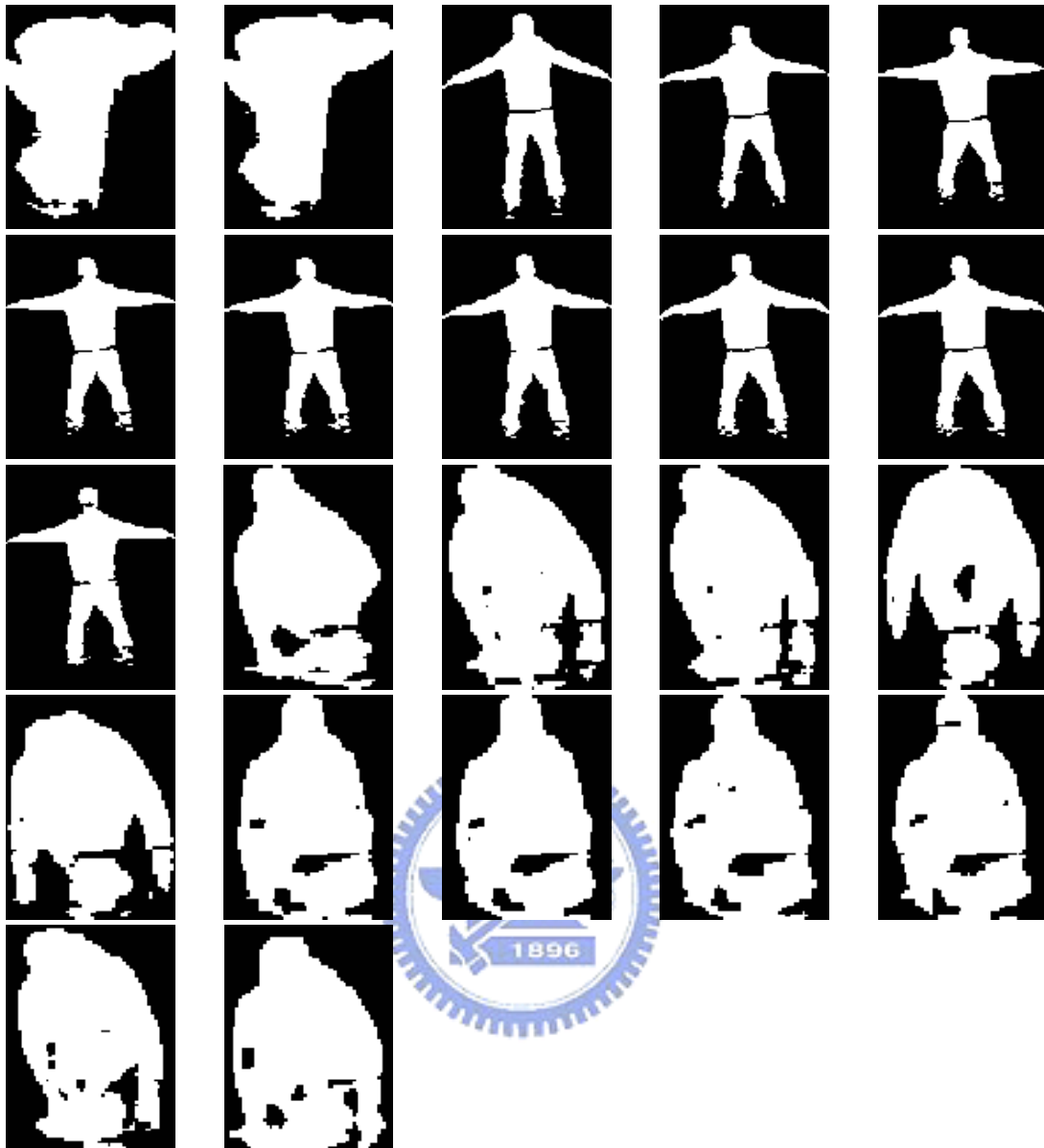


Fig. 4.8 “The new key postures,” of person 5.



Fig. 4.9 “The final new key postures,” of person 1.



Fig. 4.10 “The final new key postures,” of person 5.

TABLE VIII

THE DETECTION ACCURACY OF REAL-TIME NEW KEY POSTURES BELONGING TO THE
UNKNOWN ACTION VIDEO

	Person 1	Person 2	Person 3	Person 4	Person 5	Average
Key posture detection accuracy (%)	100.00	88.00	78.38	91.30	90.91	87.90

TABLE IX

THE DETECTION ACCURACY OF FINAL NEW KEY POSTURES BELONGING TO THE
UNKNOWN ACTION VIDEO

	Person 1	Person 2	Person 3	Person 4	Person 5	Average
Key posture detection accuracy (%)	100.00	94.12	78.05	88.45	94.75	89.56

Chapter 5 Conclusion

In this thesis, we have presented a fuzzy rule base approach to human activity recognition. In our approach, the effect of illumination variation is decreased by adopting frame ratio method. Moreover, CST and EST are used to reduce data dimensionality and optimize the class separability simultaneously, and each frame of video sequence is then converted to one of 28 key frame postures. At last, fuzzy rule base is reasoned for activity recognition. We also employ unsupervised clustering algorithm to on-line obtain the new key postures of unknown action.

Experiment results have shown that the recognition rate for eight activity classification is 84.63%. Besides, the detection accuracies of finding new key postures, that belonging to the unknown action, are 87.90% on-line, and 89.56% off-line, respectively.

To investigate further, we will further automatize the update learning of activity recognition rule base. In addition, recognition from a different viewing direction, extension of test environment, and more complicated activities are our future work.

References

- [1] 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.
- [2] 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.
- [3] 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.
- [4] I. Cohen and H. Li, "Inference of human postures by classification of 3D human body shape," in *Proc. IEEE Int. Workshop on Anal. Modeling of Faces and Gestures*, pp. 74–81, 2003.
- [5] M. Piccardi, "Background subtraction techniques: a review," in *Proc. IEEE Int. Conf. SMC.*, vol. 4, pp. 3099–3104, 2004.
- [6] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Real-time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 809–830, August 2000.
- [7] H. Saito, A. Watanabe, and S. Ozawa, "Face pose estimating system based on eigenspace analysis," in *Proc. Int. Conf. Image Processing*, vol. 1, pp. 638–642, 1999.
- [8] J. Wang, G. Yuantao, K. N. Plataniotis, and A. N. Venetsanopoulos, "Select eigenfaces for face recognition with one training sample per subject," *8th Cont., Automat. Robot. Vision Conf., ICARCV 2004*, vol. 1, pp. 391–396, 2004.

- [9] 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.
- [10] M. M. Rahman and S. Ishikawa, "Robust appearance-based human action recognition," in *Proc. the 17th Int. Conf. Pattern Recog.*, vol. 3, pp. 165–168, 2004.
- [11] 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.
- [12] M. C. Su, "A fuzzy rule-based approach to spatio-temporal hand gesture recognition," *IEEE Trans. Syst., Man Cybern.*, vol. 30, no. 2, pp. 276–281, 2000.
- [13] H. Ushida and A. Imura, "Human-motion recognition by means of fuzzy associative inference," in *Proc. Fuzzy Syst., 1994. IEEE World Congress Comput. Intell.*, vol. 2, pp. 813–818, 1994.
- [14] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," in *Proc. ICASSP*, pp. 2148–2151, 1997.
- [15] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd edition, 1300 Boylston Street Chestnut Hill, Massachusetts USA: Academic Press, 1990.
- [16] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, ch. 6. John Wiley and Sons, Inc.
- [17] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, ch. 5. Englewood Cliffs, New Jersey : Prentice Hall, 1993.
- [18] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1983.
- [19] A. M. Ferman and A. M. Tekalp, "Multiscale content extraction and representation for video indexing," in *Multimedia Storage and Archival Systems*, (Dallas, TX), Nov. 1997.

- [20] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, “Adaptive key frame extraction using unsupervised clustering, ” in *Proc. Conf. Image Processing.*, vol. 1, pp. 866-870, 1998.

