

國立交通大學

電機與控制工程研究所

碩士論文

結合時序姿態比對與模糊法則推論
於人類動作辨識

Combining Template Posture Matching and Fuzzy Rule
Inference for Human Activity Recognition

研究生：呂志濤

指導教授：張志永

中華民國九十五年七月

結合時序姿態比對與模糊法則推論

於人類動作辨識

Combining Temple Posture Matching and Fuzzy Rule

Inference for Human Activity Recognition

學 生：呂志濤

Student : Chih-Tao Lu

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



A Thesis

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

July 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年七月

結合時序姿態比對與模糊法則推論於人類動作 辨識

學生：呂志濤

指導教授：張志永博士

國立交通大學電機與控制工程研究所

摘要

人類動作辨識在自動監視系統、人機界面、居家安全照護系統和智慧型居家環境等方面的應用中佔有主要的地位。許多人類動作辨識系統僅僅利用單一張影像的姿式來辨別該動作。但是，在時間序列上，姿式狀態轉換的關係是用來辨別人類動作的重要資訊。

在此篇論文中，我們結合時序姿態比對與模糊法則的方法來完成人類動作的識別。首先，每一張影像的前景人物利用一個基於前後影像比值而建立之統計背景模型抽取出來，並將抽取出來影像轉換成二值化的影像格式；此方法可以減少照明對前景人物抽取的影響。為達到較精準與可分別度，二值化影像經由特徵空間及標準空間轉換，投影至標準空間。最後人類動作的識別在標準空間中完成。經由樣板比對的方法可將三張影像序列，此影像序列乃從動作視訊 5:1 減低抽樣獲得，轉換成轉變成一組時序姿態序列。接著，利用模糊法則的推論方法，將這組時序姿態序列分類為某一個動作類別。模糊法則，不僅能夠結合時間序列上的資訊，並且可以容忍不同人做相同動作上的差異。在我們的實驗中，我們提出的動作辨認方法比利用 HMM 的方法，辨識正確率約增加 5.4%，達到 91.8%。

Combining Temple Posture Matching and Fuzzy Rule Inference for Human Activity Recognition

STUDENT: Chih-Tao Lu

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. 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 the thesis, we combine temple 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, which is robust to illumination changes. For better efficiency and separability, 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 variation of action done by different people. In our experiment, the proposed activity recognition method has demonstrated higher recognition accuracy

of 91.87% than the HMM approach by about 5.4 %.



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, especially to Chien-Wen Cho. Finally, I would like to express my deepest gratitude to my family for their concern, support 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	2
1.3 Eigenspace and Canonical Space Transformation	3
1.4 Image Frame Classification and Activity Recognition	4
1.5 Thesis Outline	6
CHAPTER 2 FUNDAMENTALS OF EIGENSPACE AND CANONICAL SPACE TRANSFORMATION	8
2.1 Eigenspace Transformation (EST).....	9
2.2 Canonical Space Transformation (CST)	11
CHAPTER 3 HUMAN ACTIVITY RECOGNITION SYSTEM	14
3.1 Background Modeling by Frame Ratio.....	14
3.2 Extraction of Foreground Object	16
3.3 Activity template selection.....	19

3.4 Construction of Fuzzy Rules from Video Streams.....	20
3.5 Classification Algorithm	27
CHAPTER 4 EXPERIMENTAL RESULTS	30
4.1 Background Model and Object Extraction	31
4.2 Fuzzy Rule Construction for Action Recognition.....	36
4.3 The Recognition Rate of Activities Using Fuzzy Rule Base Approach.....	42
4.4 Comparison between Fuzzy Rule Base Approach and Nearest Neighbor Approach.....	45
4.5 Comparison between Fuzzy Rule Base Approach and Hidden Markov Model Approach.....	47
CHAPTER 5 CONCLUSION.....	49
REFERENCES.....	50



List of Figures

Fig. 3.1	Histogram of binary image projection in X and Y direction.	18
Fig. 3.2	The binary image of extracted foreground region.	18
Fig. 3.3	One image frame is selected as template with an interval.	19
Fig. 3.4	Common states of two different activities.	20
Fig. 3.5	The structure of classification algorithm.	29
Fig. 4.1	The environment of the classroom.	30
Fig. 4.2	The comparison between frame ratio and frame difference. (a) Background image, (b) image frame with a human, (c) frame difference, (d) frame ratio, (e) histogram of frame difference, (f) histogram of frame ratio, (g) foreground pixels of frame difference after simply taking a threshold, and (h) foreground pixels of frame ratio after simply taking a threshold.	32
Fig. 4.3	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) $k = 1.4$, (g) $k = 1.5$, and (h) $k = 1.6$	34
Fig. 4.4	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.	35
Fig. 4.5	Template selection with an interval of five frames.	36
Fig. 4.6	Some “essential templates of posture” of model A.	37

Fig. 4.7 Corresponding “essential templates of posture,” Fig. 4.6, of model B.38

Fig. 4.8 Two examples of fuzzy rules. (a) Walking from left to right (b) Climbing
down.....41

Fig. 4.9 The recognition rate of utilizing different thresholds.....44

Fig. 4.10 The structure of nearest neighbor approach.45



List of Tables

TABLE I. THE RULE NUMBERS AT DIFFERENT THRESHOLD	40
TABLE II. SOME OF THE OBTAINED FUZZY RULE BASE	41
TABLE III. THE RECOGNITION RATE OF PERSON 1 WITH DIFFERENT STARTING FRAME .	42
TABLE IV. THE FRAME NUMBER OF EACH ACTIVITY	43
TABLE V. THE RECOGNITION RATE OF EACH ACTIVITY	44
TABLE VI. THE COMPARISON OF RECOGNITION RATE BETWEEN NEAREST NEIGHBOR APPROACH AND FUZZY RULE BASE APPROACH	46
TABLE VII. THE COMPARISON OF RECOGNITION RATE BETWEEN HMM APPROACH AND FUZZY RULE BASE APPROACH	48



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. A detailed survey is introduced in 0. 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], 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 [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.

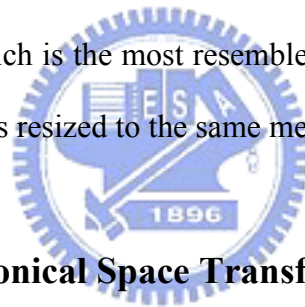
The system flowchart is illustrated in Fig 1.1. Our system can be separated into three parts. The first part is foreground subject extraction. The second part is transformation of image data in a space smaller and easier for posture recognition. The third part is posture classification of an image frame and activity recognition using frame sequences.

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 [6]

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 [8], [9]; gait analysis proposed in [10]; and action recognition proposed in [11]. 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.

The most famous method to model transition model of time-sequential data is Hidden Markov Models (HMMs). Hidden Markov Models can deal with

time-sequential data and can provide time-scale invariability for recognition. The basic concept of Hidden Markov Models is described in [12]. Hidden Markov Models have been successfully used for speech recognition because of their capability of recognizing spoken words independent of their duration [12]–[14]. Hidden Markov Models also have been used in hand gestures recognition [15], [16] and activity recognition [17], [18]. The price paid for the efficiency in this case is that we have to collect a great amount of data and a lot of time is required to estimate the corresponding parameters in HMMs.

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 [19]. Su [20] 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 [21] 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.

1.5 Thesis Outline

The thesis is organized as follows. Chapter 2 introduces the theory of eigenspace and canonical space transformation. In this chapter, we discuss the process of how to transform a large dimensional image to eigenspace and canonical space. Chapter 3 shows how we use timing information 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. In Chapter 4, we will introduce the technological processes of our recognition system. In Chapter 5, the experiment results of our recognition system are shown. At last, we conclude this thesis with a discussion in Chapter 6.

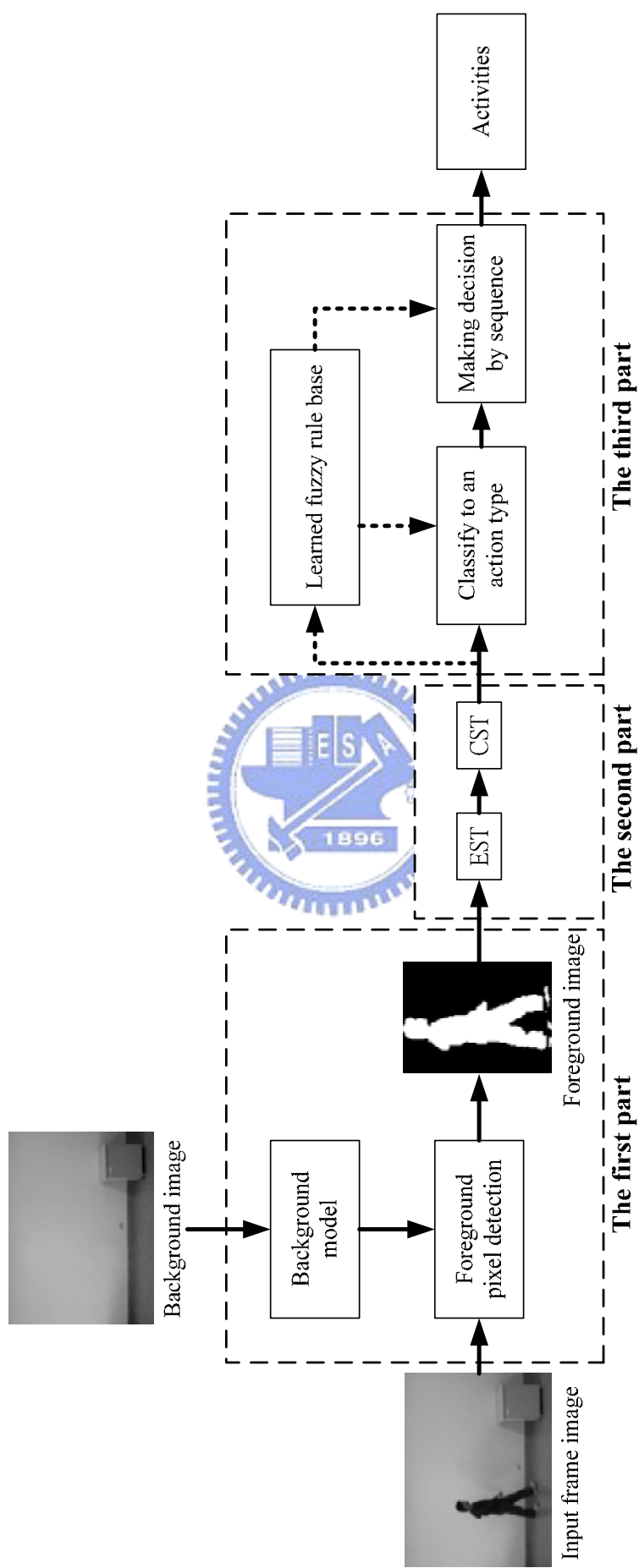


Fig. 1.1 The block diagram of human activity recognition system

Chapter 2 Fundamentals of Eigenspace and Canonical Space Transformation

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 [23] use 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. 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 (CTS). 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.

Assume that there are c classes to be learned. Each class represents a specific posture of 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 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 for 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 $\mathbf{X}\mathbf{X}^T$ is K , then K nonzero eigenvalues of $\mathbf{X}\mathbf{X}^T$,

$\lambda_1, \lambda_2, \dots, \lambda_K$, 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 $\mathbf{R} = \mathbf{X}\mathbf{X}^T$ and \mathbf{R} is a square, symmetric matrix. In order to solve (5), we need to calculate the eigenvalues and eigenvectors of the $n \times n$ matrix $\mathbf{X}\mathbf{X}^T$. But the dimensionality of $\mathbf{X}\mathbf{X}^T$ is the typical 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

$$\tilde{\mathbf{R}} = \mathbf{X}^T \mathbf{X} \quad (6)$$

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

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

where $i = 1, 2, \dots, K$.

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

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

where $i = 1, 2, \dots, c$ and $j = 1, 2, \dots, N_c$. 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.

2.2 Canonical Space Transformation (CST)

Based on canonical analysis in [22], 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

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

where $i = 1, 2, \dots, c$ and $j = 1, 2, \dots, N_i$. 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}_t denote the total scatter matrix, \mathbf{S}_w denote the within-class matrix and \mathbf{S}_b denote the between-class matrix, then

$$\begin{aligned}\mathbf{S}_t &= \frac{1}{N_T} \sum_{i=1}^c \sum_{j=1}^{N_i} (\mathbf{y}_{i,j} - \mathbf{m}_y)(\mathbf{y}_{i,j} - \mathbf{m}_y)^T \\ \mathbf{S}_w &= \frac{1}{N_T} \sum_{i=1}^c \sum_{\mathbf{y}_{i,j} \in \phi_i} (\mathbf{y}_{i,j} - \mathbf{m}_i)(\mathbf{y}_{i,j} - \mathbf{m}_i)^T \\ \mathbf{S}_b &= \frac{1}{N_T} \sum_{i=1}^c N_i (\mathbf{m}_i - \mathbf{m}_y)(\mathbf{m}_i - \mathbf{m}_y)^T\end{aligned}$$

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. That 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 \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 [22], we can solve (12) as follows

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

After solving (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

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



Chapter 3 Human Activity Recognition System

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 [7]. 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 be adaptive to varying environment. In our 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.

3.1 Background Modeling by Frame Ratio

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 propose a method utilizing frame ratio, instead of frame difference, which is proved robust to the illumination changes. Smooth changes of illumination are not

obvious, but longer duration of illumination changes are still affect object extraction.

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

$$I_i(x, y) = S_i(x, y)r_i(x, y) \quad (16)$$

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,

$$r_i(x, y) = r(x, y). \quad (17)$$

Although the reflectance is unchanged, the influence of illumination is still going on. If we shoot the background only and keep the camera stationary, the reflectance of the scene will be the same and there are only illumination changes. The problems caused by reflectance still remains if frame difference approach is adopted. Nevertheless, the influence of reflectance is eliminated in the frame ratio approach. The frame ratio between two consecutive frames is written as

$$\begin{aligned} \log\left(\frac{I_i(x, y)}{I_{i-1}(x, y)}\right) &= \log\left(\frac{S_i(x, y)r(x, y)}{S_{i-1}(x, y)r(x, y)}\right) \\ &= \log\left(\frac{S_i(x, y)}{S_{i-1}(x, y)}\right) \\ &= \log(S_i(x, y)) - \log(S_{i-1}(x, y)) \end{aligned} \quad (18)$$

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

$$\begin{bmatrix} m(x, y) \\ n(x, y) \\ d(x, y) \end{bmatrix} = \begin{cases} \begin{bmatrix} \max_i \{I_i(x, y)\} \\ \min_i \{I_i(x, y)\} \\ \max_i \{I_i(x, y)/I_{i-1}(x, y)\} \end{bmatrix} & \text{if } I_i(x, y)/I_{i-1}(x, y) \geq 1 \\ \begin{bmatrix} \max_i \{I_{i-1}(x, y)\} \\ \min_i \{I_i(x, y)\} \\ \max_i \{I_{i-1}(x, y)/I_i(x, y)\} \end{bmatrix} & \text{otherwise} \end{cases} \quad (19)$$

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

3.2 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

$$B(x, y) = \begin{cases} 0, & \text{a background pixel if } \begin{cases} I_i(x, y)/m(x, y) < kd(x, y) \\ \text{or} \\ I_i(x, y)/n(x, y) < kd(x, y) \end{cases} \\ 255, & \text{a foreground pixel otherwise.} \end{cases} \quad (20)$$

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. 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.

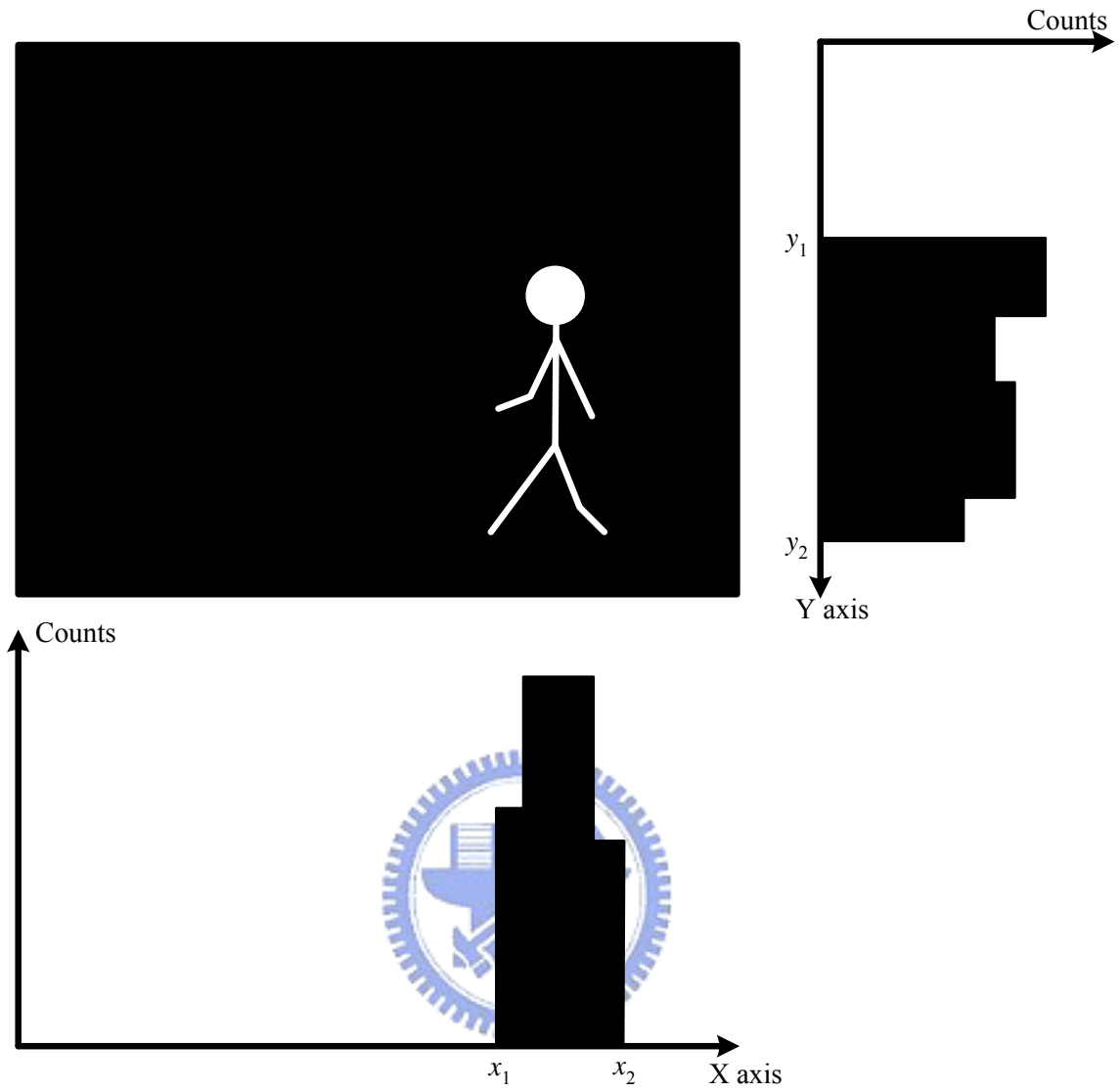


Fig. 3.1 Histogram of binary image projection in X and Y direction.



Fig. 3.2 The binary image of extracted foreground region.

3.3 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. An example is shown in Fig. 3.3. After determining the templates, each activity is represented by several essential templates.

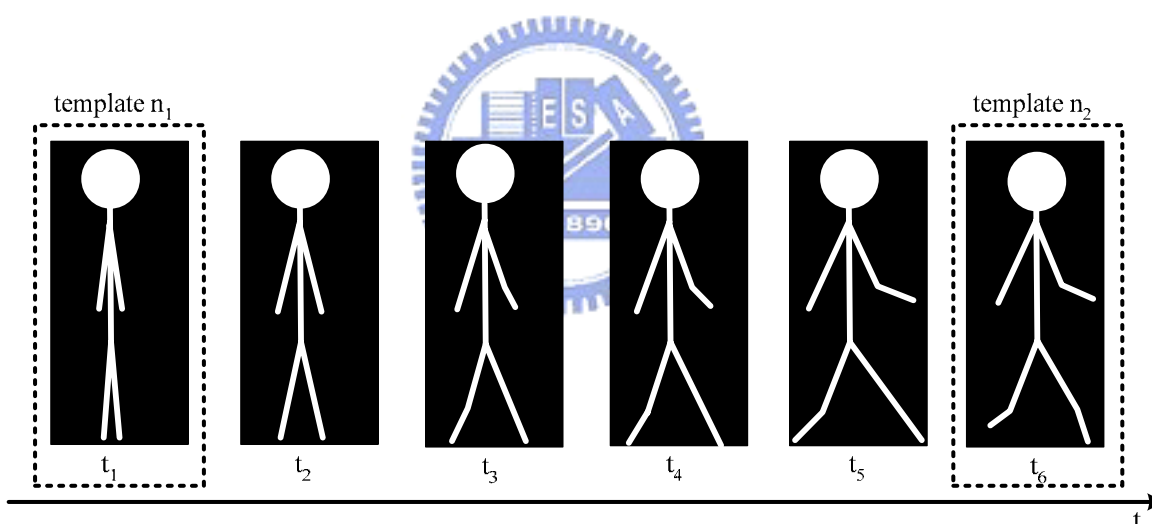


Fig. 3.3 One image frame is selected as template with an interval.

These essential templates are transformed to a new space by eigenspace transformation (EST) and canonical space transformation (CST). As described in Chapter 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 different people also barely converge to one point. Consequently, we select only essential templates rather than use all sequences for human activity recognition.

3.4 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 as shown in Fig. 3.4.

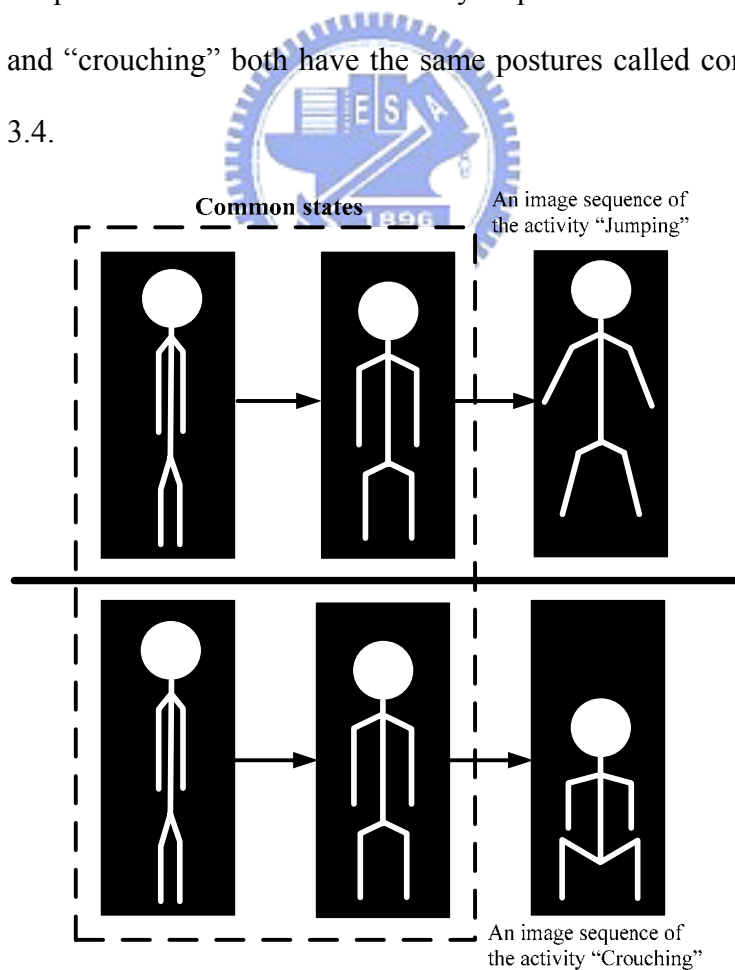


Fig. 3.4 Common states of two different activities

Human activities have lots of ambiguity, so we propose a method which can not only combines temporal sequence information for recognition but also is tolerant to variation of actions done by different people. Fuzzy rule base classification has the ability to absorb data difference by learning and has been successfully used in many applications for fusing results of two classifiers. In our system we view each transformation vectors of temporal images as a different feature. The fuzzy rule-base approach also has been proposed in gesture recognition in [20].

In our approach, EST and CST methods are used to extract features. 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 $\mathbf{g}_{i,j}$. $\mathbf{t}_{i,j}$ is computed by

$$\mathbf{t}_{i,j} = \mathbf{H} \times \mathbf{g}_{i,j}, \quad i=1, 2, \dots, c; \quad j=1, 2, \dots, n \quad (21)$$

where \mathbf{H} denote the transformation matrix combining 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, $\mathbf{t}_{i,j}$ is rewrite as

$$\mathbf{t}_{i,j} = [t_{i,j}^1, t_{i,j}^2, \dots, t_{i,j}^{c-1}]^T. \quad (22)$$

The transformation of each training model's templates is treated as a mean vector. That is,

$$\boldsymbol{\mu}_{i,j} = \mathbf{t}_{i,j} \quad (23)$$

where i is the number of template categories. The standard deviation vector of the m -th dimension is computed by

$$\sigma_m = \sqrt{\frac{\sum_{j=1}^c \sum_{i=1}^n (t_{i,j}^m - \mu_{i,j}^m)^2}{N_t - 1}} \quad (24)$$

where $m = 1, 2, \dots, c - 1$.

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

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

As the same as $\mathbf{t}_{i,j}$ in Eq. (22), \mathbf{a}_k can be rewritten as

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

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 $\boldsymbol{\Sigma}$ denote the covariance matrix of all essential template vectors and C_i denote the

i -th class of essential templates. The membership function is given by

$$\begin{aligned}
 r_{i,k} &= M(\mathbf{a}_k | C_i) \\
 &= \frac{1}{(2\pi)^{\frac{c-1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{a}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{a}_k - \boldsymbol{\mu})\right] \\
 &= \arg \max_j \left\{ \prod_{m=1}^{c-1} \frac{1}{\sqrt{2\pi}\sigma_m} \exp\left[-\frac{1}{2} \frac{(a_k^m - \mu_{i,j}^m)^2}{\sigma_m^2}\right] \right\} \quad (27)
 \end{aligned}$$

where m is the number of dimension and j is the training model number. $r_{i,k}$ denotes the grade of membership function in category i of the k -th image frame. Besides, we can obtain which category each image belongs to by

$$p_k = \arg \max_i r_{i,k} \quad (28)$$

The membership function describes the probability of which one it is like most. But it just contains the information of a temporal image. In order to include temporal information, we collect three images to form a basis for classification. If we use too many images to form a basis, the data may contain too many images of other activity. If we use too few images, it may not have enough timing information to represent an activity.

As developed by Wang and Mendel [19], 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 $[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$. An image sequence with feature vector $[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$ is associated with its output of corresponding activity. 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**.” For illustrative purpose, assume now we have c linguistic labels, each linguistic label represents a category of essential templates. Each image in the video stream can be represented by these c linguistic labels. Therefore, the feature vector $[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$ can be described by c^3 combinations of linguistic labels. Each one of the combinations represents the possible transition states of the three images. For example, an image sequence, its transformations of image 1, image 2, image 3 and belonging categories being concatenated as vector format, is given by

$$\begin{array}{c}
 \begin{array}{ccc}
 \begin{array}{|c|} \hline \text{Image 1} \\ \hline \end{array} & + & \begin{array}{|c|} \hline \text{Image 2} \\ \hline \end{array} & + & \begin{array}{|c|} \hline \text{Image 3} \\ \hline \end{array} & \longrightarrow & D_1 \\
 \end{array} \\
 \begin{array}{c}
 \mathbf{[a_1^1, a_2^1, a_3^1, D_1]} \\
 \end{array}
 \end{array} \tag{29}$$

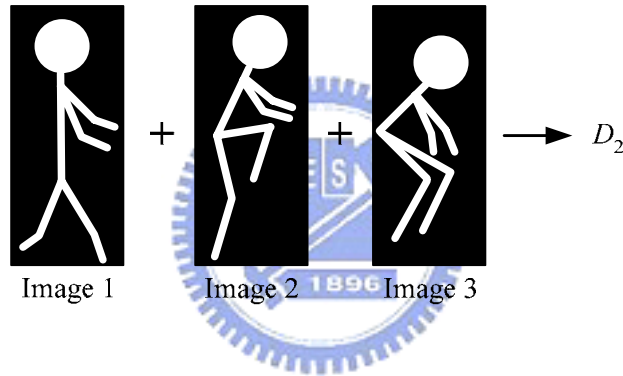
where \mathbf{a}_1^1 , \mathbf{a}_2^1 and \mathbf{a}_3^1 denote the transformation vectors of Image 1, Image 2 and Image 3 by EST and CST respectively, and D_1 is the corresponding belonging object category of the activity. Image 3 is the latest image captured by the camera of these three images. The class of each image is obtained by Eqs.(27) and (28). Suppose that Image 1, Image2 and Image 3 belong to category 1, category 2 and category 3 respectively. Therefore, we assign the image sequences, whose feature vector is $[\mathbf{a}_1^1, \mathbf{a}_2^1, \mathbf{a}_3^1]$, to the linguistic labels Posture 1, Posture 2 and Posture 3 respectively.

Finally, a rule is produced from the feature-target vector. Hence this image sequence supports the rule of

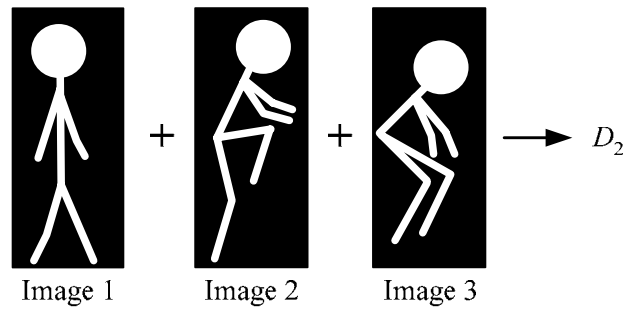
Rule 1. IF the activity's I_1 is P_1 AND its I_2 is P_2 AND its I_3 is P_3 , THEN the activity is D_1 . (30)

where I_i is Image i and P_j is Posture j . Similarly, for more examples, if we are given two feature-target vectors of input activities

$$[\mathbf{a}_1^2, \mathbf{a}_2^2, \mathbf{a}_3^2; D_2] \tag{31}$$



$$[\mathbf{a}_1^3, \mathbf{a}_2^3, \mathbf{a}_3^3; D_3] \tag{32}$$



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

$[\mathbf{a}_1^2, \mathbf{a}_2^2, \mathbf{a}_3^2; D_2]$ can imply the rule of

Rule 2. 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 D_2 . (33)

Similarly, $[\mathbf{a}_1^3, \mathbf{a}_2^3, \mathbf{a}_3^3; D_3]$ can imply the rule of

Rule 3. IF the activity's I_1 is P_2 AND its I_2 is P_{20} AND its I_3 is P_{21} , THEN the
activity is D_3 . (34)

where I_i is Image i and P_j is Posture j .

In Eq. (33), posture sequence does not appear from Posture 19 to Postures 20 in the order as our observation. However, our system is able to learn the hidden transitional modes of activities from training data. This is an advantage of our system and it will also improve the correct rate in classification. In Eq. (34), although Posture 2 is a posture of the activity D_1 but D_3 , the system still learns this sequence as the activity D_3 . We regard that Posture 2 as a common state of the two activities D_3 and D_1 . Therefore the fuzzy rules induce tolerant to some ambiguous postures of different activities and classify the image sequence to an activity more correctly.

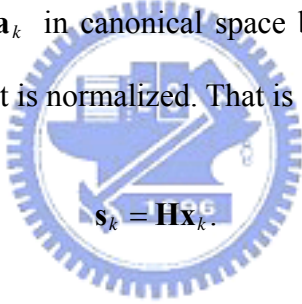
Due to a large number of training activities, some conflicting rules may be generated. The conflicting rules have the same antecedent conditions but lead to different consequent conditions. For a set of antecedent conditions, we can have only one rule to reflect it. Therefore, we have to choose one from 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.

3.5 Classification Algorithm

When a video stream is inputted for recognition, we extract image frames from the video first. Utilizing background model of Section 4.1 to extract foreground subject from the scene. The foreground object is a binary image. Suppose that we have a binary image \mathbf{x}'_k , where k is the index of the image sequence. The binary image of object \mathbf{x}'_k needs to be normalized. The binary image \mathbf{x}'_k also needs to subtract the mean vector in Eq. (3) in order to become the standardized vector \mathbf{x}_k . That is,

$$\mathbf{x}_k = \frac{\mathbf{x}'_k}{\|\mathbf{x}'_k\|} - \mathbf{m}_x. \quad (35)$$

\mathbf{x}_k is converted to a vector \mathbf{a}_k in canonical space by the transformation matrix \mathbf{H} after the binary image of object is normalized. That is



$$\mathbf{s}_k = \mathbf{H}\mathbf{x}_k. \quad (36)$$

If we assume each dimension of \mathbf{s}_k is independent, \mathbf{s}_k can be rewritten as

$$\mathbf{s}_k = [s_k^1, s_k^2, \dots, s_{i,j}^{c-1}]^T. \quad (37)$$

Let Σ denote the covariance matrix of all essential template vectors and C_i denote the i -th class of essential templates. The membership function is given by

$$\begin{aligned} r_{i,k} &= M(\mathbf{s}_k | C_i) \\ &= \frac{1}{(2\pi)^{\frac{c-1}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{s}_k - \mu)^T \Sigma^{-1}(\mathbf{s}_k - \mu)\right] \\ &= \arg \max_j \left\{ \prod_{m=1}^{c-1} \frac{1}{\sqrt{2\pi}\sigma_m} \exp\left[-\frac{1}{2} \frac{(s_k^m - \mu_{i,j}^m)^2}{\sigma_m^2}\right] \right\} \end{aligned} \quad (38)$$

where m is the number of dimension and j is the training model number. $r_{i,k}$ denotes the grade of membership function in category i of the k -th image frame. σ is the standard deviation of all essential templates. These membership functions are just the results of one image frame. We need to collect three images as a group for recognizing an activity. Therefore, we use two more transformed vector of passed image frames, which are called \mathbf{a}_{k-2} and \mathbf{a}_{k-1} . These three vectors form a feature vector $[\mathbf{a}_{k-2}, \mathbf{a}_{k-1}, \mathbf{a}_k]$. We compute the membership functions of the three vectors respectively. The procedures of calculating membership functions of \mathbf{a}_{k-2} and \mathbf{a}_{k-1} are the same as the process used for \mathbf{a}_k in Eq. (38).

We survey all fuzzy rules in our rule base and compute the similarity between the image sequence 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. (38). 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 unknown activity is classified to the activity recorded in this rule. Fig. 3.5 shows the structure of the classification algorithm.

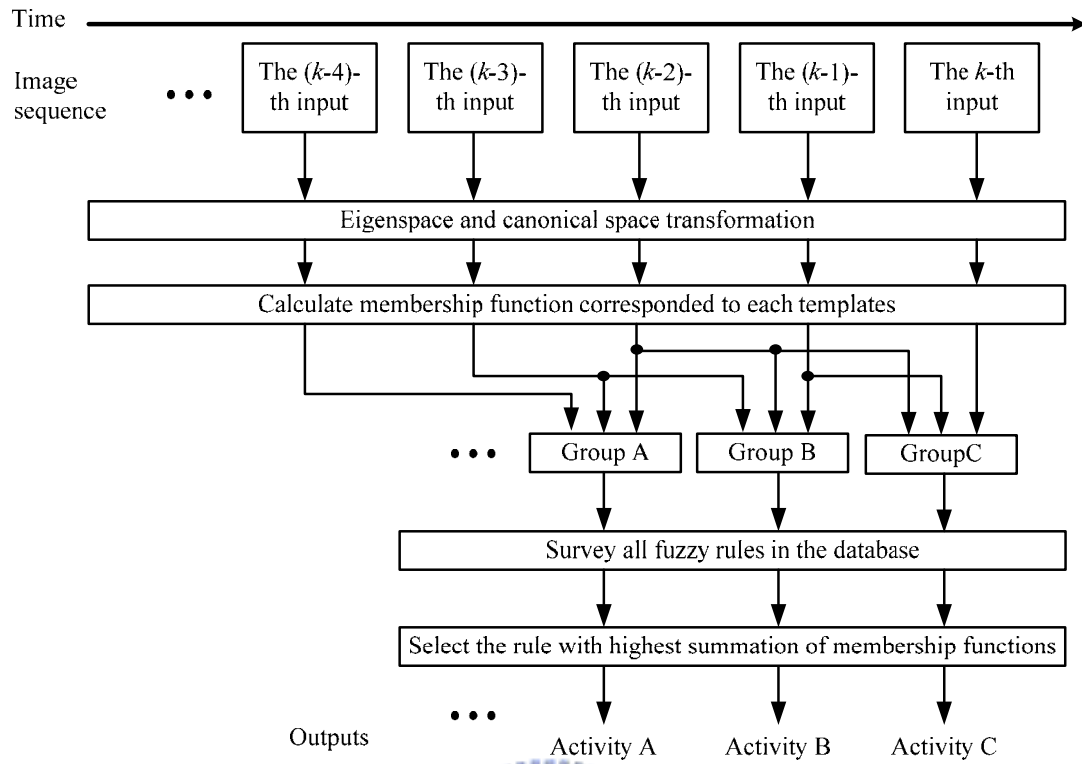


Fig. 3.5 The structure of classification algorithm.



Chapter 4 Experimental Results

In our experiment, we test our system on real temporal images. There are six model actions, which are demonstrated by the research members of our Intelligence Technology Lab of the Department of Electrical and Control Engineering at 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 uncomplicated and we equip a table in the scene. The color of clothing worn by subjects was different from the color of background. If the color of clothing and background are too similar, a moving object such as human body may not be easily segmented from image frames. 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. The environment of the classroom is shown in Fig. 4.1.



Fig. 4.1 The environment of the classroom.

Each member performed the six actions: “walking from left to right,” “walking from right to left,” “jumping,” “crouching,” “climbing up” and “climb down.” 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. Six lab members do 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 model actions is used for recognition and the other five are used for training. The video of each model is used for recognition 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 disturbance, there will be some noise or wrong segmented region left in the extracted image. Although extraction of foreground by using frame difference is the most famous method in image processing, the drawback of this method is not robust to the illumination changes. The frame ratio method can eliminate the influence of illumination variations.

Fig. 4.2 shows a comparison between frame ratio and frame difference. Fig. 4.2(a) is a background image and Fig. 4.2(b) is an image frame with a human. By using frame difference and frame ratio approach, we obtain Fig. 4.2(c) and Fig. 4.2(d), respectively. Gray level of the result images distributed from 0 to 255. Fig. 4.2(e) is the histogram of Fig. 4.3(c) and Fig. 4.2(f) is the histogram of Fig. 4.3(d). Comparing the histograms of Fig. 4.2(d) and Fig. 4.2(e), we find out that there was less noise in the region of low gray level by using frame ratio method. The Fig. 4.2(g) and Fig. 4.2(h) are the binary image of extraction images which simply took athreshold value 15 at gray level against Fig. 4.2(c) and Fig. 4.2(d). The Fig. 4.2(g) and Fig. 4.2(h) are the binary images after tacking the threshold on Fig. 4.2(c) and Fig. 4.2(d). Besides,

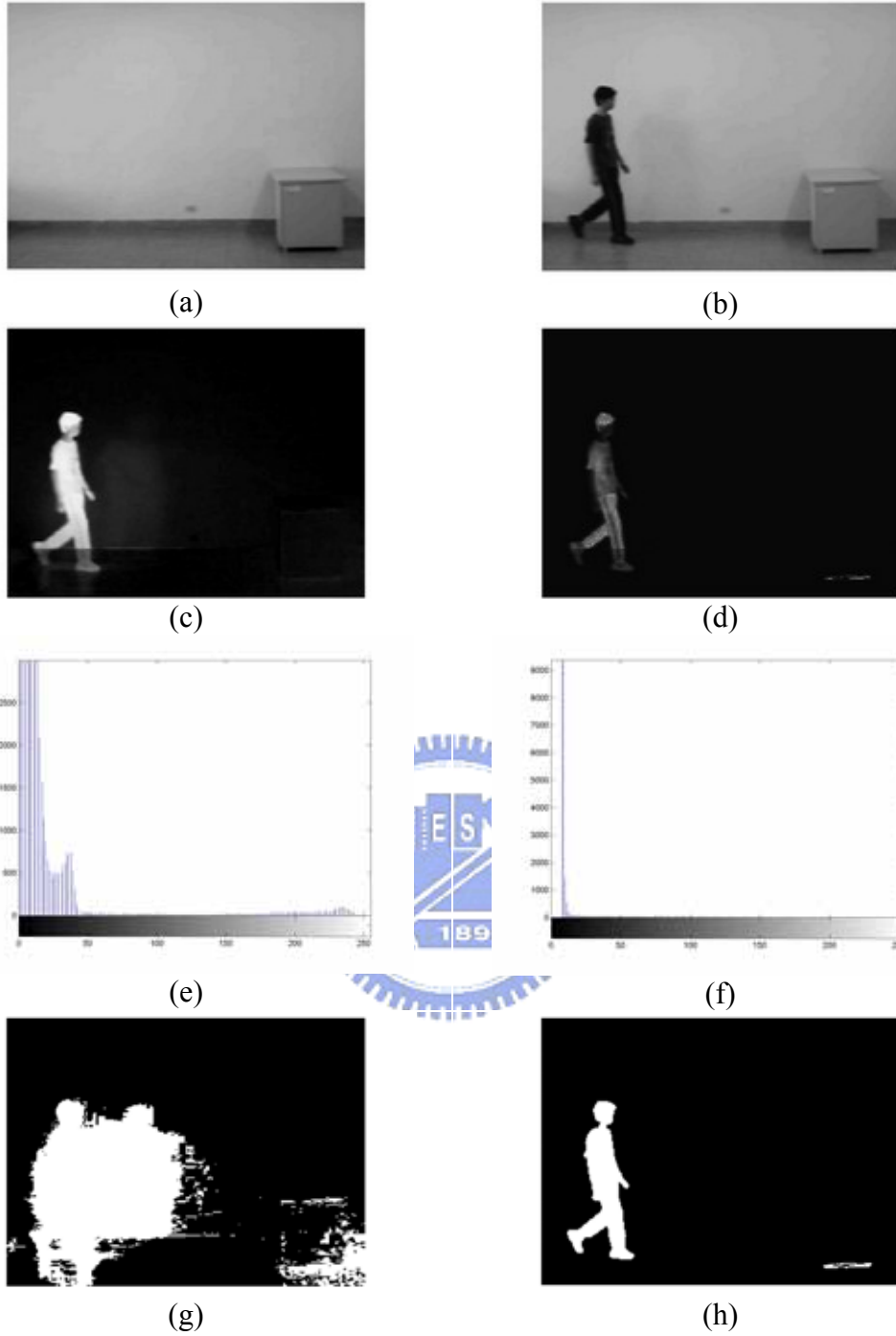


Fig. 4.2 The comparison between frame ratio and frame difference. (a) Background image, (b) image frame with a human, (c) frame difference, (d) frame ratio, (e) histogram of frame difference, (f) histogram of frame ratio, (g) foreground pixels of frame difference after simply taking a threshold, and (h) foreground pixels of frame ratio after simply taking a threshold.

frame ratio approach can centralize our concerned information at lower gray level rather than frame difference approach. We can take a lower threshold value of gray level to achieve a clean binary image of subject. Obviously, Fig. 4.2(h) has less noise and the foreground is extracted clearly.

A threshold k is applied in frame ratio approach to obtain binary image $B(x, y)$ in Eq. (20) described in Section 3.2. 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. 4.3. The threshold value one point three 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 direction. Fig. 4.4 shows an example of foreground region extraction. Fig. 4.4(a) is a image frame of the video stream. Fig. 4.4(b) is the binary image after performing background model analysis. Fig. 4.4(c) and Fig. 4.4(d) show the projection of Fig. 4.4(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 as the corner of a rectangle to extract foreground region from Fig. 4.4(b). Fig. 4.4(e) is the extracted foreground region.

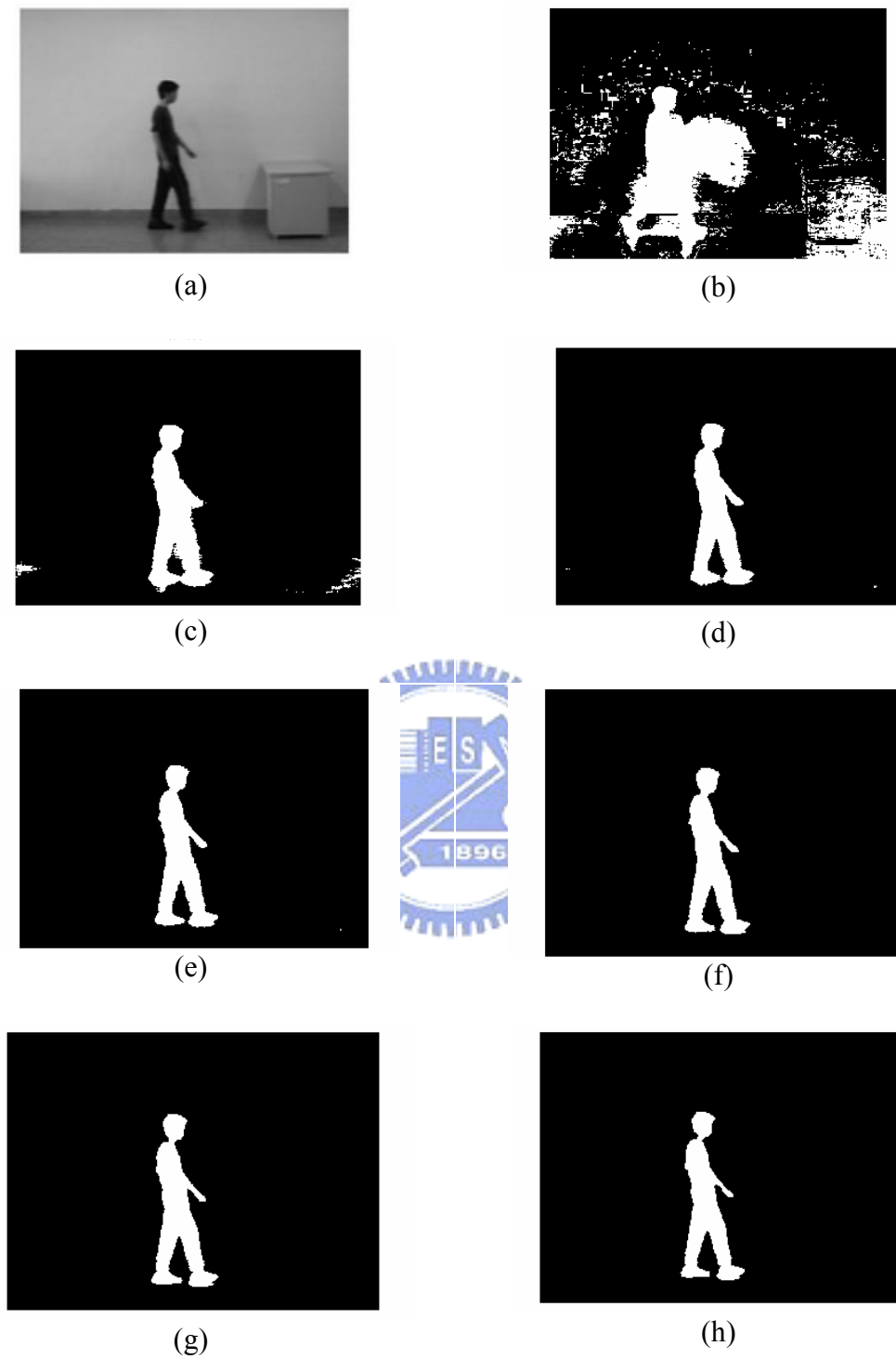


Fig. 4.3 An example of foreground region extraction at different threshold, k , values. (a) An image frame, (b) $k = 1.0$, (c) $k = 1.1$, (d) $k = 1.2$, (e) $k = 1.3$, (f) $k = 1.4$, (g) $k = 1.5$, and (h) $k = 1.6$.

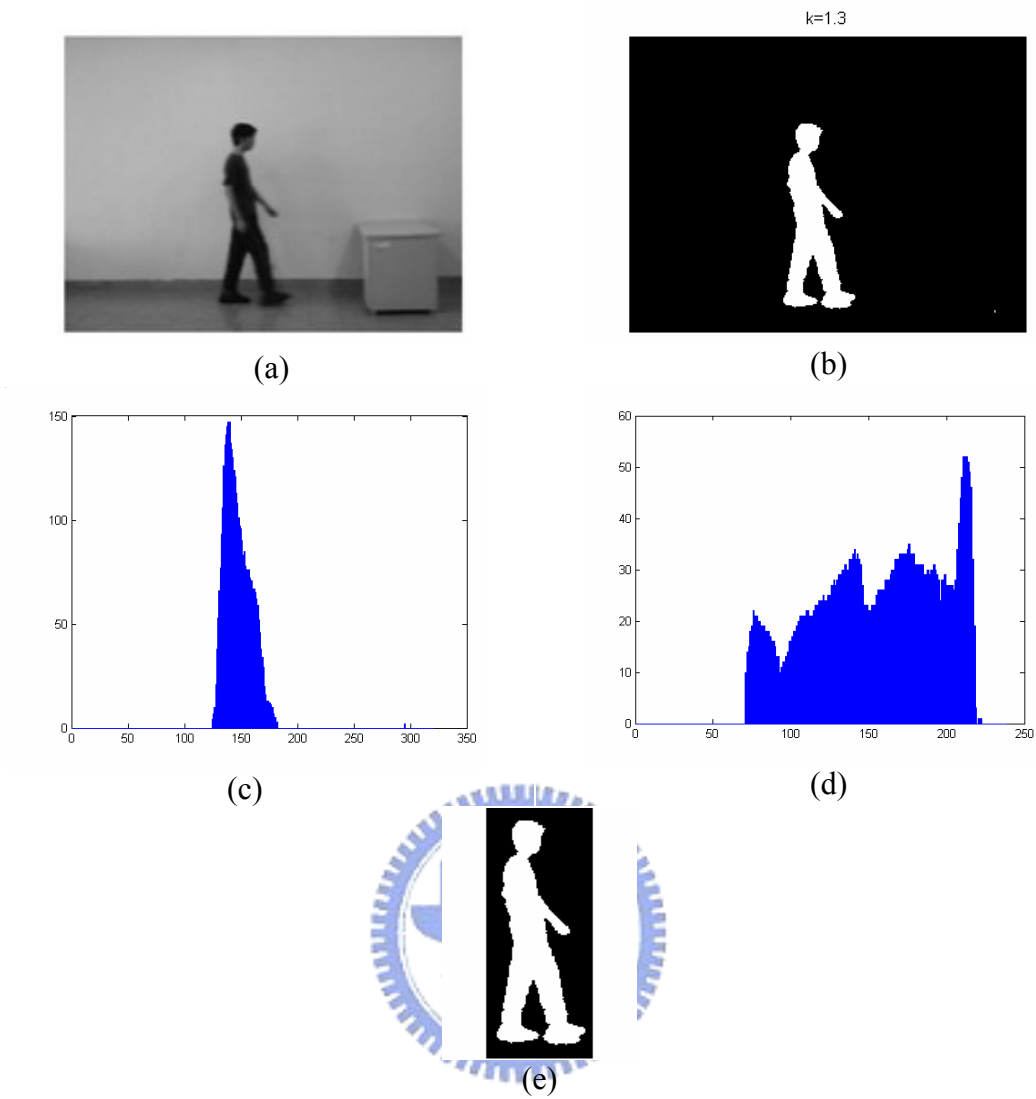


Fig. 4.4 An example of foreground region extraction. (a) An image frame, (b) binary image after background analysis, (c) projection of (b) onto X direction, (d) projection of (b) onto Y direction, (e) foreground region extracted.

4.2 Fuzzy Rule Construction for Action Recognition

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. An example is shown in Fig. 4.5. The image at the time t_1 was selected as the template n_1 and the image at time t_2 was selected as the next template n_2 .

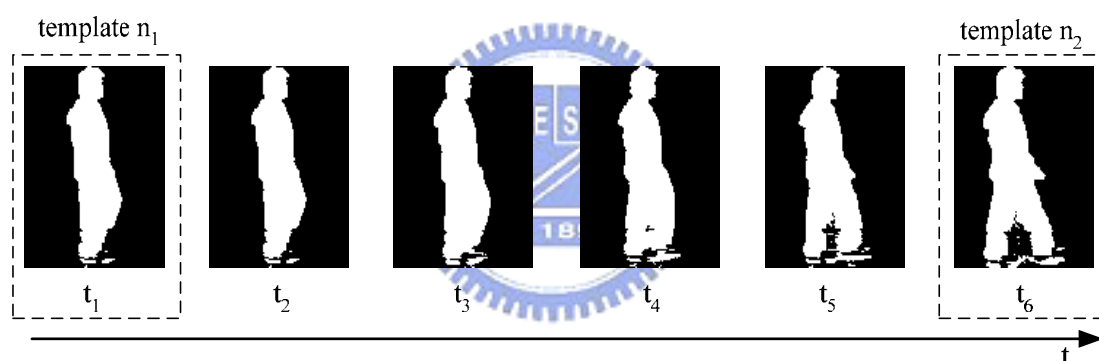


Fig. 4.5 Template selection with an interval of five frames.

We chose six kinds of 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 kinds of essential templates, and called 28 classes. The essential template numbers of each activity depend on how long it takes. Each essential template is a cluster with five template images which are from five different training person’s and have similar postures. Fig. 4.6 and Fig. 4.7 are two examples of some templates of two training

model.

In Fig. 4.6 and Fig. 4.7, if a model bend down or squat down, the bodies in template images are wider than others. It is because 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 since the ratio of image width to height is larger than the ratio of 96 to 128 and are resized by the smaller scale. On the contrary, when the height of body shape is shorter, the magnifying factor becomes larger.

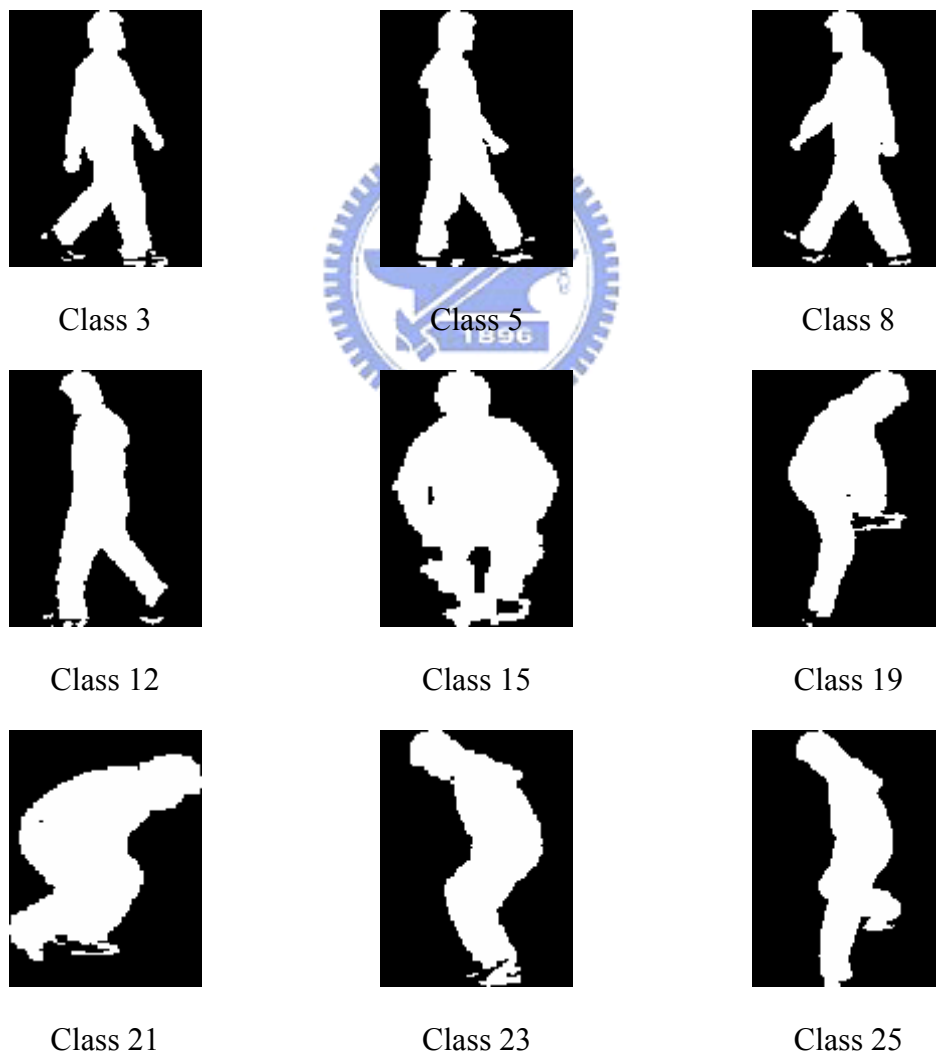


Fig. 4.6 Some “essential templates of posture” of model A.

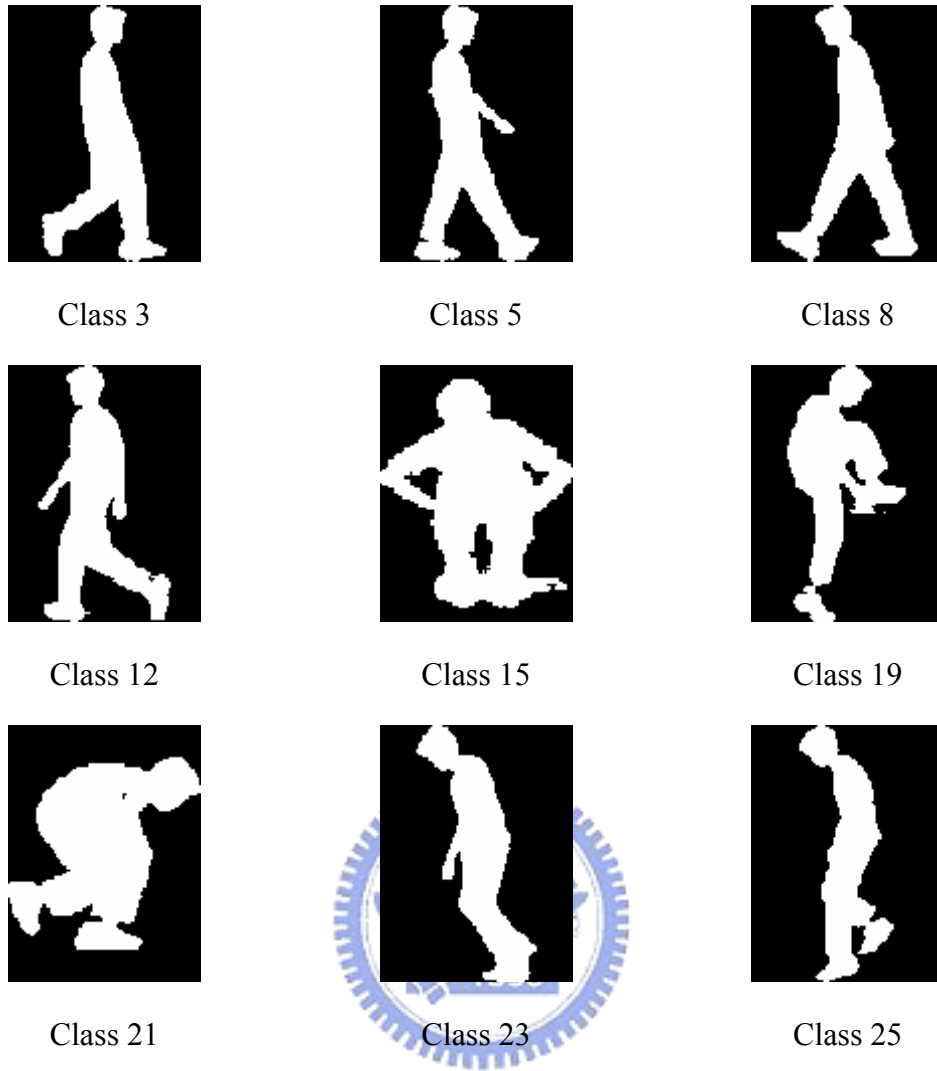


Fig. 4.7 Corresponding “essential templates of posture,” Fig. 4.6, of model B.

The template images are transformed to canonical space by the methods described in Chapter 2. The mean vectors and the standard deviation vectors of all templates were computed by Eq. (24). Each template image of a training model was treated as a center. Hence, there were 140 mean vectors because of five training models and 28 classes of templates. Besides, there were six groups of standard deviation vectors and mean vectors because of six kinds of different training models.

After determining the standard deviation vectors, the corresponding training video frames are inputted. The relationship between each image frame and each

template is calculated by using Eq. (27) in Section 3.4. We gathered three images as a group in order to include temporal information. The interval between each of these three images is five image frames which is the same as in template selection. Training is accomplished in off-line situation. Therefore, we gathered three images from different start points to train fuzzy rules. For examples: the first frame, the 6-th frame and 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; 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 are used for training fuzzy rules in our experiment, because the starting posture of testing video and of training video may not be the same. By utilizing different start points, the system is able to learn much more combinations of image frames and increase accuracy of fuzzy rules.

The group of the threes images is converted to the posture sequence which has the maximum summation of three membership function values in Eq. (27). Each posture sequence will trigger a corresponding rule one time. If the corresponding rule is not existent, a new rule is built in the form of **IF-THEN** which is 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 rules whose occurrence times of the specific sequence is relative few. The numbers of rules varies with different thresholds. Table I shows the rule numbers of different threshold values. One person video out of the six person models is chosen from the training data in order to use it as a testing datum. We can easily find out that the higher threshold we set, the fewer rules we obtained. Although higher threshold can reduce rules, fewer rules will lose the tolerance for ambiguity. If some conflicting rules are generated, we choose the rule that is supported by a maximum number of training instances.

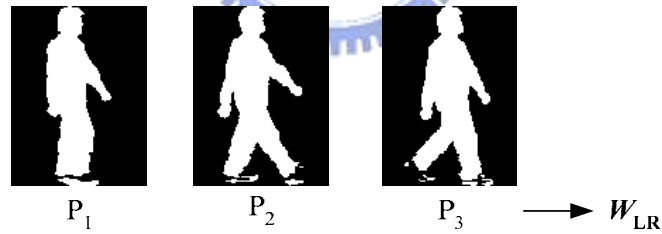
TABLE I
THE RULE NUMBERS AT DIFFERENT THRESHOLD

Training data models	Threshold = 3	Threshold = 4	Threshold = 5
Person 1	131	92	83
Person 2	130	101	80
Person 3	148	99	82
Person 4	157	105	75
Person 5	136	91	70
Person 6	150	114	87

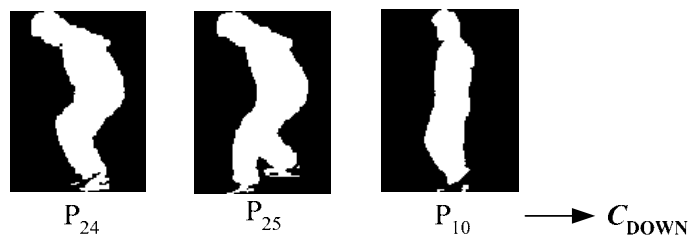
The templates and the test patterns of fuzzy rules are both sampled with a rate of five image frames. An activity should appear in proper order directly perceived through our sense. For example, P_1 through P_6 are the six linguistic labels of the activity “walking from left to right.” The activity of “walking from left to right” should 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) . We called these rules essential rules. There are totally 24 essential rules for the six activities. But there are only 18 essential rules found in our experiment if we set the threshold at three. Numbers of fuzzy rules are many more than the essential rules, because essential rules are based on the view of spatiotemporal space but fuzzy rule base is based on the view of canonical space. Fuzzy rule base is able to learn the hidden modes or replaceable existent in these actions. The appeared essential rules are less than 24 because fuzzy rule base combines some similar rules to one rule. Some of the fuzzy rules, which used the training data of person 1 and the threshold was set at three, are listed in Table II. Two of the fuzzy rules are represented in the view of template images in Fig. 4.8.

TABLE II
SOME OF THE OBTAINED FUZZY RULE BASE

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 ₃	W_{LR}
⋮	⋮	⋮	⋮	⋮
30	P ₄	P ₁₁	P ₁₂	W_{RL}
⋮	⋮	⋮	⋮	⋮
60	P ₃	P ₁₃	P ₁₄	J_{UMP}
⋮	⋮	⋮	⋮	⋮
80	P ₁₃	P ₁₆	P ₁₇	C_{ROUCH}
⋮	⋮	⋮	⋮	⋮
91	P ₂	P ₁₈	P ₁₈	C_{UP}
⋮	⋮	⋮	⋮	⋮
129	P ₂₇	P ₂₈	P ₁₀	C_{DOWN}
130	P ₂₈	P ₇	P ₇	C_{DOWN}
131	P ₂₈	P ₂₈	P ₁₀	C_{DOWN}



(a)



(b)

Fig. 4.8 Two examples of fuzzy rules. (a) Walking from left to right

(b) Climbing down

4.3 The Recognition Rate of Activities 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 fuzzy rules. Namely, we recognize the video from the first frame, the second frame, the third frame and the fourth frame, *etc.* with the sampling intervals of five frames. The testing video wasn't used for constructing templates and fuzzy rules. Hence, there are six corresponding databases for the video of the six models.

An example of recognition rate of a testing video start from different frames is shown in Table III. Where W_{LR} is the activity "walking form lest 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" and C_{DOWN} is the activity "climbing down." The threshold selecting the rule base for of this testing model is three.

TABLE III
THE RECOGNITION RATE OF PERSON 1 WITH DIFFERENT STARTING FRAME

Starting frame	Recognition rate (%)					
	W_{LR}	W_{RL}	C_{ROUCH}	J_{UMP}	C_{UP}	C_{DOWN}
From the 1 st , 6 th , ... frame	100.00	84.62	100.00	100.00	90.00	83.33
From the 2 nd , 7 th , ... frame	100.00	100.00	100.00	60.00	90.0	83.33
From the 3 rd , 8 th , ... frame	100.00	92.31	100.00	100.00	100.00	83.33
From the 4 th , 9 th , ... frame	100.00	91.67	100.00	100.00	100.00	66.67
From the 5 th , 10 th , ... frame	100.00	100.00	100.00	100.00	88.89	66.67

The frame numbers in each activity of every model are shown in Table IV. Each activity at least has one cycle. Frame numbers of J_{UMP} , C_{ROUCH} and C_{DOWN} are fewer than the other three activities because the period of these activities is much shorter. The total number of testing frames is 2016.

TABLE IV
THE FRAME NUMBER OF EACH ACTIVITY

Testing data	Frame numbers					
	W_{LR}	W_{RL}	C_{ROUCH}	J_{UMP}	C_{UP}	C_{DOWN}
Person 1	69	75	41	47	74	47
Person 2	64	67	45	44	85	45
Person 3	71	74	49	44	78	63
Person 4	60	73	37	33	57	40
Person 5	73	74	43	37	85	53
Person 6	55	39	51	28	66	30
Total numbers	392	402	266	233	445	278

Table V shows the recognition rate of our system. The threshold in constructing fuzzy rules is three. We integrated the results of the same activity but starting from different first frame to one for displaying convenience. The recognition rate of C_{DOWN} in Person 3 is relative low, because the activity looks like “jumping down” rather than “climbing down.”

TABLE V
THE RECOGNITION RATE OF EACH ACTIVITY

Testing data	Recognition rate (%)					
	W_{LR}	W_{RL}	C_{ROUCH}	J_{UMP}	C_{UP}	C_{DOWN}
Person 1	100.00	92.31	70.97	78.38	78.13	94.59
Person 2	100.00	82.46	97.14	61.76	100.00	94.29
Person 3	100.00	100.00	74.36	94.12	100.00	45.28
Person 4	100.00	93.65	100.00	91.30	93.62	76.67
Person 5	100.00	100.00	100.00	100.00	90.67	100.00
Person 6	100.00	100.00	97.56	100.00	100.00	100.00
Average	91.78					

The threshold setting in fuzzy rule construction affected the recognition rate. Fig. 4.9 compared the recognition rates of rules with different thresholds. Although recognition rate is higher when the threshold equals to two, we adopted three in our experiment. This is because that there were too few support and too many rules if the threshold was small.

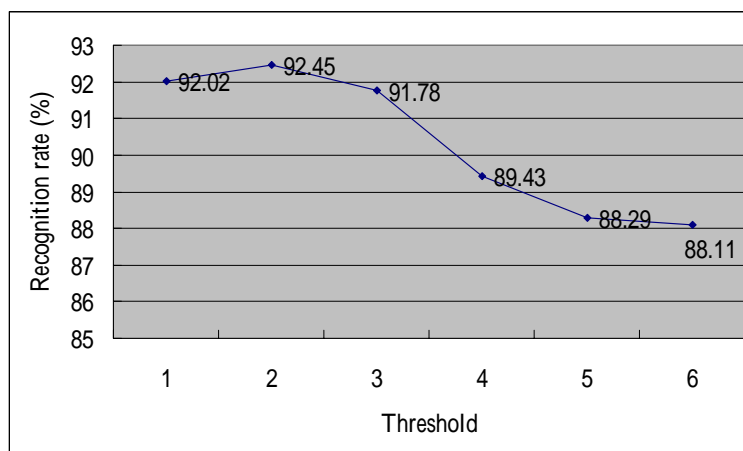


Fig. 4.9 The recognition rate of utilizing different thresholds.

4.4 Comparison between Fuzzy Rule Base Approach and Nearest Neighbor Approach

The structure of recognition utilizing nearest neighbor in canonical space is shown in Fig. 4.10. We utilize the same template images which were defined in our approach. There are 28 categories of key frame posture. Five persons' activities were used for training, and the left one person's activities were used for recognition. The activities of six models are applied to test in turn. Firstly, template images were transformed to canonical space by transformation matrix \mathbf{H} in Eq. (15). Each template image is transformed to a 27-dimensional vector. Each category has five template vectors of similar posture images done by different persons.

When a testing video is inputted, image frames were sampled with an interval of five frames as before. The sampled images were transformed to canonical space vectors which were utilized to compute the Euclidean distance. The template with the minimum distance was chosen, and the images were classified to the activity of this template.

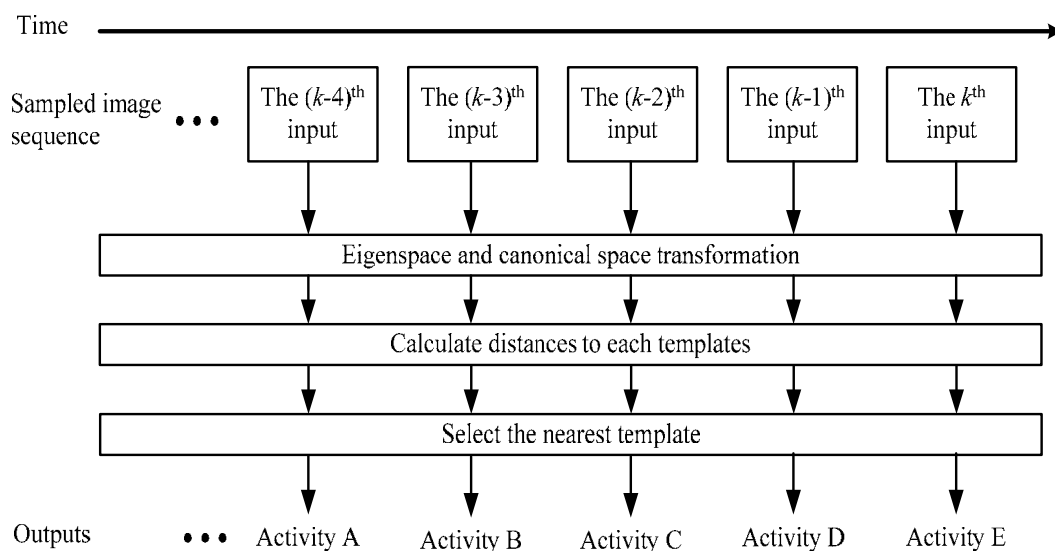


Fig. 4.10 The structure of nearest neighbor approach.

The comparison of recognition rate between nearest neighbor approach and fuzzy rule base is shown in Table VI. The training data and testing data is the same as the data in fuzzy rule base approach. The recognition rates of fuzzy rule base approach are higher than the rate of nearest neighbor approach. The total recognition rate were improved about sixteen percent. Consequently, temporal information is significant information for activity recognition.

TABLE VI
THE COMPARISON OF RECOGNITION RATE BETWEEN NEAREST NEIGHBOR
APPROACH AND FUZZY RULE BASE APPROACH

	Nearest Neighbor	Fuzzy Rule Base
Person 1	64.72	84.61
Person 2	80.00	91.03
Person 3	76.52	87.15
Person 4	54.02	93.33
Person 5	85.21	97.71
Person 6	81.78	99.52
Average	75.94	91.78

4.5 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 [17]. 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 in this section.

The first module of our HMM approach is to transform image sequences into posture sequences. The nearest neighbor approach described in the previous section was used. The template images and categories of templates are also the same as in fuzzy rule base approach. Each key frame posture was indicated by a number index. The number of HMM symbols was 28. An image in video sequence was assigned to a key posture which has the minimum distance to this image. The video sequences were transformed to posture sequences which are called observations in HMM.

The second module concerns the learning from training image sequences. In this learning phase, each HMM must be trained so that it is most likely to generate the posture patterns for its action. Training an HMM means optimizing the model parameters to maximize the probability of the observation sequence via activity video frames. The Baum-Welch algorithm is use for these estimations. In our experiment we adopted forward-chaining states and the number of states, i.e., the number of key postures, was set to 28. The length of the observation sequence was set to three.

After learning from training data, we obtain six λ_i of HMM for each activity category, where λ is a complete parameter set of the model and $i = 1, 2, \dots, 6$. Each HMM is a stochastic model for one of our six activities. To recognize an observed posture sequences of an unknown activity, we choose the model which best matches the observations from the six HMMs. When an observation sequences $O = O_1, O_2, O_3$

of three consecutive sampled image is given, we calculate $\Pr(\lambda_i | O)$ for each HMM λ_i and select λ_{c^*} , where

$$c^* = \arg \max_i (\Pr(\lambda_i | O)).$$

$\Pr(\lambda_i | O)$ is the probability that the sequence was generated by HMM λ_i . This probability was calculated by using the forward algorithm [12].

The same training and testing video images used in fuzzy rule base approach were used for HMM approach. The comparison of recognition rates between HMM approach and fuzzy rule base approach is shown in Table VII. 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 VII
THE COMPARISON OF RECOGNITION RATE BETWEEN HMM APPROACH AND FUZZY RULE BASE APPROACH

	HMM	Fuzzy Rule Base
Person 1	79.18	84.61
Person 2	90.00	91.03
Person 3	80.25	87.15
Person 4	90.00	93.33
Person 5	91.80	97.71
Person 6	88.90	99.52
Average	86.41	91.78

Chapter 5 Conclusion

In this thesis, 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 91.78% 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.

References

- [1] D. Gavrilu, "The visual analysis of human movement: a survey," *Comput. Vision and Image Understan.*, vol. 73, pp.82–98, 1999.
- [2] 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] 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.
- [6] M. Piccardi, "Background subtraction techniques: a review," in *Proc. IEEE Int. Conf. SMC.*, vol. 4, pp. 3099–3104, 2004.
- [7] 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.
- [8] 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.
- [9] J. Wang, G. Yuantao, K. N. Plataniotis, and A. N. Venetsanopoulos, "Select eigenfaces for face recognition with one training sample per subject," in *Proc. 8th Cont., Automat. Robot. Vision Conf., ICARCV 2004*, vol. 1, pp. 391–396, 2004.
- [10] 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.

- [11] 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.
- [12] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [13] E. Trentin and M. Gori, “Robust combination of neural networks and hidden Markov models for speech recognition,” *IEEE Trans. Neural Networks*, vol. 14, pp. 1519–1531, 2003.
- [14] S. Yoshizawa, N. Wada, N. Hayasaka, and Y. Miyanaga, “Scalable architecture for word HMM-based speech recognition,” in *Proc. the 2004 Int. Symposium Circuits Syst., ISCAS 2004*, vol. 3, pp. III–417–20, 2004.
- [15] L. Nianjun, B. C. Lovell, and P. J. Kootsookos, “Evaluation of HMM training algorithms for letter hand gesture recognition,” in *Proc. the 3rd IEEE Int. Symposium Signal Processing Inform. Technol., ISSPIT 2003*, pp.648–651, 2003.
- [16] A. D. Wilson and A. F. Bobick, “Recognition and interpretation of parametric gesture,” in *Proc. Sixth Int. Conf. Comput. Vision*, pp. 329–336, 1998.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] Mu-Chun Su, “A fuzzy rule-based approach to spatio-temporal hand gesture recognition,” *IEEE Trans. Sys., Man Cybern.*, vol. 30, no. 2, pp. 276–281, 2000.

- [21] 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.
- [22] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd edition, 1300 Boylston Street Chestnut Hill, Massachusetts USA: Academic Press, 1990.
- [23] K. Etemad and R. Chellappa, “Discriminant analysis for recognition of human face images,” in *Proc. ICASSP*, pp. 2148–2151.

