

國立交通大學

電控工程研究所

碩士論文

以模糊規則為基礎之日夜動作辨識及步態辨識

Fuzzy Rule Based Day-and-Night
Action Recognition and Gait Recognition

研究生：顏宏年

指導教授：張志永

中華民國一百零二年七月

以模糊規則為基礎之日夜動作辨識及步態辨識

Fuzzy Rule Based Day-and-Night
Action Recognition and Gait Recognition

學 生：顏宏年 Student : Hong- Nien Yen

指導教授：張志永 Advisor : Jyh-Yeong Chang

國立交通大學

電機工程學系

碩士論文

A Thesis

Submitted to Department of Electrical Engineering

College of Electrical Engineering

National Chiao-Tung University

in Partial Fulfillment of the Requirements

for the Degree of Master in

Electrical Control Engineering

July 2013

Hsinchu, Taiwan, Republic of China

中華民國一百零二年七月

以模糊規則為基礎之日夜動作辨識及步態辨識

學生:顏宏年

指導教授: 張志永博士

國立交通大學電控工程研究所

摘要

本篇論文實現一套自動化日夜居家監視系統，此系統為了提供良好的監控服務，著重於動作辨識和步態辨識，藉由步態辨識技術，掌控環境內每位成員的身份，其辨識動作以了解每個人的行動。本篇論文使用兩台攝影機在實驗室進行人物辨識及動作辨識。

動作辨識與步態辨識主要的資訊來自於人，擷取出人體部份為辨識的依據，為了更精確的擷取前景，使用灰階域與HSV色彩空間，建立兩種背景模型，並能有效的消除影像中陰影部分，使得擷取的前景能夠完整。接著將前景經由特徵空間轉換及標準空間轉換後，投影到維度較小的空間且能保有原影像的資訊。接著進行訓練，本方法加入時間資訊，將前景 5:1 減低抽樣取出影像，累積三張影像，建立模糊法則。辨識工作方面，使用預先學習且建立的模糊法則，進行辨識。

Fuzzy Rule Based Day-and-Night Action Recognition and Gait Recognition

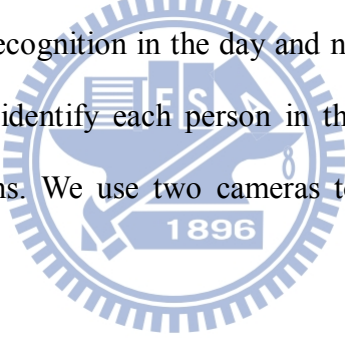
STUDENT: Hong-Nien Yen

ADVISOR: Dr. Jyh-Yeong Chang

Institute of Electrical Control Engineering
National Chiao-Tung University

ABSTRACT

In this thesis, we implement an automatic home health care system that combines action recognition and gait recognition in the day and night environments (bright and dark). Gait recognition can identify each person in the lab; action recognition can identify each person's actions. We use two cameras to recognize actions and gait, respectively.

The logo of National Chiao-Tung University is a circular seal. It features a gear-like outer border. Inside the circle, there is a central emblem with a book and a lamp. The letters 'F S A' are prominently displayed in the center. Below the emblem, the year '1896' is inscribed. The entire logo is rendered in a light blue color.

We build two background models, one in grayscale, and the other in the HSV color space, that extract the human region correctly. We also reduce the shadowing effect. For better efficiency, the binary image is transformed into a new space by eigenspace and canonical space transformation. Then we gathered three image frame sequence, 5:1 down sampling from the video, to convert to a posture sequence by template matching. The posture sequence is classified to an action or a person's gait by fuzzy rules inference, which combines temporal sequence information for 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 the people who assisted me in completing this research.

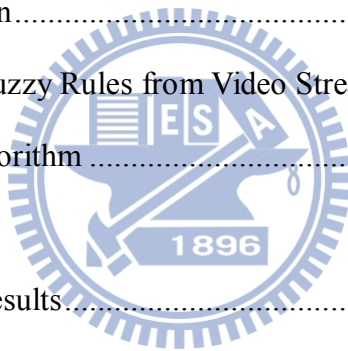
Finally, I would like to express my deepest gratitude to my family for their concern, supports and encouragements.



Contents

摘要	i
ABSTRACT	ii
ACKNOWLEDGEMENTS	iii
Contents	iv
List of Figures	vi
List of Tables	viii
Chapter 1 Introduction	1
1.1 Motivation	1
1.2 Foreground Extraction	3
1.3 Eigenspace and Canonical Space Transformation	3
1.4 Action and Gait Recognition	4
1.5 Thesis Outline	6
Chapter 2 Basic Concepts	7
2.1 Object Extraction	7
2.1.1 Background Model	7
2.1.2 Foreground Region Detection	8
2.2 The HSV Color Space	9
2.3 Eigenspace and Canonical Space Transform	12
2.3.1 Eigenspace Transformation (EST)	14
2.3.2 Canonical Space Transformation (CST)	16

Chapter 3	Activity and Human Recognition System	20
3.1	Foreground Extraction	20
3.1.1	Background Model.....	20
3.1.2	Extraction of Foreground Object	23
3.1.3	Shadow Suppression.....	25
3.1.4	Object Segmentation	27
3.1.5	Foreground Image Compensation	29
3.2	Background Update	30
3.3	Skin Color Detection.....	31
3.4	Template Selection.....	32
3.5	Construction of Fuzzy Rules from Video Stream.....	34
3.6	Classification Algorithm	38
Chapter 4	Experimental Results.....	41
4.1	Background Model and Foreground Object Extraction.....	45
4.2	Fuzzy Rule Construction for Action Recognition	50
4.3	The Action Recognition Accuracy.....	54
4.4	Fuzzy Rule Construction for Gait Recognition	57
4.5	The Recognition Rate of Gaits	58
Chapter 5	Conclusion	60
References	61



List of Figures

Fig. 1.1.	Block diagram showing the action and gait recognition system.	2
Fig. 2.1	(a) The HSV Cone. (b) Cross-section of HSV value “1.”	9
Fig. 2.2	The structure of the image analysis.	13
Fig. 3.1.	The framework we construct the background models.....	20
Fig. 3.2.	The framework we develop for foreground subject extraction.	23
Fig. 3.3.	Histogram of binary image projection in X and Y direction.....	28
Fig. 3.4.	The binary image of extracted foreground region.	28
Fig. 3.5.	(a) Foreground image. (b) Foreground image after opening and closing repair of (a).	29
Fig. 3.6.	Using 5:1 down-sampling rate to select the essential template image.	32
Fig. 3.7.	Common states of two different activities.	34
Fig. 3.8.	A fuzzy rule learned to classify action.....	37
Fig. 3.9.	The structure of action recognition algorithm.	40
Fig. 4.1.	(a) The action recognition experiment environment in the day, (b) The action recognition experiment environment in the night.....	41
Fig. 4.2.	Typical video sequences for actions of our LAB in bright environment (432 Lux).	42
Fig. 4.3.	Typical video sequences for actions of our LAB in dark environment (0.26 Lux).	43
Fig. 4.4.	(a) The gait recognition experiment environment in the day, (b) The gait recognition experiment environment in the night.	44
Fig. 4.5.	Typical video sequences for gait of our LAB in bright environments.	44
Fig. 4.6.	Typical video sequences for gait of our LAB in dark environments.	44
Fig. 4.7.	The results of foreground extraction in bright action recognition	

environment.	46
Fig. 4.8. The results of foreground extraction in the dark action recognition environment.	47
Fig. 4.9. The results of foreground extraction in the bright gait recognition environment.	48
Fig. 4.10. The results of foreground extraction in the dark gait recognition environment.	49
Fig. 4.11. Key postures of the actions of person 1.	52
Fig. 4.12. The fuzzy rule of walk from right to left	53



List of Tables

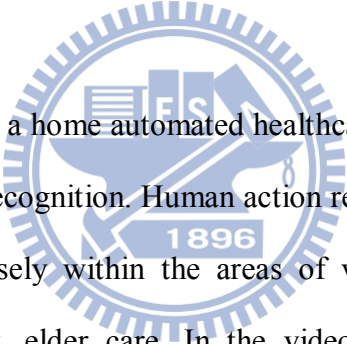
TABLE I	THE CORRECT RATE OF ACTION RECOGNITION IN THE BRIGHT ENVIRONMENT55
TABLE II	THE CORRECT RATE OF ACTION RECOGNITION IN THE DARK ENVIRONMENT56
TABLE III	THE CORRECT RATE OF PERSON RECOGNITION BY THE GAIT VIDEOS IN THE BRIGHT ENVIRONMENT.....	59
TABLE IV	THE CORRECT RATE OF PERSON RECOGNITION BY THE GAIT VIDEOS IN THE DARK ENVIRONMENT.....	59



Chapter 1 Introduction

1.1 Motivation

According to the European Union (EU) commission's projection, the number of older people will increase threefold between 2008 and 2060. This envisages great challenges towards care for older people with limited available resources [1]. Most of the home nursing care service is provided by many professional people, but human resources are limited. Therefore, the home automated healthcare system becomes a popular research area.



In this thesis, we design a home automated healthcare system which includes the action recognition and gait recognition. Human action recognition is an open problem that has been studied intensely within the areas of video surveillance, homeland security, and more recently, elder care. In the video surveillance, human action recognition system identifies each person's action. In the elder care, human action recognition identifies whether there is abnormal action, in order to ensure the health of elders. However, we can not understand who is doing the action, hence we propose to use each person's gait to identify each person.

Finally, we combine gait recognition with an action recognition system to enhance its effectiveness. We hope that the developed system can recognize a person in his home and also recognize and record his activity in the daily living environment. Our system flowchart is shown in Fig 1.1.

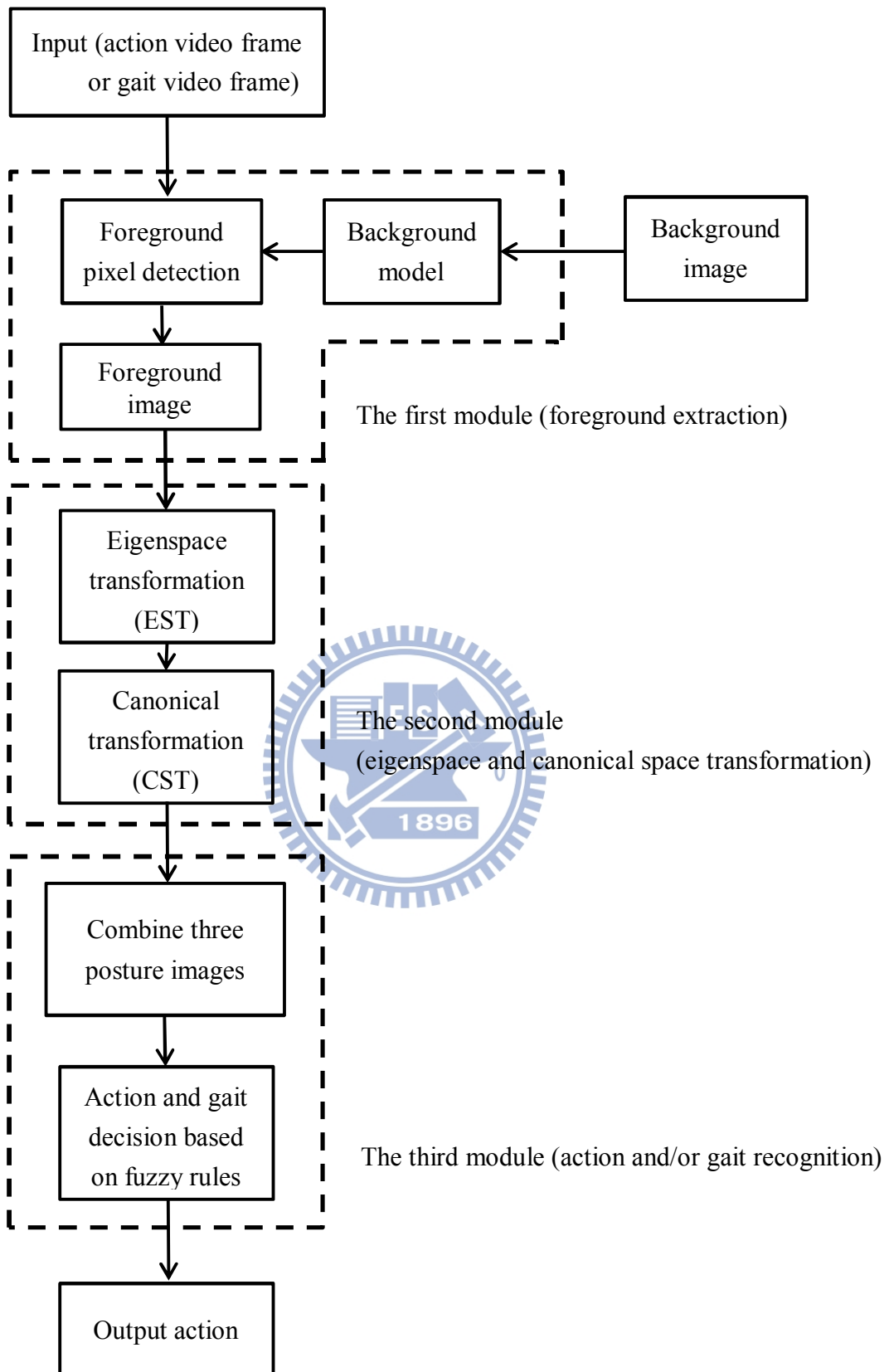


Fig. 1.1. Block diagram showing the action and gait recognition system.

1.2 Foreground Extraction

The first step of the activity recognition system is foreground extraction. We need to construct a background model. Background subtraction is a method typically used to segment moving regions in image sequences taken from a static camera by comparing each new frame to a model of the scene background [2]. There are many methods to build background models. W^4 [3] is such a popular example that using frame-difference with a threshold. In addition, foreground subject extraction is commonly affected by the additional inclusion of shadows. A lot of attempts have been developed to tackle the shadow suppression. Horprasert et al. [4] and Cucchiara et al. [5] utilized the rationale that shadows have similar chromaticity, but lower brightness than the background model. In our system, we construct two background models for more correct foreground extraction; one is based on grayscale value and the other is based on HSV color space. After subtracting each pixel value of background model from that current image frame, the resulting image is converted to a binary image by setting a threshold. Therefore, we can set a threshold in the histogram of the binary image to extract a rectangular image, which represents the shapes of a person. Then, the rectangle image is resized to the specified resolution for normalization.

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 has some redundancy. The redundancy possesses no 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 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 widely as shown below: automatic face recognition proposed in [6], [7]; gait analysis proposed in [8]; and action recognition proposed in [9]. 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 Action and Gait Recognition

In this thesis each segmented foreground in a video segmentation is transformed into an image feature vector by extracting features from images. We extract image features by using eigenspace transformation and canonical space transformation. We

have grouped three contiguous 5:1 down-sampled images and transform them to three consecutive feature vectors. Then, the three contiguous images are down-sampled and its sample rate is usually 6 frames per second. Next, 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 stage, 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 fuzzy rule, in the learned fuzzy rule based system for recognition which best matches the observed posture sequence is chosen as the recognized action category. We make use of fuzzy rule-base techniques to classify human activity and gait, not using the shape of an image. Thus our activity recognition can be tolerant of dissimilarity, uncertainty, ambiguity and irregularity exists in the data.

In our system, we propose a fuzzy rule-based approach for human activity recognition and gait recognition. Each action is represented in the form of fuzzy IF-THEN rules, extracted from the posture sequences of the training data. Each IF-THEN rule is fuzzified by employing an innovative membership function in order to represent the degree of the similarity between a pattern and the corresponding antecedent part in the training data. When our system classifies an unknown action or gait, it will be inferred by each fuzzy rule learned before using three consecutive sampled images of the video frames. The accumulated similarity measure associated with these three consecutive postures is to match the posture sequence representing an activity model or a gait model of the training database, and the unknown action or gait is classified to the one yielding the highest accumulative similarity. Our system can work in day and night (bright and dark) environments.

1.5 Thesis Outline

The thesis is organized as follows. In Chapter 2, we introduce the basic concepts of object extraction, the HSV color space, eigenspace and canonical space transform. In Chapter 3, we describe our system that includes “Foreground Extraction,” “activity recognition system” and “gait recognition system.” In Chapter 4, the experiment results of our recognition systems are shown. At last, we conclude this thesis with a discussion in Chapter 5.



Chapter 2 Basic Concepts

In this chapter, we explain the basic concepts of object extraction in the section 2.1 and the HSV color space in the section 2.2. Then in the section 2.3 we introduce the basic concepts of eigenspace and canonical space transform.

2.1 Object Extraction

The first step of human activity recognition system and human gait analysis is object extraction. We have to construct a background model for foreground 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. In this section, we will introduce W^4 how to construct the background model and detect the foreground region [3].

2.1.1 Background Model

W^4 obtains the background model even if there are moving foreground objects. It uses a two stage method based on without moving pixels from background model computation.

In the first stage, a pixel-wise median filter is applied to several seconds of video to distinguish moving pixels from stationary pixels.

In the second stage, those stationary pixels are processed to construct the background model. Let V be an array containing N consecutive images, $V^i(x)$ is the intensity of a pixel location x in the i -th image of V . $\sigma(x)$ is the standard

deviation and $\lambda(x)$ is the median value of intensities at pixel location x in all images in V . The background model for a pixel location x , $[m(x), n(x), d(x)]$, is obtained as follows:

$$\begin{bmatrix} m(x) \\ n(x) \\ d(x) \end{bmatrix} = \begin{bmatrix} \min_z \{V^z(x)\} \\ \max_z \{V^z(x)\} \\ \max \{|V^z(x) - V^{z-1}(x)|\} \end{bmatrix}, \quad (2.1)$$

where $|V^z(x) - \lambda(x)| < 2 \times \sigma(x)$. Here, $V^z(x)$ is classified as stationary pixels.

2.1.2 Foreground Region Detection

Each pixel is classified as either a background or a foreground pixel using the background model. Giving the minimum $m(x)$, maximum $n(x)$, and the median of largest interframe absolute difference d_μ images over the entire image that represent the background scene model $B(x)$, pixel x from the image I^t is a foreground pixel if:

$$B(x) = \begin{cases} 0 & \text{background} \\ 1 & \text{foreground} \end{cases} \begin{cases} |I^t(x) - m(x)| < kd_\mu \\ \text{or } |I^t(x) - n(x)| < kd_\mu \\ \text{otherwise.} \end{cases} \quad (2.2)$$

The threshold k is determined by experiment according to different environments.

2.2 The HSV Color Space

The HSV (hue, saturation and value) color space corresponds closely to the human perception of color. Conceptually, the HSV color space is a cone as shown in Fig. 2.1 (a). Fig. 2.1 (b) shows a circular and horizontal cross-section of HSV value “1,” the hue is represented by the angle of each color in the circle relative to the 0° line, which is traditionally assigned to be red. The saturation is represented as the distance from the center of the circle. The colors with high saturation are on the outer edge of the circle, whereas gray tones (which have no saturation) are at the very center. The value is determined by the color vertical position in the cone. At the point end of the cone, there is no brightness, so all colors are black. At the fat end of the cone are the bright colors.

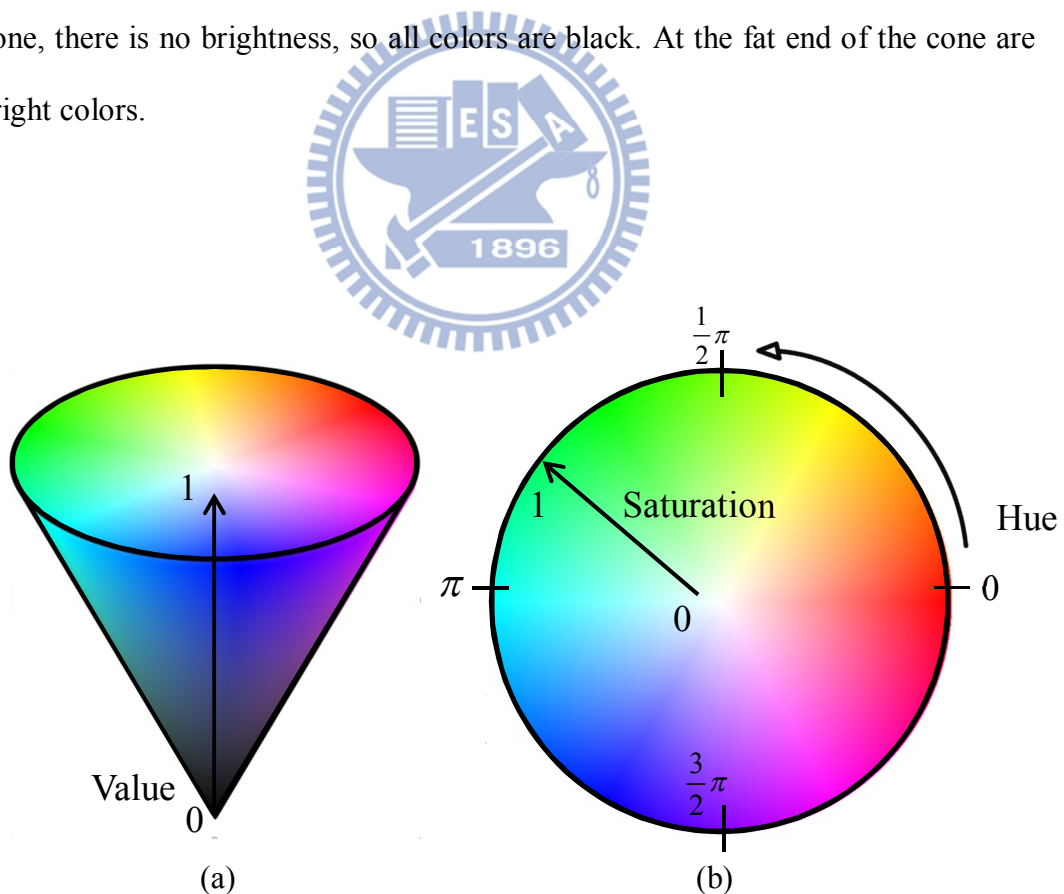


Fig. 2.1 (a) The HSV Cone. (b) Cross-section of HSV value “1.”

The formula of RGB transfers to HSV is defined as:

$$\begin{aligned}
 H &= \begin{cases} 0^\circ, & \text{if } \max^{RGB} = \min^{RGB} \\ 60^\circ \times \frac{G - B}{\max^{RGB} - \min^{RGB}} + 0^\circ, & \text{if } \max^{RGB} = R \text{ and } G \geq B \\ 60^\circ \times \frac{G - B}{\max^{RGB} - \min^{RGB}} + 360^\circ, & \text{if } \max^{RGB} = R \text{ and } G < B \\ 60^\circ \times \frac{B - R}{\max^{RGB} - \min^{RGB}} + 120^\circ, & \text{if } \max^{RGB} = G \\ 60^\circ \times \frac{R - G}{\max^{RGB} - \min^{RGB}} + 240^\circ, & \text{if } \max^{RGB} = B \end{cases} \\
 S &= \begin{cases} 0, & \text{if } \max^{RGB} = 0 \\ \frac{\max^{RGB} - \min^{RGB}}{\max^{RGB}}, & \text{otherwise} \end{cases}
 \end{aligned}$$

$$V = \max^{RGB}, \quad (2.3)$$

where $\max^{RGB} = \max(R, G, B)$ and $\min^{RGB} = \min(R, G, B)$

The hue parameter is the value which represents color information without brightness. Therefore, the hue is not affected by changing the illumination brightness and direction. Although the hue is the most useful attribute, there are three problems in using hue attribute for color segmentation: (1) hue is meaningless when the

intensity value is very low; (2) hue is unstable when the saturation is very low; and (3) saturation is meaningless when the intensity value is very low [10]. Accordingly, Ohba et al [11]. use three criteria (intensity value, saturation, and hue) to obtain the hue value reliably

■ **Intensity Threshold Value:**

If $V < V_t$, then $H = 0$, where V , V_t , and H are an intensity value, the intensity threshold value, and a hue value, respectively. If measured color is not bright enough, the color is discarded. Then, the hue value is set to a predetermined value, i.e., 0.

■ **Saturation Threshold Value:**

If $S < S_t$, then $H = 0$, where S , and H are a saturation value, the saturation threshold value, and a hue value, respectively. Using this equation, measured color close to gray is discarded in the image.

■ **Hue Threshold Value:**

If $H < \Delta P_t$ or $\|H - 2\pi\| < \Delta P_t$, then $H = 0$. The range of hue value is from 0 to 2π , and it has discontinuity at 0 and 2π . We use the phase threshold value to avoid the discontinuity effect.

2.3 Eigenspace and Canonical Space Transform

In computer vision systems need to deal with many images, dimensions of the image data are often extremely large. Because there are great deals of redundancies in images, it is common to the transform image from high-dimensional space to low-dimensional space to reduce redundancy. Many methods like Fourier Transformation, wavelet, Multi-Dimensional Scaling (MDS), Principal Component Analysis (PCA) and Eigenspace transformation (EST) has actually been demonstrated to reduce the dimension of data.

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 recognition rate of different various actions, Etemad and Chellappa [12] 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 variances. Unfortunately, this approach has high computation cost when using large images. We call this approach canonical space transformation (CST) Fig. 2.2 illustrates the processing steps that generate feature vectors by eigenspace transformation and canonical space transformation [13].

Combining EST based on PCA and CST based on CA, our approach reduces the data dimensionality and optimizes the class separability of different gait sequences and action classes. Image data in high-dimensional image space are converted to low-dimensional eigenspace using EST. The obtained vector this is further projected to a smaller canonical space using CST. Recognition is accomplished in the canonical space.

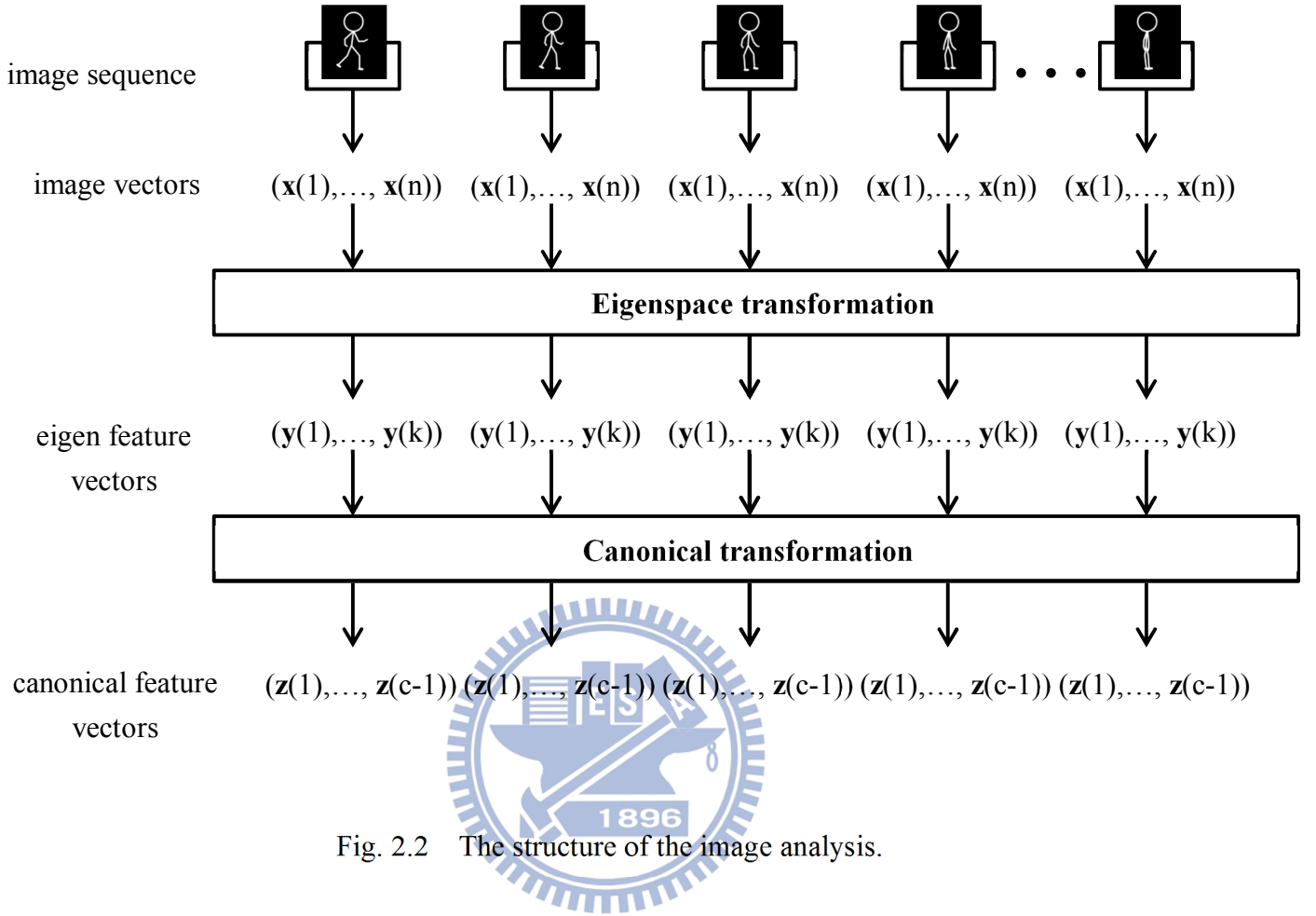


Fig. 2.2 The structure of the image analysis.

Assume that there are c training 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 the training set is $N_T = N_1 + N_2 + \dots + N_c$. This training set can be written as

$$\left[\mathbf{x}'_{1,1}, \dots, \mathbf{x}'_{1,N_1}, \dots, \mathbf{x}'_{2,1}, \dots, \mathbf{x}'_{c,N_c} \right] \quad (2.4)$$

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

At first, the brightness of each training image is normalized by

$$\mathbf{x}_{i,j} = \frac{\mathbf{x}'_{i,j}}{\|\mathbf{x}'_{i,j}\|}. \quad (2.5)$$

After normalization, we can get the mean pixel value for the full image set is given by

$$\mathbf{m}_x = \frac{1}{N_T} \sum_{i=1}^c \sum_{j=1}^{N_i} \mathbf{x}_{i,j}. \quad (2.6)$$

The training set can be rewritten as a $n \times N_T$ matrix \mathbf{X} by subtracting \mathbf{m}_x . And each image forms a column of \mathbf{X} , that is

$$\mathbf{X} = [\mathbf{x}_{1,1} - \mathbf{m}_x, \dots, \mathbf{x}_{1,N_1} - \mathbf{m}_x, \dots, \mathbf{x}_{c,N_c} - \mathbf{m}_x]. \quad (2.7)$$

2.3.1 Eigenspace Transformation (EST)

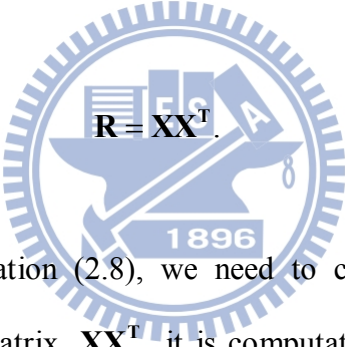
EST is widely used in the recognition of human faces and gait. Basically it is used to reduce the dimensionality of an input space by mapping the data from a correlated high-dimensional space to an uncorrelated low-dimensional space while maintaining the minimum mean-square error for the data information 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 the K nonzero eigenvalues of \mathbf{XX}^T , $\lambda_1, \dots, \lambda_K$, and their associated eigenvectors, $\mathbf{e}_1, \dots, \mathbf{e}_K$ satisfy the fundamental eigenvalue relationship

$$\lambda_i \mathbf{e}_i = \mathbf{R} \mathbf{e}_i, \quad i = 1, \dots, K, \quad (2.8)$$

where \mathbf{R} is a square, symmetric $n \times n$ matrix derived from \mathbf{X} and its transpose \mathbf{X}^T by

$$\mathbf{R} = \mathbf{XX}^T. \quad (2.9)$$


In order to solve Equation (2.8), we need to calculate the eigenvalues and eigenvectors of the $n \times n$ matrix \mathbf{XX}^T , it is computationally intractable for typical image sizes. Based on singular value decomposition theory, we can get the eigenvalues and eigenvectors by computing matrix $\tilde{\mathbf{R}}$ instead, that is

$$\tilde{\mathbf{R}} = \mathbf{X}^T \mathbf{X}, \quad (2.10)$$

in which the size of the matrix $\tilde{\mathbf{R}}$ is $N_T \times N_T$, which is much smaller than $n \times n$ of \mathbf{R} . Suppose the matrix \mathbf{R} has K nonzero eigenvalues $\lambda_1, \dots, \lambda_K$ and associated K eigenvectors $\mathbf{e}_1, \dots, \mathbf{e}_K$ which are related to those in \mathbf{R} by

$$\begin{cases} \lambda_i = \tilde{\lambda}_i \\ \mathbf{e}_i = (\tilde{\lambda}_i)^{-\frac{1}{2}} \mathbf{X} \tilde{\mathbf{e}}_i \end{cases}, \quad i = 1, \dots, K. \quad (2.11)$$

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

$$\mathbf{y}_{i,j} = [\mathbf{e}_1, \dots, \mathbf{e}_k]^T \mathbf{x}_{i,j}, \quad i = 1, \dots, c; j = 1, \dots, N_c. \quad (2.12)$$

We called this matrix $[\mathbf{e}_1, \dots, \mathbf{e}_k]^T$ the eigenspace transformation matrix. After this transformation, each 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.

2.3.2 Canonical Space Transformation (CST)

Based on canonical analysis in [14], we explain the basic concepts of CST. Suppose $\{\Phi_1, \Phi_2, \dots, \Phi_c\}$ represents the classes of transformed vectors by eigenspace transformation and $\mathbf{y}_{i,j}$ is the j -th vector in the class i . The mean

vector of entire set can be written as

$$\mathbf{m}_y = \frac{1}{N_T} \sum_i^c \sum_j^{N_i} \mathbf{y}_{i,j}, \quad (2.13)$$

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

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{\mathbf{y}_{i,j} \in \Phi_i} \mathbf{y}_{i,j}. \quad (2.14)$$

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

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

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

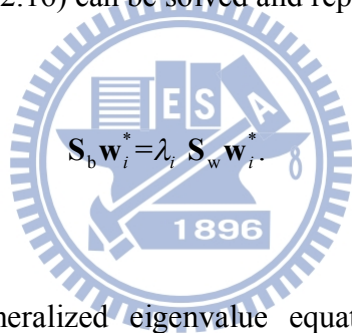
Where \mathbf{S}_w represents the mean of within-class vectors distance and \mathbf{S}_b represents the mean of between-class vectors distance. The objective is to minimize \mathbf{S}_w and maximize \mathbf{S}_b simultaneously and it is to minimize the criterion function known as the generalized Fisher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

$$\mathbf{J}(\mathbf{W}) = \frac{\mathbf{W}^T \mathbf{S}_b \mathbf{W}}{\mathbf{W}^T \mathbf{S}_w \mathbf{W}}. \quad (2.15)$$

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

$$\frac{\partial \mathbf{J}}{\partial \mathbf{W}} = 0. \quad (2.16)$$

Suppose that \mathbf{W}^* is the optimal solution where the column vector \mathbf{w}_i^* is a generalized eigenvector and corresponds to the i -th largest eigenvalue λ_i . According to the theory [14], equation (2.16) can be solved and represented as



$$\mathbf{S}_b \mathbf{w}_i^* = \lambda_i \mathbf{S}_w \mathbf{w}_i^*. \quad (2.17)$$

After we get the generalized eigenvalue equation, we can obtain $(c-1)$ nonzero eigenvalues and their corresponding eigenvectors $[\mathbf{v}_1, \dots, \mathbf{v}_{c-1}]$ that create another orthogonal basis and span a $(c-1)$ -dimensional canonical space. By using this basis, each point in eigenspace can be projected to another point in the canonical space by

$$\mathbf{z}_{i,j} = [\mathbf{v}_1, \dots, \mathbf{v}_{c-1}]^T \mathbf{y}_{i,j}, \quad (2.18)$$

where $\mathbf{z}_{i,j}$ represents the new point and the orthogonal basis $[\mathbf{v}_1, \dots, \mathbf{v}_{c-1}]^T$ is called the canonical space transformation matrix. By merging equation (2.12) and equation

(2.18), each image can be projected into one point in the new $(c-1)$ -dimensional space by

$$\mathbf{z}_{i,j} = \mathbf{H} \mathbf{x}_{i,j}, \quad (2.19)$$

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



Chapter 3 Activity and Human Recognition System

3.1 Foreground Extraction

The first step of human activity recognition and person identification system is foreground subject extraction. We extract foreground subject by using background model methods. There are many well-known background models. W^4 is such a famous example [3]. It records the maximum, minimum and maximum inter-frame difference grayscale of each pixel in background video frames. If the pixel's grayscale is in the interval between the maximum and minimum grayscale with toleration, the pixel is classified to a foreground one. But we cannot detect reliably those foreground pixels whose luminance component close to background pixel. In order to solve this problem, we build another background model in the HSV color space. We detect reliably foreground pixels in grayscale and the HSV domain.

3.1.1 Background Model

Fig. 3.1 show the framework we construct for the background models.

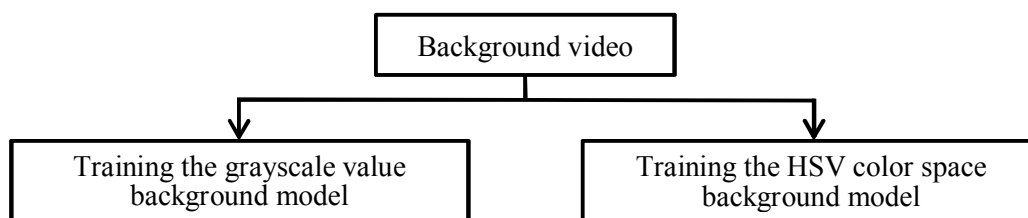
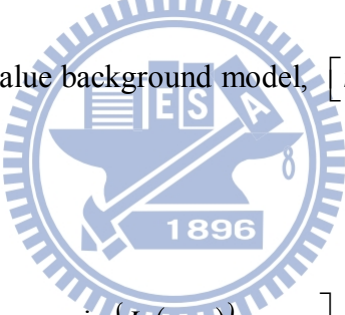


Fig. 3.1. The framework we construct the background models.

A. Grayscale Value Background Model

In the grayscale value background model, each pixel of background scene is characterized by three statistics: minimum grayscale value $m(x, y)$, maximum grayscale value $n(x, y)$ and maximum inter-frame difference $d(x, y)$ of a background video. Because these three values must be obtained through statistics, so 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. $I_i^{gray}(x, y)$ is the grayscale value of a pixel which is located at (x, y) in the i -th frame of I . The grayscale value background model, $[m(x, y), n(x, y), d(x, y)]$, of a pixel is obtained by


$$\begin{bmatrix} m(x, y) \\ n(x, y) \\ d(x, y) \end{bmatrix} = \begin{bmatrix} \min_i \{I_i(x, y)\} \\ \max_i \{I_i(x, y)\} \\ \max_i \{|I_i(x, y) - I_{i-1}(x, y)|\} \end{bmatrix}, \quad i = 1, 2, \dots, N. \quad (3.1)$$

B. HSV Color Space Background Model

Along similar line of reasoning of above, we build another background model like a grayscale value background model in each dimension of HSV (hue, saturation and value) space. Then, we also record the inter-frame ratio in the brightness information and the inter-frame different in the chromatic information. We use the

same background video to build the HSV color space background model background model. $I_i^H(x, y)$ is the pixel's hue value at (x, y) of the i -th image frame. $I_i^S(x, y)$ is the pixel's saturation value at (x, y) of the i -th image frame. $I_i^V(x, y)$ is the pixel's brightness value at (x, y) of the i -th image frame. The HSV background model of a pixel is obtained by

$$\begin{bmatrix} m^H(x, y) \\ n^H(x, y) \\ d^H(x, y) \end{bmatrix} = \begin{bmatrix} \min_i \{I_i^H(x, y)\} \\ \max_i \{I_i^H(x, y)\} \\ \max_i \{|I_i^H(x, y) - I_{i-1}^H(x, y)|\} \end{bmatrix}, \quad i = 1, 2, \dots, N. \quad (3.2)$$

$$\begin{bmatrix} m^S(x, y) \\ n^S(x, y) \\ d^S(x, y) \end{bmatrix} = \begin{bmatrix} \min_i \{I_i^S(x, y)\} \\ \max_i \{I_i^S(x, y)\} \\ \max_i \{|I_i^S(x, y) - I_{i-1}^S(x, y)|\} \end{bmatrix}, \quad i = 1, 2, \dots, N. \quad (3.3)$$

$$\begin{bmatrix} m^V(x, y) \\ n^V(x, y) \\ d^V(x, y) \end{bmatrix} = \begin{cases} \begin{bmatrix} \min_i \{I_i^V(x, y)\} \\ \max_i \{I_i^V(x, y)\} \\ \max_i \{|I_i^V(x, y) / I_{i-1}^V(x, y)|\} \end{bmatrix}, & \text{if } I_i^V(x, y) / I_{i-1}^V(x, y) \geq 1, \\ \begin{bmatrix} \min_i \{I_i^V(x, y)\} \\ \max_i \{I_i^V(x, y)\} \\ \max_i \{|I_{i-1}^V(x, y) / I_i^V(x, y)|\} \end{bmatrix}, & \text{otherwise,} \end{cases} \quad (3.4)$$

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

3.1.2 Extraction of Foreground Object

Fig. 3.2 shows the framework we apply to foreground subject extraction. Our framework of foreground subject extraction is composed of four components. The first component is foreground subject extraction. The second component is the shadow suppression. The third component is the object segmentation. And the last component is the foreground image compensation to recover the foreground pixels those are wrongly classified to the background.

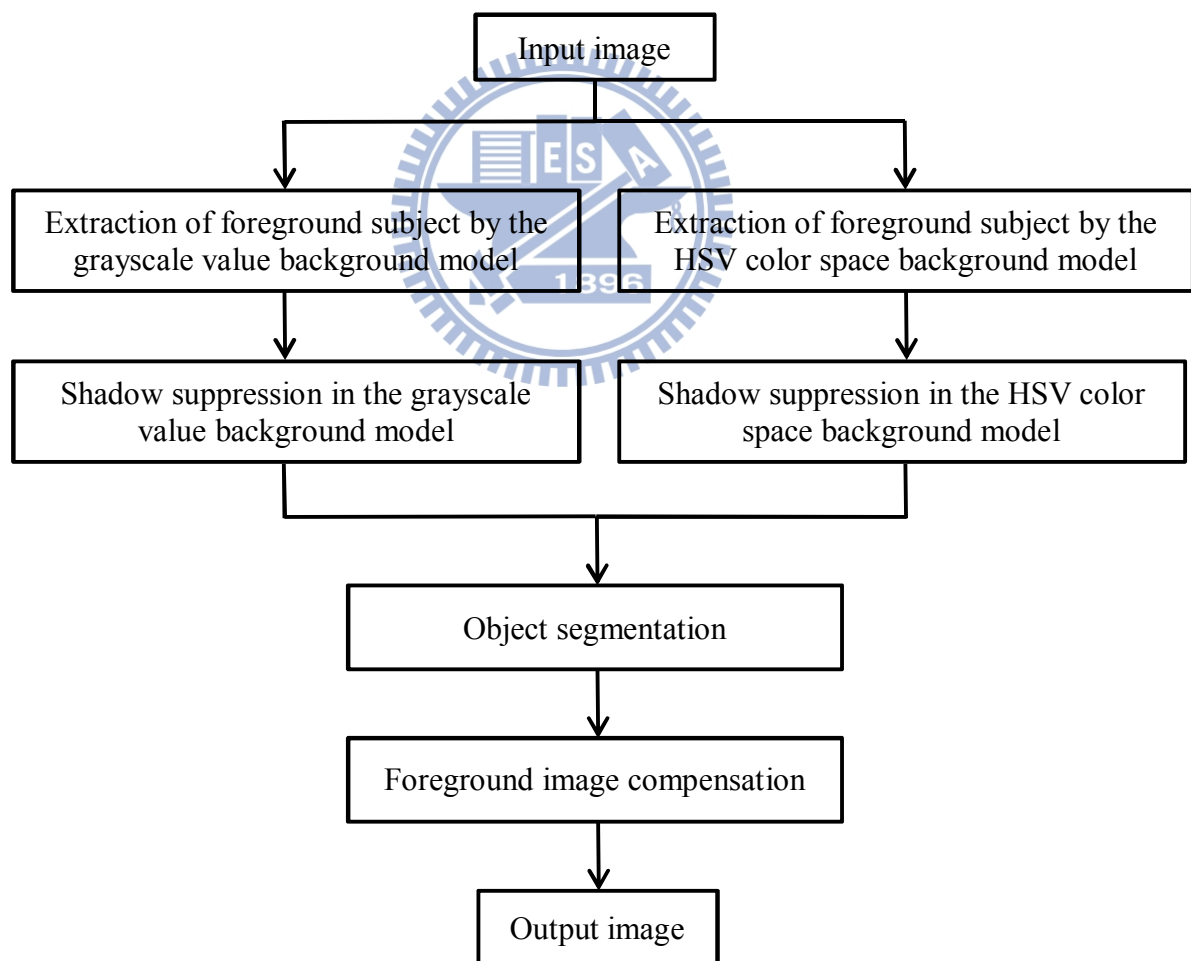


Fig. 3.2. The framework we develop for foreground subject extraction.

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, then foreground objects can be segmented from every frame of the video stream.

First, we utilize the minimum grayscale value $m(x, y)$, maximum grayscale value $n(x, y)$ and maximum inter-frame difference $d(x, y)$ of the grayscale value background model to segment a foreground by

$$I_{fg}^{gray}(x, y) = \begin{cases} 0, & \text{if } |I_i(x, y) - m(x, y)| < k\mu \\ & \text{or } |I_i(x, y) - n(x, y)| < k\mu \\ 255, & \text{otherwise,} \end{cases} \quad (3.5)$$

where $I_i(x, y)$ is the intensity of a pixel which is located at (x, y) , $I_{fg}^{gray}(x, y)$ is the gray level of a pixel in the foreground binary image, μ is the median of all $d(x, y)$, and k is determined by experiments according to different environments.

We usually use $k = 2$ in our system.

In addition to segment a foreground by Eq. (3.5), we also utilize the minimum value $m^V(x, y)$, the maximum value $n^V(x, y)$ and maximum inter-frame value ratio $d^V(x, y)$ of the HSV color space background model to segment the foreground pixel by

$$I_{fg}^{HSV}(x, y) = \begin{cases} 0, & \text{if } I_i^V(x, y) / m^V(x, y) < k_v d^V(x, y) \\ & \text{or } I_i^V(x, y) / n^V(x, y) < k_v d^V(x, y) \\ 255, & \text{otherwise,} \end{cases} \quad (3.6)$$

where $I_i^V(x, y)$ is the intensity of a pixel which is located at (x, y) , $I_{fg}^{HSV}(x, y)$ is the gray level of a pixel in a binary image, threshold k_V is determined by the light as of the scene. Threshold k_V will be reduced for in-sufficient light condition and increased otherwise.

3.1.3 Shadow Suppression

The shadows of the object are easily classified as foreground pixels in normal condition. The situation causes an object merging and object shape distortion in the binary foreground image. Therefore, we need to remove the shadow by using the shadow filter. We assume that the observed intensity of shadow pixels is directly proportional to incident light. Consequently, shadowed pixels are scaled versions (darker) of corresponding pixels in the background model.

In the first place, we utilize the estimate of Normalized Cross-Correlation (NCC) [15] to quantify the similarity between the background image and an image of the video sequence. The NCC estimate method is described as follows. Let $B(x, y)$ be the background image formed by temporal median filtering, and $I(x, y)$ be an image of the video sequence. For each pixel (x, y) belonging to the foreground, consider a 3×3 template such that $T_{xy}(m, n) = I(x + m, y + n)$, for $-1 \leq m \leq 1, -1 \leq n \leq 1$ (i.e. T_{xy} corresponds to a neighborhood of pixel (x, y)). Then, the NCC between template T_{xy} and image B at pixel (x, y) is given by:

$$NCC(x, y) = \frac{ER(x, y)}{E_b(x, y)E_{T_{xy}}}, \quad (3.7)$$

where

$$ER(x, y) = \sum_{n=-1}^1 \sum_{m=-1}^1 B(x+m, y+n) T_{xy}(m, n),$$

$$E_B(x, y) = \sqrt{\sum_{n=-1}^1 \sum_{m=-1}^1 B(x+m, y+n)^2}, \quad (3.8)$$

$$E_{T_{xy}} = \sqrt{\sum_{n=-1}^1 \sum_{m=-1}^1 T_{xy}(m, n)^2}.$$

If a pixel (x, y) is in a shadowed region, the NCC in a neighboring region T_{xy} should be large, and the energy $E_{T_{xy}}$ of this region should be lower than the energy $E_B(x, y)$ of the corresponding region in the background images. Therefore, we get

$$S^{gray}(x, y) = \begin{cases} \text{shadow,} & NCC(x, y) \geq L_{ncc} \text{ and } E_{T_{xy}} < E_B(x, y) \\ \text{foreground,} & \text{otherwise.} \end{cases} \quad (3.9)$$

Where L_{ncc} is a fixed threshold. If L_{ncc} is low, several foreground pixels may be misclassified as shadow pixels. On the other hand, selecting a large value of L_{ncc} , then the shadow pixels may not be detected.

We know that the shadow pixels have similar chromaticity but lower brightness than the background model. Therefore, we can detect the shadow region in the HSV

color space. We will build another shadow filter S^{HSV} for each (x, y) point as follows:

$$S^{HSV}(x, y) = \begin{cases} \text{shadow,} & \text{if } \frac{I_i^V(x, y)}{m^V(x, y)} < 1 \\ & \text{and } |I_i^H(x, y) - n^H(x, y)| < k_H d^H(x, y) \\ & \text{and } |I_i^S(x, y) - n^S(x, y)| < k_S d^S(x, y) \\ \text{foreground,} & \text{otherwise,} \end{cases} \quad (3.10)$$

where $I_i^H(x, y)$, $I_i^S(x, y)$, $I_i^V(x, y)$ are respectively the HSV channel of a pixel located at (x, y) . Values k_S and k_H are selected threshold values that used to measure the similarities of the hue and saturation between the background image and the current observed image. We extract the foreground objects from the two background models which is the shadow or foreground is obtained through the S^{gray} and S^{HSV} . We set a hard threshold for each background model, then we obtain the foreground objects which have less noise, but missing some foreground objects. Therefore, using the union is better than the intersection. Because of using the union can increase the foreground with less noise. Finally, the foreground subject is defined as:

$$I_{fg}(x, y) = S^{gray}(x, y) \vee S^{HSV}(x, y). \quad (3.11)$$

3.1.4 Object Segmentation

According to the binary image I_{fg} segmented by above, we extract the region

of the foreground object to minimize the image size. Foreground region extraction can be accomplished by simply introducing a threshold on the histograms in the X and Y directions. Fig. 3.3 shows an example of foreground region extraction. We utilize the binary image and project it into the X and Y directions. The interested foreground 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 four corners of a rectangle to extract a foreground region and the size of this rectangle is adjusted to 96×128 for normalization. Fig. 3.4 is the extracted foreground region.

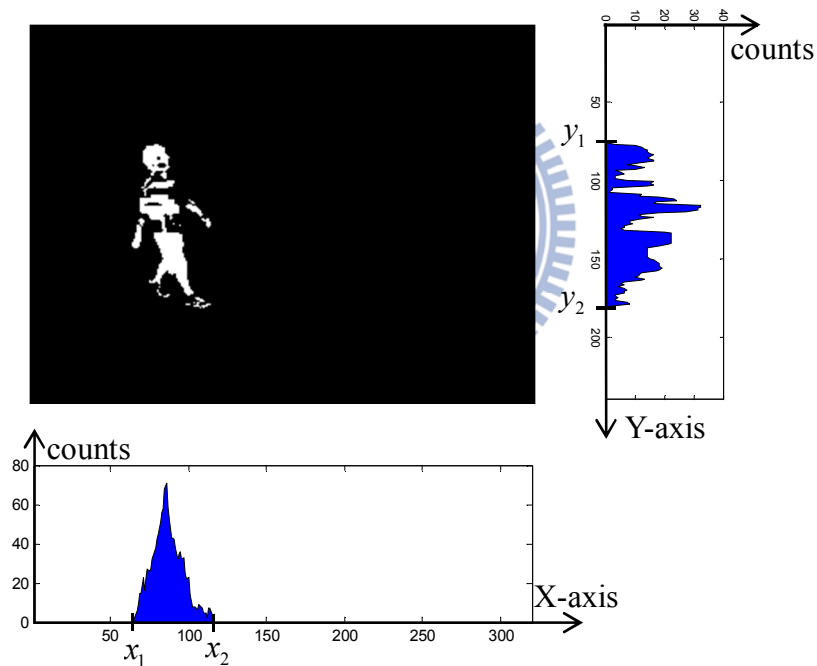


Fig. 3.3. Histogram of binary image projection in X and Y direction.



Fig. 3.4. The binary image of extracted foreground region.

3.1.5 Foreground Image Compensation

It is difficult to detect all the foreground pixels and remove all the shadows in each frame. When we want to remove shadow pixels, some foreground data will be lost and this makes the foreground image broken. In order to solve the problem, we will repair the foreground image by opening filter and closing filter [16], Fig. 3.5 shows the image which is to be repaired.

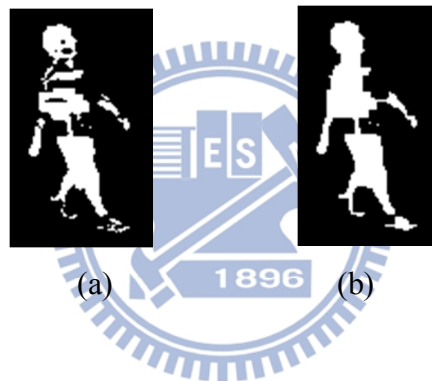


Fig. 3.5. (a) Foreground image. (b) Foreground image after opening and closing repair of (a).

3.2 Background Update

If the facilities in room are moved, they will be detected as foreground pixels of human and the human activity recognition will be misclassified. Therefore, we have to adopt a scheme that can update background models in order to avoid above situation. If the video does not vary for a long time and there is nobody in the scene the background models will be updated. By Eq. (3.12), we calculate how many times the binary values are unchanged.

$$update(x, y) = \begin{cases} update(x, y) + 1, & \text{if } I_{fg}^{t-1}(x, y) = I_{fg}^t(x, y) \\ update(x, y), & \text{otherwise,} \end{cases} \quad (3.12)$$

where $I_{fg}^t(x, y)$ is the gray level of a pixel in binary image and it is located at (x, y) . Value $update(x, y)$ is a record of how many times $I_{fg}^t(x, y)$ remains unchanged. When $update(x, y)$ exceeds a threshold, the pixel (x, y) will be included in the background model.

3.3 Skin Color Detection

By skin color detection, we can analyze whether there is a person in the scene. If a person in the scene, the background is not updated.

First, the input image is transferred into the normalized RGB color space by:

$$r = \frac{R}{R + G + B}, \quad (3.13)$$

$$g = \frac{G}{R + G + B}. \quad (3.14)$$

According to Soriano and Martinkauppi [17], a boundary condition of skin color in the r-g plane is defined as

$$f_{upper}(r) = -1.3767r^2 + 1.0743r + 0.1452, \quad (3.15)$$

$$f_{lower}(r) = -0.7760r^2 + 0.5601r + 0.1766. \quad (3.16)$$

If a pixel satisfies the following four conditions, it will be labeled as skin pixel. Therefore, we know there is a person or not from the following skin pixel marking:

$$g > f_{lower}(r) \text{ and } g < f_{upper}(r), \quad (3.17)$$

$$(r - 0.33)^2 + (g - 0.33)^2 \leq 0.0004, \quad (3.18)$$

$$R > G > B, \quad (3.19)$$

$$R - G \geq 45. \quad (3.20)$$

3.4 Template Selection

Cameras usually capture image frames in high frequency (30 frames / Sec.), but human action transforms are much slower than the camera capturing speed. There are few differences between two consecutive image frames. Therefore, we select a key frame, called as essential template image, from a sequence with a fixed interval to represent an action and gait. In our approach, we select an essential template image every 5 frames and the schematic diagram is shown in Fig. 3.7. The number of essential template image about an action and gait is dependent on the period of the action.

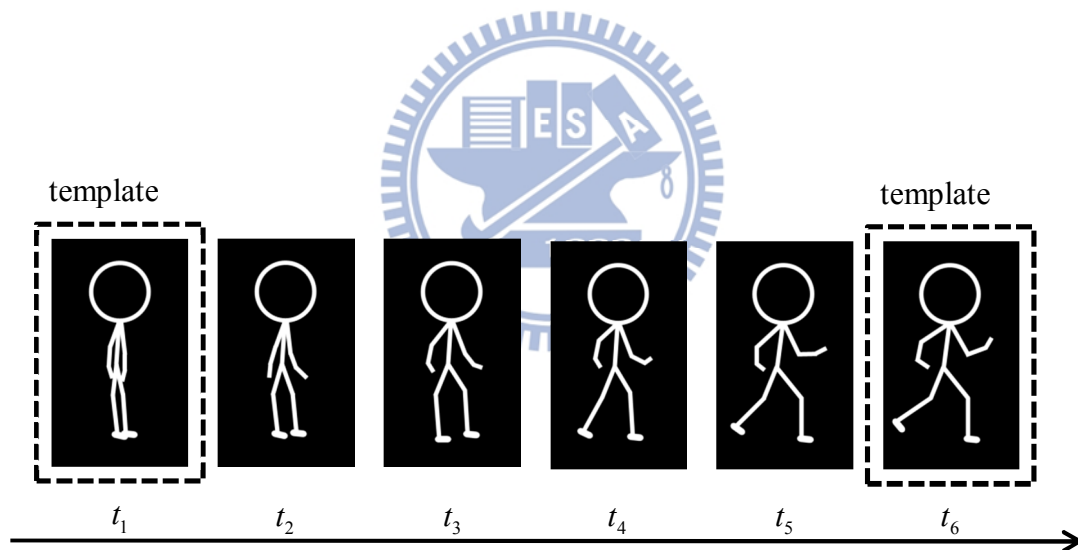


Fig. 3.6. Using 5:1 down-sampling rate to select the essential template image.

These essential templates are transformed into a new space by eigenspace transformation (EST) and canonical space transformation (CST). The approximation will lose slight information of the image with little differences, but it can decrease massive data dimensions. However, two similar image frames will converge to two

nearest points after eigenspace and canonical space transformation.

As described in Section 2.3, each image frame is transformed into a $(c-1)$ -dimensional vector by EST and CST methods [13]. 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

$$\mathbf{t}_{i,j} = \mathbf{H} \cdot \mathbf{g}_{i,j}, \quad i=1, 2, \dots, c ; j=1, 2, \dots, n \quad (3.21)$$

where \mathbf{H} denotes 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 rewritten as

$$\mathbf{t}_{i,j} = [t_{i,j}^1, t_{i,j}^2, \dots, t_{i,j}^{c-1}]^T. \quad (3.22)$$

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

$$\boldsymbol{\mu}_i = \frac{1}{n} \sum_{j=1}^n \mathbf{t}_{i,j}, \quad (3.23)$$

where i is the number of template categories.

The standard deviation vector of the m -th dimension is computed by

$$\sigma_m = \sqrt{\frac{\sum_{i=1}^c \sum_{j=1}^n (\mathbf{t}_{i,j}^m - \boldsymbol{\mu}_i^m)^2}{N_t - 1}}. \quad (3.24)$$

3.5 Construction of Fuzzy Rules from Video Stream

Transitional relationships of postures in a temporal sequence are important information for human activity classification. If we only utilize one image frame to recognize actions or gait, it may be not sufficient to obtain high correct rate because human's actions may have similar postures in two different action sequences or gait sequences. For example, the actions of "jumping" and "crouching" both have the same postures called common states as shown in Fig. 2.10. Besides, the posture sequence of each activity is dissimilar in different people.

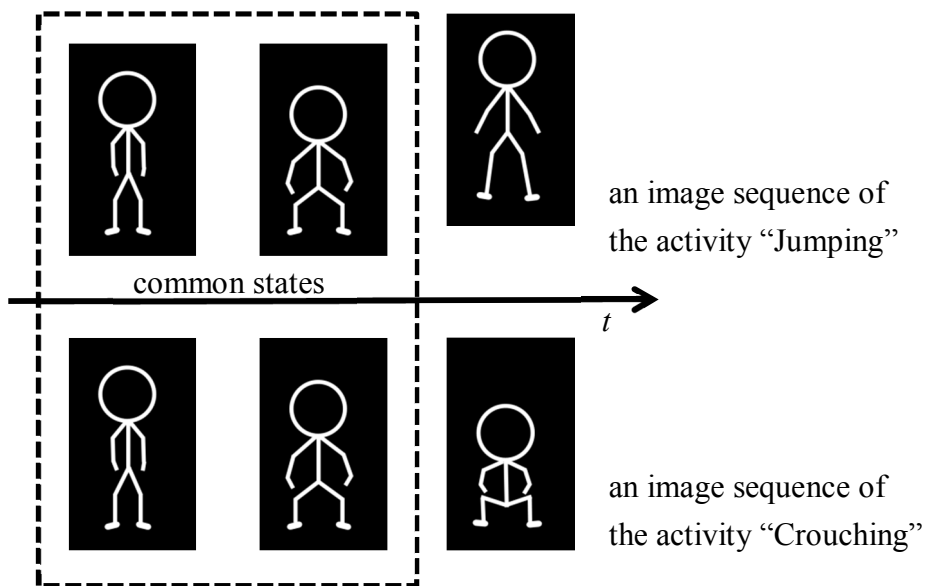



Fig. 3.7. Common states of two different activities.

Hence, we use the fuzzy rule-based approach to solve aforesaid problem. The approach combines temporal sequence information for recognition.

We use the membership function to represent the feature's possibility of each cluster. Many types of membership functions are frequently used in 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



$$\mathbf{a}_k = \mathbf{H} \mathbf{x}_{k^2} \quad (3.25)$$

where \mathbf{H} denotes the transformation matrix of EST and CST. As the same as $\mathbf{t}_{i,j}$ in Eq. (3.22), \mathbf{a}_k can be rewritten as

$$\mathbf{a}_k = [a_k^1, a_k^2, \dots, a_k^{c-1}]^T. \quad (3.26)$$

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

$$\begin{aligned}
r_{i,k} &= M(\mathbf{a}_k | C_i) \\
&= \frac{1}{(2\pi)^{\frac{c-1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\mathbf{a}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{a}_k - \boldsymbol{\mu}) \right] \\
&= \prod_{m=1}^{c-1} \frac{1}{\sqrt{2\pi} \sigma_m} \exp \left[-\frac{1}{2} \frac{(a_k^m - \mu_{i,j}^m)^2}{\sigma_m^2} \right]
\end{aligned} \tag{3.27}$$

where m is the number of dimension and j is the training model, i.e., action person, index. $r_{i,k}$ denotes the grade of membership function of category the k -th image frame. After that we can obtain which category posture the image belongs to by

$$P_k = \arg \max_i r_{i,k} \tag{3.28}$$

The membership function describes the probability of which one it is like most. But it just contains the information of a single image. Hence, we collect three images to form a basis for temporal information.

Assume we have c linguistic labels, each linguistic label represents a category of essential template. Each image frame can be represented by one of these c linguistic labels. Three contiguous images are combined as a group (I_1, I_2, I_3) in our approach. The transformation of the image group can form a feature vector $[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$. There are c^3 combinations of the feature vector. Each combination represents the possible transition states of the three images. We use Eq. (3.27) to class each image frame. Hence, we can represent the feature vector $[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$ by

linguistic label sequence $[p_1, p_2, p_3]$. An image sequence with a linguistic label sequence is associated with its output of corresponding activity.

As developed by Wang and Mendel [10], fuzzy rules can be generated by learning from training data. 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 example, an image sequence, its transformations of image 1, image 2, image 3 and belonging categories being concatenated as vector format, is given by

$$[P_1, P_2, P_3, D_1] \quad (3.29)$$

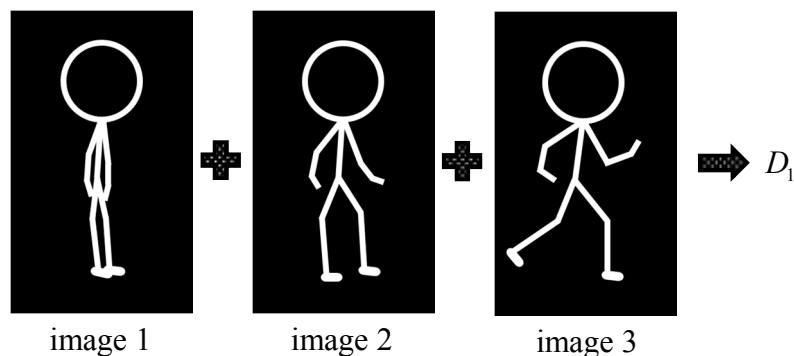


Fig. 3.8. A fuzzy rule learned to classify action.

Suppose that images 1, 2, 3 belong to categories 1, 2, 3 respectively. Therefore, the image sequence (Images 1, 2, 3) is transferred to (P_1, P_2, P_3) . Finally, a rule is supported by these three images as given by

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 action is D_1 .

After the learning step of different actions, some conflicting rules may be generated. The conflicting rules have the same image sequence but refer to different activity. Therefore, we have to choose one from conflicting rules. To this end, we choose the rule that is supported by a maximum number of training data. 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.

3.6 Classification Algorithm

After constructing the rule base, we can grade the input image sequence with each fuzzy rule by grade of membership function. Let Σ denote the covariance matrix of all essential template vectors, s_k denote the image frame transformed by EST and CST and C_i denote the i -th class of essential templates of postures. The membership function is given by

$$\begin{aligned}
r_{i,k} &= M(\mathbf{s}_k | C_i) \\
&= \frac{1}{(2\pi)^{\frac{c-1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\mathbf{s}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{s}_k - \boldsymbol{\mu}) \right] \\
&= \prod_{m=1}^{c-1} \frac{1}{\sqrt{2\pi} \sigma_m} \exp \left[-\frac{1}{2} \frac{(s_k^m - \mu_{i,j}^m)^2}{\sigma_m^2} \right]
\end{aligned} \tag{3.30}$$

where 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 as a group for recognizing an activity or gait analysis. Therefore, we use two transformed vectors of passed image frames, which are called \mathbf{s}_{k-2} and \mathbf{s}_{k-1} . These three vectors from a feature vector $[\mathbf{s}_{k-2}, \mathbf{s}_{k-1}, \mathbf{s}_k]$. We compute the membership functions of three vectors respectively.

In order to calculate the similarity between image sequence and each postural sequence in the training database, 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 categories of linguistic labels, P_{n_1} , P_{n_2} , and P_{n_3} , in the rule and have been calculated by Eq. (3.29). 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 database in the same manner. The rule, which has the highest value of similarity, is selected. Fig 3.10 shows the structure of the classification algorithm.

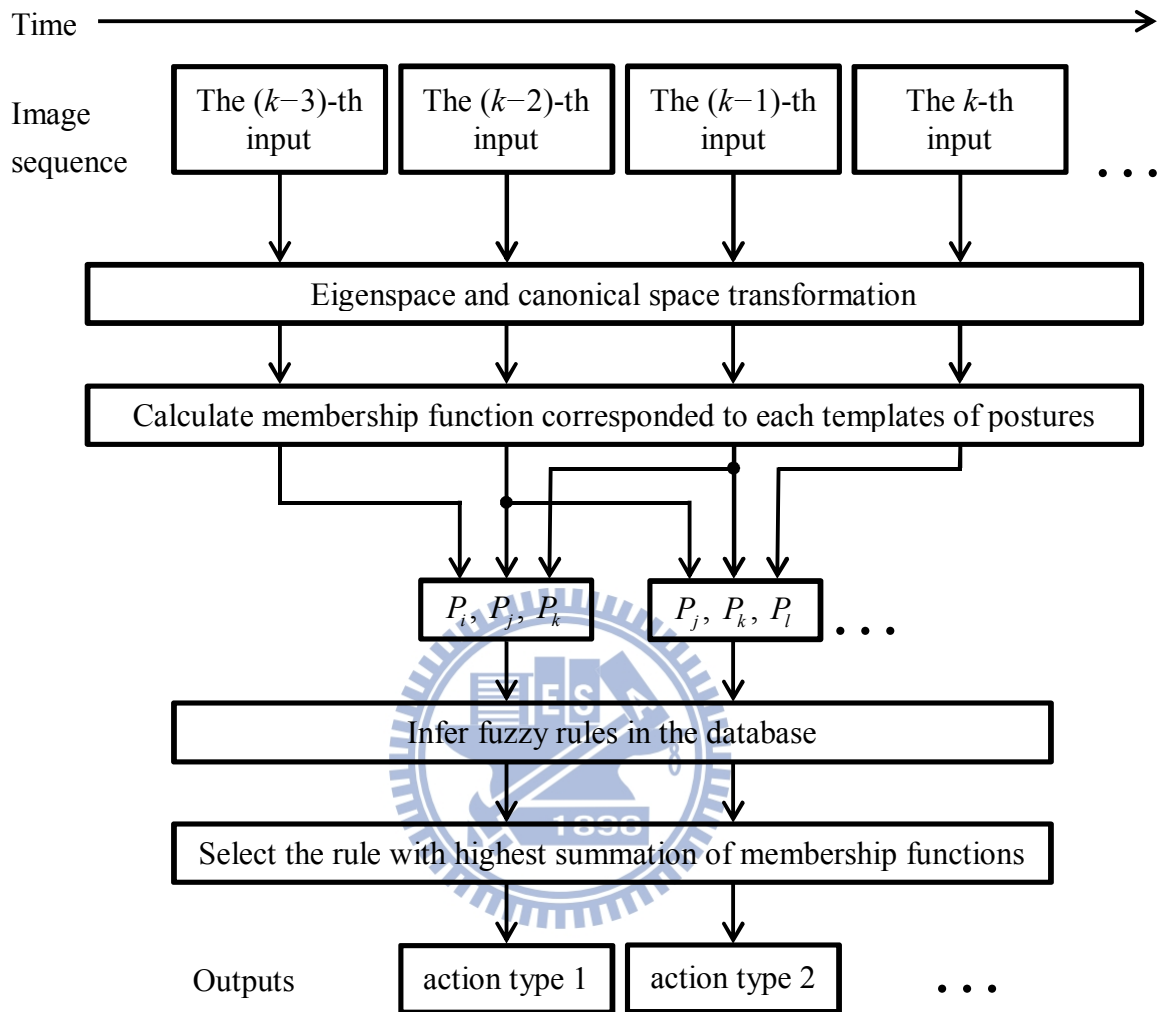


Fig. 3.9. The structure of action recognition algorithm.

Chapter 4 Experimental Results

In our experiment, we test our system on videos. The videos captured by NIR cameras in bright and dark environments. The experimental environment is in our laboratory which is in the 5-th Engineering Building on NCTU campus.

The action recognition background is the real life environment and illumination of the environment is 432 Lux in the day (bright environment) and 0.26 Lux in the night (dark environment), respectively. The NIR camera (KMT-1651N with lighting LED cells) with a lens of 4.3 mm focus is set up at the location that is far from the object about 5 meters. This camera has a frame rate of 30 frames per second and the image resolution is 320×240 pixels.

The action recognition scenes in bright and dark environments are shown in Fig. 4.1 we choose eleven actions: “walking from right to left,” “walking from left to right,” “bending,” “waving,” “sitting down on the left” “Sitting on the left” “standing up on the left” “sitting down on the right” “Sitting on the right” “standing up on the right” “walking straightly,” to recognize the action in our system. Fig. 4.2 and Fig. 4.3 shows the examples video sequence form our LAB databases.



Fig. 4.1. (a) The action recognition experiment environment in the day,
(b) The action recognition experiment environment in the night.

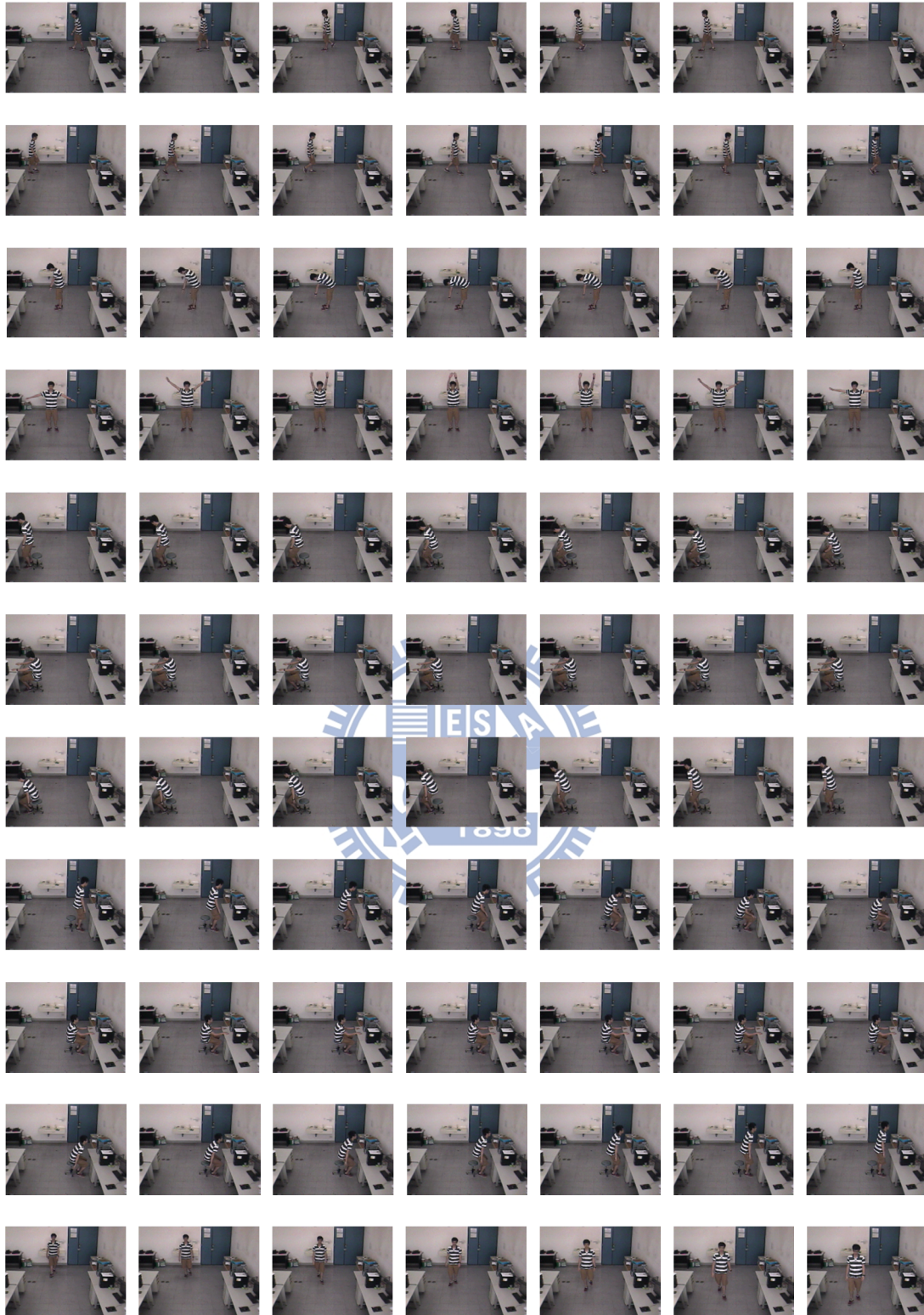


Fig. 4.2. Typical video sequences for actions of our LAB in bright environment (432 Lux). From top to bottom: “walking from right to left,” “walking from left to right,” “bending,” “waving,” “sitting down on the left,” “Sitting on the left,” “standing up on the left,” “sitting down on the right,” “Sitting on the right,” “standing up on the right,” “walking straightly.”



Fig. 4.3. Typical video sequences for actions of our LAB in dark environment (0.26 Lux). From top to bottom: “walking from right to left,” “walking from left to right,” “bending,” “waving,” “sitting down on the left,” “Sitting on the left,” “standing up on the left,” “sitting down on the right,” “Sitting on the right,” “standing up on the right,” “walking straightly.”

The gait recognition background is the real life environment and illumination of the environment is 432 Lux in the day and 0.26 Lux in the night respectively. The NIR camera with a lens of 4.3 mm focus is set up at the location that is far from the object about 4 meters. This camera has a frame rate of 30 frames per second and the image resolution is 320×240 pixel.

The gait recognition scenes in bright and dark environments are shown in Fig. 4.4 we choose “walking from left to right” to recognize the gait in our system. Fig. 4.5 and Fig. 4.6 shows the examples video sequence form our LAB databases.

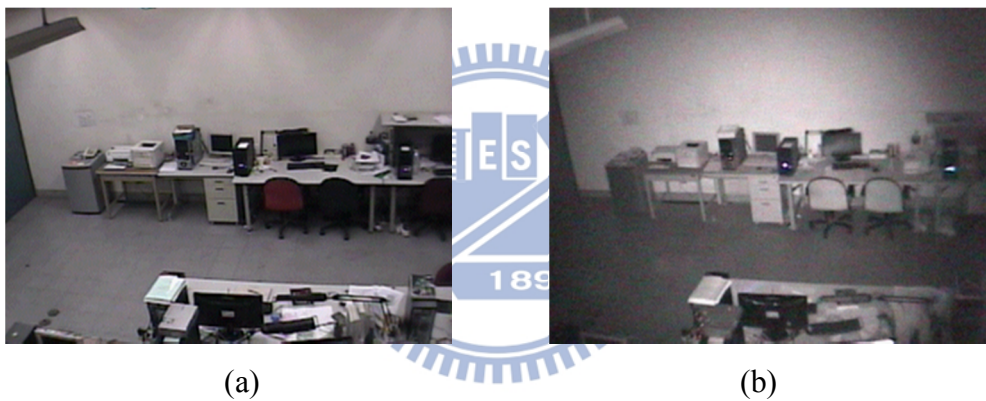


Fig. 4.4. (a) The gait recognition experiment environment in the day,
(b) The gait recognition experiment environment in the night.



Fig. 4.5. Typical video sequences for gait of our LAB in bright environments.



Fig. 4.6. Typical video sequences for gait of our LAB in dark environments.

4.1 Background Model and Foreground Object Extraction

In order to construct the background model, we first record a video of clear background about two second in bright and dark environments. We build the grayscale value and the HSV color space background models. We will extract the foreground pixel by using background models (see Eqs. 3.5 and 3.6). Then we use shadow filter to remove the shadow of the foreground (see Eq. 3.9). In order to obtain the good object extraction, we have to adjust some parameters in our system.

In the bright action recognition environment we set $k = 2.8$ for the grayscale value background model and $k_v = 1.6$ for the HSV color background model. In the dark action recognition environment we set $k = 1.7$ for the grayscale value background model and $k_v = 1.3$ for the HSV color background model; in the bright gait recognition environment we set $k = 3.1$ for the grayscale value background model and $k_v = 1.4$ for the HSV color background model. In the dark gait recognition environment we set $k = 1.4$ for the grayscale value background model and $k_v = 1.1$ for the HSV color background model.

The same parameter is used in bright and dark action recognition environment environments for shadow filter. We set $L_{ncc} = 0.95$ in the grayscale value space and $k_H = 1.3$ and $k_S = 1.3$ in the HSV color space to detect shadow pixels. Figs. 4.7 and 4.8 show the results of foreground extraction in bright and dark action recognition environment environments, respectively. Fig. 4.9 and Fig. 4.10 show the results of foreground extraction in bright and dark gait recognition environments, respectively.



(a)



(b)

(c)



(d)

Fig. 4.7. The results of foreground extraction in bright action recognition environment. (a) Background image. (b) An action image frame. (c) Binary image after foreground detection. (d) Foreground region extracted.



(a)



(b)



(c)



(d)

Fig. 4.8. The results of foreground extraction in the dark action recognition environment. (a) Background image. (b) An action image frame. (c) Binary image after foreground detection. (d) Foreground region extracted.



(a)



(b)



(c)



(d)

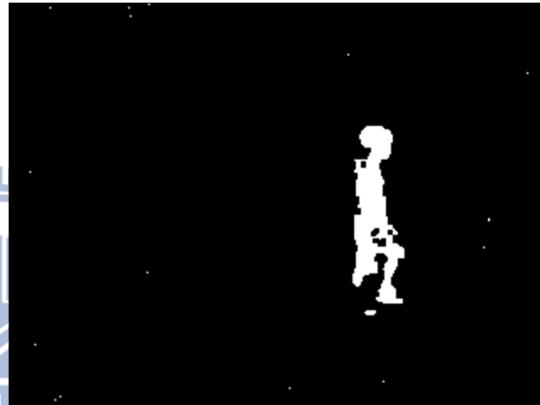
Fig. 4.9. The results of foreground extraction in the bright gait recognition environment. (a) Background image. (b) An action image frame. (c) Binary image after foreground detection. (d) Foreground region extracted.



(a)



(b)



(c)



(d)

Fig. 4.10. The results of foreground extraction in the dark gait recognition environment. (a) Background image. (b) An action image frame. (c) Binary image after foreground detection. (d) Foreground region extracted.

4.2 Fuzzy Rule Construction for Action Recognition

We construct the template model matrix and the fuzzy rule database with the training data. Firstly, we choose key posture images as essential templates from each action, and the number of each action key posture image is in proportion to its period at about 1/6 sec per key posture. Key posture images of each action for one person are shown in Fig. 4.10. We will regard each posture as one class of posture types.



(a)



(b)



(c)



(d)



(e)



(f)



(g)



(h)



(i)



(j)



(k)

Fig. 4.11. Key postures of the actions of person 1, (a) walking from right to left, (b) walking from left to right, (c) bending, (d) waving, (e) sitting down on the left, (f) Sitting on the left, (g) standing up on the left, (h) sitting down on the right, (i) Sitting on the right, (j) standing up on the right, (k) walking straightly.

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. (3.27) 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 are the same as in key posture template selection. Training is accomplished in off-line situation using recorded video. 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 to train fuzzy rule; the second frame, the 7-th frame and 12-th frame are gathered together to train another fuzzy rule; the third frame, the 8-th frame and the 13-th frame are gathered together to train another fuzzy rule, 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.

The group of the three images is converted to the posture sequence which has the maximum sum of three membership function values in Eq. (3.27). Each posture sequence will trigger a corresponding rule one time. If the corresponding rule is not exist, a new rule is built in the form of IF-THEN which is represented in Section 3.4. One of the fuzzy rule is represented in the view of template images in Fig. 4.11.

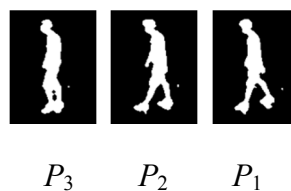


Fig. 4.12. The fuzzy rule of walk from right to left.

4.3 The Action Recognition Accuracy

In order to calculate the recognition rate of actions, we use off-line videos in our experiment. The off-line videos include five person videos and each person performed eleven actions. Then, we input the testing video from different starting frames, similar to the way for the training fuzzy rules. 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.

Table I and Table II show the recognition rate in bright and dark action environments respectively. We use leave-one-out cross-validation of five person videos. If we test these videos in Person 1, we will construct the templates and fuzzy rules by used the other three persons. That is, the testing video was not used for constructing key posture templates and fuzzy rules.

In the tables, W_{RL} is the action “walking from right to left,” W_{LR} is the action “walking from left to right,” B_{END} is the action “bending,” W_{AVE} is the action “waving,” SD_L is the action “sitting down on the left,” S_L is the action “sitting on the left,” SU_L is the action “standing up on the left,” SD_R is the action “sitting down on the right,” S_R is the action “sitting on the right,” SU_R is the action “standing up on the right,” W_S is the action “walking straight,” A_{cc}^a is the person’s action recognition accuracy in frame base.

The frame based accuracy is the total number of correct recognition divide by the total number of recognitions done. The following tables show the accuracy by using the video bases.

TABLE I

THE ACCURACY RATE OF ACTION RECOGNITION IN THE BRIGHT ENVIRONMENT

Person Action	Person 1	Person 2	Person 3	Person 4
W_{RL}	98.67% (223/226)	98.88% (177/179)	98.59% (210/213)	96.27% (155/161)
W_{LR}	98.95% (189/191)	97.28% (143/147)	100% (164/164)	95.60% (152/159)
B_{END}	100% (203/203)	100% (161/161)	100% (179/179)	92.25% (250/271)
W_{AVE}	98.48% (194/197)	90.52% (210/232)	97.55% (199/204)	95.26% (241/253)
SD_L	100% (114/114)	88.89% (120/135)	93.50% (115/123)	78.57% (99/126)
S_L	98.65% (220/223)	100% (158/158)	99.45% (182/183)	99.50% (200/201)
SU_L	98.90% (90/91)	100% (117/117)	97.46% (115/118)	88.57% (93/105)
SD_R	87.18% (102/117)	91.18% (93/102)	100% (91/91)	75.61% (93/123)
S_R	99.06% (211/213)	100% (143/143)	99.38% (161/162)	92.80% (219/236)
SU_R	91.75% (89/97)	100% (84/84)	100% (105/105)	90.32% (84/93)
W_S	94.63% (370/391)	95.53% (278/291)	97.46% (307/315)	95.08% (425/447)
A_{cc}^a	97.19% (2005/2063)	96.28% (1684/1749)	98.44% (1828/1857)	92.46% (2011/2175)
False alarm rate	0.28% (58/20630)	0.37% (65/17490)	0.16% (29/18570)	0.75% (164/21750)
Total frame based accuracy: 96.34%; Total frame based false alarm rate: 0.40%				

TABLE II

THE ACCURACY RATE OF ACTION RECOGNITION IN THE DARK ENVIRONMENT

	Person 1	Person 2	Person 3	Person 4
W_{RL}	97.62% (205/210)	98.94% (186/188)	100% (205/205)	97.08% (166/171)
W_{LR}	98.43% (188/191)	100% (158/158)	99.44% (177/178)	94.67% (160/169)
B_{END}	100% (218/218)	100% (195/195)	100% (167/167)	86.00% (221/257)
W_{AVE}	100% (218/218)	82.57% (180/218)	83.87% (182/217)	94.44% (221/234)
SD_L	92.59% (100/108)	88.89% (120/135)	100% (105/105)	79.59% (117/147)
S_L	97.96% (240/245)	92.06% (116/126)	100% (117/117)	100% (213/213)
SU_L	88.24% (90/102)	100% (96/96)	95.70% (89/93)	93.33% (84/90)
SD_R	97.41% (113/116)	95.15% (98/103)	93.00% (93/100)	77.52% (100/129)
S_R	98.64% (217/220)	100% (117/117)	100% (171/171)	92.09% (233/253)
SU_R	71.43% (75/105)	97.14% (102/105)	94.12% (80/85)	89.11% (90/101)
W_S	91.91% (375/408)	92.96% (251/270)	96.27% (258/268)	97.71% (470/481)
A_{cc}^a	95.24% (2039/2141)	94.62% (1619/1711)	96.37% (1644/1706)	92.42% (2075/2245)
False alarm rate	0.47% (102/21410)	0.54% (92/17110)	0.36% (62/17060)	0.76% (170/22450)
Total frame based accuracy: 94.54%; Total frame based false alarm rate: 0.55%				

4.4 Fuzzy Rule Construction for Gait Recognition

We construct the template model matrix and the fuzzy rule database with the training data. Firstly, we choose key posture images as essential templates from each person's gait images, and the number of each person's gait key posture image is in proportion to its period at about 1/6 sec per key posture. Key posture images of each person for one gait video are shown in Fig. 4.12. We will regard each posture as one class of posture types.

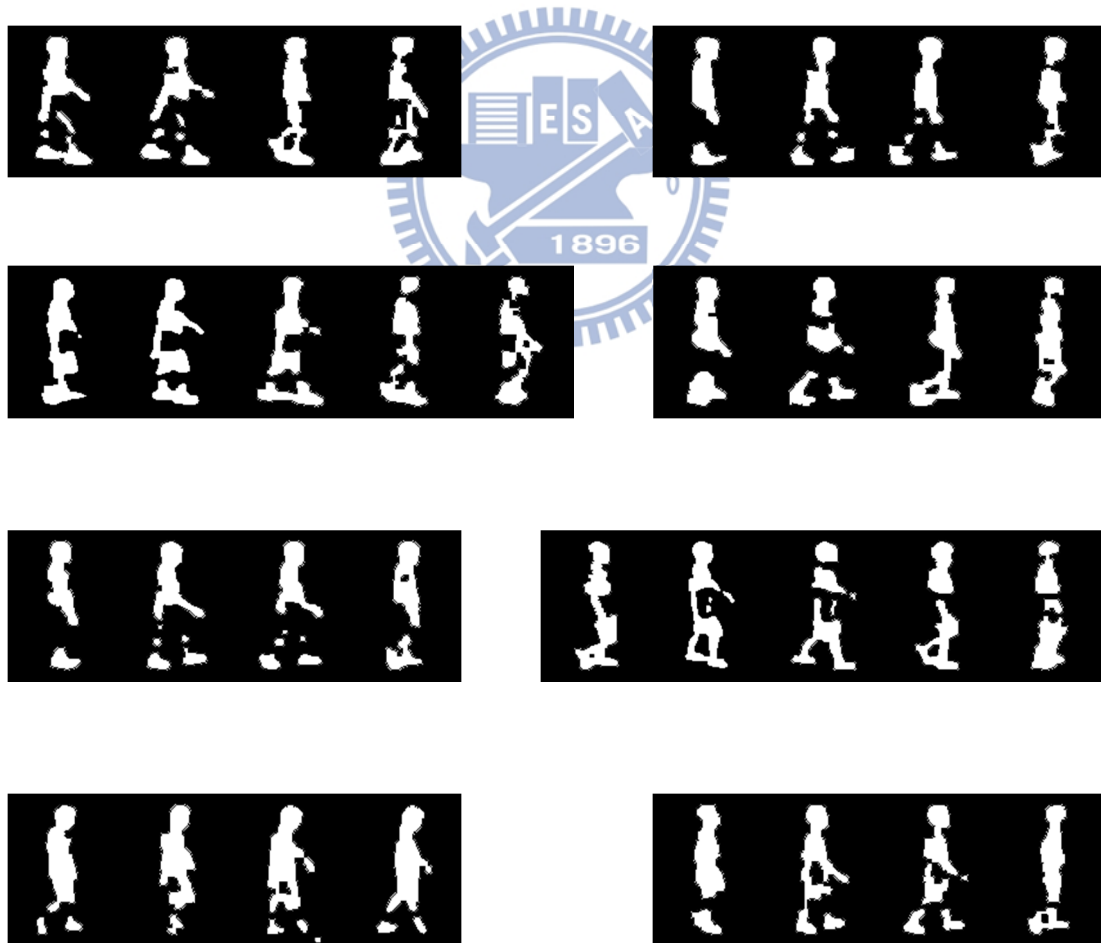


Fig. 4.12 Key postures of the eight person's gait of one gait video.

4.5 *The Recognition Rate of Gaits*

In order to calculate the recognition rate of gaits, we use off-line videos in our experiment. The off-line videos include five gait videos, every gait video has eight person's gait images. Then, we input the testing video from different starting frames which is similar to the way for the training fuzzy rules. 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.

Table III and Table IV show the recognition rate in bright and dark gait environments respectively. We use leave-one-out cross-validation of five gait videos. If we test these videos in first gait video, we will construct the templates and fuzzy rules by used the other three gait video. That is, the testing video was not used for constructing templates and fuzzy rules. A_c^a represents the person identification by the gait videos.

The frame based accuracy is the total number of correct recognition divide by the total number of recognitions done. The following tables show the accuracy by using the video bases.

TABLE III

THE ACCURACY RATE OF PERSON RECOGNITION BY THE **GAIT** VIDEOS IN THE **BRIGHT** ENVIRONMENT

condition \ outcome	Person 1	Person 2	Person 3	Person 4	Person 5	Person 6	Person 7	Person 8
Person 1	363	8	2	2	6	11	6	3
Person 2	10	325	8	6	2	15	12	2
Person 3	0	1	398	9	3	10	21	8
Person 4	4	27	15	421	11	14	16	22
Person 5	8	8	14	2	468	5	6	12
Person 6	6	5	10	11	5	355	7	1
Person 7	6	15	22	22	2	5	363	14
Person 8	18	0	9	0	5	15	17	352
A_{cc}^a	87.47%	83.55%	83.26%	89.01%	93.32%	82.56%	81.03%	85.02%
False alarm rate	1.21%	1.74%	1.69%	3.54%	1.81%	1.44%	2.77%	2.04%
Total frame based accuracy: 85.80%; Total frame based false alarm rate: 2.03%								

TABLE IV

THE ACCURACY RATE OF PERSON RECOGNITION BY THE **GAIT** VIDEOS IN THE **DARK** ENVIRONMENT

condition \ outcome	Person 1	Person 2	Person 3	Person 4	Person 5	Person 6	Person 7	Person 8
Person 1	192	0	7	0	0	3	0	0
Person 2	1	228	11	11	0	0	0	0
Person 3	6	18	282	4	0	4	3	20
Person 4	10	5	15	302	4	15	25	16
Person 5	12	0	7	6	300	11	0	6
Person 6	13	8	1	15	20	257	0	15
Person 7	3	7	9	5	3	0	323	10
Person 8	5	10	0	2	1	0	16	248
A_{cc}^a	79.34%	82.61%	84.94%	87.54%	91.46%	88.62%	88.01%	78.73%
False alarm rate	0.44%	1.04%	2.54%	4.19%	1.94%	3.27%	1.74%	1.56%
Total frame based accuracy: 85.45%; Total frame based false alarm rate: 2.08%								

Chapter 5 Conclusion

In this thesis, we implement the day and night automatic home health care system that combines the action recognition and gait recognition. The images are first extracted by background subtraction in action recognition system and gait recognition system. Then, the test images are transformed into a new space by eigenspace transform and canonical space transform for better efficiency and separability. Using three connective down-sampled images for fuzzy rule based inference system are used for action and gait recognition.

By our method, correct rate of action recognition in the bright environment is 95.97%; and the correct rate of action recognition in the dark environment by about 94.54%. The correct rate of person recognition by the gait recognition in the bright environment is 85.80%; and the correct rate of person recognition by the gait recognition in the dark environment is 85.45%.

The correct rates of action and gait recognition in dark environment are lower than those in the bright environment. This is because that the NIR image acquired in the dark environment will have little information on hue and saturation components than that in day time.

References

- [1] C. Fabien, D. Bhowmik, C. Abhayaratne, and M. S. Hawley, "Video-Based Technology for Ambient Assisted Living: A Review of the Literature," *Environments* vol. 3, no. 3, pp. 253–269, 2011.
- [2] A. Elgammal, D. Harwood, and L. Davis, "Non-Parametric Model for Background Subtraction," *Proc. Sixth European Conf. Computer Vision*, vol. II, pp. 751–767, June 2000.
- [3] 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, Aug. 2000.
- [4] T. Horprasert, D. Harwood, and L.S. Davis, "A Statistical Approach for Real-Time Robust Background Subtraction and Shadow Detection," in *Proc. IEEE ICCV'99*, 1999.
- [5] R. Cucchiara, C. Grana, M. Piccardi and A. Prati, "Improving Shadow Suppression in Moving Object Detection with HSV Color Information," in *Proc. IEEE Intelligent transportation System Conference*, pp. 334–339, 2001.
- [6] 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.
- [7] 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, Dec. 2004.
- [8] 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, Apr., 1998.
- [9] 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.
- [10] L. X. Wang and J. M. Mendel, “Generating fuzzy rules by learning from examples,” *IEEE Trans. Syst., Man Cybern.*, vol. 22, no. 6, pp. 1414–1427, Dec. 1992.
- [11] K. Ohba, Y. Sato, and K. Ikeuchi, “Appearance-based visual learning and object recognition with illumination invariance,” in *Machine Vision and Applications*, Vol. 12, No. 4, pp. 189–196, 2000.
- [12] K. Etemad and R. Chellappa, “Discriminant analysis for recognition of human face images,” in *Proc. ICASSP*, pp. 2148–2151, 1997.
- [13] 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, Apr. 1998.
- [14] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd edition, 1300 Boylston Street Chestnut Hill, Massachusetts USA: Academic Press, 1990.
- [15] J. C. S. Jacques Jr., C. R. Jung, S. R. Musse, “Background subtraction and shadow detection in grayscale video sequences,” in *Proc. SIGGRAPH*, pp. 189–196, 2005.
- [16] R. Gonzales and R. Woods, *Digital Image Processing*, 3rd ed. Pearson Education International, pp. 589–591, 2008.
- [17] Soriano M, Huovinen S, Martinkauppi B, Laaksonen M. “Using the skin locus to cope with changing illumination conditions in color-based face tracking,” in *IEEE Nordic Signal Processing Symposium, kolmarden, Sweden*, pp. 383–386,

Jun. 2000.

- [18] Y. C. Luo, "Extracting the Foreground Subject in the HSV Color space and Its Application to Human Activity Recognition System," *Master Thesis*, Elect. and Con. Eng. Dept., Chiao Tung Univ., Taiwan, 2007.

