

國立交通大學

電控工程研究所

碩士論文

以正規轉換為基礎之日夜人物辨識

Canonical Transform Based Day-and-Night Person
Identification

研究生：高仲義

指導教授：張志永

中華民國一百零二年七月

以正規轉換為基礎之日夜人物辨識
Canonical Transform Based Day-and-Night Person
Identification

學 生：高仲義

Student : Jhong- Yi Gao

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

以正規轉換為基礎之日夜人物辨識

學生：高仲義

指導教授：張志永博士

國立交通大學電控工程研究所

摘要

人物辨識系統在電腦視覺領域是很熱門的研究與應用目標。在監控系統中，最常見的方式是使用固定式攝影機，對拍攝場景的人物進行人物辨識。

本論文實現一套監控系統，此系統是在日夜環境中，分別使用多角度步態辨識系統及人臉辨識系統。本文研究對於使用兩台近紅外線攝影機進行人物辨識，一台近紅外線攝影機設置在遠處，用於擷取不同方向的步態影像，另一台近紅外線攝影機設置在近處，用於擷取人臉正面影像。

在人臉辨識系統方面，我們利用近紅外線攝影機擷取人臉影像。人臉擷取的方法是使用 Haar 疊層分類器，這是一種基於特徵運算的演算法，這種演算法比基於逐點的更快速，接著人臉影像經過特徵空間轉換與正規空間轉換後，累積五張上述人臉影像後，藉由多數決的方式，完成人物辨識。

在步態辨識系統方面，我們利用近紅外線攝影機擷取步態影像。為了擷取出完整的人體部分，本文使用背景相減法在灰階空間與 HSV 色彩空間建立背景模型，並提升消除影像中陰影部分，使得擷取前景影像能夠更完整，接著步態影像經過特徵空間轉換與標準空間轉換後，累積五張上述步態影像後，藉由多數決的方式，完成人物辨識。

Canonical Transform Based Day-and-Night Person Identification

STUDENT: Jhong- Yi Gao

ADVISOR: Dr. Jyh-Yeong Chang

Institute of Electrical Control Engineering
National Chiao-Tung University

ABSTRACT

Human recognition system is a very popular subject for research and application. Using a camera to recognize human is widely seen in surveillance system.

In this thesis, we implement the surveillance system that can recognize multi-angle human gait and human face of a person in the bright and dark environments. We use two near infrared (NIR) cameras for human recognition. One NIR camera, set in remote location, capture the gait images from different angles. And the other NIR camera, set in the vicinity, capture the face images from the person frontal view.

In human face recognition system, face region of an image is extracted based on Haar cascade classifier, which is a feature-based algorithm and works much faster than the pixel-based algorithm. Then, the face image is transformed to a new space by eigenspace and canonical space transformation for better efficiency and separability. The recognition is finally done in canonical space. Moreover, we gather five consecutive face images from video, and use majority vote to recognition the human.

In human gait recognition system, we build two background models, one in grayscale and one in HSV color space to extract the foreground image correctly. Then we reduce the shadowing effect. The gait image is then transformed to a new space by

eigenspace and canonical space transformation for better efficiency and separability. The recognition is done in the canonical space. Finally, we gather five consecutive gait images from video, and use majority vote to recognition the person.



ACKNOWLEDGEMENTS

I am grateful to my thesis advisor, Professor Jyh-Yeong Chang, who has offered me valuable suggestion in the academic studies. In the preparation of the thesis, he has spent much time reading through each draft and provided me with inspiring advice. Without his patient instruction, insightful criticism and expert guidance, the accomplishment of this thesis would not have been possible.

Finally, I would like to thank my family for their concern, supports and encouragements.



Contents

摘要.....	i
ABSTRACT.....	ii
ACKNOWLEDGEMENTS	iv
Contents	v
List of Figures.....	vii
List of Tables.....	ix
Chapter 1 Introduction.....	1
1.1 Motivation.....	1
1.2 Video Frame Preprocessing for Human Recognition	3
1.3 Video Frame Human Recognition Procedure	4
1.4 Thesis Outline	5
Chapter 2 Video Frame Preprocessing for Human Recognition	6
2.1 The HSV color space	6
2.2 Background Model Construction and Foreground Extraction.....	9
2.2.1 Background Model Construction.....	9
A. Grayscale Value Background Model.....	10
B. HSV Color Space Background Model	11
2.2.2 Background Update	12
2.2.3 Foreground Extraction	13
A. Foreground Detection	14
B. Shadow Suppression	16
C. Foreground Object Segmentation.....	18
D. Foreground Image Compensation.....	20

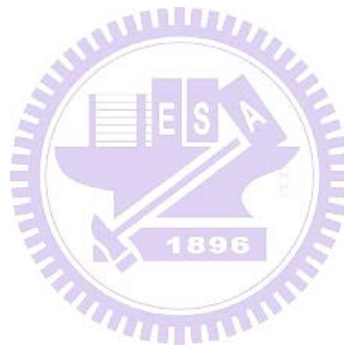
2.3 Face Extraction	21
Chapter 3 Video Frame Human Recognition Procedure	24
3.1 Human Representation.....	24
3.1.1 Eigenspace Transformation (EST).....	27
3.1.2 Canonical Space Transformation (CST).....	29
3.2 Human Recognition	31
3.2.1 Person Recognition by Gait Image Classification in a Long Distance Setting	31
3.2.2 Person Recognition by Face Image Classification in a Short Distance Setting	32
3.2.3 Majority Vote	33
Chapter 4 Experimental Results.....	34
4.1 Background Model Construction and Foreground Extraction.....	38
4.2 Experiments on our LAB Multi-Angle Gait Database	42
4.2.1 <i>Single-Angle</i> Human <i>Gait</i> Recognition	42
4.2.2 <i>Multi-Angle</i> Human <i>Gait</i> Recognition.....	46
4.3 Recognition Result on the CASIA Multi-View Gait Database.....	49
4.3.1 <i>Single-View</i> Human <i>Gait</i> Recognition	49
4.3.2 <i>Multi-View</i> Human <i>Gait</i> Recognition	53
4.4 Experiments on our LAB Face Database.....	55
4.4.1 Human <i>Face</i> Recognition	55
Chapter 5 Conclusion	62
References.....	63

List of Figures

Fig. 1.1.	The flowchart of our human recognition system.....	2
Fig. 2.1.	The HSV color space.....	7
Fig. 2.2.	The framework of background model construction.	10
Fig. 2.3.	The framework of foreground extraction.	14
Fig. 2.4.	Histogram of binary foreground image projection in X and Y direction. ..	19
Fig. 2.5.	The binary foreground image of extracted foreground region.	20
Fig. 2.6.	Foreground image after opening and closing operation.	20
Fig. 2.7.	Rectangle features shown relative to the enclosing detection window.	21
Fig. 2.8.	Sum of all pixels marked is the integral image intensity at (x, y)	22
Fig. 2.9.	The sum of pixels in rectangle D can be computed as $4+1 - (2+3)$	23
Fig. 3.1.	The structure of human recognition by gait or face image sequence.	25
Fig. 3.2.	The structure of human classification.....	33
Fig. 4.1.	(a) The scene of human gait recognition in bright environments. (b) The scene of human gait recognition in dark environments. (c) The scene of human face recognition in bright environments. (d) The scene of human face recognition in dark environments.....	35
Fig. 4.2.	Example video sequences used in our experiments. (a) and (b) are typical video sequences for gaits of LAB in bright and dark environments. From top to bottom: walking 0° , walking 45° , and walking 315° respectively...	36
Fig. 4.3.	Example video sequences used in our experiments. (a) and (b) are typical video sequences for face of LAB in bright and dark environments. From top to bottom: walking 0° , walking 45° , and walking 315° respectively...	37
Fig. 4.4.	Eleven video frames depicting person of the CASIA multi-view gait recognition database from different viewing angles.....	38

Fig. 4.5. Results of foreground detection (a) an image frame in the bright environment, (b) binary image after performing foreground detection in the bright environment, (c) projection of (b) onto X direction, (d) projection of (b) onto Y direction, (e) foreground region segmentation in the bright environment.....40

Fig. 4.6. Results of foreground detection (a) an image frame in the dark environment, (b) binary image after performing foreground detection in the dark environment, (c) projection of (b) onto X direction, (d) projection of (b) onto Y direction, (e) foreground region segmentation in the dark environment.41



List of Tables

TABLE I	THE HUMAN GAIT RECOGNITION RATES AT SPECIFIC WALKING ANGLE IN THE BRIGHT ENVIRONMENT, WITHOUT MAJORITY VOTE	43
TABLE II	THE HUMAN GAIT RECOGNITION RATES AT SPECIFIC WALKING ANGLE IN THE BRIGHT ENVIRONMENT, WITH MAJORITY VOTE OF THREE	43
TABLE III	THE HUMAN GAIT RECOGNITION RATES AT SPECIFIC WALKING ANGLE IN THE BRIGHT ENVIRONMENT, WITH MAJORITY VOTE OF FIVE	44
TABLE IV	THE HUMAN GAIT RECOGNITION RATES AT SPECIFIC WALKING ANGLE IN THE DARK ENVIRONMENT, WITHOUT MAJORITY VOTE	44
TABLE V	THE HUMAN GAIT RECOGNITION RATES AT SPECIFIC WALKING ANGLE IN THE DARK ENVIRONMENT, WITH MAJORITY VOTE OF THREE	45
TABLE VI	THE HUMAN GAIT RECOGNITION RATES AT SPECIFIC WALKING ANGLE IN THE DARK ENVIRONMENT, WITH MAJORITY VOTE OF FIVE	45
TABLE VII	THE RECOGNITION RATES OF WALKING VIDEOS IN THE BRIGHT ENVIRONMENT, WITHOUT MAJORITY VOTE	47
TABLE VIII	THE RECOGNITION RATES OF WALKING VIDEOS IN THE BRIGHT ENVIRONMENT, WITH MAJORITY VOTE OF THREE	47
TABLE IX	THE RECOGNITION RATES OF WALKING VIDEOS IN THE BRIGHT ENVIRONMENT, WITH MAJORITY VOTE OF FIVE	47
TABLE X	THE RECOGNITION RATES OF WALKING VIDEOS IN THE DARK	

ENVIRONMENT, WITHOUT MAJORITY VOTE	48
TABLE XI THE RECOGNITION RATES OF WALKING VIDEOS IN THE DARK ENVIRONMENT, WITH MAJORITY VOTE OF THREE.....	48
TABLE XII THE RECOGNITION RATES OF WALKING VIDEOS IN THE DARK ENVIRONMENT, WITH MAJORITY VOTE OF FIVE.....	48
TABLE XIII THE HUMAN RECOGNITION RATES AT SPECIFIC VIEWING ANGLE IN THE CASIA DATABASE, WITHOUT MAJORITY VOTE.....	50
TABLE XIV THE HUMAN RECOGNITION RATES AT SPECIFIC VIEWING ANGLE IN THE CASIA DATABASE, WITH MAJORITY VOTE OF THREE	51
TABLE XV THE HUMAN RECOGNITION RATES AT SPECIFIC VIEWING ANGLE IN THE CASIA DATABASE, WITH MAJORITY VOTE OF FIVE	52
TABLE XVI THE RECOGNITION RATES AT ALL VIEWING ANGLES IN THE CASIA DATABASE, WITHOUT MAJORITY VOTE	53
TABLE XVII THE RECOGNITION RATES AT ALL VIEWING ANGLES IN THE CASIA DATABASE, WITH MAJORITY VOTE OF THREE	54
TABLE XVIII THE RECOGNITION RATES AT ALL VIEWING ANGLES IN THE CASIA DATABASE, WITH MAJORITY VOTE OF FIVE	54
TABLE XIX THE RECOGNITION RATES OF HUMAN FACE VIDEOS IN THE BRIGHT ENVIRONMENT, WITHOUT MAJORITY VOTE	56
TABLE XX THE RECOGNITION RATES OF HUMAN FACE VIDEOS IN THE BRIGHT ENVIRONMENT, WITH MAJORITY VOTE OF THREE.....	57
TABLE XXI THE RECOGNITION RATES OF HUMAN FACE VIDEOS IN THE BRIGHT ENVIRONMENT, WITH MAJORITY VOTE OF FIVE	58
TABLE XXII THE RECOGNITION RATES OF HUMAN FACE VIDEOS IN THE DARK ENVIRONMENT, WITHOUT MAJORITY VOTE	59
TABLE XXIII THE RECOGNITION RATES OF HUMAN FACE VIDEOS IN THE	

DARK ENVIRONMENT, WITH MAJORITY VOTE OF THREE 60

TABLE XXIV THE RECOGNITION RATES OF HUMAN FACE VIDEOS IN THE
DARK ENVIRONMENT, WITH MAJORITY VOTE OF FIVE 61



Chapter 1 Introduction

1.1 Motivation

Human recognition plays an important role in applications such as surveillance systems, home nursing care system and security applications. Most of the security service firm is provided by professional people, such as security guard. However, the service cost is very expensive and the security guard cannot watch camera video in 24 hours. Therefore, the automatic surveillance system becomes a popular research area in recent years. For example, an automatic system will trigger an alarm condition when the automated surveillance system detects and recognizes suspicious human.

In this thesis, we implement the day-and-night (bright and dark) surveillance system that separately using multi-angle human gait and human face recognition of a person in an In-door Environment. We use two cameras for human recognition. One camera being used to capture the gait image from different angle is set in a remote location. And the other camera being used to capture the face image from the person frontal view is set in the vicinity. Fig 1.1 is illustrated our system flowchart. Our system can be separated into three components. The first component is video frame preprocessing which contain foreground subject extraction and human face extraction. The second component is the transformation of human gait image or human face image in a space smaller and easier for human recognition. The third component is the human recognition of an image frame using majority vote.

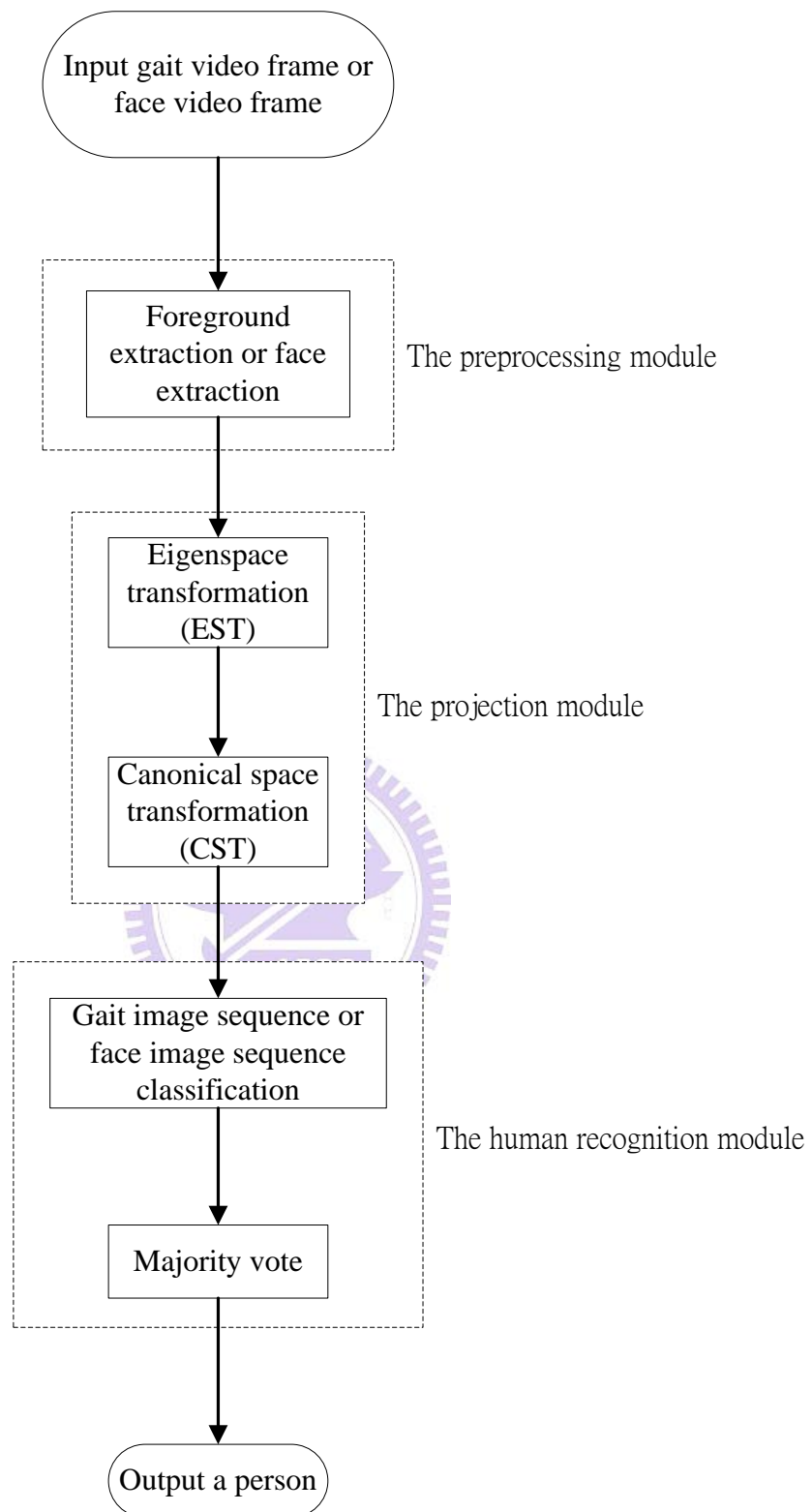


Fig. 1.1. The flowchart of our human recognition system.

1.2 Video Frame Preprocessing for Human Recognition

The first step of human gait recognition system is foreground subject extraction. We have to construct a background model for foreground subject extraction. Background subtraction is widely used for detecting moving objects from image frames of fixed cameras. The rationale of this approach is to detect the moving objects by the difference between the current image frame and a reference image frame, often called the “background model.” There are many well-known methods to build background models. A review is given in [1] where many different approaches were proposed in recent years. In our human gait recognition system, we construct two background models for more correct foreground subject extraction; one is based on grayscale value, and the other is based on HSV color space. Basically, the background image is a representation of the static scene. We have to update the background model after the subject enters the scene. After the subject leaves the scene, the background model will also be updated accordingly.

After construct two background models, we can extract foreground subject from video frames by subtracting each pixel value of background model from that current image frame. Then, the resulting image is converted to a binary image by setting a threshold. The binary image contains foreground subject and shadow. Therefore, we need to remove the shadow by using a shadow filter. Then, 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. When we want to remove shadow pixels, some foreground pixel will be lost and this makes the foreground image broken. Therefore, we will repair the rectangle image by using opening and closing operations. Finally, the rectangle image is resized to the specified resolution for normalization.

The first step of human face recognition system is human face extraction. The

purpose of face detection is to localize and extract the face region from the scene with human. We use Haar cascade classifier, proposed by Viola et al. [2], from OpenCV package [3] to detect the face regions.

1.3 Video Frame Human Recognition Procedure

In gait video or face video, the dimensions of gait image or face image are often extremely large, and these images usually contain great deals of redundancies. Hence, some space transformations are introduced to reduce redundancy of an image by reducing the 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 for human recognition, for example, wavelet transformation, Fourier transformation, Locally Linear Embedding (LLE), Multi Dimension Scaling (MDS), Principal Component Analysis (PCA), eigenspace transformation (EST), and so on. Our transformation method combines eigenspace transformation and canonical space transformation which are described as follows.

Eigenspace transformation (EST), which uses Principal Component Analysis (PCA) for dimensionality reduction, has been demonstrated to be a potent scheme used below: automatic face recognition proposed in [4], [5]; gait analysis proposed in [6]; and action recognition proposed in [7]. 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.

Due to the above classification we used nearest neighbor concept to do the human recognition in the video. There could be misclassifications in some frames; we have adopted the majority vote to conduct the human recognition, to overcome this problem.

1.4 *Thesis Outline*

The thesis is organized as follows. In Chapter 2, we introduce video frame preprocessing for human gait recognition and human face recognition. In Chapter 3, we describe our human recognition system that includes “eigenspace transform,” “canonical transform,” “human recognition,” and “majority vote.” In Chapter 4, the experiment results of our human recognition systems are shown. At last, we conclude this thesis with a discussion in Chapter 5.

Chapter 2 Video Frame Preprocessing for Human Recognition

In this chapter, we describe background model construction and foreground extraction in grayscale and the HSV color space. We also briefly introduce the basic concepts of HSV color space which transforms the coordinate system in RGB color space to HSV color space. Finally, we introduce face detection method whose the principle is based on object detection technology proposed by Viola et al. [2].

2.1 *The HSV color space*

The HSV color space stands for hue, saturation, and value, also called HSB (B for brightness), which corresponds closely to the human perception of color. Fig. 2.1 illustrates the HSV color space whose shape is like a cone. From this figure, the hue is represented by the angle of each color in the cone relative to the 0° line, which is traditionally corresponded to be red. The saturation is representing as the distance from the center of the circle. Highly saturation color is on the outer edge of the cone, whereas gray tones (which have no saturation) are at the very center. The value is determined by the colors vertical position in the cone. At the point end of the cone, there is no brightness, so all colors are blacks. At the fat end of the cone are the brightness colors.

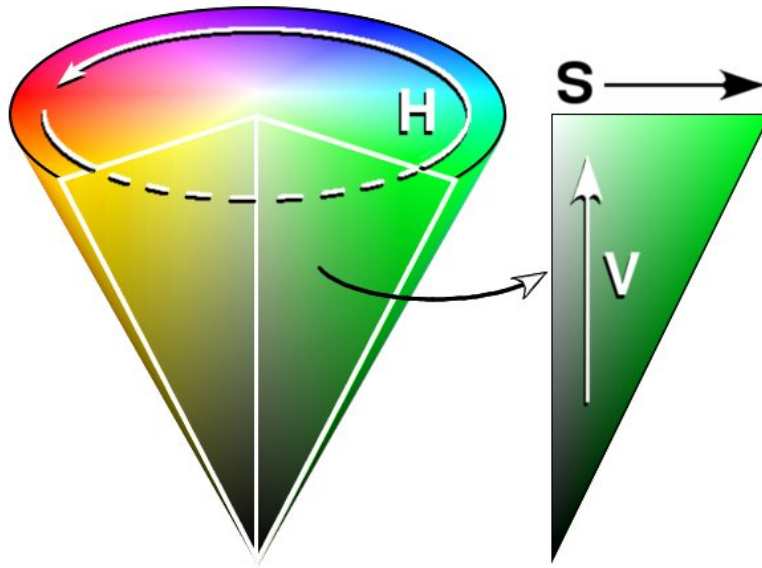


Fig. 2.1. The HSV color space.

The formula of RGB transfers to HSV is given by

$$\begin{aligned}
 H &= \begin{cases} 0^\circ, & \text{if } \max = \min \\ 60^\circ \times \frac{G-B}{\max - \min} + 0^\circ, & \text{if } \max = R \text{ and } G \geq B \\ 60^\circ \times \frac{G-B}{\max - \min} + 360^\circ, & \text{if } \max = R \text{ and } G < B \\ 60^\circ \times \frac{B-R}{\max - \min} + 120^\circ, & \text{if } \max = G \\ 60^\circ \times \frac{R-G}{\max - \min} + 240^\circ, & \text{if } \max = B \end{cases} \\
 S &= \begin{cases} 0, & \text{if } \max = 0 \\ \frac{\max - \min}{\max} = 1 - \frac{\min}{\max}, & \text{otherwise} \end{cases} \\
 V &= \max
 \end{aligned} \tag{2.1}$$

where $max = \max(R, G, B)$ and $min = \min(R, G, B)$.

The hue parameter is the value which represents color information without brightness. Therefore, the hue is not affected by change of 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) the hue is unstable when the saturation is extremely small. 2) The hue is meaningless when the intensity value is extremely small. 3) The saturation is meaningless when the intensity value is extremely small [8]. Accordingly, Ohba *et al.* [9] 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. Using this equation, the measured color close to dark 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 , S_t , 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 $H = 0$, where H and ΔP_t are a hue value, and the phase threshold value, respectively. The range of hue value is from 0 to 2π , and it has discontinuity at 0 and 2π . We use the phase threshold value ΔP_t to avoid the discontinuity effect.

2.2 Background Model Construction and Foreground Extraction

The first step of human gait recognition system is foreground extraction. We have to construct the background model for foreground extraction. There are many well-known methods to build background models. W^4 is such a typical example with some modifications [10]. It records the maximum grayscale value and the minimum grayscale value and the maximum inter-frame absolute difference of each pixel in the background video frames. Then each foreground image frame subtracts the maximum and minimum intensity value 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 as a foreground pixel. W^4 admits some rules make the background model be adaptive to varying environment.

2.2.1 Background Model Construction

If we only construct the luminance background model for foreground extraction, it cannot detect reliably those foreground pixel whose luminance component close to background pixel. In order to solve this problem, we construct another background model in HSV color space. The HSV color space corresponds closely to the human perception of color. We can have the luminance information and the chromatic information simultaneously. The hue is unreliable in some condition, so we use the three criteria (intensity value, saturation, and hue) described in sections 2.1 to obtain the hue value reliably. Fig. 2.2 shows the framework of background model construction.

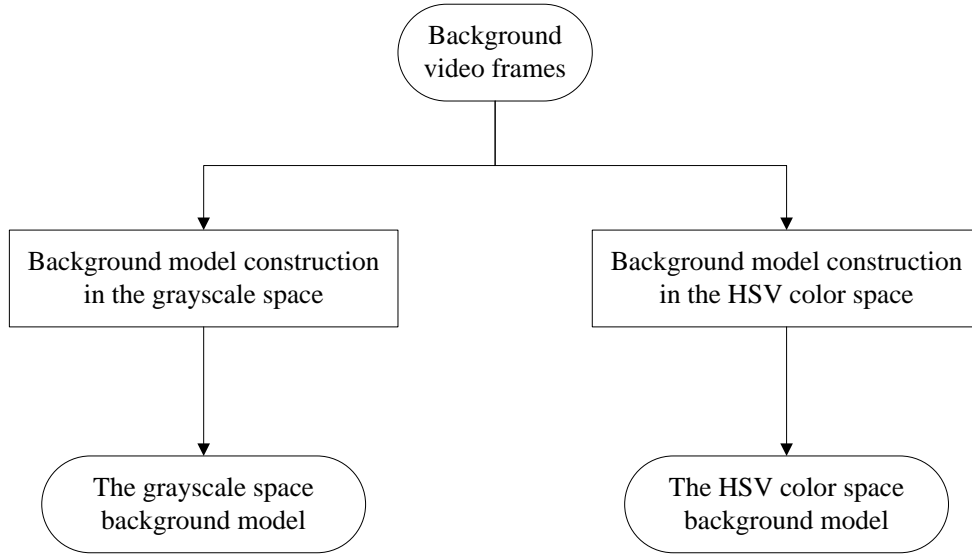


Fig. 2.2. The framework of background model construction.

A. Grayscale Value Background Model

The grayscale value background scene is modeled by representing each pixel by three values: the maximum grayscale value $n(x, y)$ and the minimum grayscale value $m(x, y)$ and the maximum inter-frame difference $d(x, y)$ of each pixel in the background video frames. Because these three values are statistical, we need a background video without any moving foreground objects for background model training. Let I be a background image frame sequence and contains N consecutive image frames. $I_i(x, y)$ be the grayscale value of a pixel location (x, y) in the i -th background image frame of I . The grayscale value background model for a pixel location (x, y) , $[n(x, y), m(x, y), d(x, y)]$, is obtained by

$$\begin{bmatrix} n(x, y) \\ m(x, y) \\ d(x, y) \end{bmatrix} = \begin{bmatrix} \max_i \{I_i(x, y)\} \\ \min_i \{I_i(x, y)\} \\ \max_i \{|I_i(x, y) - I_{i-1}(x, y)|\} \end{bmatrix}, \quad i = 1, 2, \dots, N \quad (2.2)$$

B. HSV Color Space Background Model

Along similar line of reasoning of above, we construct another background model in each dimension of HSV (hue, saturation and value) space [11]. Then, we record the maximum value $([n^H(x, y), n^S(x, y), n^V(x, y)])$ and the minimum value $([m^H(x, y), m^S(x, y), m^V(x, y)])$ and the inter-frame ratio in the brightness information and the inter-frame different in the chromatic information. Similarly, we use the same background video without any moving foreground objects for background model training. Let I be a background image frame sequence and contains N consecutive background image frames. $I_i^H(x, y)$ is the hue value of a pixel location (x, y) in the i -th background image frame of I . $I_i^S(x, y)$ is the saturation value of a pixel location (x, y) in the i -th background image frame of I . $I_i^V(x, y)$ is the brightness value of a pixel location (x, y) in the i -th background image frame of I . The HSV color space background model of a pixel is obtained by

$$\begin{bmatrix} n^H(x, y) \\ m^H(x, y) \\ d^H(x, y) \end{bmatrix} = \begin{bmatrix} \max_i \{I_i^H(x, y)\} \\ \min_i \{I_i^H(x, y)\} \\ \max_i \{|I_i^H(x, y) - I_{i-1}^H(x, y)|\} \end{bmatrix}, \quad i = 1, 2, \dots, N \quad (2.3)$$

$$\begin{bmatrix} n^s(x, y) \\ m^s(x, y) \\ d^s(x, y) \end{bmatrix} = \begin{bmatrix} \max_i \{I_i^s(x, y)\} \\ \min_i \{I_i^s(x, y)\} \\ \max_i \{|I_i^s(x, y) - I_{i-1}^s(x, y)|\} \end{bmatrix}, \quad i=1, 2, \dots, N \quad (2.4)$$

$$\begin{bmatrix} n^v(x, y) \\ m^v(x, y) \\ d^v(x, y) \end{bmatrix} = \begin{cases} \begin{bmatrix} \max_i \{I_i^v(x, y)\} \\ \min_i \{I_i^v(x, y)\} \\ \max_i \{|I_i^v(x, y) / I_{i-1}^v(x, y)|\} \end{bmatrix}, & \text{if } I_i^v(x, y) / I_{i-1}^v(x, y) \geq 1 \\ \begin{bmatrix} \max_i \{I_i^v(x, y)\} \\ \min_i \{I_i^v(x, y)\} \\ \max_i \{|I_{i-1}^v(x, y) / I_i^v(x, y)|\} \end{bmatrix}, & \text{otherwise} \end{cases} \quad i=1, 2, \dots, N \quad (2.5)$$

2.2.2 Background Update

The background model cannot be expected to stay the same for a long time. If the facilities in room are moved, they will be detected as foreground pixels of human and the human recognition will be misclassified. Therefore, we have to adopt a scheme that can update the background models in order to avoid above situation. The background models will be updated if the real-time video does not vary for a long time and there is nobody in the scene. By Eq. (2.6), we can calculate how many times the binary value remain unchanged.

$$update(x, y) = \begin{cases} update(x, y) + 1, & \text{if } I_{foreground}^t(x, y) = I_{foreground}^{t-1}(x, y) \\ update(x, y), & \text{otherwise} \end{cases} \quad (2.6)$$

where $I_{foreground}^t(x, y)$ is the grayscale value of a pixel location (x, y) in the binary image. The $update(x, y)$ value is a record of how many times $I_{foreground}^t(x, y)$ remains unchanged. When $update(x, y)$ exceeds a threshold, the pixel (x, y) will be included in the background model.

2.2.3 Foreground Extraction

The framework of foreground extraction is composed of four steps. The first step is foreground detection in the grayscale value and the HSV color space background models. The second step is the shadow suppression in the grayscale value and the HSV color space background models. The third step is the foreground object segmentation. And the final step is the foreground image compensation to recover the foreground pixels those are wrongly classified to the background. Fig. 2.3 shows the framework of foreground extraction.

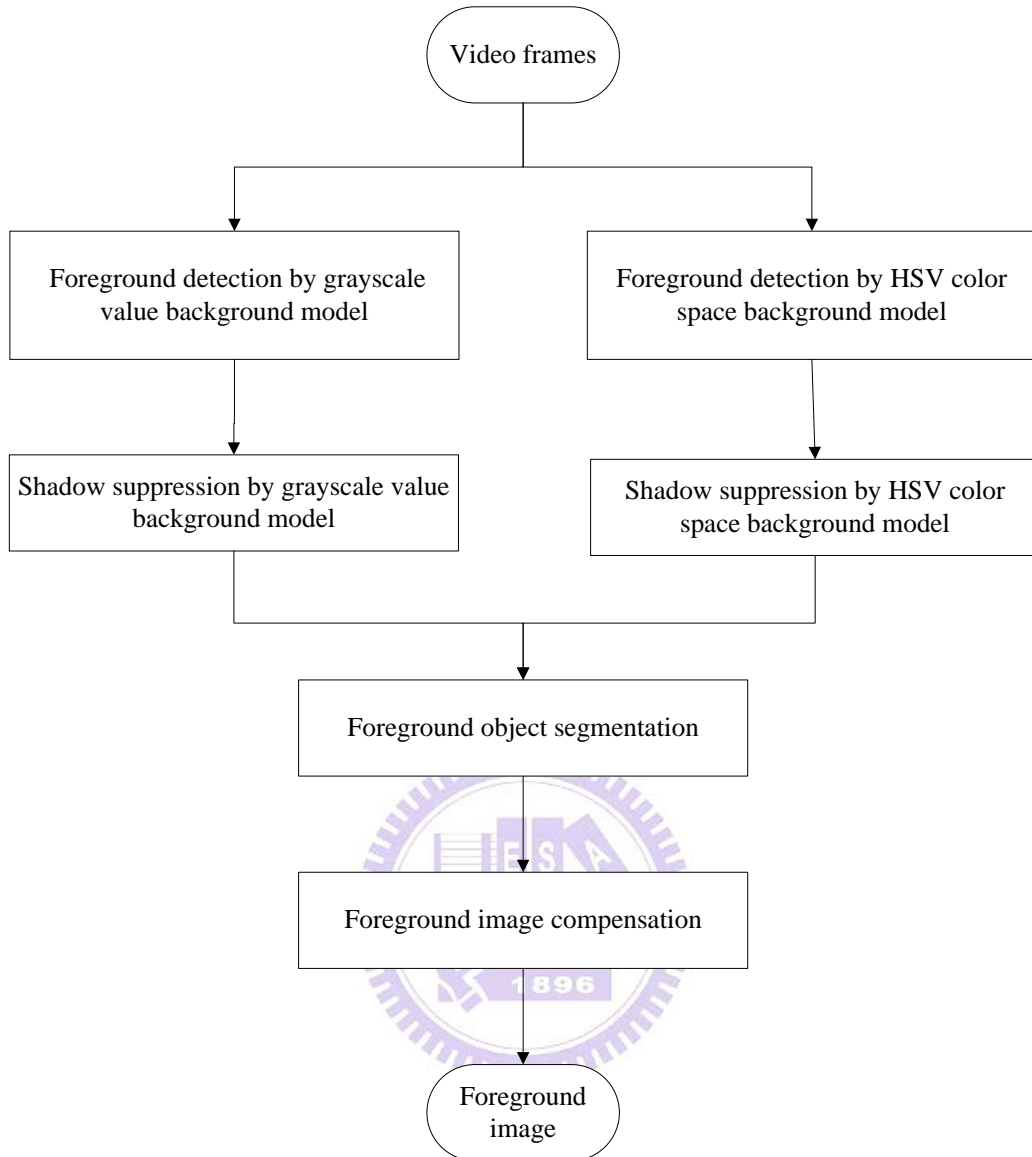


Fig. 2.3. The framework of foreground extraction.

A. Foreground Detection

The foreground objects can be detected from the background in each frame of the video sequence. Each pixel of the video frame is classified to either a foreground or a background pixel using the background model. First, we use the maximum grayscale

value $n(x, y)$ and the minimum grayscale value $m(x, y)$ and the maximum inter-frame difference $d(x, y)$ of the grayscale value background model to detect the foreground pixel by

$$I_{fg}(x, y) = \begin{cases} 0 \text{ background,} & \text{if } |I_i^{gray}(x, y) - n^{gray}(x, y)| < k_{gray}d_\mu \\ & \text{or } |I_i^{gray}(x, y) - m^{gray}(x, y)| < k_{gray}d_\mu \\ 1 \text{ foreground,} & \text{otherwise.} \end{cases} \quad (2.7)$$

where $I_i(x, y)$ is the grayscale value of a pixel location (x, y) in the i -th video frame, $I_{fg}(x, y)$ is the gray level of a pixel in the foreground binary image, d_μ is the median of all $d^{gray}(x, y)$, and k_{gray} is a threshold. Threshold k_{gray} is determined by experiments according to different environments.

On the other hand, we use the maximum value $n^V(x, y)$ and the minimum value $m^V(x, y)$ and the maximum inter-frame ratio $d^V(x, y)$ of the HSV color space background model to detect the foreground pixel by

$$I_{fg}^{HSV}(x, y) = \begin{cases} 0 \text{ background,} & \text{if } I_i^V(x, y)/n^V(x, y) < k_Vd^V(x, y) \\ & \text{or } I_i^V(x, y)/m^V(x, y) < k_Vd^V(x, y) \\ 1 \text{ foreground,} & \text{otherwise.} \end{cases} \quad (2.8)$$

where $I_i^V(x, y)$ is the intensity value of a pixel location (x, y) in the i -th video frame, $I_{foreground}^{HSV}(x, y)$ is the gray level of a pixel in binary image, and threshold k_v is determined by light as of the scene. Threshold k_v will be reduced for in-sufficient light condition and increased otherwise.

B. Shadow Suppression

The shadows of the foreground object are easily detected to foreground pixels in normal conditions. The situation causes foreground object merging and foreground object shape distortion in binary 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 [12].

First, we construct the shadow filter in the grayscale value. 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 T_{xy} such that $T_{xy}(m, n) = I(x+m, y+n)$, where $-1 \leq m \leq 1, -1 \leq n \leq 1$ (i.e. T_{xy} corresponds to a neighborhood of pixel (x, y)). Then, the NCC between templates T_{xy} and background image B at pixel (x, y) is given by

$$NCC(x, y) = \frac{ER(x, y)}{E_B(x, y)E_{T_{xy}}} \quad (2.9)$$

where

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

$$E_B(x, y) = \sqrt{\sum_{m=-1}^1 \sum_{n=-1}^1 B(x+m, y+n)^2}, \quad (2.10)$$

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

For a pixel (x, y) in a shadow region, the NCC in a neighboring region T_{xy} should be large (close to one), and the energy $E_{T_{xy}}$ of this region should be lower than lower than the energy E_B of the corresponding region in the background image. Therefore, we get

$$S^{gray}(x, y) = \begin{cases} \text{shadow,} & \text{if } NCC(x, y) \geq L_{ncc} \text{ and } E_{T_{xy}} < E_B(x, y) \\ \text{foreground,} & \text{otherwise} \end{cases} \quad (2.11)$$

where $S^{gray}(x, y)$ is the shadow mask to class the pixel in grayscale domain , and L_{ncc} is a fixed threshold. If L_{ncc} is low, several foreground pixels corresponding to moving objects may be misclassified as shadow pixels. Otherwise, selecting a larger value of L_{ncc} , then the shadow pixels may not be detected.

On the other hand, we know that the shadow pixels have similar chromaticity,

but lower brightness than the background model. Therefore, we construct another shadow filter in the HSV color space is intuitively designed as follows

$$S^{HSV}(x, y) = \begin{cases} \text{shadow,} & \text{if } I_i^V(x, y)/m^V(x, y) < 1 \\ & \text{and } |I_i^H(x, y) - n^H(x, y)| < k_H d^H(x, y) \\ & \text{and } |I_i^S(x, y) - n^S(x, y)| < k_S d^S(x, y) \\ \text{foreground,} & \text{otherwise} \end{cases} \quad (2.12)$$

where $I_i^H(x, y)$, $I_i^S(x, y)$, and $I_i^V(x, y)$ are respectively the HSV channel of a pixel location (x, y) in the i -th video frame, and $S^{HSV}(x, y)$ is the shadow mask to class the pixel in HSV domain. Values k_S and k_H are selected threshold values used to measure the similarities of the hue and the saturation between the background image and the current observed image.

In order to reduce the impact caused by shadow and noise on the foreground object, we calculate the union of $S^{gray}(x, y)$ and $S^{HSV}(x, y)$. The reason why we choose the union operator is that the union can increase the foreground with less noise. Finally, the binary image is obtained by

$$I_{fg}(x, y) = S^{gray}(x, y) \vee S^{HSV}(x, y) \quad (2.13)$$

C. Foreground Object Segmentation

According to the binary image $I_{fg}(x, y)$ segmented by above, 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 the X and Y directions. Fig. 2.4 shows an example of foreground region extraction. From this figure, we use 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 foreground region and the size of this rectangle is adjusted to 64×48 for normalization. Fig. 2.5 is the extracted foreground region.

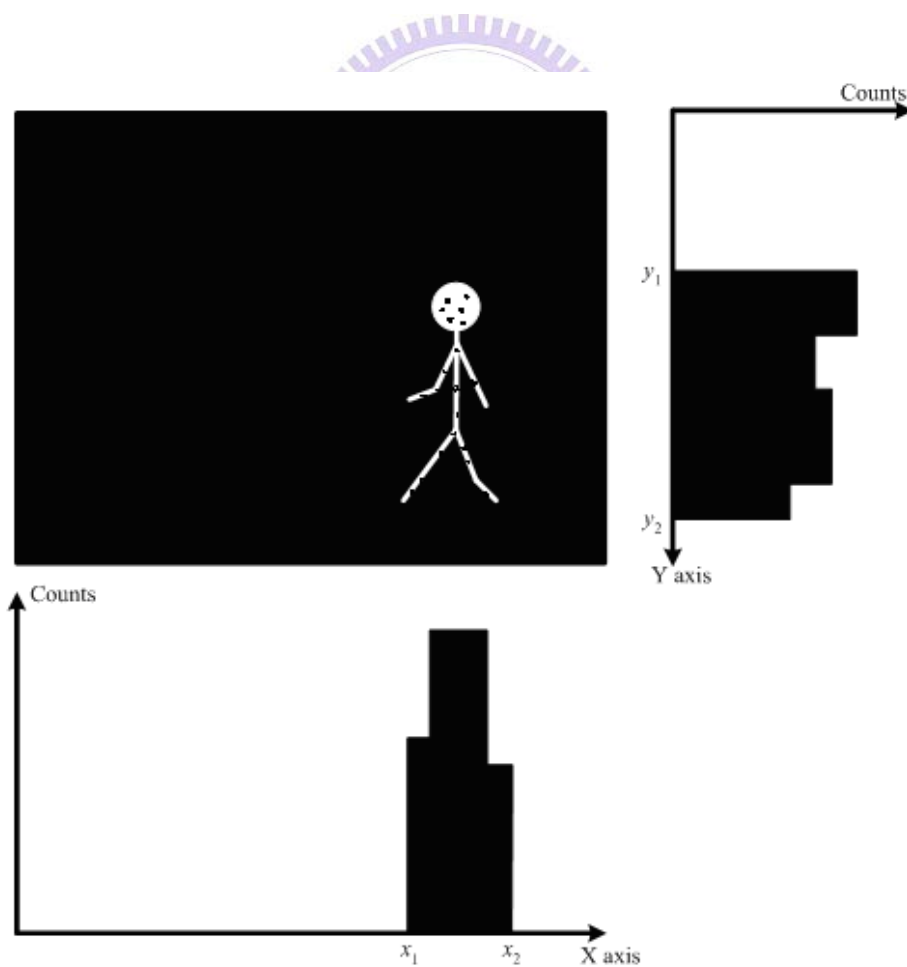


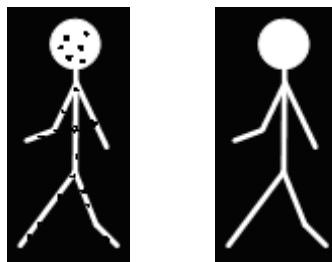
Fig. 2.4. Histogram of binary foreground image projection in X and Y direction.



Fig. 2.5. The binary foreground image of extracted foreground region.

D. Foreground Image Compensation

Detecting all foreground pixels and removing all shadows simultaneously are difficult. When we want to remove shadow pixels, some foreground pixels will be lost and this makes the foreground image be broken. Therefore, we will repair the foreground image by opening and closing operations [13]. Fig. 2.6 (a) shows all foreground pixels after shadow removal, and Fig. 2.6 (b) shows the result after applying the opening and closing operations.



(a)

(b)

Fig. 2.6. (a) Foreground image. (b) Foreground image after opening and closing of (a).

2.3 Face Extraction

The first step of human face recognition system is face extraction. We use Haar cascade classifier, proposed by Viola et al. [2], from OpenCV package [3] to detect the face regions. The classifier is based on the value of simple features. The feature-based system operates much faster than the pixel-based system. The feature-based system utilizes three kinds of features. The *two-rectangle feature*, *three-rectangle feature* and *four-rectangle feature* to classify facial region and not facial region (see Fig. 2.7). The sum of the pixels which lie within the white rectangles is subtracted from the one within the gray rectangles, and then the value is considered as a feature.

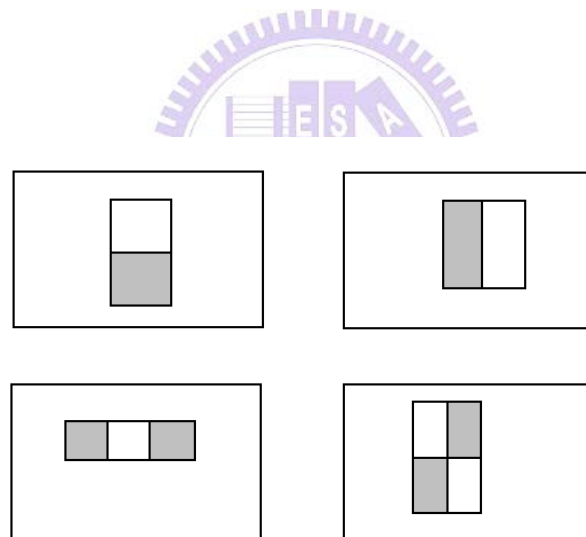


Fig. 2.7. Rectangle features shown relative to the enclosing detection window.

The cost of calculation of rectangle features can be reduced by using the integral image. The integral image intensity at location (x, y) is the sum of the pixels above and to the left of (x, y) , the mathematical description as follows:

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y') \quad (2.14)$$

where $ii(x, y)$ is the integral image and $i(x', y')$ is the original image (see Fig. 2.8)

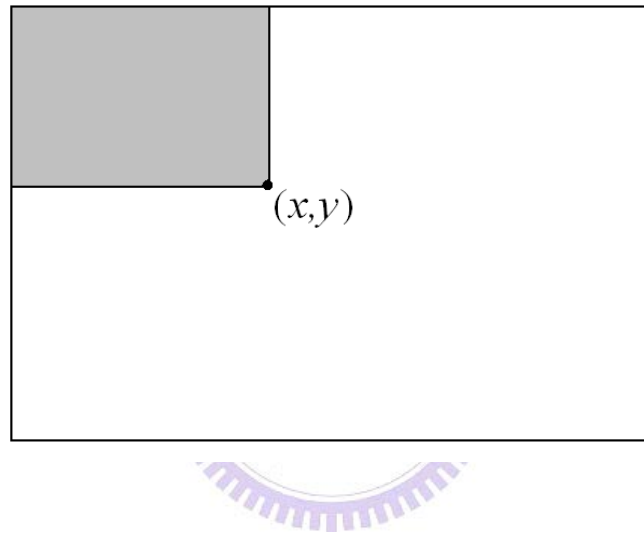


Fig. 2.8. Sum of all pixels marked is the integral image intensity at (x, y) .

The integral image can be computed in just one pass over the original image by using the following pair of recurrences:

$$s(x, y) = s(x, y-1) + i(x, y) \quad (2.15)$$

$$ii(x, y) = ii(x-1, y) + s(x, y) \quad (2.16)$$

where $s(x, y)$ is the cumulative row sum, $s(x, -1) = 0$ and $ii(-1, y) = 0$. Any

rectangular sum can be computed in four array references (see Fig. 2.9). The sum of pixels in rectangle A is the integral image intensity at location 1. The sum of A+B is at location 2, A+C is at location 3 and A+B+C+D is at location 4. Therefore, the sum of pixels in rectangle D can be computed as $4+1 - (2+3)$.

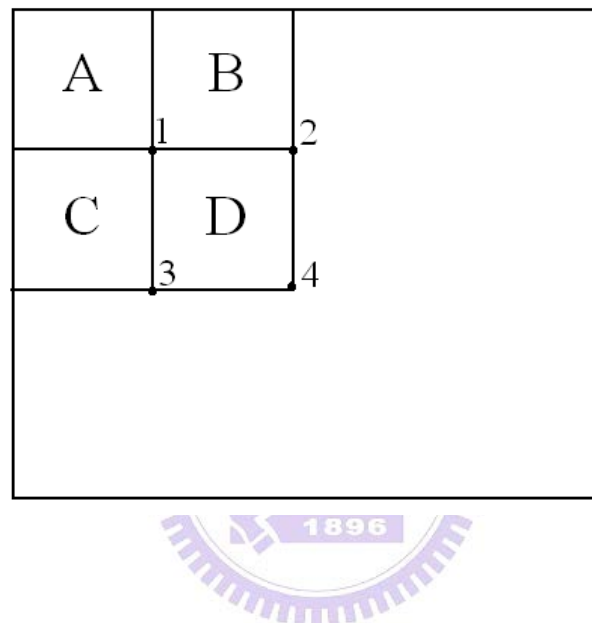


Fig. 2.9. The sum of pixels in rectangle D can be computed as $4+1 - (2+3)$.

A variant of AdaBoost is used to select the features and train the classifier. The objective of the AdaBoost algorithm is to form a stronger classifier by combining a collection of weak classification functions. If the correct rate of a weak classifier is above 50%, it is a good weak classification function. Finally, the Haar cascade classifier is built by stringing strong classifiers for detecting face region more accurately.

Chapter 3 Video Frame Human Recognition Procedure

3.1 Human Representation

In video and image processing, the dimensions of image data are often very large. Each image data is suggested to transform from high-dimensional space into low-dimensional space to obtain a small set of composite feature for human recognition. There are many well-known transformation methods for human recognition, for example, wavelet transformation, Fourier transformation, Locally Linear Embedding (LLE), Multi Dimension Scaling (MDS), Principal Component Analysis (PCA), eigenspace transformation (EST), and so on. However, PCA based on the global covariance matrix of the full set of image data is designed for efficient data representation, not sensitive to the class structure existent in the image data. In order to enhance the discriminatory power of several image features, Etemad and Chellappa [14] use linear discriminant analysis (LDA), also called Canonical Analysis [6], which can be used to optimize the class separability of different image classes and improve the classification performance. To this end, the features are obtained by maximizing between-class and minimizing within-class variations. Unfortunately, this approach has high computation cost when applying to large images. It was only tested with small images. Here we call this approach canonical space transformation (CST). Combining EST based on PCA with CST based on CA, our approach reduces the data dimensionality and optimizes the class separability of different gait sequences.

Images in high-dimensional space are first converted into low-dimensional eigenspace using EST. The obtained vector thus is further transformed to a smaller canonical space using CST. Human recognition is accomplished in the canonical

space. Fig. 3.1 shows the processing steps that generate feature vectors by eigenspace transformation and canonical space transformation each image is converted to an one-dimension canonical vector. Apparently, the reduced dimensionality results in concomitant decrease in computation cost.

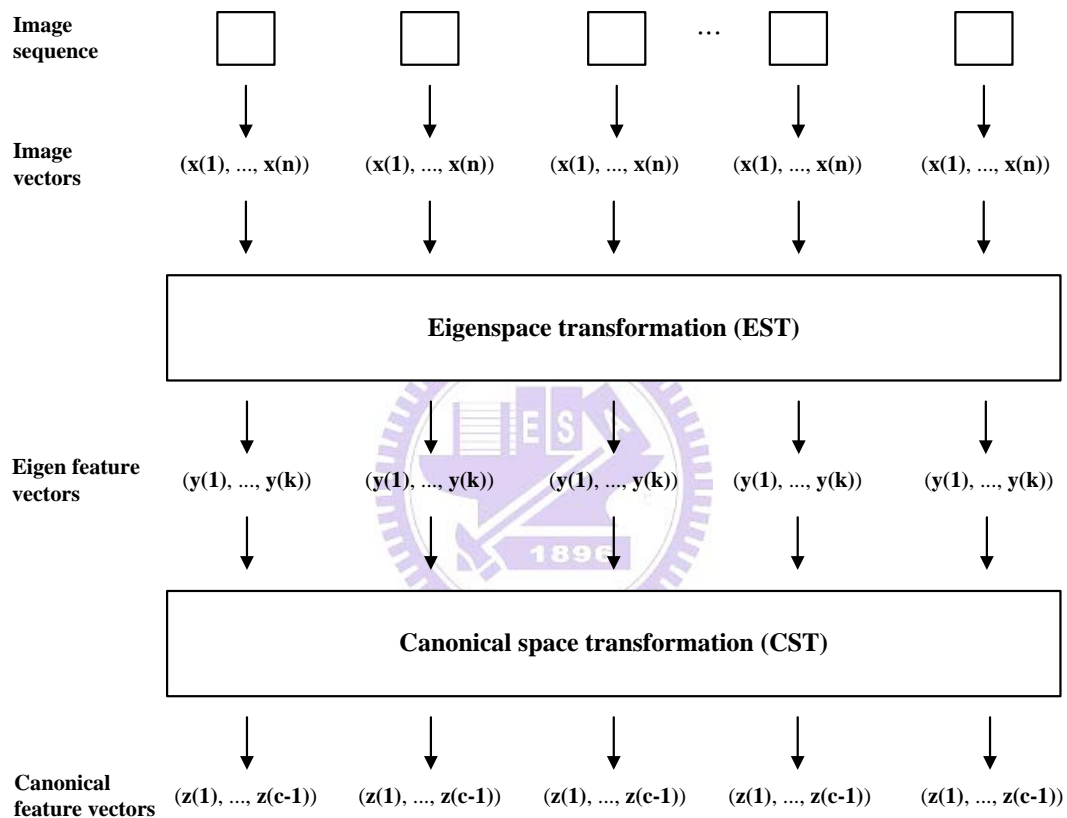


Fig. 3.1. The structure of human recognition by gait or face image sequence.


Assume that there are c classes of person to be learned. Each class is represented by a specific viewing angle gait image sequences or face image sequences of a person, which were captured and used as the training image data. Training image $\mathbf{x}'_{i,j}$ is the

j -th image in the i -th class person, and N_i is the number of images of the i -th class image sequence acquired. The total number of training image data in training set is $N_T = N_1 + N_2 + \dots + N_c$. The training set can be represented as

$$\left[\mathbf{x}'_{1,1}, \dots, \mathbf{x}'_{1,N_1}, \mathbf{x}'_{2,1}, \dots, \mathbf{x}'_{c,N_c} \right] \quad (3.1)$$

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

First, the intensity of each image data is normalized by

$$\mathbf{x}_{i,j} = \frac{\mathbf{x}'_{i,j}}{\|\mathbf{x}'_{i,j}\|} \quad (3.2)$$


Then, the mean pixel value for the training set is obtained by

$$\mathbf{m}_x = \frac{1}{N_T} \sum_{i=1}^c \sum_{j=1}^{N_i} \mathbf{x}_{i,j} \quad (3.3)$$

By subtracting the mean \mathbf{m}_x from each image data, the training set can be rewritten

as a $n \times N_T$ matrix \mathbf{X} , with each image data $\mathbf{x}'_{i,j}$ forms a column of \mathbf{X} , then

$$\mathbf{X} = \left[\mathbf{x}_{1,1} - \mathbf{m}_x, \dots, \mathbf{x}_{1,N_1} - \mathbf{m}_x, \dots, \mathbf{x}_{c,N_c} - \mathbf{m}_x \right] \quad (3.4)$$

3.1.1 Eigenspace Transformation (EST)

EST is used to reduce the dimensionality of an input space by mapping the image data from high-dimensional space into low-dimensional space while maintaining the minimum mean-square error to avoid information loss. EST uses the eigenvalues and eigenvectors generated by the image data covariance matrix to rotate the original image data coordinates along the directions of maximum variance sequentially. We can compute image data covariance matrix \mathbf{R} , then

$$\mathbf{R} = \mathbf{X}\mathbf{X}^T \quad (3.5)$$

where \mathbf{R} is a square, symmetric $n \times n$ matrix.

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

$$\lambda_i \mathbf{e}_i = \mathbf{R}\mathbf{e}_i, \quad i = 1, 2, \dots, K \quad (3.6)$$

In order to solve Eq. (3.6), we need to compute the eigenvalues and eigenvectors of the $n \times n$ matrix \mathbf{R} . But the dimensionality of \mathbf{R} is the image data size, it is often extremely large. Based on singular value decomposition, we can obtain the eigenvalues and eigenvectors by computing another image data covariance matrix $\tilde{\mathbf{R}}$ instead, that is

$$\tilde{\mathbf{R}} = \mathbf{X}^T \mathbf{X} \quad (3.7)$$

where $\tilde{\mathbf{R}}$ is a square, symmetric $N_T \times N_T$ matrix which is much smaller than $n \times n$ of \mathbf{R} .

If the rank of the matrix $\tilde{\mathbf{R}}$ is K , then the K nonzero eigenvalues of $\tilde{\mathbf{R}}$, $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_K$, and corresponding eigenvectors $\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2, \dots, \tilde{\mathbf{e}}_K$ which are related to those in \mathbf{R} by

$$\begin{cases} \lambda_i = \tilde{\lambda}_i \\ \mathbf{e}_i = \lambda_i^{-\frac{1}{2}} \mathbf{X} \tilde{\mathbf{e}}_i \end{cases}, \quad i=1, 2, \dots, K \quad (3.8)$$

These K eigenvectors are used as an orthogonal basis to span a new vector space. Each image data can be projected to a point in this K -dimensional space. Based on the theory of PCA, each image data can be approximated by taking only the k largest eigenvalues $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_k|$, $k \leq K$, and their corresponding 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 data points that are the projections of the original image data $\mathbf{x}_{i,j}$ by the equation

$$\mathbf{y}_{i,j} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k]^T \mathbf{x}_{i,j}, \quad i=1, 2, \dots, c \text{ and } j=1, 2, \dots, N_c \quad (3.9)$$

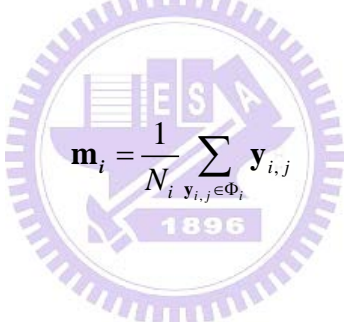
This matrix $[\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k]^T$ is called the eigenspace transformation matrix. After this transformation, each original image data $\mathbf{x}_{i,j}$ can be approximated by the linear combination of these k eigenvectors and $\mathbf{y}_{i,j}$ is a one-dimensional vector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

3.1.2 Canonical Space Transformation (CST)

According to the theory of canonical analysis [15], we suppose that $\{\Phi_1, \Phi_2, \dots, \Phi_c\}$ represents the classes of transformed vectors by eigenspace transformation and $\mathbf{y}_{i,j}$ is the j -th vector in class i . The mean vector of entire set is obtained by

$$\mathbf{m}_y = \frac{1}{N_T} \sum_{i=1}^c \sum_{j=1}^{N_i} \mathbf{y}_{i,j} \quad (3.10)$$

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



$$\mathbf{m}_i = \frac{1}{N_i} \sum_{\mathbf{y}_{i,j} \in \Phi_i} \mathbf{y}_{i,j} \quad (3.11)$$

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

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

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

where \mathbf{S}_b represents the mean of between-class vectors distance and \mathbf{S}_w represents the mean of within-class distance vectors distance. The objective is to maximize \mathbf{S}_b

and minimize \mathbf{S}_w simultaneously, which is known as the generalized Fisher linear discriminant function and obtained by

$$\mathbf{J}(\mathbf{W}) = \frac{\mathbf{W}^T \mathbf{S}_b \mathbf{W}}{\mathbf{W}^T \mathbf{S}_w \mathbf{W}} \quad (3.12)$$

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

$$\frac{\partial \mathbf{J}}{\partial \mathbf{W}} = 0 \quad (3.13)$$

We suppose the \mathbf{W}^* is the optimal solution where the column vector \mathbf{w}_i^* is a generalized eigenvector corresponding to the i -th largest eigenvalues λ_i . Based on the theory of canonical analysis [15], we can solve Eq. (3.13) as follows

$$\mathbf{S}_b \mathbf{w}_i^* = \lambda_i \mathbf{S}_w \mathbf{w}_i^* \quad (3.14)$$

After Eq. (3.14) is solved, we will obtain $c-1$ nonzero eigenvalues and associated eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{c-1}$ that create another orthogonal basis and span a $(c-1)$ -dimensional canonical space. By using these bases, each data point in eigenspace can be transformed to another data point in canonical space by

$$\mathbf{z}_{i,j} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{c-1}]^T \mathbf{y}_{i,j} \quad (3.15)$$

where $\mathbf{z}_{i,j}$ represents the new data point. This orthogonal basis $[\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{c-1}]^T$ is called the canonical space transformation matrix.

By merging Eq. (3.9) and Eq. (3.15), each image data can be transformed into a new data point in the $(c-1)$ -dimensional space by

$$\mathbf{z}_{i,j} = \mathbf{H}\mathbf{x}_{i,j} \quad (3.16)$$

$$\mathbf{C}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{z}_{i,j} \quad (3.17)$$

where $\mathbf{H} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{c-1}]^T [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k]^T$ and \mathbf{C}_i is the centroid of class i .

3.2 Human Recognition

3.2.1 Person Recognition by Gait Image Classification in a Long Distance Setting

When a video stream is inputted for human gait recognition, we extract image frames from the video first. Then we use background model of Section 2.2 to extract foreground subject from the scene. The foreground object is a binary image, also called as the gait template which is converted to low-dimensional eigenspace using EST in high-dimensional image space. The obtained vector thus is further projected to a smaller canonical space using CST. As described in Section 3.1, each gait template is transformed to a $(c-1)$ -dimensional vector by EST and CST methods. To recognize a gait template from the video frame sequence in the canonical space, the minimal *Euclidean distance* to each centroid is used. The recognized class is assigned to the class which assumes the minimal distance between a test gait template “ g ,” and

the gait class center “ \mathbf{G}_i ,” as given by

$$j = \arg \min_i \|g - \mathbf{G}_i\|, \quad i = 1, 2, \dots, c_g \quad (3.18)$$

where c_g is number of gait class and j is the result of person recognition.

3.2.2 Person Recognition by Face Image Classification in a Short Distance Setting

When a video stream is inputted for human face recognition, we extract image frames from the video first. Then we use face detection of Section 2.3 to extract human face from the scene. The human face is a grayscale image, also called as the face template which is converted to low-dimensional eigenspace using EST in high-dimensional image space. The obtained vector thus is further projected to a smaller canonical space using CST. As described in Section 3.1, each face template is transformed to a $(c-1)$ -dimensional vector by EST and CST methods. To recognize a face template from the video frame sequence in the canonical space, the minimal *Euclidean distance* to each centroid is used. The recognition class is assigned to the class which assumes the minimal distance between a test face template “ f ,” and the face class center “ \mathbf{F}_i ,” as given by

$$j = \arg \min_i \|f - \mathbf{F}_i\|, \quad i = 1, 2, \dots, c_f \quad (3.19)$$

where c_f is number of face class and j is the result of person recognition.

3.2.3 Majority Vote

Due to the above classification which use each frame to do the human recognition in the video, there may have misclassifications in some frames. To overcome this problem, we have adopted the majority vote to conduct the human recognition. Fig. 3.2 shows the structure of the human classification.

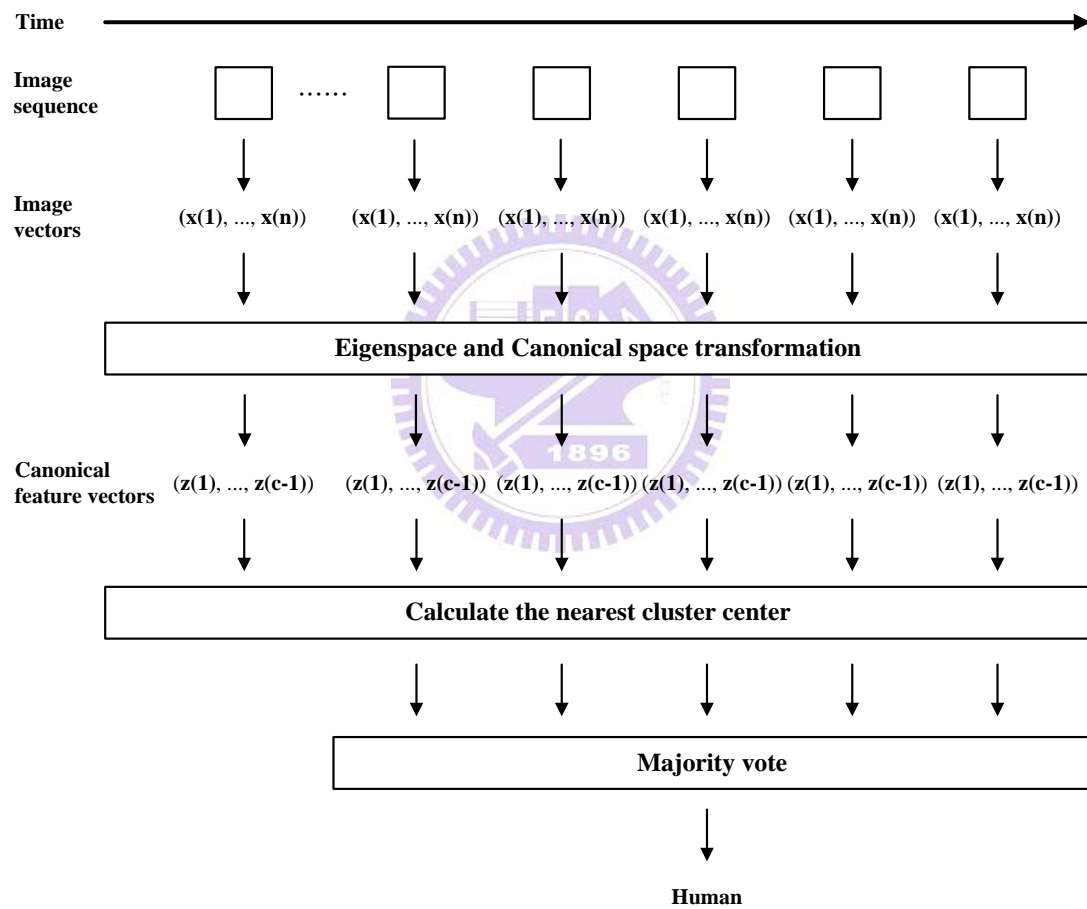


Fig. 3.2. The structure of human classification.

Chapter 4 Experimental Results

In our experiment, we tested our system on videos taken by near infrared (NIR) camera (KMT-1651N with 12 lighting led cells) in our laboratory at the 5th Engineering Building in NCTU campus. We use two cameras for human recognition from facial and walking videos. The first is *far* NIR camera with a lens focus 4.3mm is set up at the location far from the object about 6 meters, and the second is *near* NIR camera with a lens focus 6.0mm is setup at the location far from the object about 2.5 meters. These cameras have a frame rate of 30 frames per second and image resolution is 320×240 pixels. The background of the experiment environment is in real life and the illumination of the environment is 398 Lux in the bright environment and 0.26 Lux in the dark environment, respectively. Fig. 4.1(a) shows the scene of human recognition for gait videos in the bright environment. Fig. 4.1(b) shows the scene of human recognition for gait videos in the dark environment. Fig. 4.1(c) shows the scene of human recognition for face videos in the bright environment. Fig. 4.1(d) shows the scene of human recognition for face videos in the dark environment.

Our LAB gait multi-angle database consists of 32 image sequences consisting of eight persons walking in the bright and dark environments. Each person was done four times producing four sequences at three different walking angles (0° , 45° , and 315°) with respect to the person frontal view in a clockwise sense. Thus, it contains a total of $4 \times 8 \times 3 = 96$ walking video sequences for human recognition. Fig. 4.2 shows the examples video sequence form our LAB gait multi-angle databases. On the other hand, our LAB face database consists of 36 video sequences consisting of nine persons in the bright and dark environments. Moreover, each person has four face video sequences in the bright and dark environments. Fig. 4.3 shows the examples video sequence form our LAB face databases.

Furthermore, we also tested our system on CASIA database [16] which contains multi-view gait sequences. The CASIA database [17] consists of 288 image sequences depicting 48 persons. Each person is depicted in six sequences at 11 different viewing angles (0° , 18° , 36° , 54° , 72° , 90° , 108° , 126° , 144° , 162° , and 180°) with respect to the person frontal view in a counterclockwise manner. Thus, it contains a total of $6 \times 48 \times 11 = 3168$ gait sequences. Binary body image masks are provided in the CASIA database. Eleven video frames depicting person in the CASIA database from each viewing angle are illustrated in Fig. 4.4.

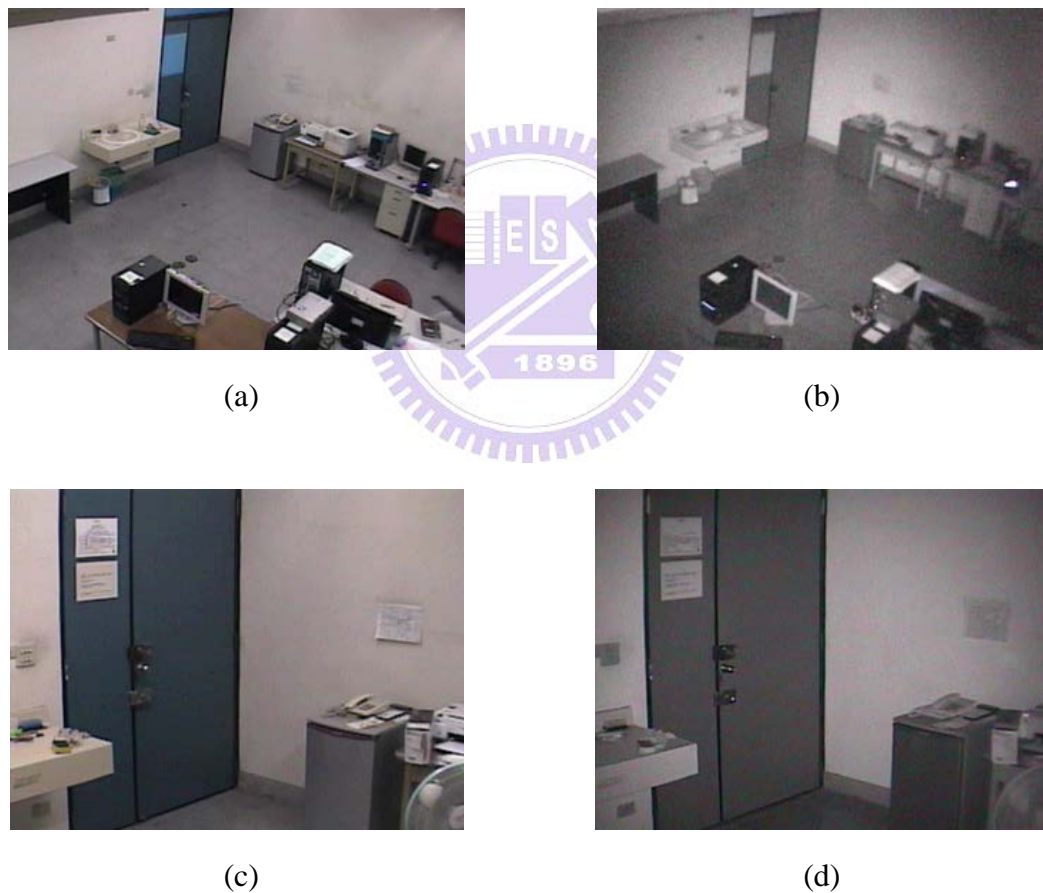


Fig. 4.1. (a) The scene of human gait recognition in the **bright** environment. (b) The scene of human gait recognition in the **dark** environment. (c) The scene of human face recognition in the **bright** environment. (d) The scene of human face recognition in the **dark** environment.



(a)



(b)

Fig. 4.2. Example video sequences used in our experiments. (a) and (b) are typical video sequences for gaits of LAB in the **bright** and **dark** environments. From top to bottom: walking 0° , walking 45° , and walking 315° , respectively.



(a)



(b)

Fig. 4.3. Example video sequences used in our experiments. (a) and (b) are typical video sequences for face of LAB in the **bright** and **dark** environments. From top to bottom: walking 0° , walking 45° , and walking 315° , respectively.

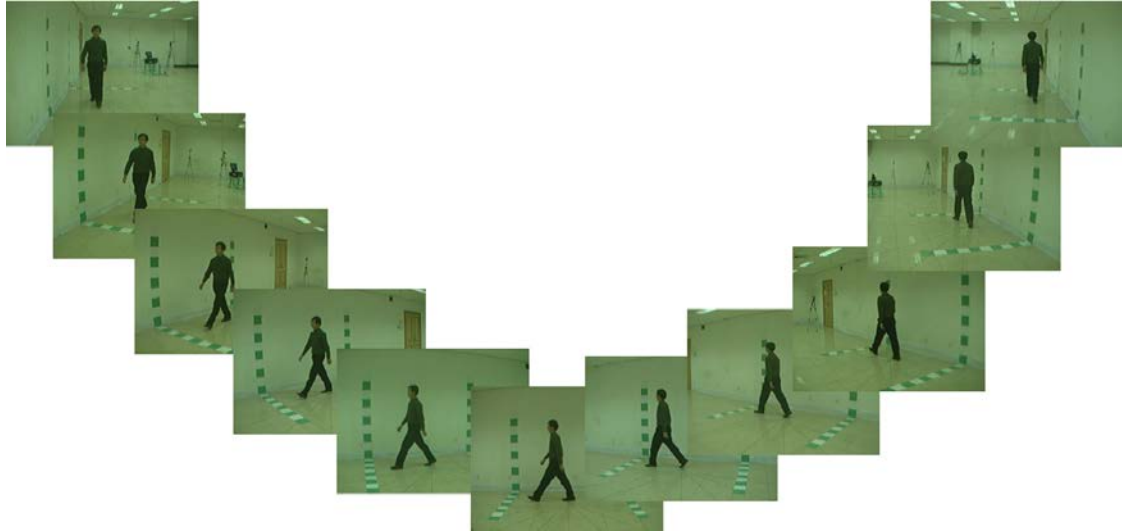


Fig. 4.4. Eleven video frames depicting person of the CASIA multi-view gait recognition database from different viewing angles.



4.1 Background Model Construction and Foreground Extraction

The background model is used for extracting the foreground object or subject. In our system, we first record a video of background (like Fig. 4.1(a) and Fig. 4.1(b)) about two seconds in the bright and dark environments to build the background models. After building the grayscale value and the HSV color space background models, we will detect the foreground pixels by using Eq. (2.7) and Eq. (2.8) in Section 2.2.3. Then, we continue to process the foreground image by using the shadow filter, the opening and the closing operations.

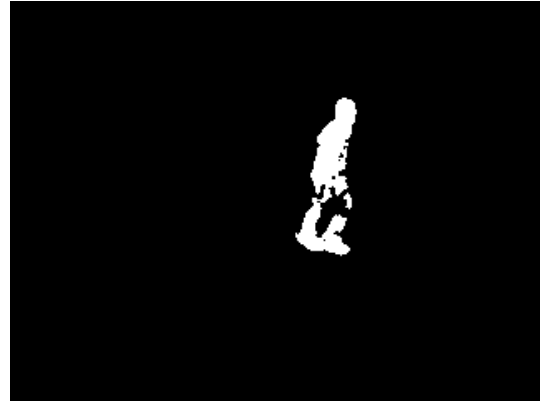
In order to get the optimal result of foreground detection, we have to adjust some threshold in our system. We set $k_{gray} = 3.0$ and $k_{gray} = 2.0$ for the grayscale value background models and $k_v = 1.6$ and $k_v = 1.1$ for the HSV color background

models in the bright and dark environments, respectively. The same threshold is used in the bright and dark environments for shadow filter. We set $L_{ncc} = 0.965$ in the grayscale value space and $k_H = 1.5$ and $k_S = 1.5$ in the HSV color space to detect shadow pixels. Then, we simply introduce a threshold on the histograms in X and Y directions to determine the minimal size of foreground images, and then resize the images to 64×48 for normalization.

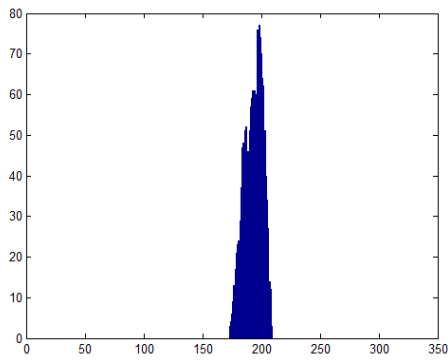
Fig. 4.5(a) shows an image frame in the bright environment. Fig. 4.5(b) shows the binary image after performing foreground detection in the bright environment. Figs. 4.5(c) and 4.5(d) show the projection of Fig. 4.5(b) onto the X and Y directions, respectively. We can find the boundary coordinates of X and Y direction by observing the projection histogram. We used these boundary coordinates to define a rectangle to segment foreground region from Fig. 4.5(b). Fig 4.5(e) shows the result of foreground region segmentation in the bright environment. Fig. 4.6(a) shows an image frame in the dark environment. Fig. 4.6(b) shows the binary image after performing foreground detection in the dark environment. Figs. 4.6(c) and 4.6(d) show the projection of Fig. 4.6(b) onto the X and Y directions, respectively. We can find the boundary coordinates of X and Y direction by observing the projection histogram. We used these boundary coordinates to define a rectangle to segment foreground region from Fig. 4.6(b). Fig 4.6(e) shows the result of foreground region segmentation in the dark environment.



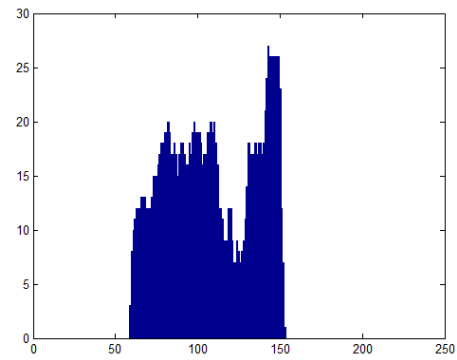
(a)



(b)



(c)



(d)

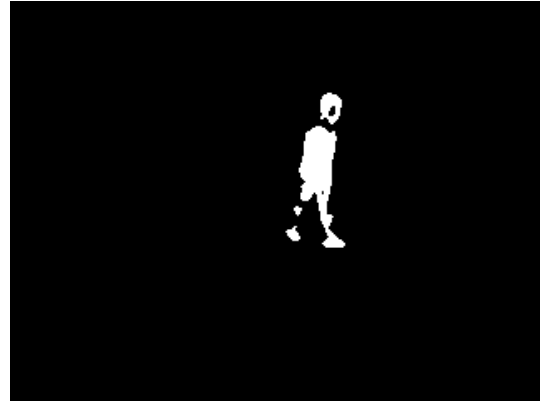


(e)

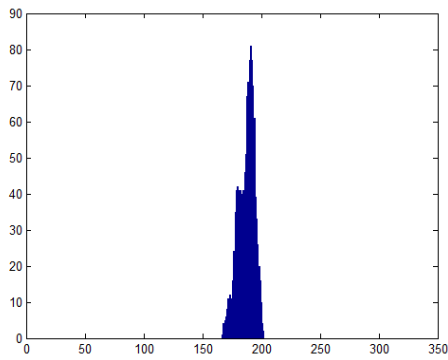
Fig. 4.5. Results of foreground detection. (a) an image frame in the **bright** environment, (b) binary image after performing foreground detection in the **bright** environment, (c) projection of (b) onto X direction, (d) projection of (b) onto Y direction, (e) foreground region segmentation in the **bright** environment.



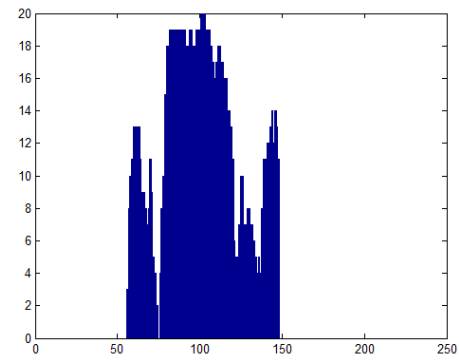
(a)



(b)



(c)



(d)



(e)

Fig. 4.6. Results of foreground detection. (a) an image frame in the **dark** environment, (b) binary image after performing foreground detection in the **dark** environment, (c) projection of (b) onto X direction, (d) projection of (b) onto Y direction, (e) foreground region segmentation in the **dark** environment.

4.2 Experiments on our LAB Multi-Angle Gait Database

In our LAB multi-angle gait database, we use two methods to test our system. First method is single-angle human recognition by walking videos taken in our LAB database. Second method is multi-angle human recognition by walking videos taken in our LAB database.

4.2.1 Single-Angle Human Gait Recognition

In this experiment, activity videos depicting three walk sequences at one specific walking angle performed by eight persons in our LAB database were used for training. The recognition rate is measured based on leave-one-out strategy. Table I shows the human recognition rates at each walking angle without majority vote in the bright environment. Table II shows the human recognition rates at each walking angle with majority vote of three in the bright environment. Table III shows the human recognition rates at each walking angle with majority vote of five in the bright environment. Table IV shows the human recognition rates at each walking angle without majority vote in the dark environment. Table V shows the human recognition rates at each walking angle with majority vote of three in the dark environment. Table VI shows the human recognition rates at each walking angle with majority vote of five in the dark environment. In these tables, W_{0° represents the case of classification of person in 0° walking angle, W_{45° represents the case of classification of person in 45° walking angle, and W_{315° represents the case of classification of person in 315° walking angle.

TABLE I

THE HUMAN **GAIT** RECOGNITION RATES AT SPECIFIC WALKING ANGLE IN THE
BRIGHT ENVIRONMENT, WITHOUT MAJORITY VOTE

	W_{0°	W_{45°	W_{315°
Accuracy	94.51% (2581/2731)	91.60% (2867/3130)	89.92% (2846/3165)
False alarm rate	0.78% (150/19117)	1.20% (263/21910)	1.44% (319/22155)
Average Accuracy	91.89% (8294/9026)		
Average False alarm rate	1.16% (732/63182)		

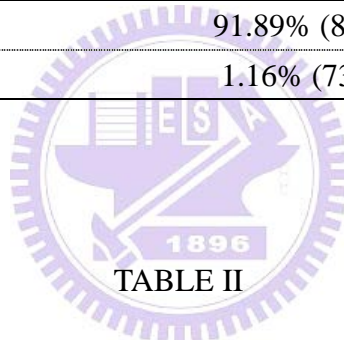


TABLE II

THE HUMAN **GAIT** RECOGNITION RATES AT SPECIFIC WALKING ANGLE IN THE
BRIGHT ENVIRONMENT, WITH MAJORITY VOTE OF THREE

	W_{0°	W_{45°	W_{315°
Accuracy	95.54% (2548/2667)	93.80% (2876/3066)	92.42% (2866/3101)
False alarm rate	0.64% (119/18669)	0.89% (190/21462)	1.08% (235/21707)
Average Accuracy	93.84% (8290/8834)		
Average False alarm rate	0.88% (544/61838)		

TABLE III

THE HUMAN **GAIT** RECOGNITION RATES AT SPECIFIC WALKING ANGLE IN THE
BRIGHT ENVIRONMENT, WITH MAJORITY VOTE OF FIVE

	W_{0°	W_{45°	W_{315°
Accuracy	96.27% (2506/2603)	95.97% (2881/3002)	95.23% (2892/3037)
False alarm rate	0.53% (97/18221)	0.58% (121/21014)	0.68% (145/21259)
Average Accuracy	95.80% (8279/8642)		
Average False alarm rate	0.60% (363/60494)		

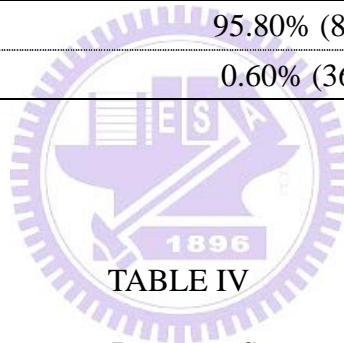


TABLE IV

THE HUMAN **GAIT** RECOGNITION RATES AT SPECIFIC WALKING ANGLE IN THE
DARK ENVIRONMENT, WITHOUT MAJORITY VOTE

	W_{0°	W_{45°	W_{315°
Accuracy	94.37% (2783/2949)	80.25% (2596/3235)	84.64% (2701/3191)
False alarm rate	0.80% (166/20643)	2.82% (639/22645)	2.19% (490/22337)
Average Accuracy	86.19% (8080/9375)		
Average False alarm rate	1.97% (1295/65625)		

TABLE V

THE HUMAN **GAIT** RECOGNITION RATES AT SPECIFIC WALKING ANGLE IN THE **DARK** ENVIRONMENT, WITH MAJORITY VOTE OF THREE

	W_{0°	W_{45°	W_{315°
Accuracy	95.46% (2754/2885)	84.11% (2667/3171)	88.04% (2753/3127)
False alarm rate	0.65% (131/20195)	2.27% (504/22197)	1.71% (374/21889)
Average Accuracy	89.01% (8174/9183)		
Average False alarm rate	1.57% (1009/64281)		

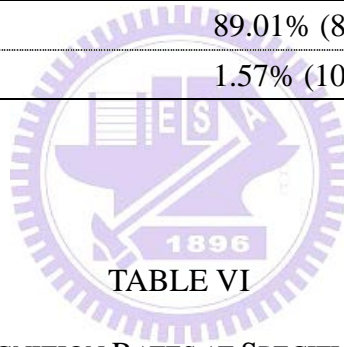


TABLE VI

THE HUMAN **GAIT** RECOGNITION RATES AT SPECIFIC WALKING ANGLE IN THE **DARK** ENVIRONMENT, WITH MAJORITY VOTE OF FIVE

	W_{0°	W_{45°	W_{315°
Accuracy	96.35% (2718/2821)	88.90% (2762/3107)	92.23% (2825/3063)
False alarm rate	0.52% (103/19747)	1.59% (345/21749)	1.11% (238/21441)
Average Accuracy	92.37% (8305/8991)		
Average False alarm rate	1.09% (686/62937)		

4.2.2 *Multi-Angle Human Gait Recognition*

In this experiment, activity videos depicting three walk sequences at three walking angle performed by eight persons in our LAB database were used for training. The recognition rate is measured based on leave-one-out strategy. Table VII shows the recognition rates at all walking angle without majority vote in the bright environment. Table VIII shows the recognition rates at all walking angle with majority vote of three in the bright environment. Table IX shows the recognition rates at all walking angle with majority vote of five in the bright environment. Table X shows the recognition rates at all walking angle without majority vote in the dark environment. Table XI shows the recognition rates at all walking angle with majority vote of three in the dark environment. Table XII shows the recognition rates at all walking angle with majority vote of five in the dark environment. In these tables, *PA* represents the case of classification of person and walking angle as well. *A* stands for the case of classification of walking angle in all walking angle and *P* is the classification of person in all walking angle.

TABLE VII

THE RECOGNITION RATES OF WALKING VIDEOS IN THE **BRIGHT** ENVIRONMENT,
WITHOUT MAJORITY VOTE

Case	<i>PA</i>	<i>A</i>	<i>P</i>
Average Accuracy	86.98% (7851/9026)	98.49% (8890/9026)	87.54% (7901/9026)
Average False alarm rate	0.57% (1175/207598)	0.75% (136/18052)	1.78% (1125/63182)

TABLE VIII

THE RECOGNITION RATES OF WALKING VIDEOS IN THE **BRIGHT** ENVIRONMENT,
WITH MAJORITY VOTE OF THREE

Case	<i>PA</i>	<i>A</i>	<i>P</i>
Average Accuracy	89.89% (7941/8834)	98.98% (8744/8834)	90.32% (7979/8834)
Average False alarm rate	0.44% (893/203182)	0.51% (90/17668)	1.38% (855/61838)

TABLE IX

THE RECOGNITION RATES OF WALKING VIDEOS IN THE **BRIGHT** ENVIRONMENT,
WITH MAJORITY VOTE OF FIVE

Case	<i>PA</i>	<i>A</i>	<i>P</i>
Average Accuracy	92.96% (8034/8642)	99.21% (8574/8642)	93.36% (8068/8642)
Average False alarm rate	0.31% (608/198766)	0.39% (68/17284)	0.95% (574/60494)

TABLE X

THE RECOGNITION RATES OF WALKING VIDEOS IN THE **DARK** ENVIRONMENT,
WITHOUT MAJORITY VOTE

Case	<i>PA</i>	<i>A</i>	<i>P</i>
Average Accuracy	79.02% (7408/9375)	96.31% (9029/9375)	80.23% (7522/9375)
Average False alarm rate	0.91% (1967/215625)	1.85% (346/18750)	2.82% (1853/65625)

TABLE XI

THE RECOGNITION RATES OF WALKING VIDEOS IN THE **DARK** ENVIRONMENT,
WITH MAJORITY VOTE OF THREE

Case	<i>PA</i>	<i>A</i>	<i>P</i>
Average Accuracy	81.89% (7520/9183)	97.01% (8908/9183)	83.15% (7636/9183)
Average False alarm rate	0.79% (1663/211209)	1.50% (275/18366)	2.41% (1547/64281)

TABLE XII

THE RECOGNITION RATES OF WALKING VIDEOS IN THE **DARK** ENVIRONMENT,
WITH MAJORITY VOTE OF FIVE

Case	<i>PA</i>	<i>A</i>	<i>P</i>
Average Accuracy	87.15% (7836/8991)	97.58% (8773/8991)	88.17% (7927/8991)
Average False alarm rate	0.56% (1155/206793)	1.21% (218/17982)	1.69% (1064/62937)

4.3 Recognition Result on the CASIA Multi-View Gait Database

In the CASIA multi-view gait database, we use two methods to test our system. First method is single-view human recognition in the CASIA database. Second method is multi-view human recognition in the CASIA database.

4.3.1 Single-View Human Gait Recognition

In this experiment, activity videos depicting five walk sequences at one specific viewing angle performed by 48 persons in the CASIA database were used for training. The recognition rate is measured based on leave-one-out strategy. Table XIX shows the human recognition rates at each viewing angle without majority vote in the CASIA database. Table XX shows the human recognition rates at each viewing angle with majority vote of three in the CASIA database. Table XXI shows the human recognition rates at each viewing angle with majority vote of five in the CASIA database. In these tables, W_{0° represents the case of classification of person in 0° viewing angle, W_{18° represents the case of classification of person in 18° viewing angle, W_{36° represents the case of classification of person in 36° viewing angle, W_{54° represents the case of classification of person in 54° viewing angle, W_{72° represents the case of classification of person in 72° viewing angle, W_{90° represents the case of classification of person in 90° viewing angle, W_{108° represents the case of classification of person in 108° viewing angle, W_{126° represents the case of classification of person in 126° viewing angle, W_{144° represents the case of classification of person in 144° viewing angle, W_{162° represents the case of

classification of person in 162° viewing angle, and W_{180° represents the case of classification of person in 180° viewing angle.

TABLE XIII

THE HUMAN RECOGNITION RATES AT SPECIFIC VIEWING ANGLE IN THE CASIA DATABASE, WITHOUT MAJORITY VOTE

	W_{0°	W_{18°	W_{36°
Accuracy	95.27% (27530/28896)	85.38% (26422/30946)	80.39% (24386/30334)
False alarm rate	0.10% (1366/1358112)	0.31% (4524/1454462)	0.42% (5948/1425698)
	W_{54°	W_{72°	W_{90°
Accuracy	80.75% (22486/27845)	90.77% (18413/20286)	90.47% (16941/18726)
False alarm rate	0.41% (5359/1308715)	0.20% (1873/953442)	0.20% (1785/880122)
	W_{108°	W_{126°	W_{144°
Accuracy	89.01% (17898/20108)	88.75% (21732/24487)	91.26% (23057/25266)
False alarm rate	0.23% (2210/945076)	0.24% (2755/1150889)	0.19% (2209/1187502)
	W_{162°	W_{180°	
Accuracy	92.72% (23209/25030)	95.81% (24771/25853)	
False alarm rate	0.15% (1821/1176410)	0.09% (1082/1215091)	
Average Accuracy	88.86% (246845/277777)		
Average False alarm rate	0.24% (30932/13055519)		

TABLE XIV

THE HUMAN RECOGNITION RATES AT SPECIFIC VIEWING ANGLE IN THE CASIA
DATABASE, WITH MAJORITY VOTE OF THREE

	W_{0°	W_{18°	W_{36°
Accuracy	96.64% (27368/28320)	87.96% (26712/30370)	83.19% (24756/29758)
False alarm rate	0.07% (952/1331040)	0.26% (3658/1427390)	0.36% (5002/1398626)
	W_{54°	W_{72°	W_{90°
Accuracy	84.11% (22937/27269)	93.97% (18521/19710)	93.58% (16984/18150)
False alarm rate	0.34% (4332/1281643)	0.13% (1189/926370)	0.14% (1166/853050)
	W_{108°	W_{126°	W_{144°
Accuracy	92.06% (17981/19532)	91.98% (21993/23911)	93.73% (23141/24690)
False alarm rate	0.17% (1551/918004)	0.17% (1918/1123817)	0.13% (1549/1160430)
	W_{162°	W_{180°	
Accuracy	94.49% (23106/24454)	96.87% (24485/25277)	
False alarm rate	0.12% (1348/1149338)	0.07% (792/1188019)	
Average Accuracy	91.36% (247984/271441)		
Average False alarm rate	0.18% (23457/12757727)		

TABLE XV

THE HUMAN RECOGNITION RATES AT SPECIFIC VIEWING ANGLE IN THE CASIA
DATABASE, WITH MAJORITY VOTE OF FIVE

	W_{0°	W_{18°	W_{36°
Accuracy	97.97% (27181/27744)	92.36% (27518/29794)	89.57% (26137/29182)
False alarm rate	0.04% (563/1303968)	0.16% (2276/1400318)	0.22% (3045/1371554)
	W_{54°	W_{72°	W_{90°
Accuracy	90.75% (24225/26693)	97.08% (18576/19134)	96.89% (17027/17574)
False alarm rate	0.20% (2468/1254571)	0.06% (558/899298)	0.07% (547/825978)
	W_{108°	W_{126°	W_{144°
Accuracy	96.15% (18226/18956)	95.76% (22345/23335)	96.39% (23243/24114)
False alarm rate	0.08% (730/890932)	0.09% (990/1096745)	0.08% (871/1133358)
	W_{162°	W_{180°	
Accuracy	96.55% (23055/23878)	97.77% (24150/24701)	
False alarm rate	0.07% (823/1122266)	0.05% (551/1160947)	
Average Accuracy	94.94% (251683/265105)		
Average False alarm rate	0.11% (13422/12459935)		

4.3.2 Multi-View Human Gait Recognition

In this experiment, activity videos depicting five walk sequences at eleven viewing angle performed by 48 persons in the CASIA database were used for training. The recognition rate is measured based on leave-one-out strategy. Table XXII shows the recognition rates at all viewing angle without majority vote in the CASIA database. Table XXIII shows the recognition rates at all viewing angle with majority vote of three in the CASIA database. Table XXIV shows the recognition rates at all viewing angle with majority vote of five in the CASIA database. In these tables, *PA* represents the case of classification of person and viewing angle as well. *A* stands for the case of classification of viewing angle in all viewing angle and *P* is the classification of person in all viewing angle.



THE RECOGNITION RATES AT ALL VIEWING ANGLES IN THE CASIA DATABASE,
WITHOUT MAJORITY VOTE

Case	<i>PA</i>	<i>A</i>	<i>P</i>
Average Accuracy	60.42% (167819/277777)	72.17% (200464/277777)	78.03% (216740/277777)
Average False alarm rate	0.08% (109958/146388479)	2.78% (77313/2777770)	0.47% (61037/13055519)

TABLE XVII

THE RECOGNITION RATES AT ALL VIEWING ANGLES IN THE CASIA DATABASE,
WITH MAJORITY VOTE OF THREE

Case	<i>PA</i>	<i>A</i>	<i>P</i>
Average Accuracy	62.21% (168866/271441)	74.51% (202247/271441)	80.63% (218854/271441)
Average False alarm rate	0.07% (102575/143049407)	2.55% (69194/2714410)	0.41% (52587/12757727)

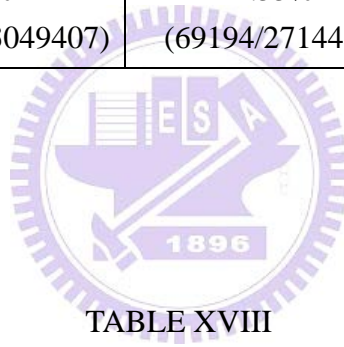


TABLE XVIII

THE RECOGNITION RATES AT ALL VIEWING ANGLES IN THE CASIA DATABASE,
WITH MAJORITY VOTE OF FIVE

Case	<i>PA</i>	<i>A</i>	<i>P</i>
Average Accuracy	70.00% (185573/265105)	78.03% (206870/265105)	87.46% (231866/265105)
Average False alarm rate	0.06% (79532/139710335)	2.20% (58235/2651050)	0.27% (33239/12459935)

4.4 Experiments on our LAB Face Database

4.4.1 Human Face Recognition

In this experiment, a specific video randomly chosen from the four repeated videos is used for recognition and the other three are used for training, and this procedure is repeated in turn for four times. It is to be noted that all the facial recognition from video will be mostly effective only when the person is far from the *near* camera during the range 1.8–2.5m. In this distance range, the face images captured for the camera were large and clear enough for face recognition purpose. Hence, all the experiment data shown below, all the face images were captured in the distance range of 1.8–2.5m from the *near* camera. Table XIII shows the recognition rates of human face without majority vote in the bright environment. Table XIV shows the recognition rates of human face with majority vote of three in the bright environment. Table XV shows the recognition rates of human face with majority vote of five in the bright environment. Table XVI shows the recognition rates of human face without majority vote in the dark environment. Table XVII shows the recognition rates of human face with majority vote of three in the dark environment. Table XVIII shows the recognition rates of human face with majority vote of five in the dark environment.

TABLE XIX

THE RECOGNITION RATES OF HUMAN FACE VIDEOS IN THE **BRIGHT** ENVIRONMENT,
WITHOUT MAJORITY VOTE

	Person 1	Person 2	Person 3
Accuracy	96.01% (530/552)	94.99% (739/778)	98.22% (496/505)
False alarm rate	1.12% (59/5252)	0.22% (11/5026)	1.26% (67/5299)
	Person 4	Person 5	Person 6
Accuracy	87.15% (631/724)	97.20% (626/644)	92.23% (665/721)
False alarm rate	0.00% (0/5080)	0.93% (48/5160)	0.45% (23/5083)
	Person 7	Person 8	Person 9
Accuracy	98.11% (725/739)	95.62% (546/571)	98.95% (564/570)
False alarm rate	0.71% (36/5065)	0.46% (24/5233)	0.27% (14/5234)
Average Accuracy	95.14% (5522/5804)		
Average False alarm rate	0.61% (282/46432)		

TABLE XX

THE RECOGNITION RATES OF HUMAN FACE VIDEOS IN THE **BRIGHT** ENVIRONMENT,
WITH MAJORITY VOTE OF THREE

	Person 1	Person 2	Person 3
Accuracy	97.59% (527/540)	96.60% (738/764)	99.20% (493/497)
False alarm rate	0.60% (31/5157)	0.16% (8/4933)	1.02% (53/5200)
	Person 4	Person 5	Person 6
Accuracy	89.49% (630/704)	97.96% (623/636)	95.62% (677/708)
False alarm rate	0.00% (0/4993)	0.69% (35/5061)	0.30% (15/4989)
	Person 7	Person 8	Person 9
Accuracy	99.31% (723/728)	97.50% (546/560)	99.82% (559/560)
False alarm rate	0.28% (14/4969)	0.25% (13/5137)	0.23% (12/5137)
Average Accuracy	96.82% (5516/5697)		
Average False alarm rate	0.40% (181/45576)		

TABLE XXI

THE RECOGNITION RATES OF HUMAN FACE VIDEOS IN THE **BRIGHT** ENVIRONMENT,
WITH MAJORITY VOTE OF FIVE

	Person 1	Person 2	Person 3
Accuracy	97.01% (520/536)	96.98% (739/762)	99.80% (488/489)
False alarm rate	0.74% (38/5124)	0.29% (14/4898)	1.06% (55/5171)
	Person 4	Person 5	Person 6
Accuracy	88.70% (628/708)	98.57% (619/628)	95.89% (676/705)
False alarm rate	0.00% (0/4952)	0.62% (31/5032)	0.22% (11/4955)
	Person 7	Person 8	Person 9
Accuracy	99.31% (718/723)	97.30% (540/555)	99.46% (551/554)
False alarm rate	0.30% (15/4937)	0.10% (5/5105)	0.24% (12/5106)
Average Accuracy	96.80% (5479/5660)		
Average False alarm rate	0.40% (181/45280)		

TABLE XXII

THE RECOGNITION RATES OF HUMAN FACE VIDEOS IN THE **DARK** ENVIRONMENT,
WITHOUT MAJORITY VOTE

	Person 1	Person 2	Person 3
Accuracy	96.97% (448/462)	98.85% (771/780)	95.84% (507/529)
False alarm rate	1.15% (60/5198)	0.43% (21/4880)	0.58% (30/5131)
	Person 4	Person 5	Person 6
Accuracy	97.63% (617/632)	98.07% (660/673)	90.79% (651/717)
False alarm rate	0.06% (3/5028)	0.90% (45/4987)	0.10% (5/4943)
	Person 7	Person 8	Person 9
Accuracy	89.13% (492/552)	97.87% (598/611)	89.35% (629/704)
False alarm rate	0.55% (28/5108)	1.49% (75/5049)	0.40% (20/4956)
Average Accuracy	94.93% (5373/5660)		
Average False alarm rate	0.63% (287/45280)		

TABLE XXIII

THE RECOGNITION RATES OF HUMAN FACE VIDEOS IN THE **DARK** ENVIRONMENT,
WITH MAJORITY VOTE OF THREE

	Person 1	Person 2	Person 3
Accuracy	98.00% (442/451)	99.48% (768/772)	98.08% (510/520)
False alarm rate	0.96% (49/5101)	0.23% (11/4780)	0.42% (21/5032)
	Person 4	Person 5	Person 6
Accuracy	99.03% (615/621)	99.10% (658/664)	92.60% (651/703)
False alarm rate	0.04% (2/4931)	0.49% (24/4888)	0.04% (2/4849)
	Person 7	Person 8	Person 9
Accuracy	91.21% (488/535)	98.84% (594/601)	91.53% (627/685)
False alarm rate	0.38% (19/5017)	1.31% (65/4951)	0.12% (6/4867)
Average Accuracy	96.42% (5353/5552)		
Average False alarm rate	0.45% (199/44416)		

TABLE XXIV

THE RECOGNITION RATES OF HUMAN FACE VIDEOS IN THE **DARK** ENVIRONMENT,

WITH MAJORITY VOTE OF FIVE

	Person 1	Person 2	Person 3
Accuracy	98.43% (439/446)	100.00% (764/764)	97.86% (502/513)
False alarm rate	1.12% (57/5070)	0.27% (13/4752)	0.36% (18/5003)
	Person 4	Person 5	Person 6
Accuracy	99.68% (614/616)	99.09% (651/657)	92.72% (650/701)
False alarm rate	0.02% (1/4900)	0.64% (31/4859)	0.00% (0/4815)
	Person 7	Person 8	Person 9
Accuracy	89.93% (482/536)	98.66% (587/595)	89.53% (616/688)
False alarm rate	0.32% (16/4980)	1.46% (72/4921)	0.06% (3/4828)
Average Accuracy	96.17% (5305/5516)		
Average False alarm rate	0.48% (211/44128)		

Chapter 5 Conclusion

In this thesis, we implement the surveillance system that can recognize multi-angle human gait and human face of a person in the bright and dark environments. The human gait recognition system can be applied more easily than the human face recognition system, which is seriously restricted by obtaining frontal, large, clear face to recognize.

By our method, the recognition rates of walking videos in the bright environment, with majority vote of five is 93.36%; and the recognition rates of walking videos in the dark environment, with majority vote of five is 88.17%. The recognition rates of human face videos in the bright environment, with majority vote of five is 96.80%; and the recognition rates of human face videos in the dark environment, with majority vote of five is 96.17%.

It is difficult to recognize a person in the dark environment obtaining frontal, large, clear face to recognize. This research of human walking recognition provides a promising solution to recognize a person in the bright and dark, i.e., all day, environment.

References

- [1] M. Piccardi, "Background subtraction techniques: a review," in *Proc. IEEE Int. Conf. SMC.*, vol. 4, pp. 3099–3104, Oct. 2004.
- [2] P. Viola and M. Jones, "Robust Real-Time Face Detection," *Int. Journal Computer Vision*, vol. 57, no. 2, pp. 137–154, Mar. 2004.
- [3] "OpenCV 2.4, Open Source Computer Vision Library," <http://www.intel.com/technology/computing/opencv/>, 2012.
- [4] 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.
- [5] 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.*, vol. 1, pp. 391–396, Dec. 2004.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] K. Ohba, Y. Sato, and K. Ikeuchi, "Appearance-based visual learning and object recognition with illumination invariance," *Machine Vision and Applications*, Vol.

- 12, No. 4, pp. 189–196, 2000.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] R. Gonzalez and R. Woods, *Digital Image Processing*, 2nd ed. Pearson Education International, pp. 528–532, 2008.
- [14] K. Etemad and R. Chellappa, “Discriminant analysis for recognition of human face images,” in *Proc. ICASSP*, pp. 2148–2151, 1997.
- [15] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd edition, Massachusetts USA: Academic Press, 1990.
- [16] S. Yu, D. Tan, and T. Tan, “A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition,” in *Proc. IEEE 18th Int. Conf. Pattern Recognition*, 2006, vol. 4, pp. 441–444.
- [17] A. Iosifidis, A. Tefas, and I. Pitas, “Activity-based person identification using fuzzy representation and discriminant learning,” in *IEEE Transactions on Information Forensics and Security*, vol. 7, 2012, 530–542.