國 立 交 通 大 學

電 機 與 控 制 工 程 研 究 所

碩 士 論 文

HSV 色彩空間前景物體抽取及其於人體動作辨識系統應用

Extracting the Foreground Subject in the HSV Color Space and Its Application to Human Activity Recognition System

研 究 生：駱 易 辰

指 導 教 授：張 志 永

中 華 民 國 九十六 年 七 月

HSV 色彩空間前景物體抽取及其於人體動作辨識系統應用

Extracting the Foreground Subject in the HSV Color Space and Its Application to Human Activity Recognition System

學　　生：駱易辰　　Student : Yi-Chen Luo

指導教授：張志永　　Advisor : Jyh-Yeong Chang

國立交通大學

電機與控制工程學系

碩士論文

A Thesis

Submitted to Department of Electrical and Control Engineering

College of Electrical Engineering and Computer Science

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of Master in

Electrical and Control Engineering

July 2007

Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 六 年 七 月

HSV 色彩空間前景物體抽取
及其於人體動作辨識系統應用

學生: 駱易辰　　指導教授: 張志永博士

國立交通大學電機與控制工程研究所

摘要

利用串流影像資訊於人類行動辨識能在許多地方應用，如:人機介面、安全監控、居家安全照護等系統，本論文的提出一個可以自動監控、追蹤辨識人類動作的系統。在一般前、後景色彩深淺差別大時，可以簡單的使用亮度的資訊將前後景分離，但當前後景亮度接近時，例如; 當辨識的目標穿著和背景相似的衣服時，若只使用灰階影像並無法將完整的前景資訊分離，因此我們使用 HSV 色彩空間加入像素點色彩成分的考慮建立背景模型，達到前、後景的分離，且能對陰影的問題加以消除改進。但是使用 HSV 色彩空間必須先解決色調一些不穩定的問題，所以我們在色調不穩定的區域加以限制，以增加抽取前景影像的準確性。

將抽取的影像以二值化，再將經過特徵空間以及標準空間轉換，投影至標準空間。經由樣板比對的方法將三張影像合為一個姿態變化序列，此影像序列乃從動作視訊 5:1 減低抽樣獲得。接著，利用模糊法則的推論方法，將這組時序姿態序列分類為某一個動作類別。跟單用亮度成分的方法比較，實驗證明，HSV 色彩空間不但在前景影像抽取有明顯的改進,而且在人體動作辨識結果也有顯著的改進。

Extracting the Foreground Subject in the HSV Color Space and Its Application to Human Activity Recognition System

STUDENT: Yi-Cheng Luo ADVISOR: Dr. Jyh-Yeong Chang

Institute of Electrical and Control Engineering
National Chiao-Tung University

ABSTRACT

Human activity recognition from video streams has a wide range of application such as human-machine interface, security surveillance, home care system, etc. The objective of this thesis is to provide a human-like system to auto-survey and then to track people and identify their activities. When the foreground color is different from the background color, the foreground subject can be extracted easily by the luminance component. When the foreground color is similar to the background color, we cannot extract the foreground image completely by the luminance component. To solve this, we utilize the HSV color space to build the background model, in line with similar spirit of W^4 segmentation algorithm, which can not only extract foreground image but also be helpful to shadow removal. Since H and S component are not reliable in some conditions, we make use of three criteria to obtain reliable and static hue values.

A foreground subject is first converted to a binary image and transformed to a new space by eigenspace and canonical space transformations. Recognition is done in canonical space. A three image frame sequence, 5:1 down sampling from the video, is converted to a posture sequence by template matching. The posture sequence is classified to an action by fuzzy rules inference. In our experiment, extracting the foreground image in the HSV space improves not only the accuracy of foreground image but also human activity recognition accuracy.

ACKNOWLEDGEMENTS

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

Content

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Motivation

Human activity recognition from video streams has many applications such as home care system, human-machine interface, and automatic surveillance, etc. However, there is no rigid syntax and well-defined structure in human action recognition system; therefore, it makes human activity recognition a very challenging task.

Several human activity recognition methods have been proposed in the past few years. Yamato *et al.* [1] turn image frames into a symbol sequence and use HMM to recognize human action. Bobick and Davis [2] recognize human activities by comparing motion-energy and motion-history of template images with temporal images. Cohen and Li [3] use a view-independent 3-D shape description for classifying and identifying human activity using SVMs. There have been some significant projects on detecting, tracking people and recognizing their activities. W^4 [4] is one of them. W^4 can detect people (single person or people in group) by adopting an adaptive background model and identify the activities by finding the body parts on the silhouette boundary.

The objective of this thesis is to provide a human-like system to auto-surveillance and to track people and identify their activities. This system can tell where the foreground subject is in an image, and what the subject is doing.

The system flowchart is illustrated in Fig. 1.1 Our system can be separated into three components. The first component is foreground subject extraction. The second component is the transformation of image data in a space smaller and easier for

posture recognition. The third component is the posture classification of an image frame and activity recognition using frame sequences.

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

1.2 Foreground subject extraction

Foreground subject extraction is an important step of the vision-based human activity recognition system. Many authors have developed methods of detecting people in images. Park and Aggarwal subtracted foreground pixels from background by computing Mahalanobis distance in each pixel in the HSV color model [5]. Leung and Yang built a human body outline labeling system [6]. Jabri and Duric [7] used color and edge information to improve the quality and reliability of the results. They all try to find out the real poses a human did by human body outline or by silhouettes.

Background subtraction is widely used for detecting moving objects from image frames of static cameras. Most of this work has been based on background subtraction using color or luminance information. In these approaches, difference between the coming frame and the background image is performed to detect foreground objects.

If we only use the luminance information to do background subtraction, we cannot detect a foreground pixel correctly when it is similar to the background pixel. To make fully use of the spectrum of a pixel, it is imperative to do the segmentation in the color domain. To the end, foreground subject extraction is done in the HSV color space. We can have both the luminance information and the chromatic information in the background subtraction task.

Background subtraction is extremely sensitive to dynamic scene changes due to illumination change. In order to solve the effect of varying luminance conditions, we develop a method which is robust to the illumination changes. The method utilizes frame ratio rather than frame difference in luminance component.

Furthermore, the moving cast shadows mostly exhibit a challenge for accurate foreground subject detection. A lot of attempts have been developed to tackle the shadow suppression [8]−[13] encountered in background subtraction. Horprasert *et al.* [8] and Cucchiara *et al.* [9] utilized the rationale that shadows have similar chromaticity, but lower brightness than the background model. Under the proposed frame work in the HSV color space, we can effectively identify the shadow existent in our detected foreground subject.

After building a background model, we can extract foreground subjects from video frames by subtracting each pixel value of background model from that of current image frame. The resulting image is converted to a binary one by setting a threshold. The binary image mainly contains foreground subjects with only little noise. Therefore, we can set a threshold in the histogram of the binary image to extract a

rectangle image, which is a good representation resemble shape of a person, of the target subject. The rectangle image is normalized to a uniform benchmark .

1.3 Eigenspace and Canonical Space Transformation

In most 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 reduction often transforms an image from spatiotemporal space to another data space. The transformation can use fewer dimensions to approximate the original image. There are many well-known transformation methods such as Fourier Transformation, Wavelet Transformation, Principal Component Analysis and so on. Our transformation method combines eigenspace transformation and canonical space transformation which are described as follows.

Eigenspace transformation (EST), based on Principal Component Analysis, has been demonstrated to be a potent scheme used widely as shown below: automatic face recognition proposed in [14], [15]; gait analysis proposed in [16]; and action recognition proposed in [17]. The subsequent transformation, Canonical space transformation (CST) based on Canonical Analysis, is mainly 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 and reduce the dimension as well. Thus each image can be projected from a high-dimensional spatiotemporal space to a low-dimensional canonical space. In this new space the recognition of

human activities becomes much simpler and easier.

1.4 Image frame classification and activity recognition

In this thesis each in a video segmentation, images are transformed into an image feature vector by extracting features from images. We utilize eigenspace and canonical space transformation used to extract image features. If we only adopt the shape-based features to recognize an activity, many activities remain unidentified since the temporal information is discarded. Hence, we group three consecutive image feature vectors from three contiguous, but down-sampled images. Consequently, the time-sequential images are converted to a posture sequence by using these three feature vectors. The posture sequence is dignified by the index number of the posture template. In the learning phase, we build a transition model in terms of three consecutive posture sequences which are the category symbols of the posture template. For human action recognition, the model which best matches the observed posture sequence is chosen as the recognized action category.

The most famous method to model transition model of time-sequential data is Hidden Markov Models (HMMs). Hidden Markov Models can deal with time-sequential data and can provide time-scale invariability for recognition. The basic concept of Hidden Markov Models is described in [18]. Hidden Markov Models have been successfully used for speech recognition because of their capability of recognizing spoken words independent of their duration [18]–[20]. Hidden Markov Models also have been used in hand gestures recognition [21] and activity recognition [1]. The price paid for the efficiency in this case is that we have to collect a great amount of data and a lot of time is required to estimate the corresponding parameters

in HMMs.

After transforming image frames to eigenspace and canonical space domain, some data information have been omitted. By using fuzzy rule-base techniques, the activity analysis task is tolerant to uncertainty, ambiguity and irregularity. Relevant articles using the fuzzy theory are described as follows. Wang and Mendel [22] proposed that fuzzy rules to be generated by learning from examples. Su [23] presented a fuzzy rule-based approach to spatio-temporal hand gesture recognition. He employed a powerful method based on hyperrectangular composite neural networks (HRCNNs) for selecting templates.

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

1.5 Thesis outlines

The thesis is organized as follows. Before introducing the technique of our human activity recognition system, the basic concepts concerning the HSV color space, eigenspace transform, and canonical space transform are introduced in

Chapter2. In this chapter, we introduce the HSV color space and discuss the process of how to transform a high dimensional image to eigenspace and canonical space. Chapter 3 describes our human activity recognition system in detail. In Chapter 4, the experiment results of our recognition system are shown. At last, we conclude this thesis with a discussion in Chapter 5.

Chapter 2 BASIC CONCEPT

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

2.1 Fundamentals of Eigenspace and Canonical Space Transform

In video and image processing, the dimensions of image data are often extremely large. There are many well-known transformation methods to reduce the size of data such as Fourier transformation, wavelet, principal component analysis (PCA), eigenspace transformation (EST) and so on. However, PCA based on the global covariance matrix of the full set of image data is not sensitive to the class structure existent in the data. In order to increase the discriminatory power of various activity features, Etemad and Chellappa [24] used linear discriminant analysis (LDA), also called canonical analysis (CA), which can be used to optimize the class separability of different activity classes and improve the classification performance. The features are obtained by maximizing between-class and minimizing within-class variations. Here we call this approach canonical space transformation (CST). Combining EST with CST, our approach reduces the data dimensionality and optimizes the class separability among classes.

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

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

$$\left[\mathbf{x}'_{1,1}, \cdots, \mathbf{x}'_{1,N_1}, \cdots, \mathbf{x}'_{2,1}, \cdots, \mathbf{x}'_{c,N_c} \right] \tag{1}$$

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

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

$$\mathbf{x}_{i,j} = \frac{\mathbf{x}'_{i,j}}{\left\| \mathbf{x}'_{i,j} \right\|}. \tag{2}$$

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

$$\mathbf{m}_{\mathrm{x}} = \frac{1}{N_T} \sum_{i=1}^{c} \sum_{j=1}^{N_i} \mathbf{x}_{i,j}. \tag{3}$$

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

$$\mathbf{X} = \left[\mathbf{x}_{1,1} - \mathbf{m}_{\mathrm{x}}, \cdots, \mathbf{x}_{1,N_1} - \mathbf{m}_{\mathrm{x}}, \cdots, \mathbf{x}_{c,N_c} - \mathbf{m}_{\mathrm{x}} \right]. \tag{4}$$

2.1.1 Eigenspace Transformation (EST)

Basically EST is widely used to reduce the dimensionality of an input space by mapping the data from a correlated high-dimensional space to an uncorrelated low-dimensional space while maintaining the minimum mean-square error to avoid information loss. EST uses the eigenvalues and eigenvectors generated by the data covariance matrix to retain the original data coordinates along the directions of maximal variance sequentially.

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

$$\lambda_i \mathbf{e}_i = \mathbf{R}\mathbf{e}_i, \qquad i = 1, 2, \cdots, K \tag{5}$$

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

$$\tilde{\mathbf{R}} = \mathbf{X}^{\mathrm{T}}\mathbf{X} \qquad \mathbf{X}: \text{ data matrix} \tag{6}$$

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

$$\begin{cases} \lambda_i = \tilde{\lambda}_i \\ \mathbf{e}_i = \left(\tilde{\lambda}_i\right)^{-\frac{1}{2}} \mathbf{X}\, \tilde{\mathbf{e}}_i \end{cases} \qquad i = 1, 2, \cdots, K \tag{7}$$

These *K* eigenvectors are used as an orthogonal basis to span a new vector space. Each image can be projected to a point in this *K*-dimensional space. Based on the theory of PCA, each image can be approximated by taking only the largest eigenvalues $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_k|$, $k \leq K$, and their associated eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$. This partial set of *k* eigenvectors spans an eigenspace in which $\mathbf{y}_{i,j}$ are the points that are the projections of the original images $\mathbf{x}_{i,j}$ by the equation

$$\mathbf{y}_{i,j} = [\, \mathbf{e}_1,\ \mathbf{e}_2, \cdots,\ \mathbf{e}_k \,]^{\mathrm{T}}\, \mathbf{x}_{i,j} \qquad i = 1,2,...,c \ ;\ j = 1,2,...,N_c \tag{8}$$

We called this matrix $[\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k]^{\mathrm{T}}$ 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.

2.1.2 Canonical Space Transformation (CST)

Based on canonical analysis in [25], we suppose that $\{\phi_1, \phi_2, \cdots, \phi_c\}$ represents the classes of transformed vectors by eigenspace transformation and $\mathbf{y}_{i,j}$ is the *j*-th vector in class *i*. The mean vector of entire set can be written as

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

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

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

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

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

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

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

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

$$\frac{\partial \mathbf{J}}{\partial \mathbf{W}} = 0. \tag{12}$$

Suppose that $\mathbf{W}^*$ is the optimal solution where the column vector $\mathbf{w}_i^*$ is a generated eigenvector corresponding to the *i*-th largest eigenvalues λ_i. According to the theory presented in [25], we can solve Eq. (12) as follows

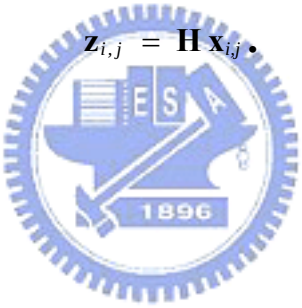
$$\mathbf{S}_b \mathbf{w}_i^* = \lambda_i \mathbf{S}_w \mathbf{w}_i^*. \tag{13}$$

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

$$\mathbf{z}_{i,j} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{c-1}]^{\mathrm{T}} \mathbf{y}_{i,j} \tag{14}$$

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

$$\mathbf{z}_{i,j} = \mathbf{H}\,\mathbf{x}_{i,j}. \tag{15}$$



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. Viewed from the circular side of the cone, the hues are represented by the angle of each color in the cone relative to the 0^o line, which is traditionally assigned to be red. The saturation is represent as the distance from the center of the circle. Highly saturation color are on the outer edge of the cone, whereas gray tones (which have no saturation) are at the very center. The brightness 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.

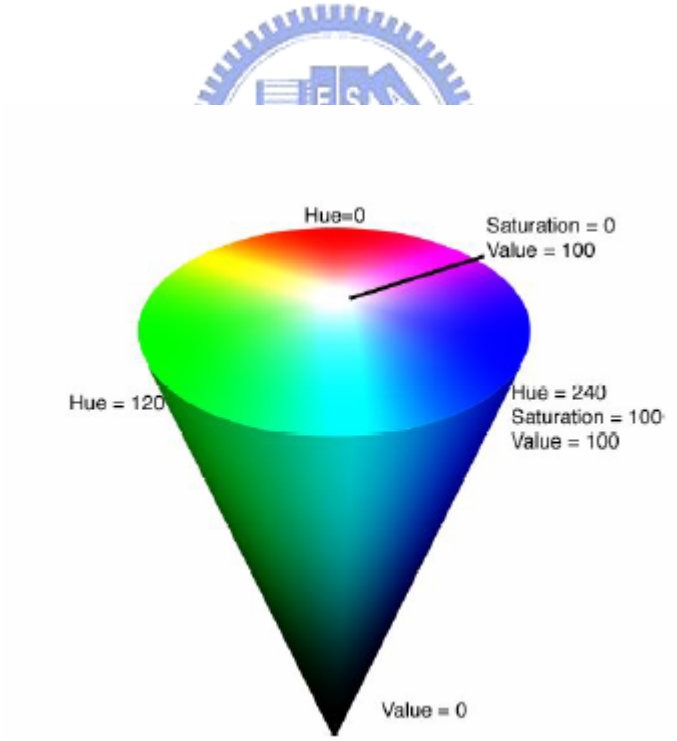


Fig. 2.1 The HSV Cone

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 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 [11]. Accordingly, Ohba *et al*. [26] 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, S_t, and H are an 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 ΔP_t to avoid the discontinuity effect.

Chapter 3 Human Activity Recognition System

3.1 Object extraction

3.1.1 The intensity of the image

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

$$I_i(x,y) = S_i(x,y)r_i(x,y), \tag{16}$$

where I_i is the intensity of the image, S_i is the spatial distribution of source illumination, r_i is the distribution of scene reflectance, (x,y) is the location of a pixel in the image, and i is the image sequence index. Now we can compare the difference caused by illumination change between frame difference and frame ratio. If we hold the camera still with no foreground subjects pass by, the reflectance of this background should be the same at any time. That is,

$$r_i(x,y) = r(x,y). \tag{17}$$

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

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

$$\log\left(\frac{I_i^r(x,y)}{I_{i-1}^r(x,y)}\right)=\log\left(\frac{S_i^r(x,y)r(x,y)}{S_{i-1}^r(x,y)r(x,y)}\right)$$
$$=\log\left(\frac{S_i^r(x,y)}{S_{i-1}^r(x,y)}\right) \tag{19}$$
$$=\log\left(S_i^r(x,y)\right)-\log\left(S_{i-1}^r(x,y)\right),$$

where I^d is the intensity of scene captured by camera of frame difference, S^d is the spatial distribution of source illumination of frame difference, and I^r and S^r is of frame ratio. Comparing Eqs. (18) and (19), we can find that the problems cause by reflectance still remains in the frame difference approach; nevertheless, the influence of reflectance is eliminated in the frame ratio approach.

Fig.3.1 shows a comparison between frame ratio and frame difference. Fig.3.1(a) is a background image and Fig. 3.1(b) is an image frame with a human. By using frame difference and frame ratio approach, we obtain Fig. 3.1(c) and Fig. 3.1(d), respectively. Gray level of the resulting images distributed from 0 to 255. Fig. 3.1(e) is the histogram of Fig. 3.1(c) and Fig. 3.1(f) is the histogram of Fig. 3.1(d). Comparing the histograms of Fig. 3.1(d) and Fig. 3.1(e), we find out that there was less noise in the region of low gray level by using frame ratio method. The Fig. 3.1(g) and Fig. 3.1(h) are the binary image of extraction images which simply took a threshold value 15 at gray level against Fig. 3.1(c) and Fig. 3.1(d).

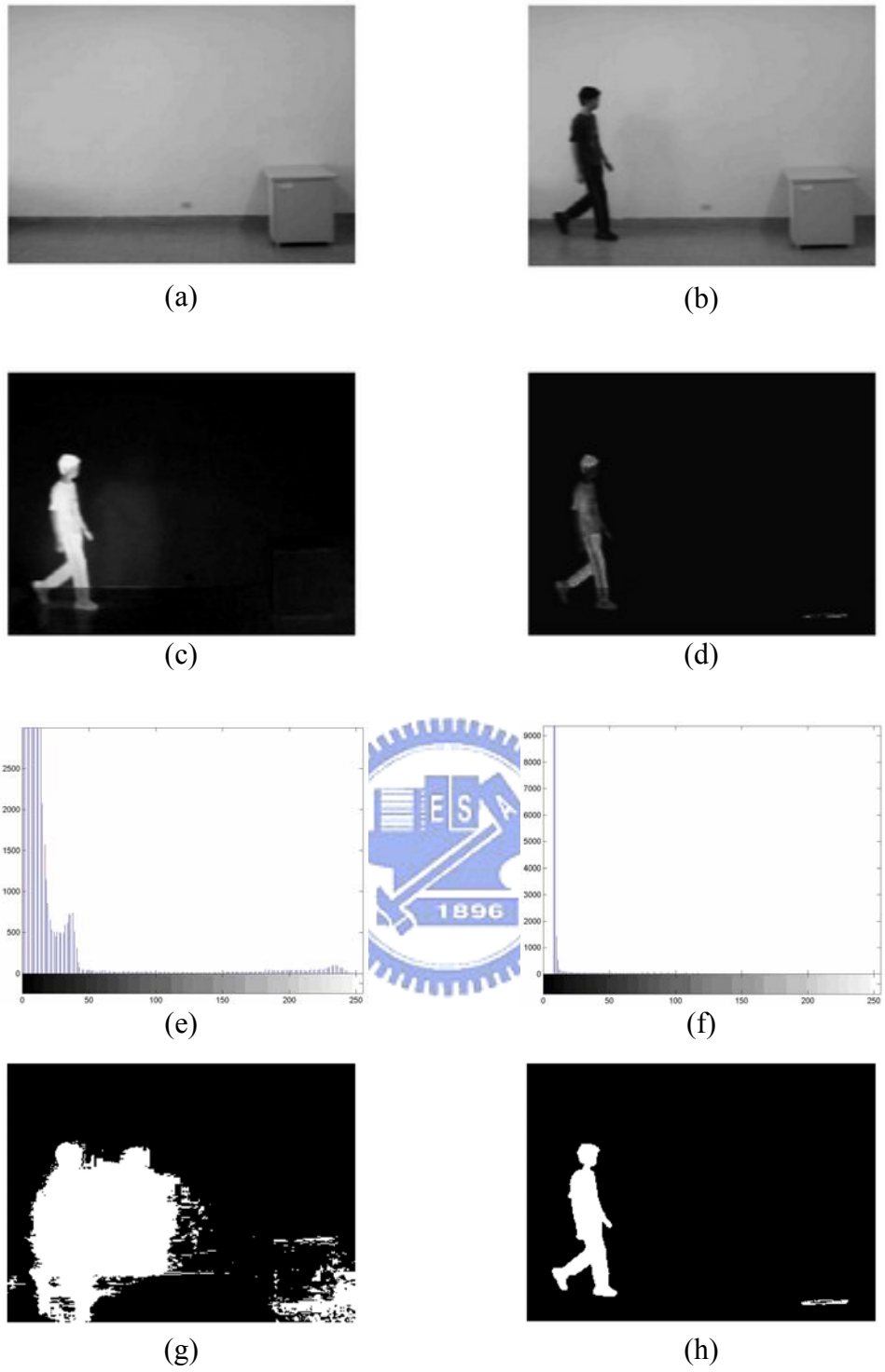


Fig. 3.1 The comparison between frame ratio and frame difference. (a) Background image, (b) image frame with a human, (c) frame difference, (d) frame ratio, (e) histogram of frame difference, (f) histogram of frame ratio, (g) foreground pixels of frame difference after simply taking a threshold, and (h) foreground pixels of frame ratio after simply taking a threshold

3.1.2 Background model

If we only use the luminance component to do background subtraction, we cannot detect reliably those foreground pixel whose luminance component close to background pixel. In order to solve this problem, we build our background model in the 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. Hue is unreliable in some condition, so we use the three criteria (*intensity value*, *saturation*, and *hue*) described in Chapter 2 to obtain the hue value reliably.

In the previous section, we have seen the advantage of using frame ratio approach to counter the luminance change. Hence, we propose to utilize the frame ratio to build the background model in the luminance component. We build our background model with the minimum value ($[n^H(x,y), n^S(x,y), n^V(x,y)]$) and maximum value ($[m^H(x,y), m^S(x,y), m^V(x,y)]$) in each HSV domain. Besides, we also record the inter-frame ratio in the brightness information and the inter-frame different in the chromatic information.

We need a background video, without any moving objects, for background model training. Suppose the observed image frame sequence contains *N* consecutive images. $I_i^H(x,y)$ be the pixel's hue value at (x,y) of the *i*-th image frame. $I_i^S(x,y)$ be the pixel's saturation value at (x,y) of the *i*-th image frame. $I_i^V(x,y)$ be the pixel's brightness value at (x,y) of the *i*-th image frame. The background model of a pixel is obtained by

$$\begin{bmatrix} m^H(x,y) \\ n^H(x,y) \\ d^H(x,y) \end{bmatrix} = \begin{bmatrix} \max_i \{I_i^H(x,y)\} \\ \min_i \{I_i^H(x,y)\} \\ \max_i \{|I_i^H(x,y) - I_{i-1}^H(x,y)|\} \end{bmatrix} \tag{20}$$

$$\begin{bmatrix} m^S(x,y) \\ n^S(x,y) \\ d^S(x,y) \end{bmatrix} = \begin{bmatrix} \max_i \{I_i^S(x,y)\} \\ \min_i \{I_i^S(x,y)\} \\ \max_i \{|I_i^S(x,y) - I_{i-1}^S(x,y)|\} \end{bmatrix} \tag{21}$$

$$\begin{bmatrix} m^V(x,y) \\ n^V(x,y) \\ d^V(x,y) \end{bmatrix} = \begin{cases} \begin{bmatrix} \max_i \{I_i^V(x,y)\} \\ \min_i \{I_i^V(x,y)\} \\ \max_i \{I_i^V(x,y)/I_{i-1}^V(x,y)\} \end{bmatrix} & \text{if } I_i^V(x,y)/I_{i-1}^V(x,y) \geq 1 \\ \begin{bmatrix} \max_i \{I_i^V(x,y)\} \\ \min_i \{I_i^V(x,y)\} \\ \max_i \{I_{i-1}^V(x,y)/I_i^V(x,y)\} \end{bmatrix} & \text{otherwise} \end{cases} \tag{22}$$

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

3.1.3 Foreground subject extraction and shadow detection

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 by luminance. The second component is the shadow suppression. The third component is the object segmentation. And the finally component is the color compensation to recover the foreground pixels wrongly classified to the background due to their high luminance

similarly.

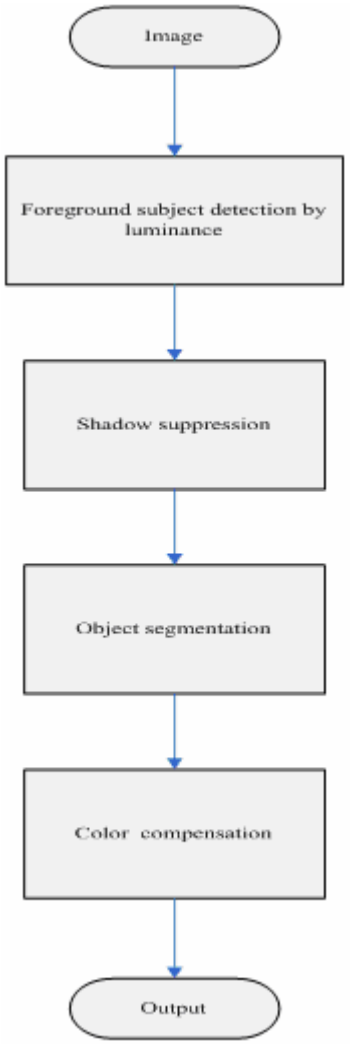


Fig.3.2 The framework we apply to foreground subject extraction

A. Foreground subject detection by luminance

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

maximum inter-frame luminance ratio $d^V(x, y)$ of the training background model to segment the foreground pixel by

$$B(x,y)=\begin{cases} 0, & \text{if } I_i^V(x,y)/m^V(x,y) < k_V d^V(x,y) \\ & \text{or } I_i^V(x,y)/n^V(x,y) < k_V d^V(x,y) \\ 255, & \text{otherwise} \end{cases} \tag{23}$$

where $I_i^V(x, y)$ is the intensity of a pixel which is located at (x,y), $B(x, y)$ is the gray level of a pixel in a binary image, and k_V is a threshold, determined by light sufficiency of the scene. The value of k_V is normally set to 1.3 for normal light condition, and k_V will be reduced for in-sufficient light condition and increased otherwise.

B. Shadow suppression

The pixels of the moving cast shadows are easily detected as the foreground pixel in normal condition. Because the shadow pixels and the object pixels share two important visual features: motion model and detectability. For this reason, the moving shadows cause object merging and object shape distortion. Horprasert *et al.* [8] and Cucchiara *et al.* [9] utilize the rationale that shadows have similar chromaticity, but lower brightness than the background model. Hence, we can detect the shadow from foreground subject in the HSV color space. We analyze only points belonging to possible moving object that are detected in step A. We define a shadow mask S for each (x, y) point as follows:

$$S(x,y)=\begin{cases}\text{shadow}, & \text{if } I_i^V(x,y)-n^V(x,y)<0 \\ & \text{and } \left|I_i^H(x,y)-m^H(x,y)\right|<k_H d^H(\text{x,y}) \\ & \text{and } \left|I_i^S(x,y)-m^S(x,y)\right|<k_S d^S(\text{x,y}) \\ \text{object}, & \text{otherwise}\end{cases} \tag{24}$$

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 located at (x,y), and $S(x,y)$ is the shadow mask to class the pixel in the moving cast shadow. Values k_S and k_H are selected threshold values used to measure the similarities of the hue and saturation between the background image and the current observed image. We can utilize the shadow mask $S(x,y)$ to change the shadow pixels into background in $B(x,y)$.

C. Object segmentation

According to the binary image B 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 X and Y direction. Fig. 3.3 shows an example of foreground region extraction. We utilize the binary image and project it to X and Y directions. The interested section has higher counts in the histogram. We obtain the boundary coordinates x_1, x_2 of X axis and y_1, y_2 of Y axis from the projection histogram. We can use these boundary coordinates as the corner of a rectangle to extract foreground region (B_s). Fig. 3.4 is the extracted foreground region.

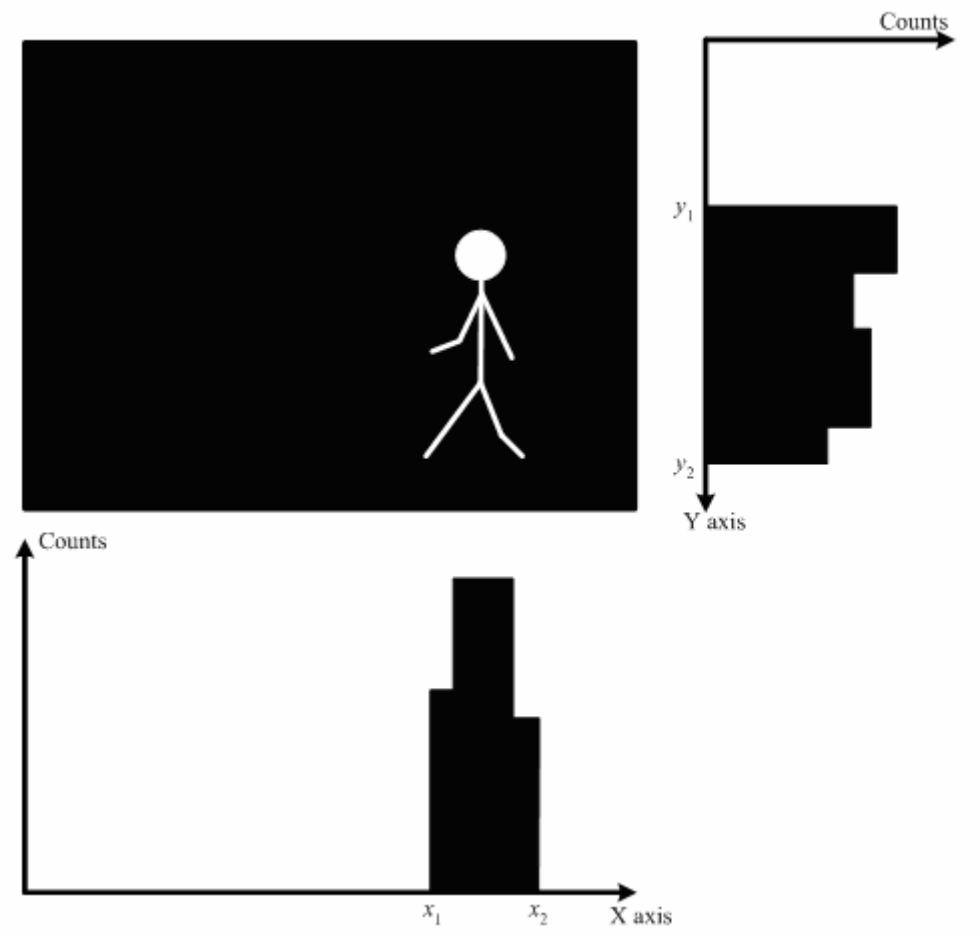


Fig. 3.3 Histogram of binary image projection in X and Y direction.

Fig. 3.4 The binary image of extracted foreground region.

D. Color compensation

Some colors such as yellow, pink, and light blue have similar luminance value. If we only use the luminance component to do background subtraction, we cannot detect foreground pixel correctly when its luminance is similar to that of a background pixel. In order to improve detectability, background subtraction is computed by taking into account not only a point's luminance, but also its chromaticity. We want to use the chromaticity to enhance the accuracy of the foreground object. We only analyze the region B_s obtained in subsection C above. Based on the amount of the chromaticity change, we reanalyze its background in B_s to be changed to a foreground of object, by

$$B_f(x,y)=\begin{cases}255, & \text{if } \left|I_i^S(x,y)-m^S(x,y)\right|>k_S d^S(\text{x,y}) \\ & \text{or } \left|I_i^H(x,y)-m^H(x,y)\right|>k_H d^H(\text{x,y}) \\ 0, & \text{otherwise}\end{cases} \tag{25}$$

where $I_i^H(x,y)$ and $I_i^S(x,y)$ are respectively the hue and saturation components of a pixel at (x,y), k_S and k_H are selected threshold values. B_f is the final foreground object after the refined step of Eq. (25).

3.2 Activity template selection

A human body is a rigid body, thus has its natural frequency; namely, it has restriction on action speed when doing some specific actions. Because cameras usually capture image frames in a high frequency, there are few differences between two postural image frames in a short interval. Therefore, we select some key frames from a sequence to represent an activity. In our approach, we select one image frame, called as the essential template image, with a fixed interval instead of each image. An example is shown in Fig. 3.5. After determining the templates, each activity is represented by several essential templates.

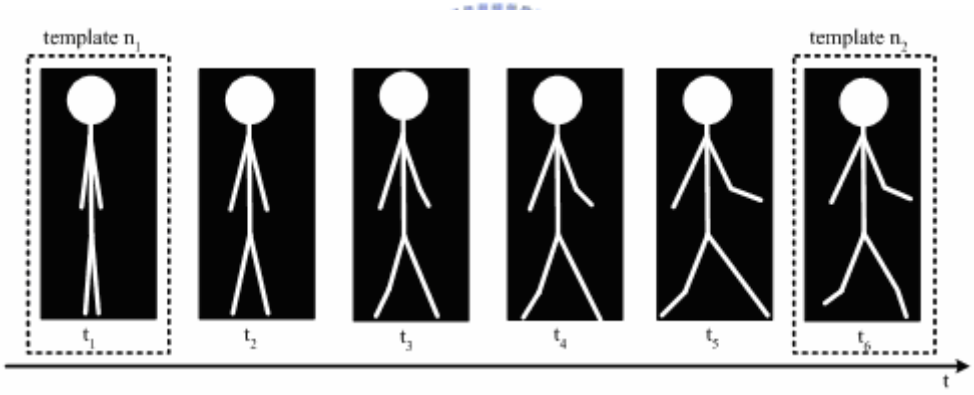


Fig. 3.5 One image frame is selected as template with an interval.

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

recognition.

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

$$\mathbf{t}_{i,j} = \mathbf{H} \times \mathbf{g}_{i,j}, \qquad i = 1,\ 2,\ \cdots,\ c;\ \ j = 1,\ 2,\ \cdots, n \tag{26}$$

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

$$\mathbf{t}_{i,j} = \left[t_{i,j}^{1},\ t_{i,j}^{2}, \cdots,\ t_{i,j}^{c-1} \right]^{\mathrm{T}}. \tag{27}$$

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

$$\boldsymbol{\mu}_{i,j} = \mathbf{t}_{i,j} \tag{28}$$

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

$$\sigma_m = \sqrt{\frac{\sum_{j=1}^{c}\sum_{i=1}^{n}\left(t_{i,j}^{m} - \mu_{i,j}^{m}\right)^2}{N_t - 1}} \tag{29}$$

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

3.3 Construction of fuzzy rules form 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 classify the action, classification result may be failed easily because human's actions may have similar postures in two different activity sequences. For example, the action of "jumping" and "crouching" both have the same postures called common states as shown in Fig. 3.6. Besides, the posture sequence of each activity is dissimilar in different people.

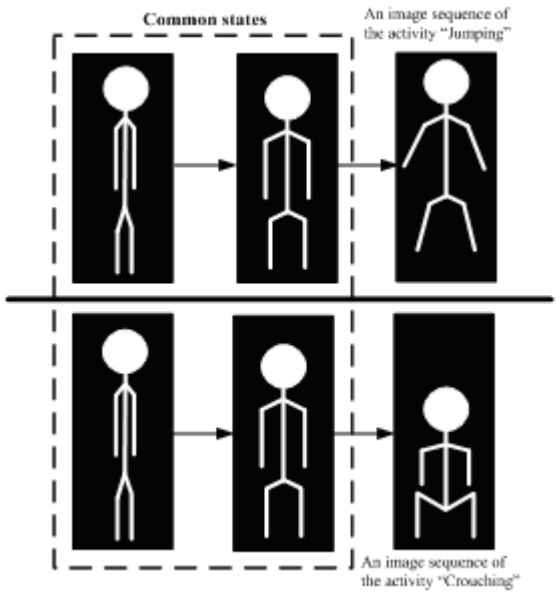


Fig. 3.6 Common states of two different activites.

Hence, we propose a method which not only combines temporal sequence

information for recognition but also is tolerant to variations of different people. We use the fuzzy rule-base approach to design our system. The fuzzy rule-base approach also has been proposed in gesture recognition in [23]; it has ability to absorb data difference by learning.

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

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

$$\mathbf{a}_k = \mathbf{H}\,\mathbf{x}_k. \tag{30}$$

As the same as $\mathbf{t}_{i,j}$ in Eq.(27), $\mathbf{a}_k$ can be rewritten as

$$\mathbf{a}_k = \left[a_k^1, a_k^2, \cdots, a_{i,j}^{c-1}\right]^{\mathrm{T}}. \tag{31}$$

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

$$
\begin{aligned}
r_{i,k} &= M(\mathbf{a}_k \mid C_i) \\
&= \frac{1}{(2\pi)^{\frac{c-1}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} \mathbf{a}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{a}_k \right] \\
&= \arg \max_j \left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2} \frac{\left(a_k - \mu_{i,j}\right)^2}{\sigma^2} \right] \right\}
\end{aligned}
\tag{32}
$$

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. Besides, we can obtain which category each image belongs to by

$$
p_k = \arg \max_i r_{i,k} \tag{33}
$$

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. If we use too many images to form a basis, the data may contain too many images of other activity. If we use too few images, it may not have enough timing information to represent an activity.

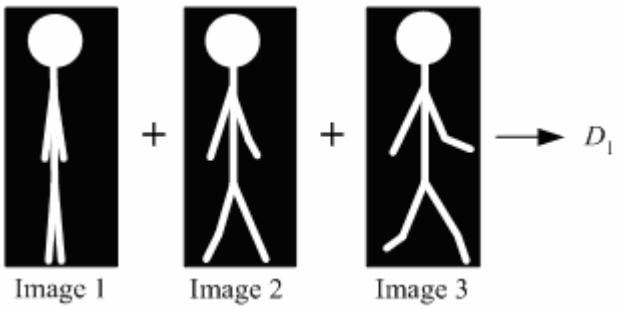
Assume we have c linguistic labels, each linguistic label represent a category of essential template. Each image frame can be represented by one of these c linguistic labels. In our approach, we combine three contiguous images to a group (I_1, I_2, I_3). The transformation of the image group can form a feature vector $[a_1, a_2, a_3]$. There are c^3 combinations of the feature vector. Each combination represents the possible transition states of the three images. We use Eqs. (32) and (33) to class each image frame. Hence, we can represent the feature vector ($[a_1, a_2, a_3]$) by linguistic label sequence ($[a_1^i, a_2^i, a_3^i]$). An image sequence with linguistic label sequence is associated with its output of corresponding activity.

As developed by Wang and Mendel [22], fuzzy rules can be generated by learning from examples. Such image sequence constitutes an input-output pair to be learned in the fuzzy rule base. In this setting, the generated rules are a series of associations of the form

"**IF** antecedent conditions hold, **THEN** consequent conditions hold."

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

$$\left[\mathbf{a}_1^1, \mathbf{a}_2^1, \mathbf{a}_3^1; D_1\right] \tag{34}$$



Suppose that Image 1, Image2 and Image 3 belong to category 1, category 2 and category 3 respectively. Therefore, we assign the image sequences, whose feature vector is $[\mathbf{a}_1^1,\ \mathbf{a}_2^1,\ \mathbf{a}_3^1]$, to the linguistic labels Posture 1, Posture 2 and Posture 3 respectively. Finally, a rule is produced from the feature-target vector. Hence this image sequence supports the rule of

Rule 1. IF the activity's I_1 is P_1 AND its I_2 is P_2 AND its I_3 is P_3, THEN the activity is D_1. (35)

where I_i is Image i and P_j is Posture j.

Our system is able to learn the hidden transition modes of activities from data. This is an advantage of our system and it will also improve the correct rate in classification. For example, the Posture 1 is a posture of activity D1 but D4, the system still learn a sequence with Posture 1 as the activity D4. We regard Posture 1 as a common state of the two activities D1 and D4. Therefore the fuzzy rules induce tolerant to some ambiguous postures of different activities and classify the image sequence to an activity more correctly.

Sometimes conflicting rules may be generated; they have the same image sequence but refer to different activity. Therefore, we have to choose one from the two or more conflicting rules from each qualified cluster. To this end, we choose the rule that is supported by a maximum number of examples. Furthermore, to prune redundant or inefficient fuzzy rules, if the supporting actions of a rule are less than a threshold, the rule is excluded from defining an **IF-THEN** rule.

3.4 Classification algorithm

After constructing the rule base, we can grade the input image sequence with each fuzzy rule by grade of membership function. Let $\mathbf{\Sigma}$ denote the covariance matrix of all essential template vectors, C_i denote the i-th class of essential templates and $\mathbf{s}_k$ denote the image frame transformed by EST and CST. The membership function is given by

$$
\begin{aligned}
r_{i,k} &= \mathrm{M}\left(\mathbf{s}_k \middle| C_i\right) \\
&= \frac{1}{(2\pi)^{\frac{c-1}{2}} \left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\mathbf{s}_k - \mu\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\left(\mathbf{s}_k - \mu\right)\right] \\
&= \arg \max_{\mathrm{j}} \left\{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{\left(s_k - \mu_{i,j}\right)^2}{\sigma^2}\right]\right\}
\end{aligned} \tag{36}
$$

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 for recognizing an activity. Therefore, we use two more transformed vector of passed image frames, which are called $\mathbf{a}_{k-2}$ and $\mathbf{a}_{k-1}$. These three vectors form a feature vector $[\mathbf{a}_{k-2}, \mathbf{a}_{k-1}, \mathbf{a}_k]$. We compute the membership functions of the three vectors respectively. The procedures of calculating membership functions of $\mathbf{a}_{k-2}$ and $\mathbf{a}_{k-1}$ are the same as the process used for $\mathbf{a}_k$ in Eq. (36).

In order to calculate the similarity between image sequence and each postural sequence in the training data base, we take out the membership functions r_{k-2,n_1}, r_{k-1,n_2} and r_{k,n_3} which are corresponding to the three category of linguistic labels, P_{n_1}, P_{n_2} and P_{n_3}, in the rule and have been calculated by Eq. (36). The summation of r_{k-2,n_1}, r_{k-1,n_2} and r_{k,n_3} is the similarity between current image sequence and the postural sequence of this rule. We can obtain the similarity related to all fuzzy rules of training data base in the same manner. The rule, which has the highest value of similarity, is selected and the unknown activity is classified to the activity recorded in this rule. Fig. 3.7 shows the structure of the classification algorithm.

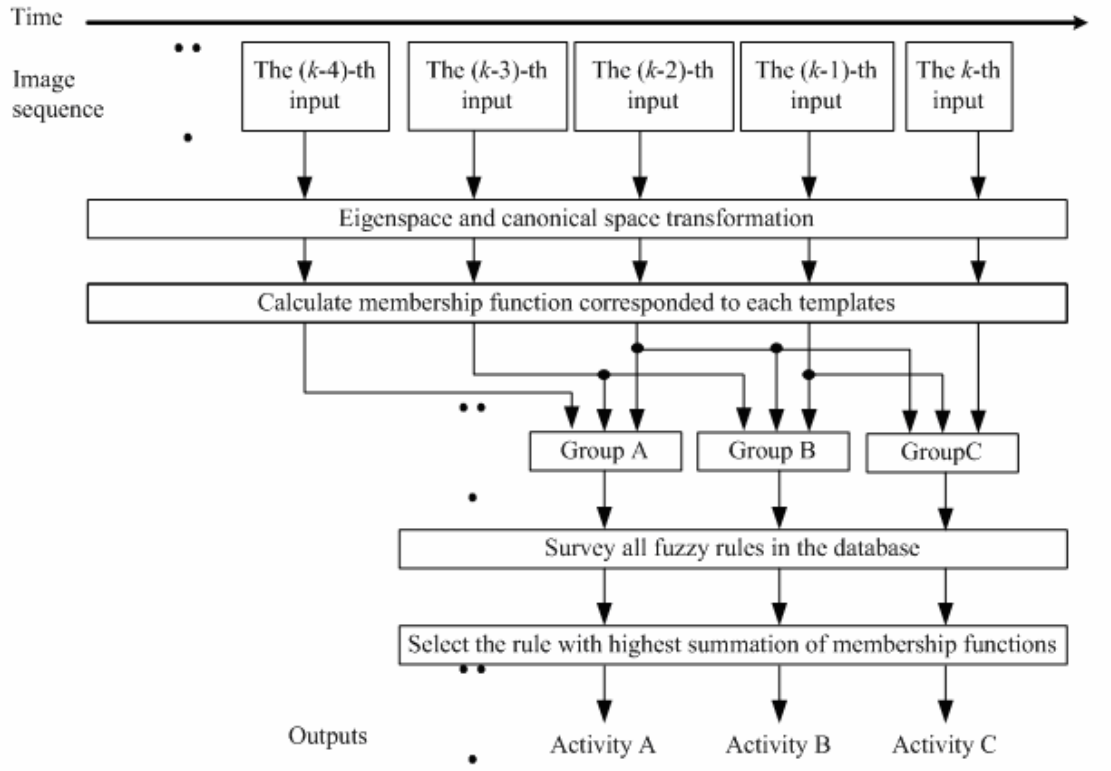


Fig. 3.7 The structure of classification algorithm.

Chapter 4 Experimental Result

In our experiment, we tested our system on videos taken by digital camera. We took the video in our laboratory at the 5th Engineering Building in NCTU campus. The camera has a frame rate of thirty frames per second and image resolution is 320×240 pixels. The experimental environment is shown in Fig. 4.1.

Fig. 4.1 The experimental environment.

The background is not complex and we equipped a table in the scene. The light source is fluorescent lamps and is stable. Each person performed six actions: "walking from left to right," "walking from right to left," "jumping," "crouching," "climbing up" and "climb down." The action "climbing up" is to climb up on the table from the ground. The action "climbing down" is to climb down to the ground from the table.

We test the foreground detection capability and then the action recognition accuracy in two cases depending on the color of clothing worn by action subjects. That the action subject wore the clothing with color different from that of background is first case. And the second case is that the action subject wore the clothing with color similar to that of background. When the color of clothing and

background are similar in the second case, a moving object, such as human body, may not be segmented easily from image frame. We compare the result in these two cases and the color compensation in our action recognition system demonstrates eminent improvement in the segmentation quality. We classify our model into two groups: Group A has six models in which the subject wears clothing with color different from the background, and Group B has three models in which the subject wears clothing similar to the background. Fig. 4.2 shows our models in the experiment.

Fig. 4.2 Various images of our models.

4.1 Background model construction

We built the background model in the HSV color space. The value of H or S or V is between 0 and 255. Figs. 4.3(a), 4.3(b), and 4.3(c) show the background image in the H, S, and V component, respectively. We can find from these three figures that the hue value is relatively unstable when the saturation is close to zero. We make an experiment to test the changes in the HSV components in constructing the background model. Fig. 4.4 represents the H, S, and V variations of two pixels at coordinates $(x, y) = (10, 10)$ and $(x, y) = (120, 160)$ during the first 300 frames in the background video. From Fig. 4.4, we can see that V component is most stable of the background model. H and S components are less stable than V. Hence, we need to solve this problem.

(a) (b) (c)

Fig. 4.3 Background images. (a) Background image in the H component, (b) Background image in the S component, and (c) Background image in the V component.

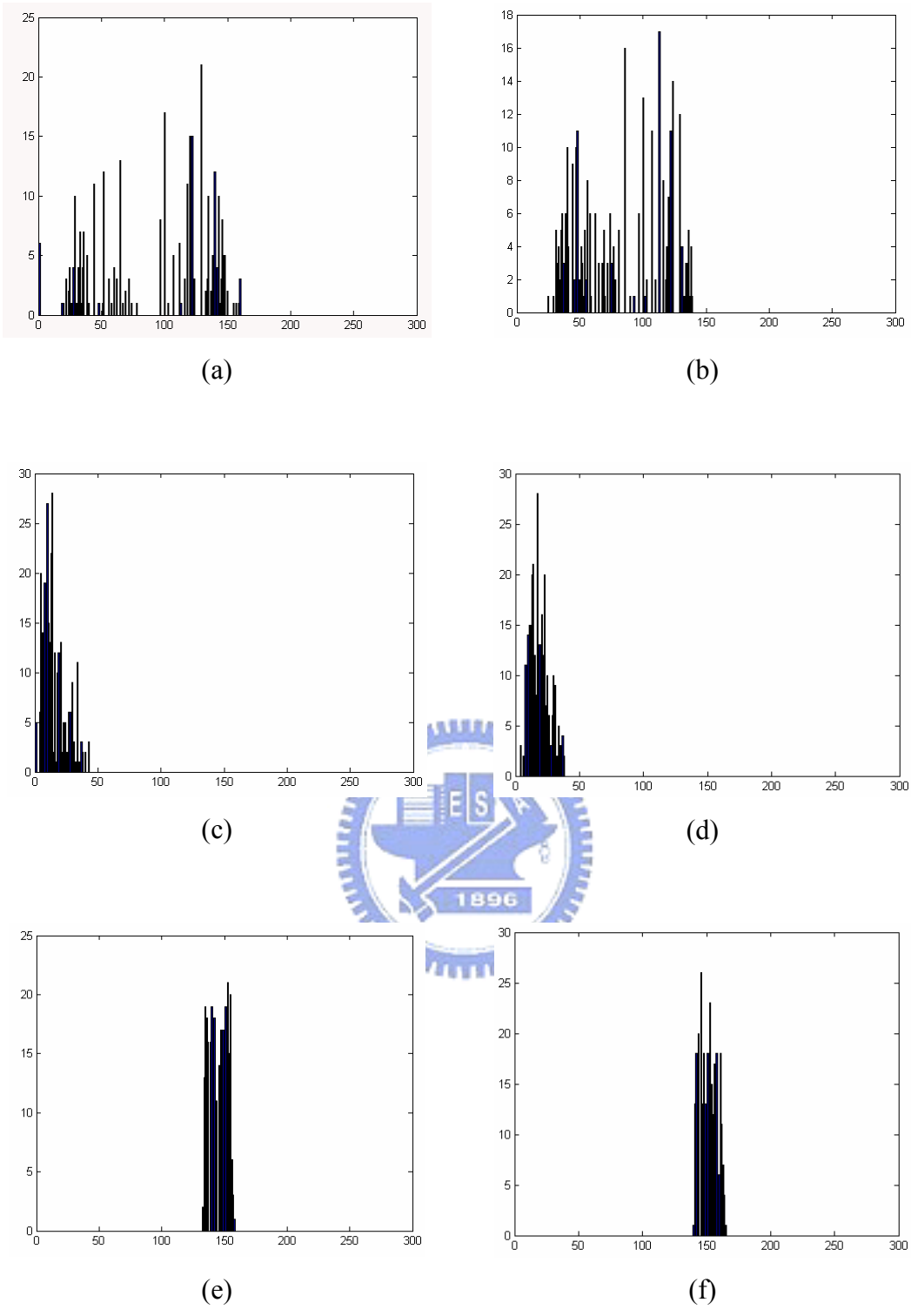


Fig. 4.4 H, S, and V variations versus frame index of background video frome frame 1 to frame 300. (a) H at (10, 10), (b) H at (120, 160), (c) S at (10, 10), (d) S at (10, 10), (e) V at (10, 10), and (f) V at (10, 10).

In Sec. 2.2, we know that hue is unreliable when the color is close to the gray tones. Hence, we use three criteria (V_t, S_t, H_t) to obtain the hue value reliably in building the background model. In our experiment, we set three criteria by

$$V_t = 50,\ S_t = 50, \text{ and } H_t = 25$$

to make hue value reliably.

Fig. 4.5 shows the background image in the H color components after we use criterion to redefine it. We can find that the hue values in the background image are almost be set to zero. The reason is that our background is simple and the color is similar to the gray tones.

Fig. 4.5 Background image in the redefined H color components.

4.2 Foreground subjects extraction

The V color component is stable and reliable, but it has two drawbacks: the illumination change make it change and the similar color such as yellow, pink, and light blue has the similar value in it. In normal condition, the subjects wear the clothing with the color different from the background, so we can do background subtraction with good performance in the V color component.

In the first step, we use the frame ration in the V color component to get the binary image $B(x, y)$ in Eq. (23) described in Sec. 3.1.3. The value k_V is chosen by experiments and varies with different trials. Hence, we ran a series of experiments to determine the optimal threshold k_V . Fig. 4.6 shows the binary image $B(x, y)$ got by different k_V i with subject's clothing color different from the background. Fig. 4.7 shows the binary image $B(x, y)$ got by different k_V with subject's clothing color similar to the background. Comparing Figs. 4.6 and 4.7, we can find that if the color is different from the background, we can use the threshold value k_V to get a good foreground subject extraction. But we cannot adjust k_V to get a complete and noise-free foreground subject when the clothing color is similar to the background. After the experiment, we set $k_V = 1.3$ in our system.

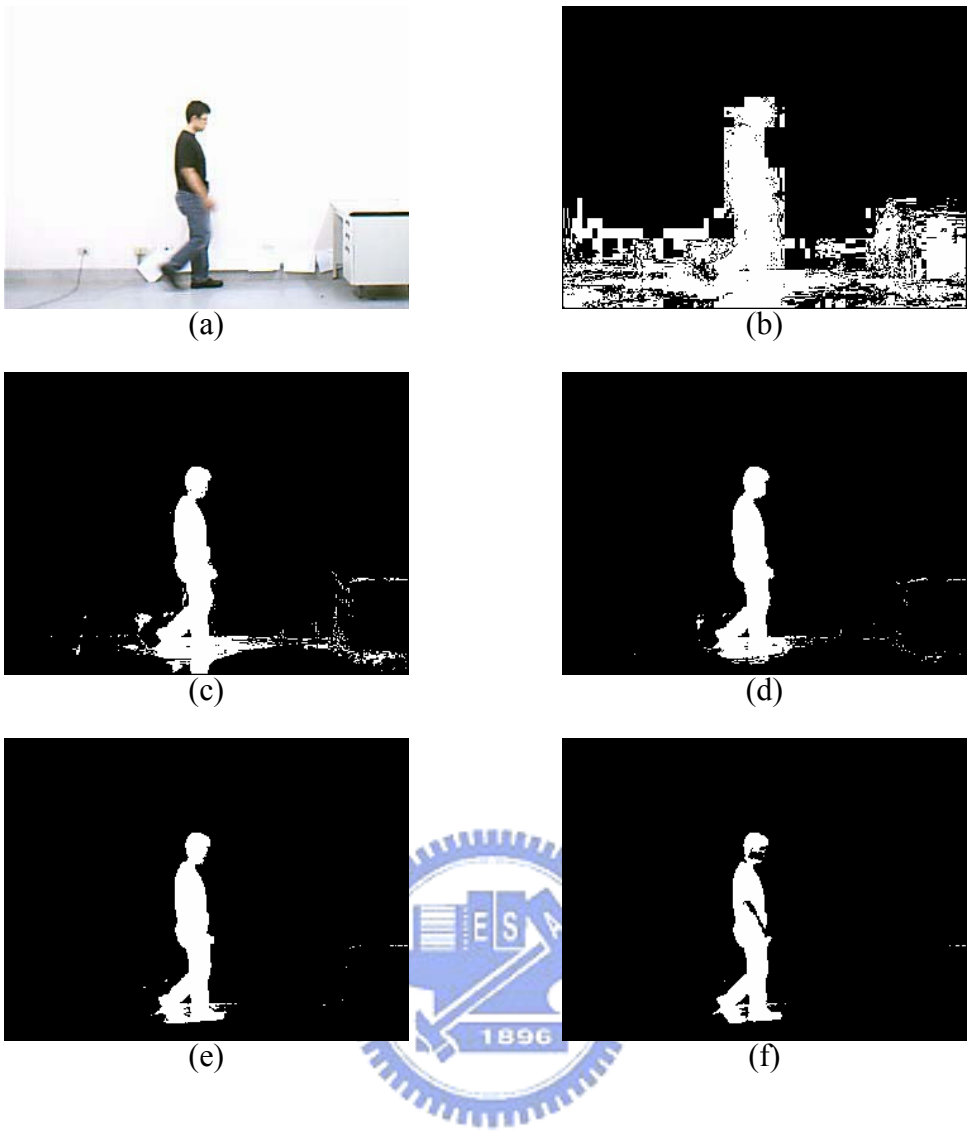


Fig. 4.6 An example of foreground extraction at different k_V thresholds.

(a) An image frame with subject's clothing color different from the background, (b)−(f) foreground detected images, (b) $k_V = 1.0$, (c) $k_V = 1.1$, (d) $k_V = 1.2$, (e) $k_V = 1.3$, and (f) $k_V = 1.4$

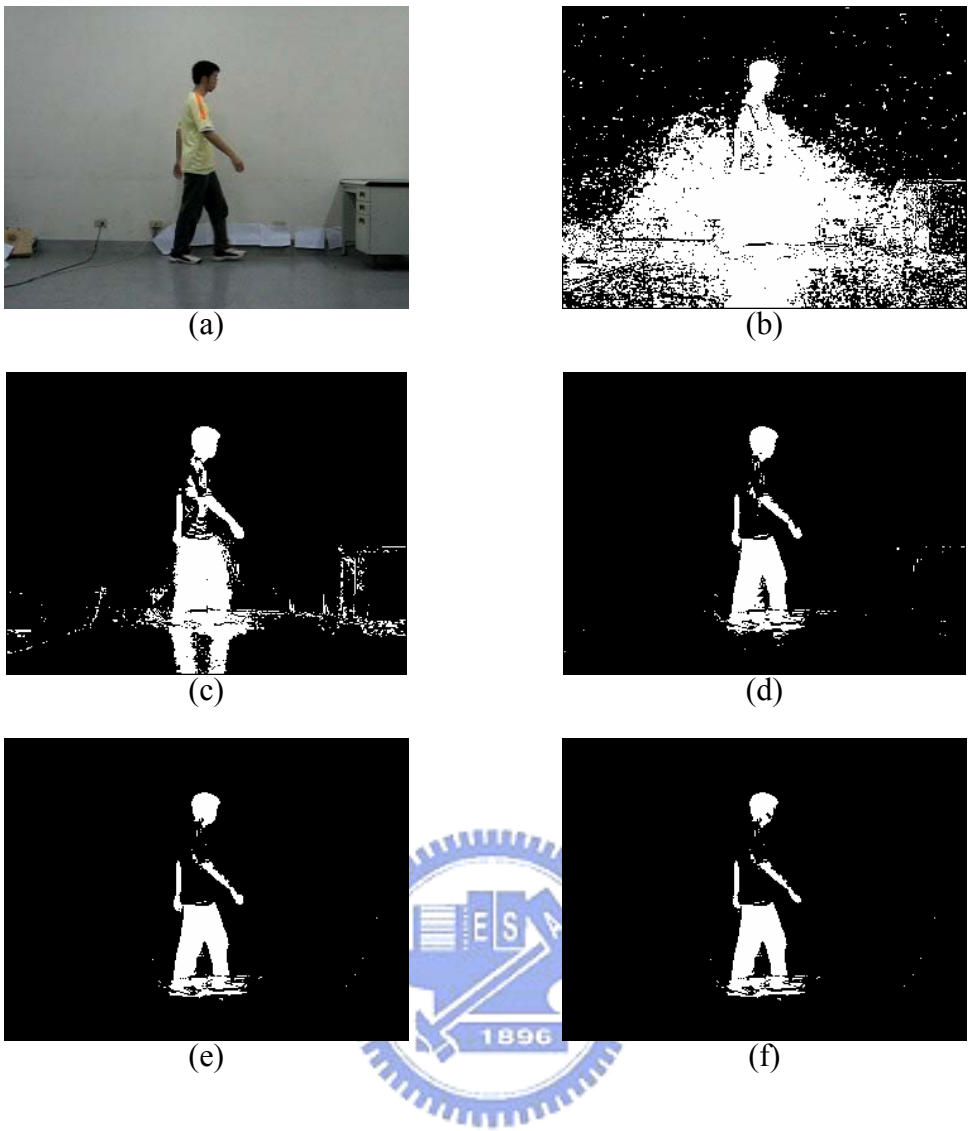


Fig. 4.7 An example of foreground region extraction at different k_V threshold. (a) An image frame with subject's clothing color similar to the background, (b)−(f) foreground detected images, (b) $k_V = 1.0$, (c) $k_V = 1.1$, (d) $k_V = 1.2$, (e) $k_V = 1.3$, and (f) $k_V = 1.4$,

The influence of shadow makes the foreground subjects distort and influence the recognition result. We use the shadow mask in Eq. (24) described in Sec. 3.1.3 to classify the pixels whether it is a shadow point or not. Fig. 4.8 shows the process result in shadow suppression. Fig. 4.8 (a) and (b) are two input images. Fig. 4.8 (c) and (d) are the foreground subject without shadow detection. The foreground subject

with shadow detection is shown in Fig. 4.8 (e) and (f).

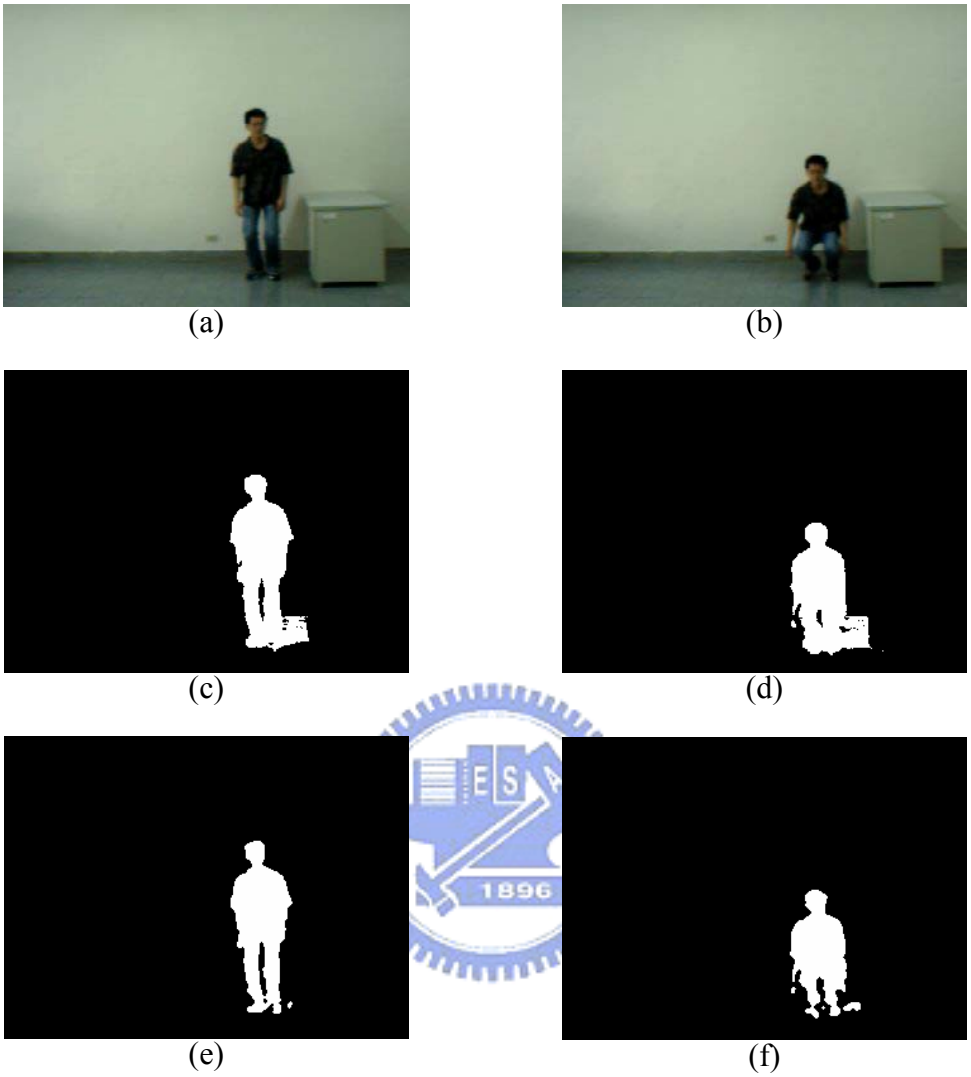


Fig. 4.8 The example of the shadow detection.

The model in Group B contains three action subjects wear light blue clothing, yellow clothing, and pink clothing, respectively. In the previous experiment, we cannot adjust k_V to get a complete and clean foreground subject in Group B. Hence, we do the color compensation in Eq. (25) described in Sec. 3.1.3. In what follows, the effectiveness of color compensation in obtaining a more accurate foreground is described. In Fig. 4.9, the left column contains input images; the middle column contains the resulting foreground images, without color com-

pensation step; and the right column is the foreground images detected with color compensation step. From the Fig.4.9, we have found that we can get good compensation when the clothing color is light blue and yellow, but cannot obtain good compensation when the clothing color is pink. The reason is that when pink color pixels are transformed from RGB color space to HSV color space, the saturation of pink is lower than the set criterion S_t. Hence, we cannot recover those pixels from background to foreground for such small chromaticity difference in this space.


(a)


(a1)


(a2)


(b)


(b1)


(b2)


(c)


(c1)


(c2)

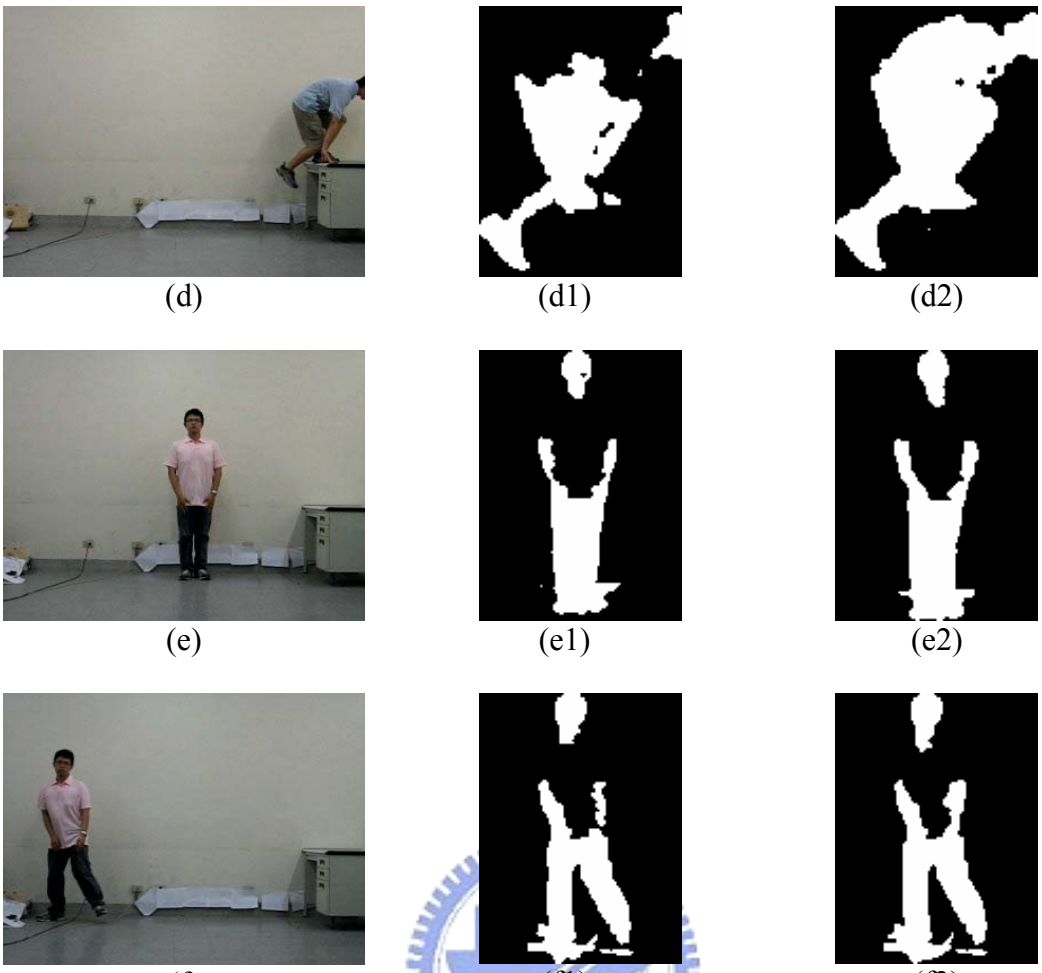


Fig. 4.9. Foreground detection without and with color compensation. (a)−(f) is the input images, (a1)−(f1) the foreground images, without color compensation, (a2)−(f2) the foreground images detected with color compensation.

We randomly selected 100 frames from the video sequence of the model with a subject wearing clothing similar to the background color. The " foreground subject ground truths" of these 100 frames were generated manually. Let A be a detected foreground subject region and B be the corresponding "ground truth." Then we test the pixel accuracy by the following two metrics. Metric 1 is a measure concerning whole segmented region pixels relative to these pixels in A the same with in B. To this end, we calculate the accuracy rate by

$$\text{Accuracy rate}_1 = \frac{N_s}{N_{total}} \times 100\%, \tag{37}$$

where N_{total} is the pixel number of segmented foreground image, and N_s is the pixel number that the pixel in A is the same as that in B, i.e., such of true positive and false negative pixels of A relative to B. Metric 2 is adopted from [27] by

$$\text{Accuracy rate}_2 = \frac{A \cap B}{A \cup B} \times 100\%, \tag{38}$$

This measure counts the percentage of the mutual positive pixels to expanded positive pixels. Table I shows the accuracy rate in metric 1 and metric 2 of 100 frames, and demonstrates the improvement of color compensation over that without color compensation.

TABLE I

COMPARISON RESULT OF THE PIXEL ACCURACY RATES OVER 100 IMAGES

	Without color compensation	With color compensation
Metric 1	78.81%	89.13%
Metric 2	59.61%	81.23%

4.3 Fuzzy rule construction for action recognition

We construct the template model matrix and the fuzzy rule database with the Group A. We chose six kinds of essential templates for “walking from left to right,” “walking from right to left,” and “climb down,” respectively; five for “climbing up,” three for “crouching” and two for “jumping.” There are total 28 kinds of essential templates, and called 28 classes. The essential template numbers of each activity

depend on how long it takes. Each essential template is a cluster with five template images which are from five different training person's and have similar postures. Fig 4.10 and Fig. 4.11 are two examples of some templates of two training model.

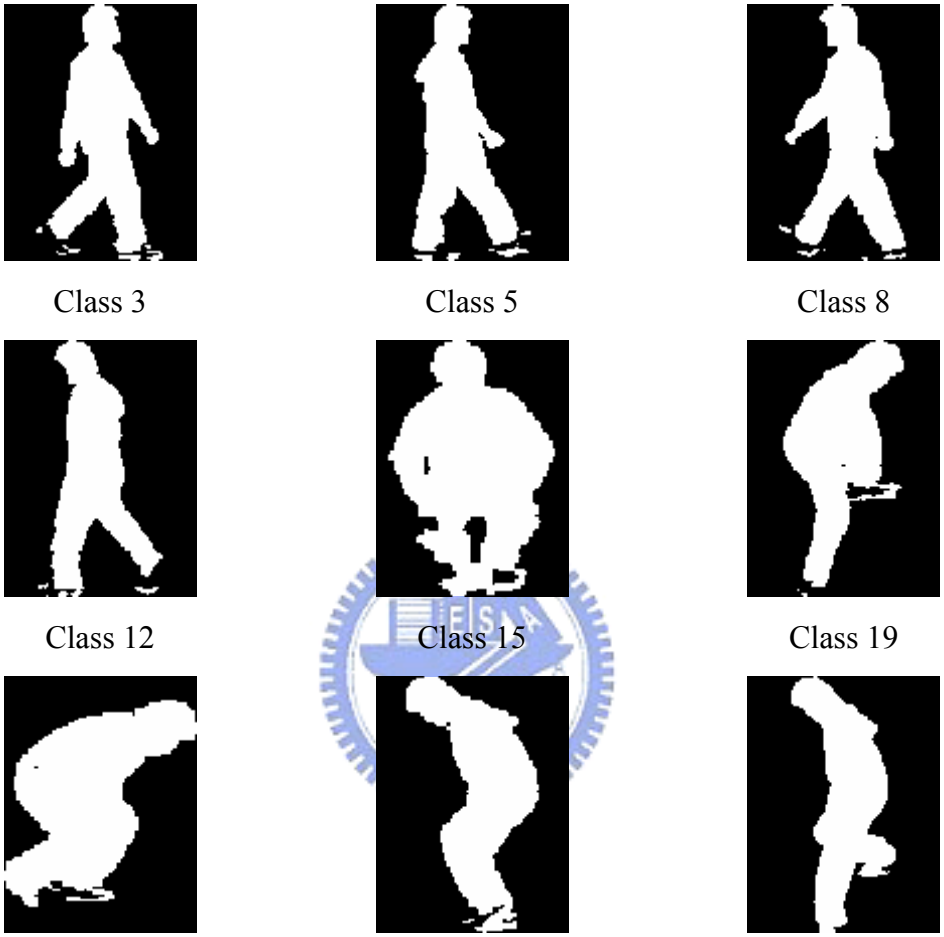


Fig. 4.10 Some "essential templates of posture" of model A.

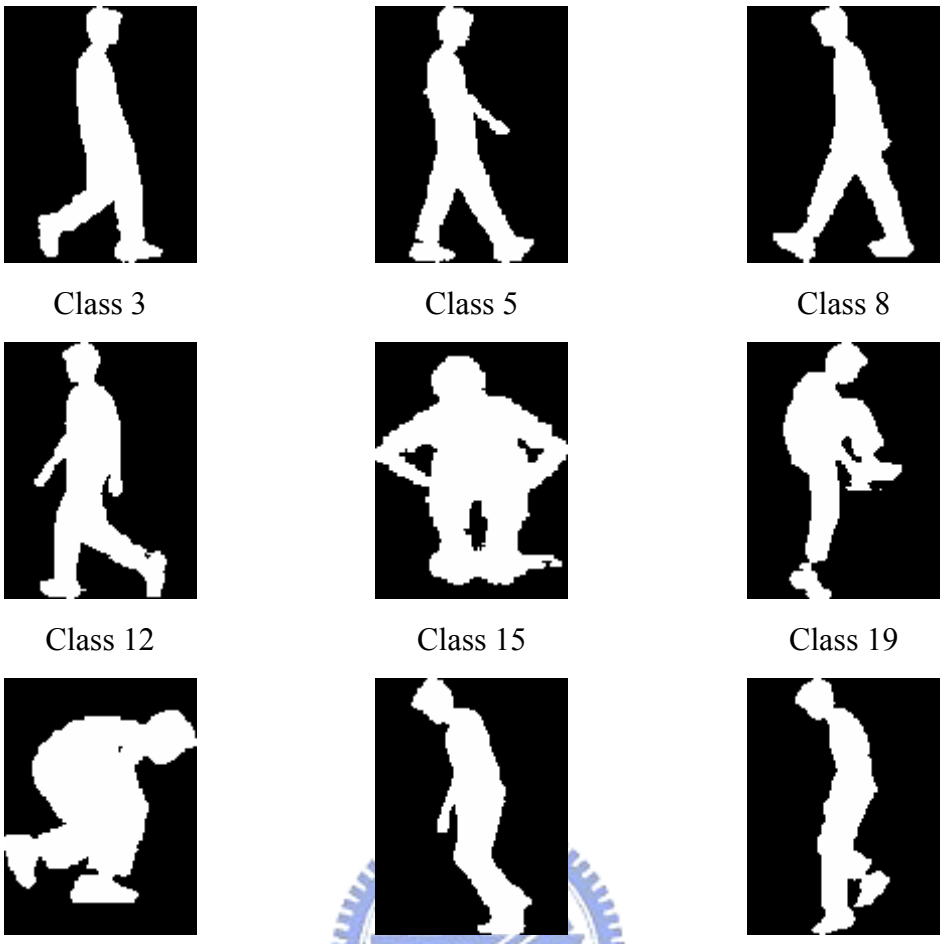


Fig. 4.11 Corresponding "essential templates of posture," Fig. 4.10, of model B.

The template images are transformed to canonical space by the methods described in Chapter 2. The mean vectors and the standard deviation vectors of all templates were computed by Eq. (29). Each template image of a training model was treated as a center. Hence, there were 140 mean vectors because of five training models and 28 classes of templates. Besides, there were six groups of standard deviation vectors and mean vectors because of six kinds of different training models.

After determining the standard deviation vectors, the corresponding training video frames are inputted. The relationship between each image frame and each template is calculated by using Eq. (32) in Section 3.3. We gathered three images as

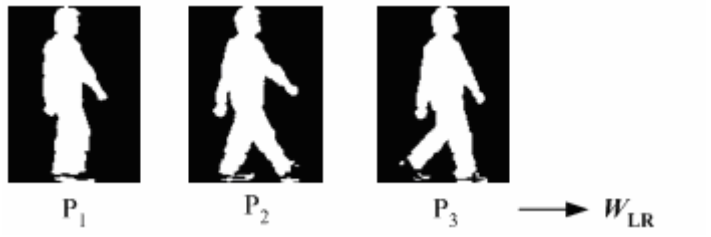
a group in order to include temporal information. The interval between each of these three images is five image frames which is the same as in template selection. Training is accomplished in off-line situation. Therefore, we gathered three images from different start points to train fuzzy rules. For examples: the first frame, the 6-th frame and 11-th frame are gathered together as an input training data; the second frame, the 7-th frame and 12-th frame are gathered together as another input training data; the third frame, the 8-th frame and the 13-th frame are gathered together as an other input training data *etc*. Different start points of image frames are used for training fuzzy rules in our experiment, because the starting posture of testing video and of training video may not be the same. By utilizing different start points, the system is able to learn much more combinations of image frames and increase accuracy of fuzzy rules.

The group of the three images is converted to the posture sequence which has the maximum summation of three membership function values in Eq. (32). Each posture sequence will trigger a corresponding rule one time. If the corresponding rule is not existent, a new rule is built in the form of **IF-THEN** which is represented in Section 3.3. Table II shows some fuzzy rules in the experiment. Where W_{LR} is the activity "walking form lest to right," W_{RL} is the activity "walking from right to left," J_{UMP} is the activity "jumping," C_{ROUCH} is the activity "crouching," C_{UP} is the activity "climbing up" and C_{DOWN} is the activity "climbing down." $P_1, P_2, ..., P_{28}$ are the linguistic labels that represent the templates of the activities. Two of the fuzzy rules are represented in the view of template images in Fig. 4.12.

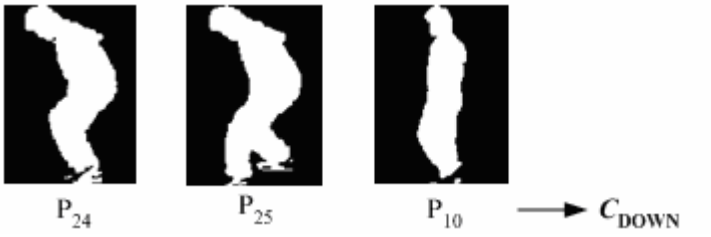
TABLE II

SOME OF THE OBTAINED FUZZY RULE BASE

Number	Image 1	Image 2	Image 3	Class
1	P_1	P_1	P_1	W_{LR}
2	P_1	P_1	P_2	W_{LR}
3	P_1	P_1	P_3	W_{LR}
⋮	⋮	⋮	⋮	⋮
30	P_4	P_{11}	P_{12}	W_{RL}
⋮	⋮	⋮	⋮	⋮
60	P_3	P_{13}	P_{14}	J_{UMP}
⋮	⋮	⋮	⋮	⋮
80	P_{13}	P_{16}	P_{17}	C_{ROUCH}
⋮	⋮	⋮	⋮	⋮
91	P_2	P_{18}	P_{18}	C_{UP}
⋮	⋮	⋮	⋮	⋮
129	P_{27}	P_{28}	P_{10}	C_{DOWN}
130	P_{28}	P_7	P_7	C_{DOWN}
131	P_{28}	P_{28}	P_{10}	C_{DOWN}



(a)

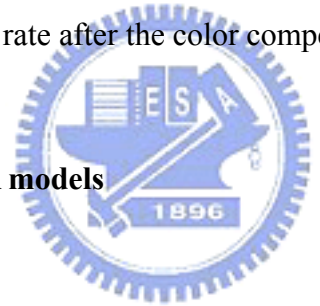


(b)

Fig. 4.12 Two examples of fuzzy rules (a) Walking from left to right and (b) Climbing down.

4.4 The recognition rate of activities

The activity recognition system in our experiment is off-line presented and tested; therefore, the testing video is not done in real time phase. We input the testing video from different starting frames which is similar to the way for the training fuzzy rules. Namely, we recognize the video from the first frame, the second frame, the third frame and the fourth frame, *etc.* with the sampling intervals of five frames. We experimented in Group A and Group B. When experimenting in Group A, the testing video was not used for constructing templates and fuzzy rules. Hence, there are six corresponding databases for the video of the six models. But in Group B, we constructed the templates and fuzzy rules by used the six model in Group A and compared the recognition rate after the color compensation.



- **Experiment in Group A models**

The frame numbers in each activity of every model are shown in Table III. The total number of testing frames is 2685. Table IV. shows the recognition rate in Group A, and the average action accuracy rate is 93.51%.

TABLE III.

THE FRAME NUMBER OF EACH ACTIVITY IN GROUP A MODELS

Testing data	Frame numbers					
	W_{LR}	W_{RL}	C_{ROUCH}	J_{UMP}	C_{UP}	C_{DOWN}
Model A1	90	86	74	80	70	50
Model A2	83	93	94	78	67	55
Model A3	84	84	81	76	65	50
Model A4	87	84	80	75	59	60
Model A5	79	78	87	76	62	61
Model A6	81	84	75	76	68	53

TABLE IV

THE RECOGNITION RATE OF EACH ACTIVITY IN GROUP A MODELS

Testing data	Recognition rate (%)					
	W_{LR}	W_{RL}	C_{ROUCH}	J_{UMP}	C_{UP}	C_{DOWN}
Model A1	100.00	100.00	100.00	77.5	97.56	90.97
Model A2	100.00	82.46	97.14	85.76	100.00	94.29
Model A3	100.00	85.58	74.36	94.12	100.00	74.43
Model A4	100.00	93.65	86.17	91.30	93.62	76.67
Model A5	100.00	97.06	100.00	97.89	90.67	97.75
Model A6	100.00	95.31	92.59	96.84	100.00	100.00
Average	93.51					

- **Experiment in Group B models**

Table V shows the recognition rate of the model wear yellow clothing in Group B. Table VI shows the recognition rate of the model wear light blue clothing in Group B. Table VII shows the recognition rate of the model wear pink clothing in Group B. In these three tables, we can find that the recognition rate can be improve when the color compensable.

TABLE V

THE RECOGNITION RATE WITH THE MODEL WEARING YELLOW CLOTHING

	Recognition rate (%)	
	Without color compensation	With color compensation
W_{LR}	93.55	100
W_{RL}	31.43	75.71
C_{ROUCH}	51.72	96.13
J_{UMP}	0	78.49
C_{UP}	56.94	78.78
C_{DOWN}	63.89	72.78
Average	48.10	84.01

TABLE VI

THE RECOGNITION RATE WITH THE MODEL WEARING LIGHT BLUE CLOTHING

	Recognition rate (%)	
	Without color compensation	With color compensation
W_{LR}	84.44	95.56
W_{RL}	11.65	73.55
C_{ROUCH}	71.82	84.76
J_{UMP}	0	75.36
C_{UP}	90.77	93.84
C_{DOWN}	90.24	97.56
Average	53.13	85.18

TABLE VII

THE RECOGNITION RATE WITH THE MODEL WEARING PINK CLOTHING

	Recognition rate (%)	
	Without color compensation	With color compensation
W_{LR}	81.54	86.15
W_{RL}	6.25	14.63
C_{ROUCH}	39.43	50
J_{UMP}	0	0
C_{UP}	50.76	55.69
C_{DOWN}	71.01	76.32
Average	42.63	45.86

Chapter 5 Conclusion

In this thesis, we proposed the foreground subject extraction in the HSV color space to improve the human activity recognition system and define three criteria to reduce the hue instability effects on the chromatic channels of pixels. In the HSV color space, we can utilize not only the luminance component but also the chromatic component existent in the background image. In this way, we can reliably extract the foreground subject, even when the foreground luminance is similar to that of the background. Experimental results have shown that we get good results in the foreground subject extraction.

In our action recognition system, CST and EST are used to reduce data dimensionality and optimize the class separability simultaneously. Fuzzy rule base for activity recognition is obtained by learning from three temporal postures extracted and down sample from tracing video. In the testing phase, a three posture sequences is processed by fuzzy rule base, and the recognition result is determined as the action which best matches the posture sequence in the fuzzy rules.

Experiment results have shown that extracting the foreground image in the HSV space improves not only the pixel accuracy of the foreground image segmented but also the recognition accuracy of human activity.

Some subjects wearing light color clothing, e.g., pink, still cannot be extracted well, which deserves to be investigated further. In addition, recognition from a different viewing direction, extensions of various test environments, more complicated surrounding, and more complicated activity are our future work.

References

[1] J. Yamato, J. Ohya, K. Ishii, "Recognizing Human Action in Time-SequentialImages using Hidden Markov Model," In *Proc. IEEE CVPR*, pp. 379－385, 1992.

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

[3] I. Cohen and H. Li, "Inference of human postures by classification of 3D human body shape," in *Proc. IEEE Int. Workshop on Anal. Modeling of Faces and Gestures*, pp. 74－81, 2003.

[4] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: Real-Time Surveillance of People and Their Activities," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809－830, 2000.

[5] S. Park and J. K. Aggarwal, "Segmentation and Tracking of Interacting Human Body Parts under Occlusion and Shadowing," in *Proc. of the Workshop on Motion and Video Computing*, pp.105－111, 2002.

[6] M. K. Leung and Y. H. Yang, "First sight: A human-body outline labeling system," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, no. 4, pp. 359－377,1995.

[7] S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfeld, "Detection and Location of People in Video Images Using Adaptive Fusion of Color and Edge Information," in *Proc. Int. Conf. Pattern Recognition*, pp. 627 – 630, 2000.

[8] 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.

[9] 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.

[10] A. Prati, I. Mikic, M. Trivedi and R. Cucchiara, "Detecting Moving Shadows:Algorithms and Evaluation," in *Proc. IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 918 – 923, 2003.

[11] B. Chen and Y. Lei, "Indoor and Outdoor People detection and Shadow Suppression by Exploiting HSV Color Information," *Fourth International Conference on Computer and Information Technology*, pp. 137 – 142, 2004.

[12] S. Vitabile, G. Pilato, G. Pollaccia, and F. Sorbello, "Road Signs Recognition Using a Dynamic Pixel Aggregation Technique in the HSV Color Space," in *Proc. 11th International Conference on Image Analysis and Processing*, pp. 572 – 577, 2002.

[13] R. Cucchiara, M. Piccardi and A. Prati, "Detecting Moving Objects, Ghosts, and Shadows in Video Streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1337–1342, 2003.

[14] 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.

[15] J. Wang, G. Yuantao, K. N. Plataniotis, and A. N. Venetsanopoulos, "Select eigenfaces for face recognition with one training sample per subject," *8th Cont., Automat. Robot. Vision Conf.*, *ICARCV 2004*, vol. 1, pp. 391–396, 2004.

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

[17] 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.

[18] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[19] E. Trentin and M. Gori, "Robust combination of neural networks and hidden Markov models for speech recognition," *IEEE Trans. Neural Networks*, vol.

14, pp. 1519－1531, 2003.

[20] S. Yoshizawa, N. Wada, N. Hayasaka, and Y. Miyanaga, "Scalable architecture for word HMM-based speech recognition," in *Proc. the 2004 Int. Symposium Circuits Syst., ISCAS 2004*, vol. 3, pp. III–417－20, 2004

[21] L. Nianjun, B. C. Lovell, and P. J. Kootsookos, "Evaluation of HMM training algorithms for letter hand gesture recognition," in *Proc. the 3rd IEEE Int. Symposium Signal Processing Inform. Technol., ISSPIT 2003*, pp.648－651, 2003

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

[23] Mu-Chun Su, "A fuzzy rule-based approach to spatio-temporal hand gesture recognition," *IEEE Trans. Sys., Man Cybern*, vol. 30, no. 2, pp. 276－281, 2000

[24] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," in *Proc*. *ICASSP*, pp. 2148－2151, 1997.

[25] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd edition, 1300 Boylston Street Chestnut Hill, Massachusetts USA: Academic Press, 1990.

[26] 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.

[27] L. Li, W. Huang, I. Y. H. Gu, and Q. Tian, "Statistical Modeling of Complex Backgrounds for Foreground Object Detection," *IEEE Transactions of Image Process,* vol.13, no.11, pp.1459 – 1472, 2004.