國 立 交 通 大 學

電 控 工 程 研 究 所

碩 士 論 文

室內環境之人物追蹤與連續動作辨識

Person Tracking and Continuous Activity Recognition in an In-door Environment

研 究 生：許 懷 顥

指 導 教 授：張 志 永

中 華 民 國 九十九 年 七 月

室內環境之人物追蹤與連續動作辨識

Person Tracking and Continuous Activity Recognition in an In-door Environment

學　　生：許懷顥　　Student : Huai-Hao Syu

指導教授：張志永　　Advisor : Jyh-Yeong Chang

國立交通大學

電控工程研究所

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 and Control Engineering

July 2010

Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 九 年 七 月

室內環境之人物追蹤與連續動作辨識

學生：許懷顥　　指導教授：張志永 博士

國立交通大學 電控工程研究所

摘要

人體動作辨識系統在電腦視覺領域一直是很熱門的研究與應用目標。在居家監控系統中最常見的方式是，使用固定式的攝影機，對室內的人物進行追蹤與動作辨識。為了達到即時監控之目標，處理的演算法必須快速，而且又必須能夠有效的分析影像。

在本論文中，動作辨識的目標是人體，為了更正確的擷取出人體部份，我們同時使用灰階域與 HSV 色彩空間，建立兩個背景模型，提升消除影像中陰影部分之效果，使得前後景之分離結果能夠更完整。我們以 5：1 降低取樣頻率，取得即時影像，擷取出的前景部份，經過特徵空間轉換與標準空間轉換後，累積三張上述降頻取樣動作影像後，藉由預先學習而建立之模糊法則與時序動作姿態比對，完成人體動作之辨識。當人在室內活動時，系統能夠依據在 YCbCr 色彩空間中，藉由建立之衣服色彩模型，進行人物之辨識與追蹤。此系統可以追蹤某人於何時進入或離開此房間及辨識某人在此期間的動作。

Person Tracking and Continuous Activity Recognition in an In-door Environment

STUDENT: Huai-Hao Syu ADVISOR: Dr. Jyh-Yeong Chang

Institute of Electrical Control Engineering
National Chiao-Tung University

ABSTRACT

Human activity recognition system is now a very popular subject for research and application. Using a fixed camera to track a person and recognize his (her) activity is widely seen in home surveillance. For real-time surveillance, the embedded algorithmic of the process must be efficient and fast to meet the real-time constraint.

In the thesis, a new person tracking and continuous activity recognition is proposed. To extract the human correctly, we build two background models, in grayscale and HSV color space as well, which could also reduce the shadowing effect well. For better efficiency and separability, the binary image is firstly transformed to a new space by eigenspace and then canonical space transformation, and the recognition is finally done in canonical space. A three image frame sequence, 5:1 down sampling from the video, is converted to a posture sequence by template matching. The posture sequence is classified to an action by fuzzy rules inference. Fuzzy rule approach can not only combine temporal sequence information for recognition but also be tolerant to variation of action done by different people while absorbing deviation due to down sample at the same time. According to the generated clothes color models in the YCbCr color space, we can identify and track the person in the room. The system can track someone who enter or leave the room and recognize his (her) activity.

致謝

本論文承蒙指導教授 張志永博士的細心指導與督促，在每一次的討論與教導當中，皆讓我獲益良多。如果沒有老師的協助，論文中的研究方法與實驗流程可能無法順利完成。

也非常感謝我們實驗室的同學：天健、品宏、嘉臨、泳發、竣超、忠憲以及家杰學長，除了協助我拍攝動作姿態影片進行實驗與測試外，也適時的給我支持鼓勵與參與討論，提供我許多寶貴的意見。

最後將此論文獻給我親愛的家人與朋友，謝謝大家給予鼓勵與支持，讓我有動力可以順利完成論文。

目次

圖次

表次

第一章　研究介紹

1.1　研究動機

許多的居家安全照護系統、自動監控系統的核心部分就是動作辨識系統。例如有了動作辨識的功能，居家安全照護系統才能夠來判斷需要照護的目標物是否有異常的動作或行為出現，決定是否發出警告通知；自動監控系統藉由行為模式，辨識目標物是否為可疑目標，決定是否要加以錄影存證或追蹤。但是人體的行為模式、動作分析，無法由一個精確的程式語法來做定義，這是動作辨識的困難之處。

動作辨識系統在電腦視覺領域一直是很熱門的研究與應用目標，例如：Yamato *et al.* [1] 將影像序列轉換成一個象徵序列，使用 HMM(Hidden Markov Model)來做動作辨識； W^4 [2] 應用在偵測與追蹤人並且辨識其動作。很多的辨識系統多半是在一個比較單純的背景下作分析，或是處理演算法過於複雜，導致處理時間過長因而無法達到及時處理的效果。如果要應用在實際面，這些問題就必須要克服，例如背景條件可能就必須選擇稍加複雜的環境，以接近於日常生活的背景條件；而當讀入即時影像時，處理的演算法必須簡化，而且又必須能夠有效的分析影像，這些都是此實驗的難題處。

我們系統的目標需要自動的監控人並且辨識出此人的行為動作。除了能夠辨識人體動作外，當有人進入此監控的空間中時，系統會建立出此人的衣服顏色模型，這是為了當有兩人以上在影像中時，能夠依據衣服的色彩模型，辨識出是何人進行何種動作。

此系統的流程圖如下，大略可以分成三部分。第一部分是前景的萃取；第二部分是為了降低計算量加快處理速度，把影像的資料轉換到一個維度較低的空間中，我們稱之為標準空間轉換；第三部分是根據由事前建立之模糊法則來辨識動作姿態，結束後繼續重覆以上步驟。

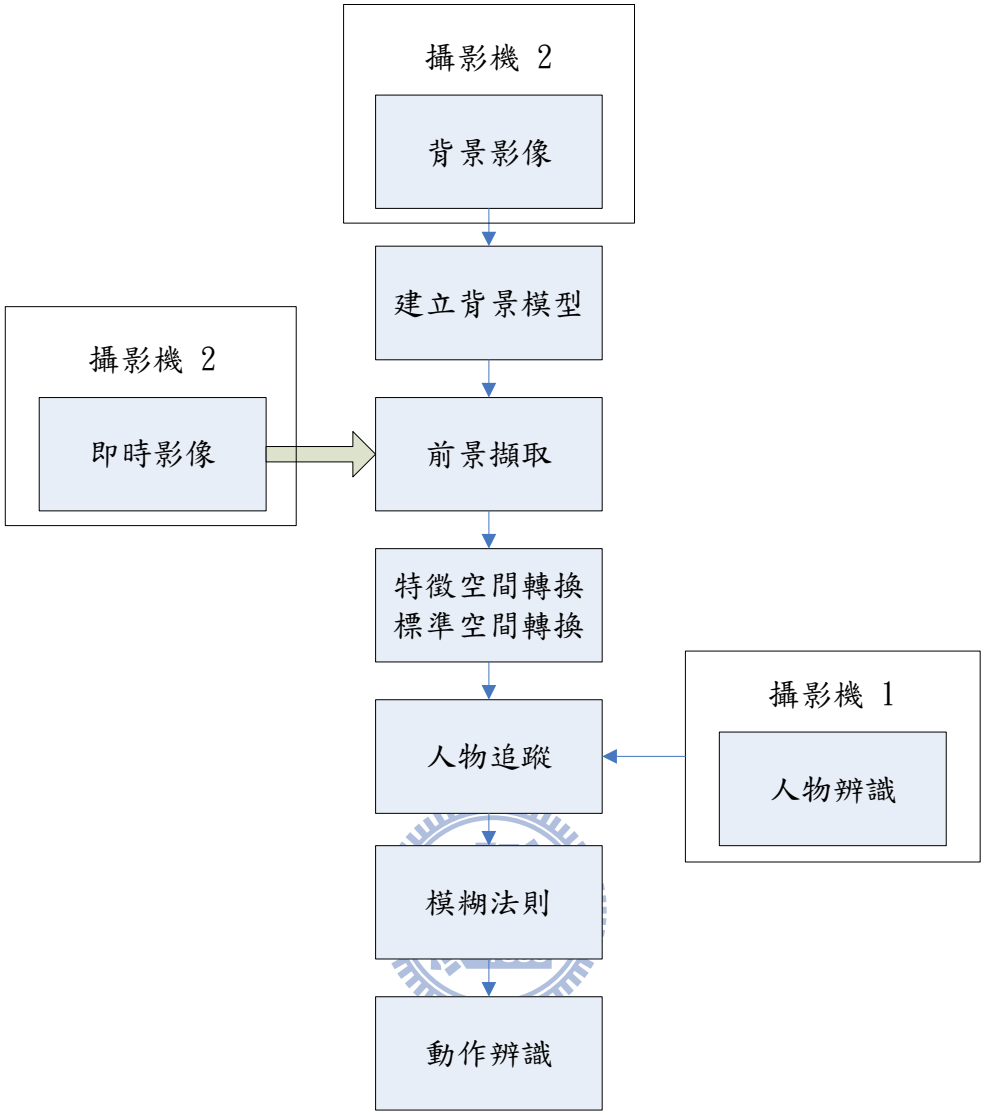


圖 1.1 系統流程圖

1.2 前後景分離

建立一個背景模型是進行前後景分離的首要目標，背景模型的建立可以藉由一張影像或是一段影片，我們選取後者可以避免因為取單張影像當背景，如果影像中有雜訊而造成背景模型建立錯誤，而導致擷取前景錯誤的情況發生。

在一個固定攝影機的環境下，我們可以很容易取得一段無人的單純背景影片。利用此背景影片建立一個背景模型，每當攝影機畫面更新時，可以隨時進行一次前後景分離即得到一張二元影像。其中在原始影像中所出現的陰影部份，經

過前後景分離後，常會被標示為前景部分，造成動作辨識效果上的誤判，文獻上已經有很多方法，都是用來濾除此陰影部份的[3]-[7]。將影像中的前景部分以矩形方式剪裁並且將剪裁後的影像縮放至一固定大小，此為一標準化程序。

因為我們設備是使用彩色攝影機，因此可以不需侷限在灰階域中做分析。在很多參考文獻中，其作者已設計出偵測人的方法。除了使用灰階域做背景模型之外，我們加入了 HSV 色彩空間做背景模型執行前後景分離。

1.3 特徵空間轉換與標準空間轉換

如果直接接讀取到的影像做處理，裡面包含有許多多餘的資料，如果可以藉由某些空間轉換，消去多餘的資料保留重要的影像特徵，將可以降低影像資料量的大小，而且處理速度也可以加快。我們需要一個空間轉換能夠降低影像的維度，並且轉換後的影像資料，保留了原始影像的重要部分。例如：傅立葉轉換(Fourier transformation)、小波轉換(wavelet transformation)、主要元素分析(principal component analysis, PCA)。其中我們使用的是特徵空間轉換(eigenspace transformation, EST)與標準空間轉換(canonical transformation, CST)。

特徵空間轉換(EST)是根據主要元素分析(PCA)為基礎，應用在很多方面例如：自動人臉辨識 [8]、[9]，動作辨識 [10]；標準空間轉換是根據標準分析(canonical analysis)，將座標軸轉換資料至一個基底，它能盡量增加不同群組間距離，縮小相同群組資料間的的聚集度，可以降低資料量的維度大小最大量，維持資料裡不同群間的特性。但是如果資料量很大，例如影像或視訊，如果直接進行標準空間轉換，計算量將會相當可觀。因此我們整合特徵空間轉換與標準空間轉換，就可以達到降低資料的維度、提升分群的效能。在這新的空間中，動作辨識將會更加簡單化，處理速度也可以加快。

1.4 動作辨識

我們降低取樣頻率為 5：1，從影像序列中取出間隔五張的影像，經過特徵空間轉換(EST)與標準空間轉換(CST)後，可以萃取出動作影像的特徵。在固定的間隔時間影像序列中取得三張接序之 5：1 降低取樣頻率的影像，分別經過上述步驟處理後，可以由已經建立的模糊法則來判斷為何種動作。

HMM 在動作辨識與語音辨識上是相當有名的方法，是將時間序列上的每個暫態動作或語音視為一模型，基本觀念可以參考 [11]、[12]。

影像在經過特徵空間轉換(EST)與標準空間轉換(CST)後，資料維度降低造成有些許的影像資料遺失。在我們的系統中，是使用模糊法則為依據來辨識人的動作姿態，並非單就影像圖形的形狀來做辨識，因為模糊法則的特色就是可以在某些模糊或沒有明確定義的條件下作判斷，例如：當不同人進行同一種動作時，動作影像略有差異；同一個人，進行同種動作，但在不同環境下，前後景分離後的結果，也會有差異；當影像經過空間轉換時，造成影像資料遺失等因素，而模糊法則可以吸收每個人之間動作姿態差異性，並不會對系統的動作辨識結果造成太大的影響。我們先選取出每種動作的樣版動作影像進行訓練，產生出代表每個動作的 IF-THEN 法則。藉由這些模糊法則，計算出這三張連續影像與訓練模型中之何者樣版影像最近似，就可以辨識出動作。

1.5 論文架構

我們將在第二章介紹特徵空間轉換與標準空間轉換，說明如何將較高維度的資料轉換至較低維度的資料，以利運算加快處理速度並提升分群的效能。在第三章中，將詳細逐步說明動作辨識系統的流程細節與架構。第四章將列出實驗結果與動作辨識正確率之數據。最後，將在第五章作上述之結論。

第二章 基礎觀念介紹

2.1 特徵空間轉換與標準空間轉換介紹

在影像處理處理方面，影像的資料內容維度通常都很龐大，因此我們需要藉由空間的轉換，來降低維度減少計算量並加快處理速度。在眾多的空間轉換方式中，有許多方法都可以降低資料的維度，例如：傅立葉轉換(Fourier transformation)、小波轉換(wavelet transformation)、主要元素分析(principal component analysis, PCA)。其中我們使用的 PCA 是根據整張影像資料陣列的共變數矩陣(global covariance matrix)，但是要找出資料中存在的分群結構並不容易。為了加強區分出分群結構的能力，Etemad 和 Chellappa [13] 使用線性區分分析，(linear discriminant analysis, LDA)，也被稱作標準分析(canonical analysis, CA)，此方式可以將分群最佳化與提升分群的效果，也就是達到群與群之間的距離最大化，同一群中元素的距離最小化，我們稱作標準空間轉換(canonical transformation, CST)。結合 EST 與 CST 兩種方式的優點，我們將可以降低資料的維度並且使分群的效果最佳化。

影像資料的處理過程先經由 PCA，將資料由高維度的空間轉換到低維度的空間，再經過 CST 在低維度空間中提升分群效果，也加快運算速度。

假設我們在系統中需要學習訓練共有 C 群，每一群都各有代表的姿態，$\mathbf{x}'_{i,j}$ 代表第 i 群中的第 j 張影像，N_i 代表第 i 群中的影像個數，所以訓練影像的集合中總共有 $N_T = N_1 + N_2 + \cdots + N_c$ 張影像，訓練影像集合可以表示成

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

首先要先將每張影像做正規化

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

接著我們可以求得訓練影像像素的平均值

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

因此訓練影像的集合可以改寫成 $n \times N_T$ 的矩陣 $\mathbf{X}$ 。每張影像 $\mathbf{x}_{ij}$ 構成矩陣 $\mathbf{X}$ 中的一行，表示方式如下

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

2.1.1 特徵空間轉換(EST)

EST 最大的特色就是能夠將資料由原始高維度的空間轉換到低維度的空間，並且保有最小的均方誤差(mean-square error)，如此可以避免在轉換的過程中造成資料遺失。EST 就是由影像資料的共變異矩陣(covariance matrix)的特徵值與特徵向量，依據最大的變異方向，來旋轉原始影像資料的座標。

假設 $\mathbf{XX^T}$ 矩陣的秩(rank)是 K ，因此 $\mathbf{XX^T}$ 矩陣的 K 個非零特徵值是 $\lambda_1, \lambda_2, \cdots, \lambda_K$ ，相對應的特徵向量是 $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_K$ ，即關係式為

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

其中 $R = XX^{\mathrm{T}}$ 且 $\mathbf{R}$ 對稱且方正矩陣。由上式關係式，我們可以求得大小為 $n \times n$ 的 $\mathbf{XX^T}$ 矩陣的特徵值與其對應的特徵向量，但是 $\mathbf{XX^T}$ 矩陣的維度就是影像的維度，如果要直接計算是相當困難的。因此我們使用奇異值分解(singular value decomposition)理論，由另一個取代矩陣 $\tilde{\mathbf{R}}$ ，來計算特徵值與特徵向量

$$\tilde{\mathbf{R}} = \mathbf{X}^{\mathrm{T}}\mathbf{X} \qquad \mathbf{X} \text{ 為資料矩陣} \tag{6}$$

其中矩陣 $\tilde{\mathbf{R}}$ 的大小是 $N_T \times N_T$，而且其維度比矩陣大小為 $n \times n$ 的 $\mathbf{R}$ 小。假設矩陣 $\tilde{\mathbf{R}}$ 有 K 個非零特徵值為 $\tilde{\lambda}_1, \tilde{\lambda}_2, \cdots, \tilde{\lambda}_K$，相對應的特徵向量是 $\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2, \cdots, \tilde{\mathbf{e}}_K$。矩陣 $\mathbf{R}$ 與矩陣 $\tilde{\mathbf{R}}$ 的特徵值和特徵向量關係可以表示成

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

利用上式求得的 K 個特徵向量，作為正交基底(orthogonal basis)來延展出一個新的向量空間，每一張影像都可以被投影到此維度大小為 K 之空間中一點。根據 PCA 定理，每張影像都可以由前 k 個最大的特徵值 $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_k|$ 與相對應的特徵向量 $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$ 來近似表示。由這 k 個特徵向量 $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$，可以延展出一個特徵空間，由此特徵空間可以將原始的影像 $\mathbf{x}_{i,j}$ 投影至另一空間中以 $\mathbf{y}_{i,j}$ 表示，其方程式可以表示成

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

我們稱此矩陣 $[\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k]^{\mathrm{T}}$ 為特徵空間轉換矩陣(eigenspace transformation matrix)。原始影像 $\mathbf{x}_{i,j}$ 經由此特徵空間轉換後，就可以由這 k 個特徵向量作線性組合表示，$\mathbf{y}_{i,j}$ 是維度為一的向量，其中共有 k 個元素，分別代表相對應的特徵向量作線性組合時的係數。

2.1.2 標準空間轉換(CST)

根據標準分析(canonical analysis)，我們假設$\{\phi_1,\phi_2,\cdots,\phi_c\}$是經過特徵空間轉換後的類別分群，$\mathbf{y}_{i,j}$代表第 i 群中的第 j 個向量。所有群集合中的平均向量可以表示成

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

第 i 群集合中的平均向量可以表示成

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

令 $\mathbf{S}_\mathrm{w}$ 表示同一群中向量平均距離的矩陣，$\mathbf{S}_\mathrm{b}$ 表示群與群之間向量平均距離的矩陣，方程式可以表示成

$$\mathbf{S}_\mathrm{w} = \frac{1}{N_T}\sum_{i=1}^{c}\sum_{\mathbf{y}_{i,j}\in\varphi_i}\left(\mathbf{y}_{i,j}-\mathbf{m}_i\right)\left(\mathbf{y}_{i,j}-\mathbf{m}_i\right)^\mathrm{T} \tag{11}$$

$$\mathbf{S}_\mathrm{b} = \frac{1}{N_T}\sum_{i=1}^{c}N_i\left(\mathbf{m}_i-\mathbf{m}_y\right)\left(\mathbf{m}_i-\mathbf{m}_y\right)^\mathrm{T}$$

此時我們的目標就是要同時使 $\mathbf{S}_\mathrm{w}$ 最小化與使 $\mathbf{S}_\mathrm{b}$ 最大化，即可以使分群的效果最佳化。我們可以使用 generalized Fisher linear discriminant function 如下

$$\mathbf{J}(\mathbf{W}) = \frac{\mathbf{W}^\mathrm{T}\mathbf{S}_\mathrm{b}\mathbf{W}}{\mathbf{W}^\mathrm{T}\mathbf{S}_\mathrm{w}\mathbf{W}}. \tag{12}$$

若滿足下式，即可以藉由選定的特徵轉換 $\mathbf{W}$，使得新空間中的變異比例值最大化

$$\frac{\partial \mathbf{J}}{\partial \mathbf{W}} = 0. \tag{13}$$

假設 $\mathbf{W}^*$是最佳解，其中行向量$\mathbf{w}_i^*$即是第 i 大的特徵值λ_i所對應的特徵向量。根據標準分析定理可以求解上式方程式

$$\mathbf{S}_{\mathrm{b}}\mathbf{w}_i^* = \lambda_i \mathbf{S}_{\mathrm{w}}\mathbf{w}_i^* \tag{14}$$

求解(12)式後，我們可以求得 c–1 個非零的特徵值與其相對應的特徵向量$[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_c]$，由這些特徵向量可以當作新正交基底，延展出一維度為 c–1 的標準空間。因此，特徵空間中的每一個點皆可以投影到標準空間的一個點，表示式如下

$$\mathbf{z}_{i,j} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{c-1}]^{\mathrm{T}}\mathbf{y}_{i,j} \tag{15}$$

其中$\mathbf{z}_{i,j}$代表經過投影後的點，而由正交基底構成的矩陣$[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{c-1}]^{\mathrm{T}}$即為標準空間轉換矩陣。經由兩(13)、(14)方程式，每張影像皆可以投影到一維度為 c—1 的空間中之一點。

$$\mathrm{z}_{i,j} = \mathrm{H} \cdot \mathrm{x}_{i,j} \tag{16}$$

其中 $\mathrm{H} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{c-1}]^{\mathrm{T}}[\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k]^{\mathrm{T}}$.

2.2 HSV 色彩空間介紹

HSV(hue 色調、saturation 飽和度和 value 亮度)色彩空間是比較接近人類對色彩的定義。在 HSV 色彩空間中，色彩的分佈成一圓錐體。由此圓錐體的圓形切面討論，定義一條0°線代表純紅色，而色調就是在圓面上與0°線的夾角角度；飽和度就是圓切面上某點到圓心的距離，越接近圓心表示飽和度越低，越接近圓周表示飽和度越高；亮度就是某點的圓切面到圓錐體錐點切面的距離，在圓錐體錐點亮度為 0。

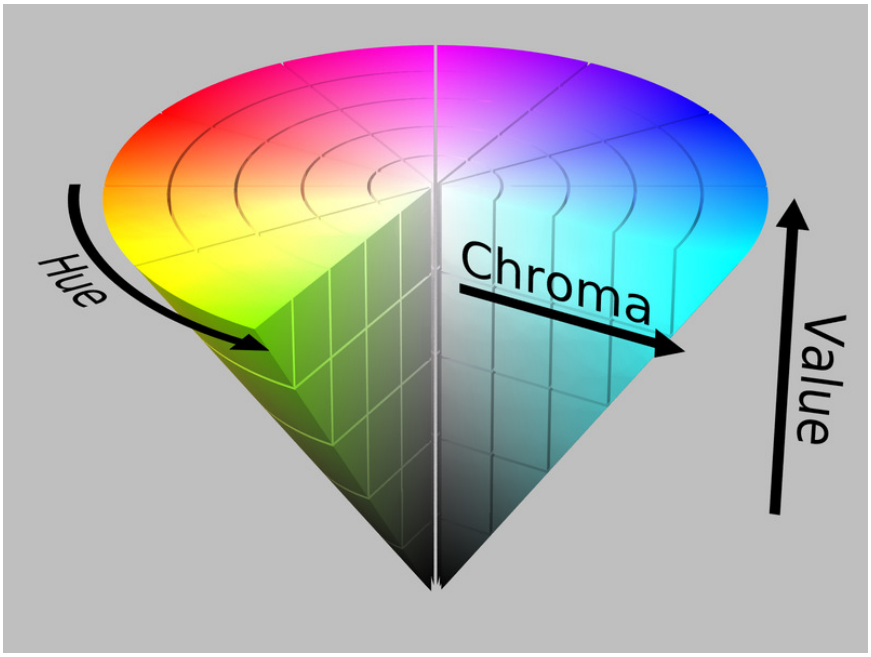


圖 2.1 HSV 色彩空間圓錐體

色調(hue)中不包含亮度成份，因此不會因為環境光線的亮或暗造成影響，這特色對我們系統在建立背景模型，進行前後景分離時有很大幫助。例如在進行前後景分離時，不會因為攝影機擷取的即時影像與背景模型亮度有所差異時，造成前後景分離錯誤。但是有幾個關鍵因素，會造成問題：(1)當亮度極低時，即接近圓錐體錐點處，除了色調無法使用，而且飽和度也無法使用。(2)當飽和度極低時，即接近圓切面的圓心處時，色調無法使用 [14]。根據 Ohba *et al.* [15]，為了避開上述的極端條件使得能夠正確區別色調，我們設定三個門檻參數解決此

問題。

(1)因為色調在 HSV 色彩空間中，即為圓錐體圓切面上與0°線的夾角角度，在夾角角度為0與2π時無法定義為何種色調，為了避開此區域，我們設定若H在0與H_t之區間或$2\pi - H_t$與2π之區間時，令色調$H = 0$。(2)當飽和度極低時，即接近圓切面的圓心處時，色調無法使用，所以若$S < S_t$時，令色調$H = 0$。(3)當亮度極低時，即接近圓錐體錐點處，色調無法使用，所以若$V < V_t$時，令色調$H = 0$。由以上設定的參數門檻可以解決在特殊條件下無法正確區分色調的問題。

第三章 動作辨識系統

當我們要進行人們動作辨識的第一步驟就是要進行前後景分離，將前景擷取出來，使用此部分來作辨識。在前後景分離的這一部分之研究已有許多可以參考的文獻，例如 W^4 [2] 就是利用一段單純背景的影片，由統計的方式找出影片序中灰階值的最大值與最小值，以及連續影片中相同位置灰階差值最大的值，由此建立出一個在灰階域中的背景模型。除此之外，我們使用類似 W^4 在灰階域的統計方式，在 H、S、V 中分別計算影片序中相同位置像素最大值與最小值，與影片序中前後張影像相同位置像素的比例之最大值。由此背景模型，我們可以藉由前景的判斷式找出前景部份，並做影像的雜訊濾除與補償後，就可以藉由建立的模糊法則來做動作辨識。

3.1 建立背景模型

在系統中，我們將使用兩個背景模型。在建立背景模型中，我們參考了 W^4 的文獻，利用灰階建立背景模型。因為 W^4 的文獻中，強調的是不需使用彩色攝影機。但我們系統是使用彩色攝影機，而且之後為了擷取更完整的前景影像，所使用的陰影濾波器，在 HSV 色彩空間相較於 RGB 色彩空間較容易區分出陰影區塊，因此我們增加了在另一個背景模型於 HSV 色彩空間中，相似於 W^4 在灰階域的統計方式，各別應用在 H、S、V 中，來建立另一個背景模型。因此在系統中，我們共使用兩個背景模型，分別為灰階與 HSV 色彩空間雙背景模型，來進行前後景的分離。灰階背景模型如下：

$I_i^{gray}(x, y)$ 代表第 i 張影像在 (x, y) 位置的灰階值

$$\begin{bmatrix} m^{gray}(x,y) \\ n^{gray}(x,y) \\ d^{gray}(x,y) \end{bmatrix} = \begin{bmatrix} \max_i \left\{ I_i^{gray}(x,y) \right\} \\ \min_i \left\{ I_i^{gray}(x,y) \right\} \\ \max_i \left\{ \left| I_i^{gray}(x,y) - I_{i-1}^{gray}(x,y) \right| \right\} \end{bmatrix} \tag{17}$$

其中 $i = 1, 2, ..., N$

HSV 色彩空間背景模型如下：

$I_i^H(x,y)$ 代表第 i 張影像在 (x,y) 位置的色調值(hue value)，$I_i^S(x,y)$ 代表第 i 張影像在 (x,y) 位置的飽和度(saturation value)，$I_i^V(x,y)$ 代表第 i 張影像在 (x,y) 位置的亮度值(hue value)。

$$\begin{bmatrix} m^H(x,y) \\ n^H(x,y) \\ d^H(x,y) \end{bmatrix} = \begin{bmatrix} \max_i \left\{ I_i^H(x,y) \right\} \\ \min_i \left\{ I_i^H(x,y) \right\} \\ \max_i \left\{ \left| I_i^H(x,y) - I_{i-1}^H(x,y) \right| \right\} \end{bmatrix} \tag{18}$$

$$\begin{bmatrix} m^S(x,y) \\ n^S(x,y) \\ d^S(x,y) \end{bmatrix} = \begin{bmatrix} \max_i \left\{ I_i^S(x,y) \right\} \\ \min_i \left\{ I_i^S(x,y) \right\} \\ \max_i \left\{ \left| I_i^S(x,y) - I_{i-1}^S(x,y) \right| \right\} \end{bmatrix} \tag{19}$$

$$\begin{bmatrix} m^V(x,y) \\ n^V(x,y) \\ d^V(x,y) \end{bmatrix} = \begin{cases} \begin{bmatrix} \max_i \left\{ I_i^V(x,y) \right\} \\ \min_i \left\{ I_i^V(x,y) \right\} \\ \max_i \left\{ \left| I_i^V(x,y) / I_{i-1}^V(x,y) \right| \right\} \end{bmatrix} & \text{if } I_i^V(x,y) / I_{i-1}^V(x,y) \geq 1 \\ \begin{bmatrix} \max_i \left\{ I_i^V(x,y) \right\} \\ \min_i \left\{ I_i^V(x,y) \right\} \\ \max_i \left\{ \left| I_{i-1}^V(x,y) / I_i^V(x,y) \right| \right\} \end{bmatrix} & \text{if } I_i^V(x,y) / I_{i-1}^V(x,y) < 1 \end{cases} \tag{20}$$

其中 $i = 1, 2, ..., N$

3.2 前後景分離

當讀入一段欲進行前後景分離的影片時，可以由上一步驟建立的兩個背景模型，分別在灰階域與 HSV 色彩空間中各別進行前景影像的擷取。在灰階域中參考 W^4 文獻中的前景判斷式，$I^t(x,y)$ 代表第 t 時讀入即時影像，在 (x,y) 位置的灰階值，若滿足下列判斷式就可以在灰階域中擷取出前景部分。

$$I^1{}_{foreground}(x,y)=\begin{cases}255, & I^t(x,y)>(m^{gray}(x,y)-k\mu) \\ & \text{and}\quad I^t(x,y)<(n^{gray}(x,y)+k\mu) \\ 0, & \text{otherwise}\end{cases} \tag{21}$$

其中 $I^1{}_{foreground}(x,y)$ 是二元影像，$m^{gray}(x,y)$ 與 $n^{gray}(x,y)$ 由 (17) 式所得，μ 是 $d^{gray}(x,y)$ 陣列的中位數，k 是可調參數，在系統中我們設定 k 值為 2。

另外在 HSV 色彩空間中，若滿足以下之前景判斷式，即可擷取出前景部分。$I_i^V(x,y)$ 是讀入影像在 HSV 色彩空間中的亮度部份(value)，$m^V(x,y)$、$n^V(x,y)$ 與 $d^V(x,y)$，由(20)式所得。

$$I^2{}_{foreground}(x,y)=\begin{cases}0, & \text{if } I_i^V(x,y)/m^V(x,y)<k_V d^V(x,y) \\ & \text{or } I_i^V(x,y)/n^V(x,y)<k_V d^V(x,y) \\ 255, & \text{otherwise}\end{cases} \tag{22}$$

其中 $I^2{}_{foreground}(x,y)$ 是二元影像，我們在系統中設定 k_V 為 1.6。若在光線較弱的地方，可以調低 k_V 值，反之調高 k_V 值，來達到較完整的前景影像擷取效果。

3.3 陰影濾波器

觀察前後景分離結束之影像，發現除了分離出來的人體前景之外，在背光處還有許多雜訊出現，此雜訊即為陰影部分。因為陰影的出現，我們無法正確的取出人體動作的部分，我們參考了W[4]的文獻，因此加入了陰影濾波器，消除陰影部份，擷取出更完整的前景影像。陰影濾波器的原理敘述如後，陰影部分的像素灰階值相較於臨近的像素灰階值大，而且與光線的亮度成一定的比例關係。因此，我們使用一量化公式 normalized cross-correlation (NCC)，判斷此分離後的前景部分是否為陰影部分，減少因為陰影所造成的前景誤判。令 $B(x,y)$ 是由背景影片中，任取得的一張影像，經過大小為3×3的中位數濾波器(median filter)處理後，所得的結果。而 T_{xy} 代表在原始影像轉灰階後，以像素 (x,y) 為中心，選擇大小為3×3的視窗。

$$NCC(x,y)=\frac{ER(x,y)}{E_B(x,y)E_{T_{xy}}} \tag{23}$$

其中

$$\begin{aligned} ER(x,y)&=\sum_{n=-N}^{N}\sum_{m=-N}^{N}B(x+n,y+m)T_{xy}(n,m)\\ E_B(x,y)&=\sqrt{\sum_{n=-N}^{N}\sum_{m=-N}^{N}B(x+n,y+m)^2}\\ E_{T_{xy}}&=\sqrt{\sum_{n=-N}^{N}\sum_{m=-N}^{N}T_{xy}(n,m)^2} \end{aligned} \tag{24}$$

若 (x,y) 像素點滿足下列式子，則此像素為陰影。

$$S^1(x,y)=\begin{cases}\text{shadow}, & NCC(x,y)\ge L_{ncc}\ \text{and}\ E_{T_{xy}}<E_B(x,y)\\ \text{foreground}, & \text{otherwise}\end{cases} \tag{25}$$

其中 $S^1(x,y)$ 是二元影像，L_{ncc} 是一個固定的設定值，調整的依據如後，如果 L_{ncc} 設定太低的，會導致原來是正確前景的部分被誤判成陰影部分，而造成前景過於殘破而遺失前景資訊；如果 L_{ncc} 設定太高，陰影部分的濾除效果就不佳，所以在此系統中，我們設定 L_{ncc} 為 0.995。

同樣地，在 HSV 的色彩空間中，加入陰影濾波器消除陰影部份，方程式如下：

$$S^2(x,y)=\begin{cases}\text{shadow} & \text{if } I_i^V(x,y)-n^V(x,y)<0 \\ & \text{and} \left|I_i^H(x,y)-m^H(x,y)\right|<k_H d^H(x,y) \\ & \text{and} \left|I_i^S(x,y)-m^S(x,y)\right|<k_S d^S(x,y) \\ \text{foreground} & \text{otherwise}\end{cases} \tag{26}$$

其中 $S^2(x,y)$ 是二元影像，$I_i^H(x,y)$ 、$I_i^S(x,y)$ 與 $I_i^V(x,y)$ 是讀入影像在 HSV 色彩空間中的色調(hue)、飽和度(saturation)與亮度部份(value)，$m(x,y)$ 、$n(x,y)$ 與 $d(x,y)$ ，由(20)式所得。

由雙背景模型各別擷取出來的前景部分，取聯集部分可以消去因為前景判斷式中參數設定過於嚴格，造成本來應該是前景卻未正確擷取出的部分。其中前景判斷式中的參數設定較嚴格是為了能夠更精確區分出前後景，因此部分接近背景模型的像素會無法被擷取出來，但因為我們系統是使用雙背景模型擷取後取聯集部分，所以此效應會削弱許多，誤判的程度可以降低。

$$I_{foreground}(x,y)=S^1(x,y)\vee S^2(x,y) \tag{27}$$

3.4 前景影像補償與處理

要將陰影部分完全濾除而且保留所有正確的前景部分是相當困難的，因此在濾除陰影時總是會遺失些許的前景資料，所以會造成前景影像有些許破碎的情況發生。為了解決這個問題，我們加入了 morphological 濾波器中的 *opening* 與 *closing filter*，來修補前景影像。

完成了前景雜訊的濾除與補償處理之後，就可以將原始影像大小的前景圖。首先分別統計 x 軸方向之前景像素個數與 y 軸方向之前景像素個數，如圖 3.1 所示，標示出欲進行前景裁減的 x 軸方向區域為 x_1 到 x_2 與 y 軸方向區域為 y_1 到 y_2 後，將僅有前景的部分剪下，並且縮放至大小為 128×96 的影像，將影像大小正規化才可以進行訓練與辨識。

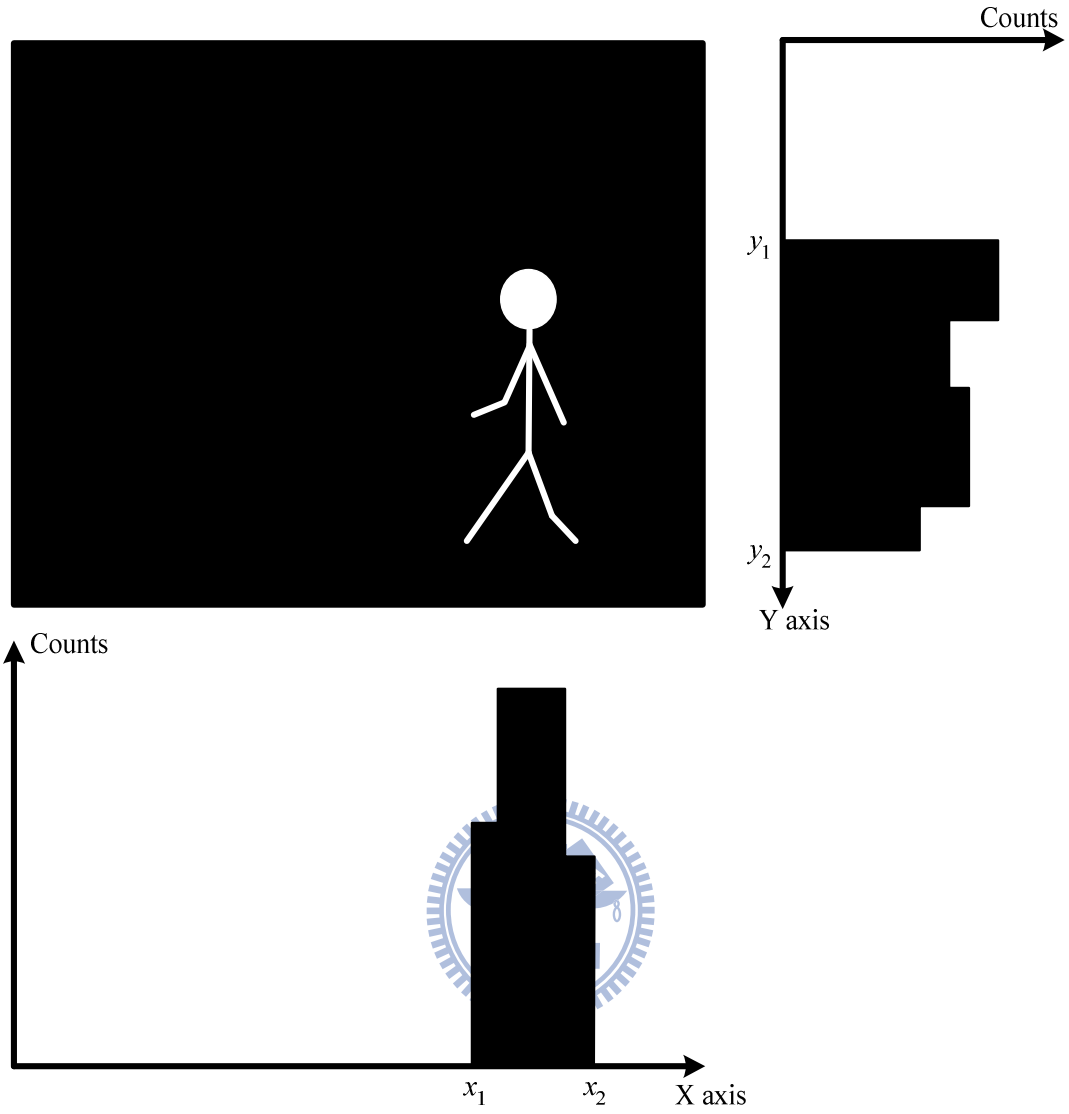


圖 3.1 二元圖 X 軸方向與 Y 軸方向之統計長條圖

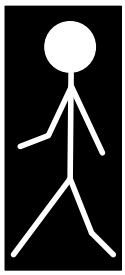


圖 3.2 裁剪後的前景圖

3.5 背景模型更新

為了確保能夠正確的擷取出人體部分，避免取出的前景部分為非人體部分，例如物品等。我們在系統中設定，當場景中為無人狀態時，即可以更新背景模型，避免室內物品設施，因為被移動而成為非人體的前景影像，造成動作辨識誤判。其背景模型的更新條件為：(1)當讀入的即時影像，經過一段時間均無變化。(2)場景中為無人狀態。由下列式子，用來統計讀入的即時影像，經後前後景分離後，其二元值固定無變化之次數。

$$update(x,y)=\begin{cases} update(x,y)+1, & \text{if } I^{t-1}_{foreground}(x,y)=I^{t}_{foreground}(x,y) \\ update(x,y), & \text{otherwise} \end{cases} \tag{28}$$

其中 $I^{t}_{foreground}(x,y)$ 陣列是前後景分離後，像素位置 (x,y) 在第 t 時之二元值。$update(x,y)$ 是用來統計 $I^{t}_{foreground}(x,y)$ 二元值固定無變化之次數。

若要判別此監控環境中是否為無人狀態，可以使用膚色偵測的方式來判斷是否有人存在。首先先將讀取到的即時影像，轉換至正規劃 RGB 色彩空間中

$$r=\frac{R}{R+G+B} \tag{29}$$

$$g=\frac{G}{R+G+B} \tag{30}$$

根據 Soriano 和 Martinkauppi [17]，所提出的在 r-g 平面上之膚色區域邊界條件式如下：

$$f_{upper}(r)=-1.3767r^2+1.0743r+0.1452 \tag{31}$$

$$f_{lower}(r)=-0.7760r^2+0.5601r+0.1766 \tag{32}$$

若符合下列四式判別式，即可以標示出此像素是否為膚色部分，進而偵測出此監

控環境中是否有人存在

$$g > f_{lower}(r) \ \ and \ \ g < f_{upper}(r) \tag{33}$$

$$(r-0.33)^2 + (g-0.33)^2 \le 0.0004 \tag{34}$$

$$R > G > B \tag{35}$$

$$R - G \ge 45 \tag{36}$$

3.6 選擇樣板動作

通常人體的動作表現都有一定的頻率，而我們使用的攝影機設備擷取影像的時間相當快速，其擷取影像的速度為每秒中擷取 30 張影像，因此任意兩張連續影像中僅有些微差距。因此我們從每一個動作序列中，間隔相同的時間差，選取幾張可以代表此動作的影像，當作樣版動作(template image)。

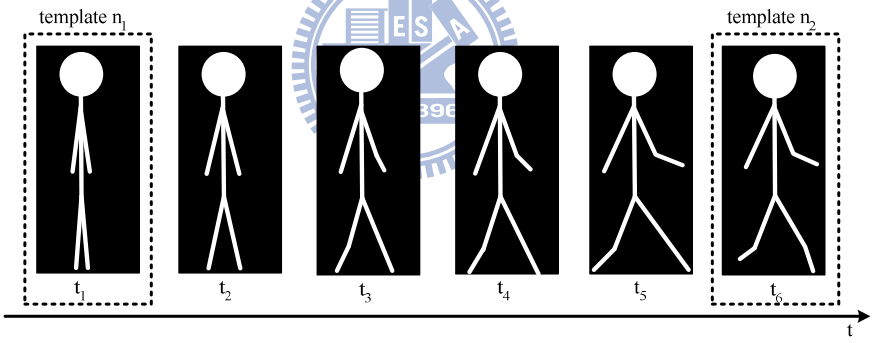


圖 3.3 間隔五張影像取得的樣版影像

這些被選取的樣版動作，藉由特徵空間轉換(EST)與標準空間轉換(CST)，轉換至一個新的標準空間。轉換後因為將影像資料維度降低的關係，會造成些許的資料遺失。但是對任意兩張相似的影像而言，當影像經過轉換後，在新的空間中仍舊會收斂至接近的點，對於辨識效果來說，不會造成太大問題。而且系統的特色就是可以針對同一種動作但由不同的人進行時，此時的影像必定與樣版動作略有差異，但是當影像經過轉換後，在新的空間中仍舊會收斂至接近的點，所以系統仍舊可以正確辨識。也因此我們在選擇樣版動作時，不必選取此動作中的所有

影像，而只需選取部份的樣版動作即可。

當影像經過 EST 與 CST 轉換後，即轉換成維度是 $c-1$ 的向量，用來代表動作特徵。假設系統中有 n 個訓練者模型，每個訓練者模型所有動作總共有 c 張代表影像，表示有 c 個群集合，因此系統中總共有 $n\times c$ 個樣版動作影像。令樣版動作影像中第 i 個分類第 j 個訓練模型表示成向量 $\mathbf{g}_{i,j}$，$\mathbf{g}_{i,j}$ 經過轉換後的向量為 $\mathbf{t}_{i,j}$，方程式可以表示為

$$\mathbf{t}_{i,j}=\mathbf{H}\cdot\mathbf{g}_{i,j},\qquad i=1,2,\cdots,c;\ \ j=1,2,\cdots,n \tag{37}$$

其中 $\mathbf{H}$ 為結合 EST 與 CST 轉換的矩陣，n 代表每一個動作中，選取的樣版動作總個數。$\mathbf{t}_{i,j}$ 是一個維度為 $c-1$，而且其中每個維度均獨立，可以表示成

$$\mathbf{t}_{i,j}=\left[t_{i,j}^{1},\ t_{i,j}^{2},\cdots,\ t_{i,j}^{c-1}\right]^{\mathrm{T}}. \tag{38}$$

3.7 由影像串流建立模糊法則與動作辨識

在動作辨識中，如果僅用單張圖片判斷為何種動作，會有很大的機會發生誤判，例如當人欲進行彎腰與坐下的動作時，都是由站立為起始動作，僅用單張影像判斷時，雖欲進行不同動作但因為有相似的動作出現，會造成誤判發生。因此，我們使用三張間隔相同時間的影像，來完成一次動作的辨識 [18]，這樣可以避免僅用單張影像造成的誤判結果。而且我們希望此系統能夠，同一種動作姿態，如果是由不同人進行的話，仍舊能夠正確辨識。

因此，我們在影像序列中間隔相同時間取一張為樣版影像，進行多次上述步驟後構成一樣版影像序列。我們使用模糊法則作為辨識系統的設計基礎，因為模糊法則的特色就是可以針對不同的資料間的差異加以學習訓練。

在我們的系統中，我們先將每一張做為樣板動作的影像作轉換，所得的轉換後之向量視為不同的動作特徵，可以建立出模糊法則資料庫，用來辨識動作姿態。即可以由不同的影像資料來作學習訓練。

當影像經過 EST 與 CST 轉換後，即轉換成維度是 $c-1$ 的向量，用來代表動作特徵。為了區分這些經過轉換後的影像是屬於何者樣版影像，我們使用高斯分布關係函數(Gaussian membership function)來代表不同動作的特徵。首先，當第 k 個訓練影像 $\mathbf{x}_k$ 輸入後，即可以求得特徵動作的代表向量 $\mathbf{a}_k$

$$\mathbf{a}_k = \mathbf{H} \cdot \mathbf{x}_k . \tag{39}$$

其中 $\mathbf{a}_k$ 是一個維度為 $c-1$，而且其中每個維度均獨立，可以表示成

$$\mathbf{a}_k = \left[a_k^1, a_k^2, \cdots, a_k^{c-1} \right]^{\mathrm{T}} \tag{40}$$

我們假設所有動作特徵的向量彼此都是獨立的，我們可以計算出輸入即時動作影像的向量與樣版動作向量的關係。將 $\mathbf{\Sigma}$ 定義為樣版動作向量的共變異矩陣(covariance matrix)，C_i 定義為第 i 種動作姿態的樣版影像，其關係函數可以表示成

$$\begin{aligned} r_{i,k} &= \mathrm{M}(\mathbf{a_k} | C_i) \\ &= \frac{1}{(2\pi)^{\frac{c-1}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} (\mathbf{a_k} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{a_k} - \mu) \right] \\ &= \arg\max_{\mathrm{j}} \left\{ \prod_{m=1}^{c-1} \frac{1}{\sqrt{2\pi}\sigma_m} \exp\left[-\frac{1}{2} \sum_{m=1}^{c-1} \frac{(a_k^m - \mu_{i,j}^m)^2}{\sigma_m^2} \right] \right\} \end{aligned} \tag{41}$$

其中 j 代表訓練模型的個數，$r_{i,k}$ 代表第 k 張影像在 i 類中的相關程度。之後，我們要找出此張影像是屬於何種動作，可以由下式方程式

$$p_k = \arg\max_{\mathrm{i}} r_{i,k} \tag{42}$$

由以上的關係函式可以判斷讀入的即時動作影像是屬於何種動作，但是這僅是單張影像的結果，還未符合我們系統所設定的由三張影像做動作辨識的理念。

因此，我們要選取三張影像來結合組成樣版動作的資訊。假設我們若選取過多影像時間拉長，則此影像序列可能包含其他動作；若我們若選取若少的影像，則此影像序列資料量過少，不足以代表任何動作。以上兩種情況均會造成，無法針對一動作明確辨識。

假設我們總共有 c 種語言上標示定義(linguistic labels)，每個語言上標示的定義均代表一種樣版動作，因此每張讀入的影像都可以由這 c 種語言上標示定義來描述。我們在系統中，整合三張影像組成一群 (I_1, I_2, I_3)，藉由前文提過的轉換方式，可以將此影像群組轉換成動作特徵向量 $[a_1, a_2, a_3]$，再藉由(32)式將此動作特徵向量 $[a_1, a_2, a_3]$ 以語言上標示定義動作 $[P_1, P_2, P_3]$，再根據已建立的模糊法則來判斷語言上標示定義動作 $[P_1, P_2, P_3]$ 屬於何種動作行為。

參考文獻由 Wang 和 Mendel [18]，所發展的藉由例子之學習建立模糊法則判別式中，影像序列當輸入藉由模糊法則為依據，經過學習訓練後輸出一的動作名稱。其訓練後產生的模糊法則格式如下

若(IF)前述條件成立，**則(THEN)**結果成立

前述條件的個數即為此動作特徵的個數，彼此條件之間用交集**(AND)**做連結。其中 P_j^i 代表三張影像序列中的第 j 張影像，單張影像判斷結果是屬於第 i 種動作，而 D_i 代表結合以上三張影像的辨識結果，此影像序列的判斷結果是屬於第 i 種動作。

$$[P_1^1, P_2^1, P_3^1; D_1]$$

舉例來說，當輸入一個三張影像序列經過 CST 的轉換後的動作特徵向量 $[\mathbf{a}_1^1, \mathbf{a}_2^1, \mathbf{a}_3^1]$，各自代表原三張影像序列的動作特徵，也就是代表動作 $[P_1, P_2, P_3]$。

我們根據這三個代表動作是屬於何種動作的群集合中，就可以辨識出三張影像是屬於何種動作。其判斷式如下

若影像 I_1 屬於動作 P_1^1 **且**影像 I_2 屬於動作 P_2^1 **且**影像 I_{3_1} 屬於動作 P_3^1 **則**動作 D_1 。

$$[P_1^1, P_2^1, P_3^1; D_1]$$

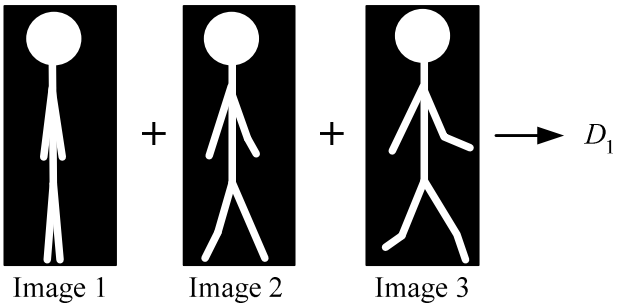


圖 3.4 由三張影像進行動作辨識

第四章 實驗結果

我們選擇在國立交通大學工程五館 807 實驗室中架設彩色攝影機 PTZ 進行實驗，此攝影機擷取畫面速度為每秒鐘取出 30 張影像，其影像解析度為 320×240 個像素。實驗環境光源為日光燈，光線充足且穩定。我們系統設計理念是為了滿足實際生活應用上的需求，因此我們在一個背景稍加複雜的環境下進行實驗，而不是單純的背景環境。下圖 4.1 為固定攝影機的情況下，所擷取到的畫面

圖 4.1 實驗環境

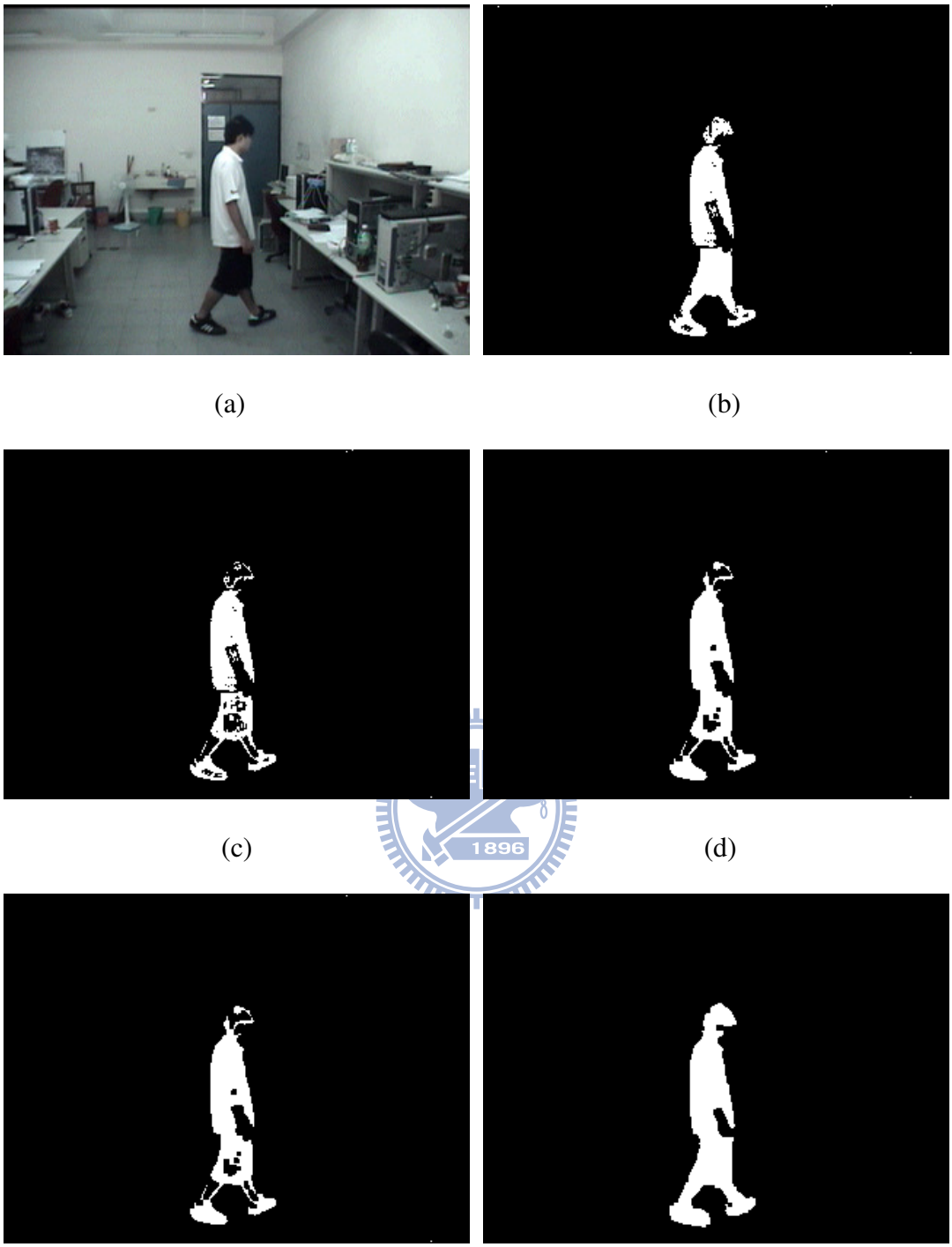
我們已於事前建立模糊法則，進行訓練的動作如下：可辨識的動作有「右側位置讀書」、「左側位置讀書」、「右側位置使用電腦」、「左側位置使用電腦」、「右側位置起立」、「左側位置起立」、「右側位置坐下」、「左側位置坐下」、「進入房間」、「離開房間」、「揮手」、「彎腰」、「對著鏡頭走近或走遠」、「對著鏡頭橫向行走」，總共有十四個動作。

4.1 背景模型建立與前景擷取

首先使用已架設固定式的攝影機，先擷取 100 張連續影像其時間長度約為 3.3 秒，用來訓練建立背景模型。分別在灰階域與 HSV 色彩空間建立模型後，可以由 W4 灰階域與 HSV 色彩空間的前景判斷式(21)、(22)式，找出前景後進行雜訊濾除與前景補償，即可以擷取出大小為 96×128 的前景影像，等待累積至三張影像後，隨後進行動作辨識。

系統中的參數需要調整，以達到最佳的前後景分離效果。在第二章 HSV 色彩空間中，為了避開無法正確區分出色調的區域，所設定的門檻參數為 $H_t = 25$、$S_t = 40$、$V_t = 40$；在 HSV 色彩空間中的前景判斷式參數，如(22)式所示，設定 $k_v = 1.6$；灰階域中的陰影濾波器，判別是否為陰影部份的參數，如(25)式所示，設定 $L_{ncc} = 0.995$；HSV 色彩空間中的陰影濾波器，判別是否為陰影部份的參數，如(26)式所示，$k_H = 1.3$、$k_S = 1.3$。

下圖 4.2 是進行動作「對著鏡頭橫向行走」，依第三章所述之前景擷取的每個步驟之結果。



(a) (b)

(c) (d)

(e) (f)

圖 4.2 (a)原始影像 (b)雙背景模型下擷取之前景 (c)經過陰影濾波器後之結果 (d)前景補償處理 (e)*closing* (f)*opening*

完成由原始影像擷取的前景圖形之後，即進行前景剪圖。首先分別統計 x 軸方向前景像素個數與 y 軸方向前景像素個數，就可以標示出欲進行前景裁減的區域，將此區域標示出後將此圖縮放至大小為 96×128 後，即完成前景圖的剪裁。下圖 4.3 為裁減前景的過程

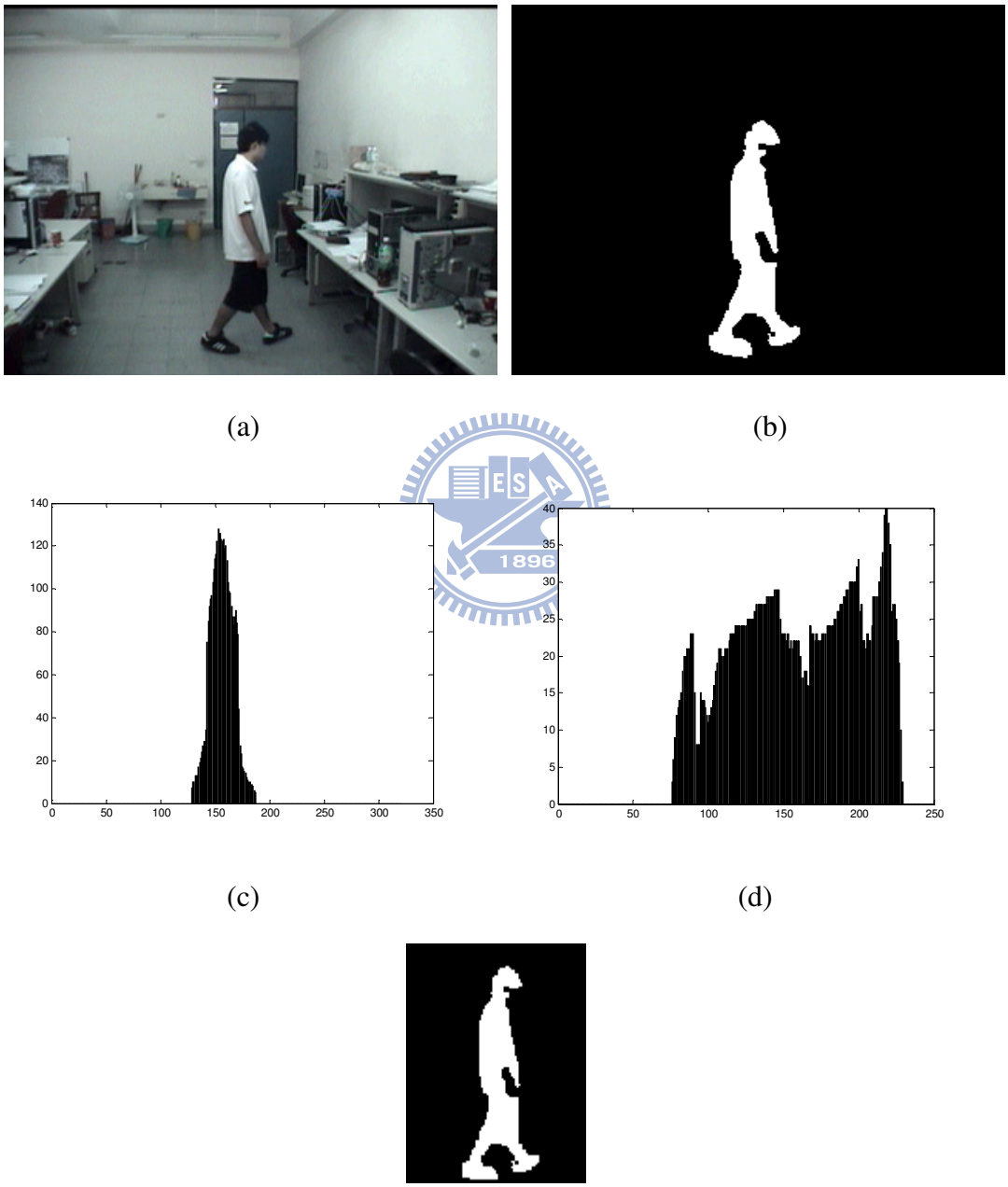


圖 4.3 (a)原始影像 (b)雜訊濾除與補償後結果 (c)統計 x 軸方向前景像素個數 (d)統計 y 軸方向前景像素個數 (e)前景剪圖

若影像中有兩個人時，此時條件是兩人在影像中的位置不重疊。先完成由原始影像擷取的前景影像之後，如圖 4.4(b)所示，再進行前景剪圖。首先統計 x 軸方向前景像素個數，如圖 4.4(c)所示，就可以標示出在 x 軸方向，欲進行前景裁減的兩個區域。在已標示 x 軸方向的左側區域(64~135)內，統計區域內 y 軸方向前景像素個數，如圖 4.4(e)所示，可以標示出在 y 軸方向，欲進行前景裁減的區域，將此區域 x 與 y 座標標示出後，將此圖縮放至大小為 96×128，即完成前景圖的剪裁；在已標示 x 軸方向的右側區域(205~239)內，作法同上述方式。下圖 4.4 為裁減前景的過程

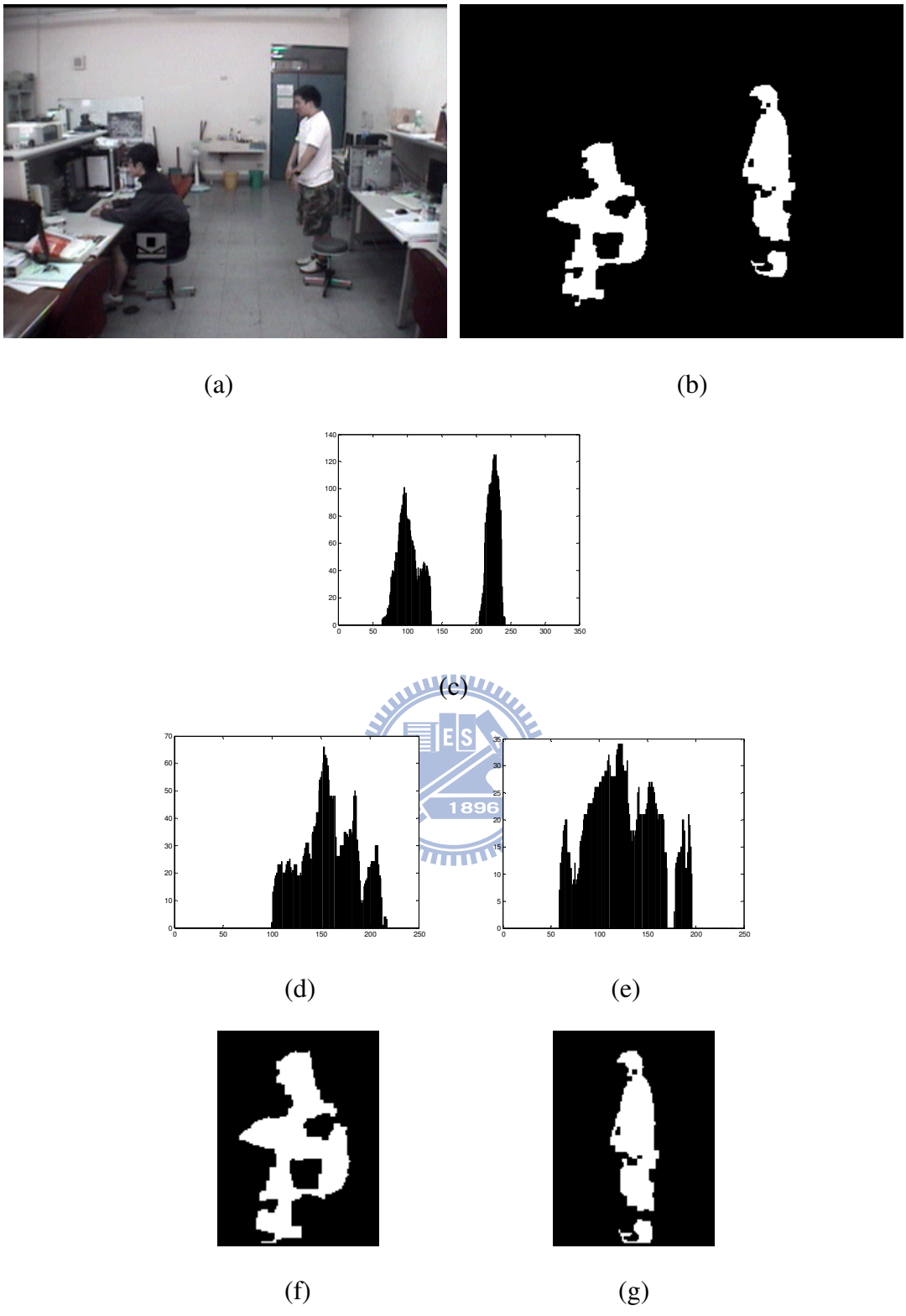


圖 4.4 (a)原始影像 (b)雜訊濾除與補償後結果 (c)統計 x 軸方向前景像素個數 (d)統計左側區域內 y 軸方向前景像素個數 (e)統計右側區域內 y 軸方向前景像素個數 (f)左側前景剪圖 (g)右側前景剪圖。

若有兩個人依序進入房間時，若要辨識進行動作是誰，我們根據衣服的顏色來辨識。當開始第一個人進入房間時，我們設定系統當此人進入室內約三秒後，建立衣服的色彩模型於 YCbCr 色彩空間中，可以在每次辨識動作後，接續辨識是誰進行此動作。若當第二個人進入房間時，此時影像中會出現兩個人，因此我們必須先辨識，何人是第一位進入房間後，再建立另一個人的衣服色彩模型於 YCbCr 色彩空間中，系統中就擁有兩個人的衣服色彩模型，可以辨識為何人進行何種動作。

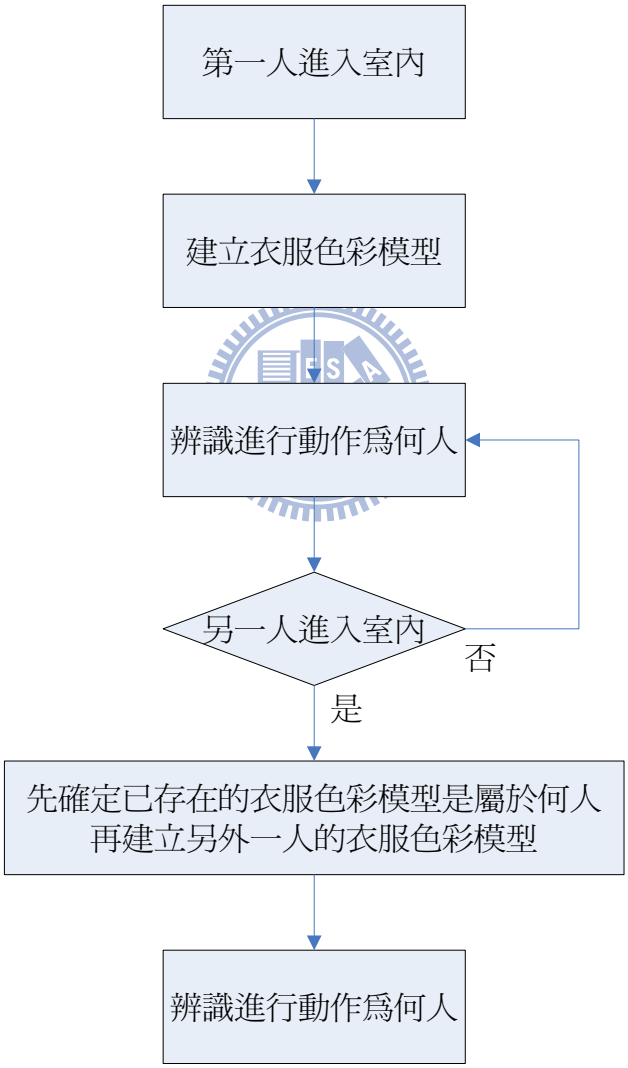


圖 4.5 衣服色彩模型建立與辨識流程圖

動作名稱	原始影像	前景剪裁後之圖形
左側位置讀書		
右側位置讀書		
彎腰		
揮手		
左側位置坐下		

動作名稱	原始影像	前景剪裁後之圖形
右側位置坐下		
平行鏡頭方向走		
垂直鏡頭方向走		
左側位置用電腦		
右側位置用電腦		

動作名稱	原始影像	前景剪裁後之圖形
左側位置起立		
右側位置起立		

圖 4.6 原始影像與前景剪裁後之圖形

依據我們在第三章敘述的選擇樣版影像的理論，在影像序列中所選擇的樣版影像必須間隔五張影像，其間隔時間約為 0.165 秒，例如取影像序列中的第 1 張影像、第 6 張影像與第 11 張影像。當開始進行即時影像擷取時，攝影機擷取到第 1 張影像，我們利用與稍後欲取的第 6 張影像之間隔時間來進行第一張影像的前景雜訊濾除與剪裁，完成時間約為 0.14 秒，小於 0.165 秒，因此在容許的範圍內。同理，攝影機擷取到第 6 張影像，我們利用與稍後欲取的第 11 張影像之間隔時間來進行第二張影像的前景雜訊濾除與剪裁。最後，擷取到第 11 張影像之後，即進行前景雜訊濾除與剪裁並且以三張動作影像，套用模糊法則來辨識出為何種動作，所花費時間為前景雜訊濾除與剪裁時間約為 0.14 秒和進行標準空間轉換(CST)後動作辨識時間約為 0.235 秒，所以第三步驟所花費時間約為 0.375 秒。因此，我們的動作辨識系統，完成單次動作的時間約為 0.705 秒。完成動作辨識後，攝影機繼續擷取欲下次動作辨識的第 1 張影像，繼續上述步驟。

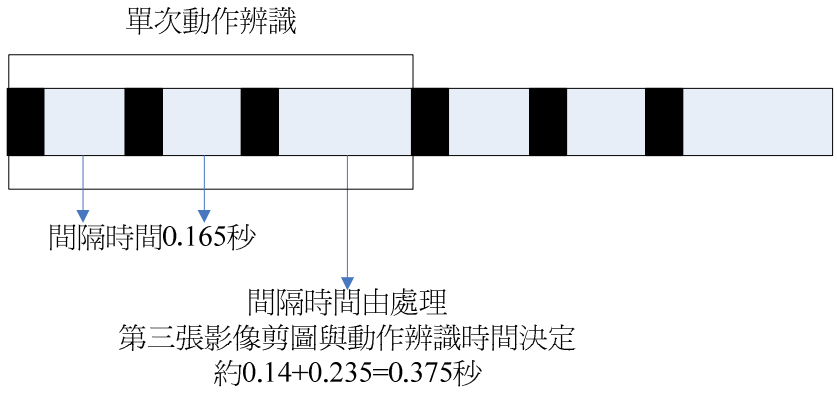


圖 4.7 單人動作辨識之處理時序

處理雙人影像動作辨識時，若同時處理同一張影像中，兩個人的各別動作時，前景影像的裁剪會耗時約兩倍處理單人動作的時間，在處理即時影像時，無法保持取得影像序列中，三張接序之 5：1 降低取樣頻率的影像。為了避免此問題，我們取得影像序列中三張接序之 5：1 降低取樣頻率的影像後，皆先處理辨識影像中左側人體的動作；之後，再另取得影像序列中三張接序之 5：1 降低取樣頻率的影像後，處理辨識影像中右側人體的動作。攝影機繼續擷取欲進行下次動作辨識的影像，重複上述步驟辨識影像中左側人體的動作。

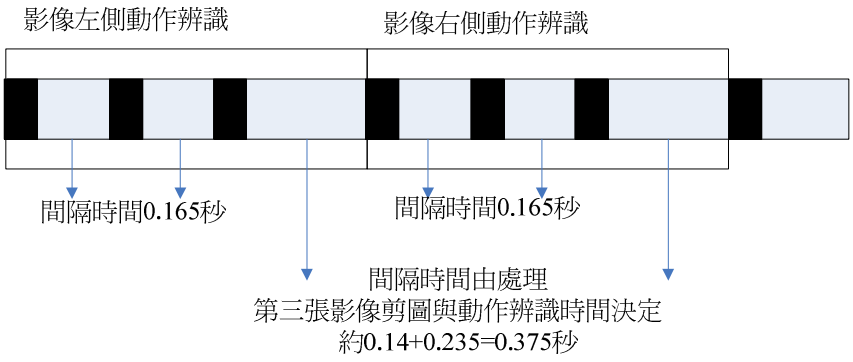


圖 4.8 雙人動作辨識之處理時序

完成動作辨識後，有些辨識的結果會出現錯誤。為了解決這個問題，我們在最後動作辨識結果輸出前加上多數決投票(majority vote)。若動作辨識結果已達五次以上，則可以在第五次的動作辨識輸出結果，與前四次的辨識結果，進行多數決投票決定何種動作為最後輸出結果。如圖 4.7 所示，我們採取的是滑動窗罩(sliding window)觀念，原始的輸出結果仍須保留，參與下次多數決投票。

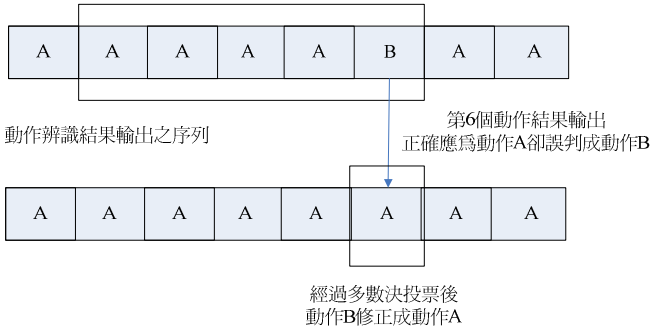


圖 4.9 多數決投票方式修正辨識結果

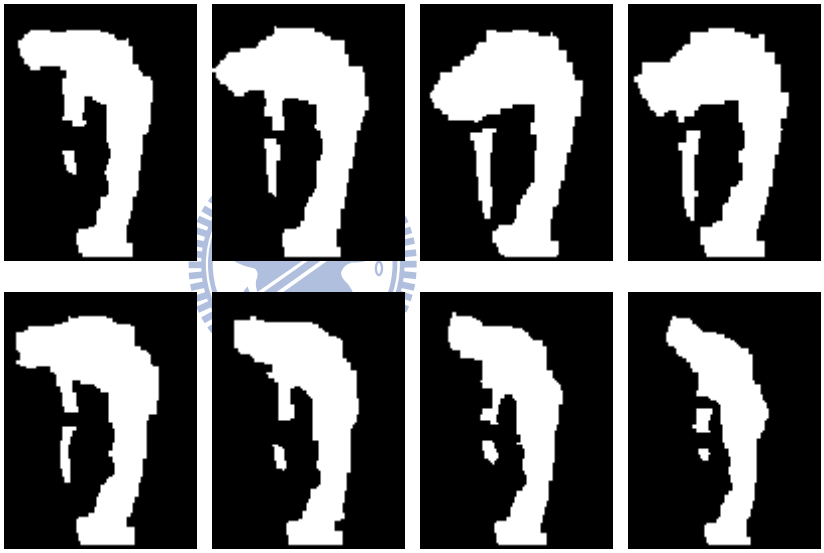
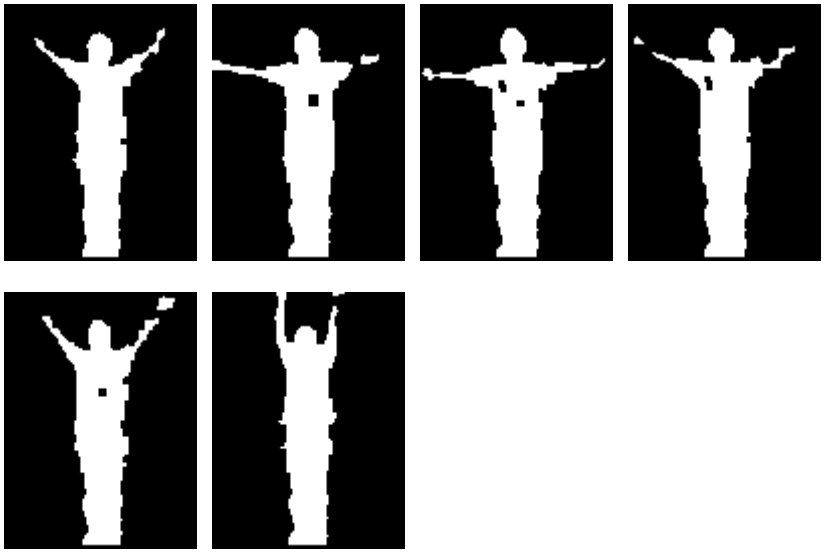
4.2 建立辨識動作之模糊法則

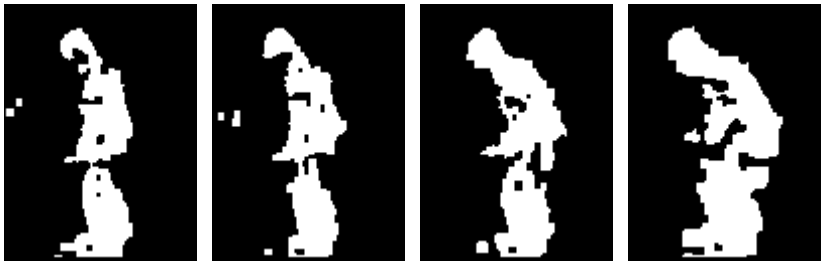
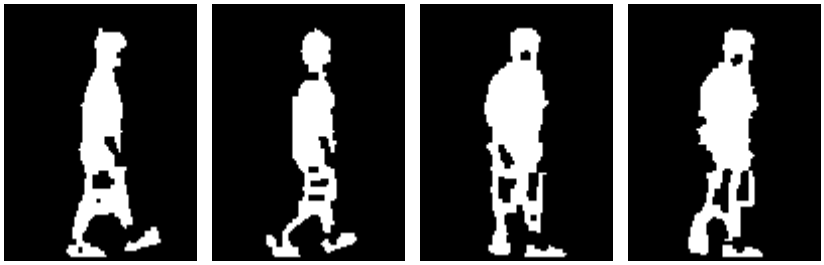
為了建立糊模法則以提供動作辨識使用，在每種動作的影像序列中我們必須挑選出具有代表性的動作影像作為樣版動作，用來學習訓練產生糊模法則。總共有 60 個樣板動作，下表為每種動作所挑選樣板動作的個數。

表 1 每種動作所挑選樣板動作的個數

動作姿態	樣版影像個數	動作姿態	樣版影像個數
右側位置讀書	1	左側位置讀書	1
右側位置使用電腦	1	左側位置使用電腦	1
右側位置起立	4	左側位置起立	4
右側位置坐下	4	左側位置坐下	4
進入房間	4	離開房間	4
對著鏡頭走近或走遠	5	對著鏡頭橫向行走	5
揮手	6	彎腰	8

選取樣版影像的個數，是以完成一個動作所需的時間為依據，例如讀書或是使用電腦的動作，因為在進行這兩種動作時，其姿態幾乎是固定沒有改變的，因此僅需幾張樣版影像代表即可；進行彎腰動作時，相較之下所花的時間就稍微長些，因此需要取 8 張樣版影像來作代表。以下列舉出幾個動作的樣版影像

動作姿態	訓練樣版影像
左側位置讀書	
右側位置讀書	
彎腰	
揮手	

動作姿態	訓練樣版影像
左側位置坐下	
右側位置坐下	
平行鏡頭方向走	 
垂直鏡頭方向走	 

動作姿態	訓練樣版影像
左側位置用電腦	
右側位置用電腦	
左側位置起立	
右側位置起立	

圖 4.10 樣版動作影像

這些欲進行學習訓練的樣版影像，需經過標準空間轉換(CST)。在我們系統中，有 4 個人進行上述的所有動作，因此總共有 4 個訓練者模型與 52 個樣板動作。也就是說總共有 208 個平均向量(mean vector)。

建立糊模法則是此系統的前置作業，無法在即時的情況下完成。我們必須先降低影像的取樣頻率 5：1，之後取三張已降低影像取樣頻率所取得的影像，組

成一集合。例如取第 1 張影像、第 6 張影像、第 11 張影像組合成一待輸入的集合後；取第 2 張影像、第 7 張影像、第 12 張影像組合成另一待輸入的集合；取第 3 張影像、第 8 張影像、第 13 張影像組合成另一待輸入的集合，以此類推。上述我們在同一動作中選取不同的起始影像是為了能夠將一動作，以更多的影像集合來代表，增進糊模法則的數量，能夠更精確的辨識動作。而且我們在測試辨識動作時，每個動作的進行有可能不是重頭開始。例如進行彎腰動作時，不一定是由挺直身體到彎曲身體，再到挺直身體。可能由微彎身體到彎曲身體，再到挺直身體。如此一來，使用上述的訓練方式，就可以克服此辨識問題。

4.3 動作辨識正確率

我們的動作辨識系統可以處理即時影像或是預錄的影片，但是為了統計辨識的正確率，我們僅能預錄一段已知動作影片，統計每一次的辨識輸出結果是否為該動作。動作辨識正確率表格建立方式為以測試者模型 2、測試者模型 3 與測試者模型 4 進行學習訓練後所產生的模糊法則，以測試者模型 1 進行動作辨識測試，其結果如動作辨識正確率表中第一欄所示。其餘欄位數據建立方式同理可得，其中數據表示方式為：辨識正確率(正確辨識輸出數/總共辨識輸出數)

表 2 動作辨識正確率

	測試者模型 1	測試者模型 2	測試者模型 3	測試者模型 4
左側位置讀書	100(64/64)	100(59/59)	100(53/53)	100(53/53)
右側位置讀書	82.7(48/58)	100(53/53)	87.2(41/47)	100(53/53)
彎腰	94.0(47/50)	92.1(57/62)	100(53/53)	100(59/59)
揮手	70.5(62/88)	92.2(71/77)	66.2(43/65)	90.1(64/71)
左側位置坐下	82.8(24/29)	65.2(15/23)	63.0(17/27)	75.9(22/29)
右側位置坐下	100(12/12)	73.3(11/15)	63.2(12/19)	100(13/13)
平行鏡頭方向行走	65.7(46/70)	81.7(58/71)	96.6(86/89)	84.4(65/77)
垂直鏡頭方向行走	98.4(61/62)	91.5(54/59)	100(77/77)	100(70/70)
離開房間	93.1(67/72)	93.4(57/61)	94.9(56/59)	90.2(37/41)
進入房間	97.2(74/76)	95.4(62/65)	93.1(54/58)	95.2(59/62)
左側位置使用電腦	100(47/47)	80.9(38/47)	93.2(55/59)	69.5(41/59)
右側位置使用電腦	63.4(26/41)	84.9(45/53)	100(59/59)	100(59/59)
左側位置起立	100(11/11)	100(15/15)	100(14/14)	100(17/17)
右側位置起立	100(13/13)	94.1(16/17)	100(15/15)	100(19/19)

我們統計上表的整體平均正確率有兩個方式：(1)直接取所有測試者模型與各種動作辨識正確率之平均值，我們稱之為整體平均正確 1，如方程式(33)所示；因為每段動作影片的長度不一，因此辨識輸出數目也不一。若在測試時某些預錄時間較長的影片，其辨識輸出數也較多，因此其辨識正確率會影響正確率，造成

統計不客觀。(2)所以我們另一個統計方式為，表中所有之正確辨識輸出數為分子，所有辨識輸出數為分母，來表示辨識正確率，我們稱之為整體平均正確率 2，如方程式(34)所示。

$$\text{整體平均正確率 1} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}}{m \times n} \tag{33}$$

其中 a_{ij} 代表第 i 個測試者模型第 j 個動作正確率，由(33)式可得，整體平均正確率 1 為 90.3%

$$\text{整體平均正確率 2} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} b_{ij}} \tag{34}$$

其中 c_{ij} 代表第 i 個測試者模型第 j 個動作正確辨識輸出數，b_{ij} 代表第 i 個測試者模型第 j 個動作所有辨識輸出數，由(34)式可得，整體平均正確率 2 為 90 . 4%

表 2 中的「進入房間」與「離開房間」兩動作，起初我們採用動作姿態訓練方式，產生模糊法則來提供辨識使用。雖有不錯的正確率，但是房間外的背景環境，我們無法掌握，例如在我們實驗環境中，多半呈現深暗黑色，若此時進入或離開房間的人身穿黑色或暗色系服裝，直接採用第三章所述的步驟，擷取出的前景極為破碎，甚至無法偵測出前景影像。為了解決這個問題，我們改採當有人進入房間時，前景影像中的前景像素增加；當有人離開房間時，前景影像中的前景像素減少的特性，來辨識「進入房間」與「離開房間」兩動作。使用此特性，我們可以精確的辨識是否有人進出房間。

但在監控環境中，若已經有一人在室內，另一個人進出時，則無法如同上述方式，由前景數目的增減來判斷是否有人進入或離開監控空間，因此我們利用上述 3.4 步驟，先由前景影像切圖之座標來判斷在監控空間中有幾個人存在，若監

控環境中，僅有一人存在時，前後張前後景分離影像之切圖座標會很相近，由此我們可以推論出此監控環境中僅有一人，如下圖 4.11 所示；若監控環境中，有兩個人同時存在時，若不考慮人物重疊的情況下，前後張前後景分離影像之切圖座標會有差距，如下圖 4.12 所示。藉由監控環境中人數的變化，進而推論出是否有人進出此監控空間。由下列判斷式，可以判斷監控環境中之人數。在系統中，我們設定 $threshold = 40$ 。

$$human\ count = \begin{cases} 1, \text{ if } |A-P|+|B-Q|+|C-R|+|D-S| < threshold \\ 2, \text{ if } |A-P|+|B-Q|+|C-R|+|D-S| > threshold \end{cases} \tag{35}$$

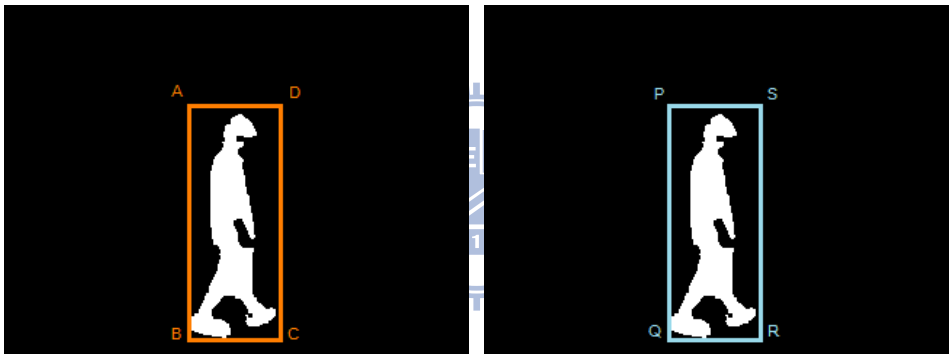


圖 4.11 單人前後景分離影像之切圖座標標示

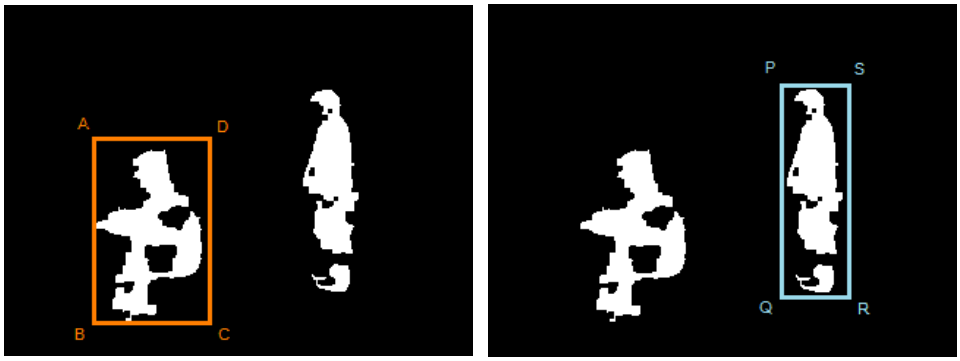


圖 4.12 雙人前後景分離影像之切圖座標標示

第五章 結論

我們在系統中建立雙背景模型，即包含灰階域與 HSV 色彩空間，在各模型中設定較嚴格的擷取前景參數，可以避免擷取出非人體部分的前景，如陰影，但也會有部分是人體前景的部分未擷取出來。但因為有雙背景模型，可以將兩個模型所擷取出來的前景部分作聯集處理，可以改善上述的缺點，得到較完整的前景圖形。EST 與 CST，降低資料維度加快處理速度且提升前景人體姿態的效果。藉由 5：1 降低的取樣頻率取出影像序列中的 3 個序列影像資料，作為樣版動作學習訓練後，建立動作辨識模糊法則。攝影機的三張降低抽樣頻率影像資料經過上述處理程序，依據所建立的動作辨識模糊法則，判斷為何種動作。

實驗結果顯示出，我們建立的十四種動作中，其判斷正確率為 90.3%。我們希望未來能夠使系統能在更加複雜的背景下正確工作，進一步，如果人的姿態動作不屬於模糊法則資料庫中的任一動作時，系統能夠自動將此新的動作姿態加以訓練學習，產生新的模糊法則加入到資料庫中。

參考資料

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

[2] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4 : Real-Time Surveillance of People and Their Activities," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809－830, 2000.

[3] 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.

[4] R. Cucchiara, C. Grana, M. Piccardi and A. Prati, "Improving Shadow Suppression in Moving Object Detection with HSV Color Information," in *Proc. IEEE Intelligence Transportation System Conference*, pp. 334－339, 2001.

[5] 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

[6] A. Prati, I. Mike, M. Trivedi and R. Cucchiara, "Detection Moving Shadow: Algorithms and Evaluation," in *Proc. IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 918－923, 2003.

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

[8] H. Saito, A. Watanabe, and S. Ozawa, "Face pose estimating system based on

eigenspace analysis," in *Proc. Int. Conf. Image Processing*, vol 1, pp. 638 – 642, 1999.

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

[10] M. M. Rahman and S. Ishikawa, "Robust appearance-based human action recognition," in *Proc. The 17th Int. Conf. Pattern Recog.*, Vol. 3, pp. 165 – 168, 2004.

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

[12] 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.

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

[14] 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.

[15] K. Ohba, Y. Sato, and K. Ikeuchi, "Appearance-based visual learning and object recognition with illumination invariance," *Machine Vision and Application*, vol. 12, no. 4, pp. 189 – 196, 2000.

[16] Soriano M, Huovinen S, Martinkauppi B, Laaksonen M. "Using the skin locus to cope with changing illumination conditions in color-based face tracking," in *IEEE Nordic Signal Processing Symposium, kolmarden, Sweden*, pp. 383 – 6,

2000.

[17] 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.

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