

# 國立交通大學

## 多媒體工程研究所

### 碩 士 論 文

根據姿勢與外貌整合的影像人臉註記

Clustering People in Video by Combining Appearance and Pose  
Information

研 究 生：蘇裕傑

指導教授：王才沛 教授

中 華 民 國 一 百 年 八 月

根據姿勢與外貌整合的影像人臉註記  
**Clustering People in Video by Combining Appearance and Pose  
Information**

研究生：蘇裕傑

Student : Yu-Chieh Su

指導教授：王才沛

Advisor : Tsai-pei Wang



Computer Science

August 2011

Hsinchu, Taiwan, Republic of China

中華民國一百年八月

# 整合影像資訊以改良影像人臉註記的準確性

學生：蘇裕傑

指導教授：王才沛

國立交通大學多媒體工程研究所 碩士班

## 摘 要

在本篇論文中，所展示的是一個完整的系統流程，以論文名稱直接解釋之，即為將影片中的重要腳色所出現的時間軸位置紀錄，不但可以讓使用者了解腳色出現的時間點，也可供使用個直接索引所想要的腳色。本流程從分鏡切割，人臉偵測、追蹤，人臉特徵空間的投影，最後展示各種不同參數下的分群結果。追蹤人臉時僅用前後影格人臉的座標及大小作為製作人臉串列的依據，搭配上分鏡變化做切割，得到正確的人臉串列。在人臉特徵向量選取區塊，我不只僅比較不同方法如：PCA, 2DPCA，還加入了人臉姿勢的相似值，作為此兩臉之間相似的可信賴度。在本文實驗中取了固定的值作為信賴度所用的閾值。最後展示了在不同參數及分群方式下的實驗結果。

## 誌 謝

今天能完成這篇論文，除了要感謝王教授以外，也要感謝上一屆的學長與同學、學弟們。教授不但細心且有耐心的指導我們，對於我們錯誤的地方也不厭其煩的一步步引導我們。在剛進碩一時，對於什麼都不懂的我們，傾囊相授。同學間在課堂上也會互相扶持，面對難題時，也一同研究、解決。就算是不同科目的問題也會盡力而為，尤其是我們在大學部時所學的東西不甚相同，遇到個別的強項，也不會有所保留，互相討論。在作此篇論文的實驗時，遇到的瓶頸也常常是在大家一起討論時找到突破。也因此沒有以上各位，也就無法完成此篇論文。在共同進行競賽時更能發揮各別的專長，齊力完成作品，也是一個非常寶貴的經驗，對於日後研究的方向也很有幫助。平時閒暇之餘，也能再球場上較勁。在此，除了感謝大家在課業上的幫助外，也要感謝大家樂觀開朗的態度，使得實驗室中充滿著愉悅的氣氛，並且保持的上進的心。



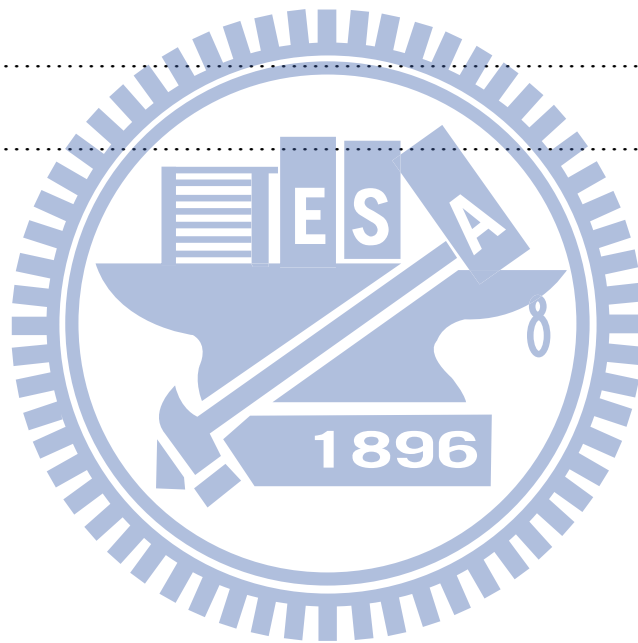
# 目 錄

摘要	.....	III
誌謝	.....	IV
目錄	.....	V
表目錄	.....	VII
圖目錄	.....	VIII
一、	簡介 .....	1
1.1	研究動機 .....	1
1.2	章節概要 .....	2
二、	文獻探討 .....	3
2.1	人臉偵測 .....	3
2.2	人臉辨識 .....	3
2.3	人臉校正 .....	3
2.4	人臉分群 .....	4
三、	實驗方法 .....	5
3.1	前置作業 .....	6
3.1.1	畫面擷取 .....	6
3.1.2	人臉偵測 .....	6
3.1.3	膚色偵測 .....	6
3.1.4	影像前處理 .....	7

3.2	演員串列建立 .....	9
3.2.1	分鏡偵測 .....	10
3.2.2	人臉追蹤 .....	12
3.3	結合身體與臉部相似資訊 .....	15
3.3.1	臉部相似度 .....	15
3.3.2	姿勢相似值計算 .....	16
3.3.3	身體相似度 .....	20
3.3.4	身體與臉部資訊合併 .....	21
3.3.5	臉部與姿勢資訊整合 .....	22
3.4	演員串列分群 .....	23
3.4.1	分群方法 .....	23
3.4.2	分群限制 .....	25
四、	實驗結果 .....	26
4.1	評估工具 .....	26
4.1.1	Adjusted RAND index (ARI) .....	26
4.1.2	Classification via Clustering (CVC) .....	27
4.2	測試資料介紹 .....	27
4.3	不同條件下的實驗比較 .....	28
4.3.1	各種環境變數設計 .....	39
4.3.2	臉部與身體資訊結合參數比較 .....	32
五、	結論與未來展望 .....	39
	參考文獻 .....	41

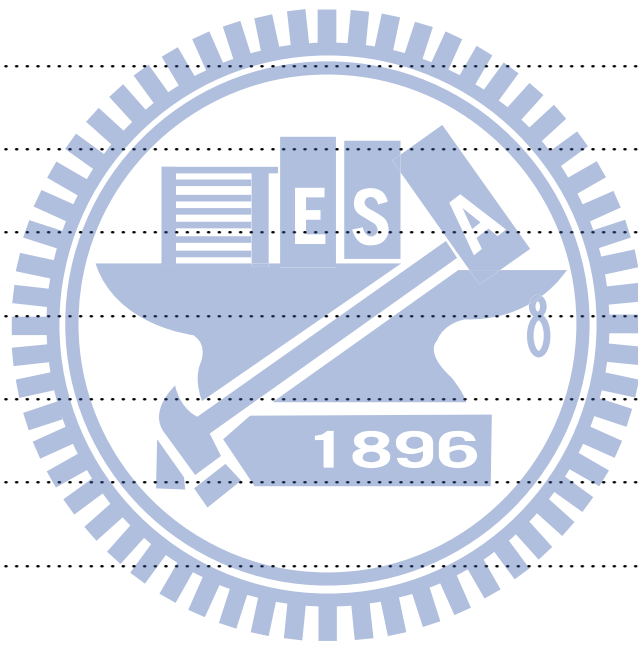
# 表 目 錄

表 3-1	.....	13
表 4-1	.....	29
表 4-2	.....	30
表 4-3	.....	30
表 4-4	.....	31
表 4-5	.....	33
表 4-6	.....	35
表 4-7	.....	37



# 圖 目 錄

圖 3-1	.....	5
圖 3-2	.....	7
圖 3-3	.....	8
圖 3-4	.....	9
圖 3-5	.....	11
圖 3-6	.....	14
圖 3-7	.....	14
圖 3-8	.....	14
圖 3-9	.....	17
圖 3-10	.....	18
圖 4-1	.....	27
圖 4-2	.....	28
圖 4-3	.....	34
圖 4-4	.....	36
圖 4-5	.....	38





# 第一章 簡介

## 1.1 研究動機

從三十多年前開始，人臉辨識一直是人們持續探討的問題。無論是從心理學的觀點，或是圖學的角度，研究學者們對於人類如何辨識不同人臉一直抱持著很大的興趣。但就算是現今，依舊沒有一個擁有完美的辨識率的方法被提出，應該說，這議題是逐漸改良，與我們學術上處理問題一樣，可以說遇到了瓶頸，難以有突破性發展。而現今資訊爆炸的時代，多媒體儲存格式重心從圖片轉移到影像。在影像中就能提供更多的資訊以提供作為人臉辨識的參考依據。最近十年來，由於資訊爆炸及科技的進步，不但影片的數量龐大，硬體及軟體的功能也大幅提升，進而人們對於資訊的使用也希望能更加方便、自動化。因此針對影像內容分析以及分類的需求也就萌生，有些學者研究如何從影片中得到額外的訊息，如聲音或文字稿等。有些學者著重於加強人臉辨識的技術，並且應用在現有的整合技術上。而前者涵蓋了一些不同領域的技術，技術門檻較高，後者則大多是稍稍提升了人臉辨識的準確度。

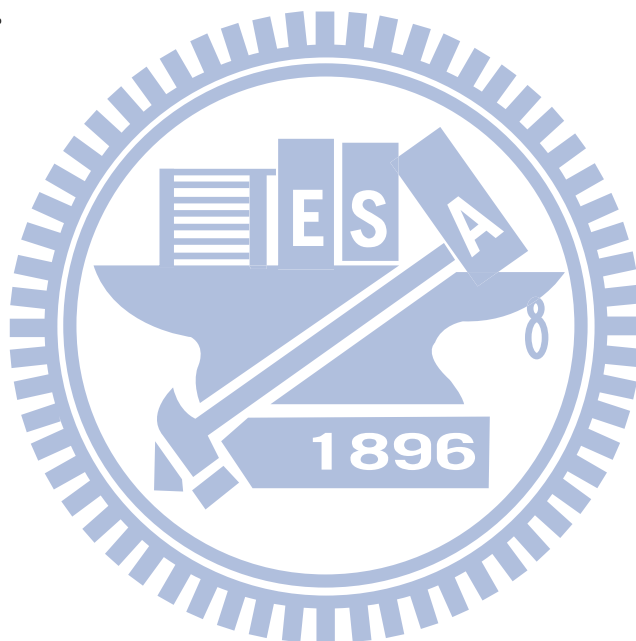
本論文是採用後者，希望以提升辨識正確率的人臉辨識合併至前者現有的整合技術。如今已經有許多學者的文章顯示，透過校正人臉可以大幅度的提高人臉辨識的結果，但一般的學者朝向正規化的校正方式，如對每一張人臉都經過眼睛、嘴巴偵測來完成旋轉角度與大小的校正，甚至使用 3D 模型貼圖在進行辨識比對。此方法的缺點是需要大量計算，且無法預測錯誤偵測對後續辨識結果的影響。而有另一群學者，希望利用影片中人臉可能包含多種角度姿勢的特性來達到類似校正的效果，也就是在影片中同一個人極有可能有多種角度，在做人臉比對辨識時便可選取姿勢較為接近的做比較。

本論文的概念就是源至此，因此我尋找著能夠以最簡單方式分辨此兩張臉的角度相似關係。最後我採取了 Gabor 的紋理擷取方式，有許多文章已經提出在此紋理提取後，可以透過 PCA 降低維度，得到姿勢的分佈空間。因此我根據此方法在人臉中試圖找出可以評量兩張臉姿勢的計算公式。接著透過此姿勢相似值整合前面提到的影片中額外資訊，如軀體，整合與臉部相似的關係，以提高最後分群的正確性。最後可以得到標記完成的影片，可以做到將所有角色出現的影片時間點列出，可提供使用者尋找特定的人物或是特定的互動分鏡，或是秀出特定人物出現的所有時間點，也就是俗稱的“recall”。在

本篇論文中，我研究的重點是如何去提升人臉辨識，進而提升系統整體的辨識率。於是我希望能以較簡易的方式，取得臉與臉之間不同角度的相似資訊，透過處於相似姿勢下的人臉才拿來做比較的簡單演算法，來試圖達到前面複雜演算法所達到的結果。

## 1.2 章節概要

在第二章會依序介紹在各個階段的方法以及其改良，與其他文章中所使用的方法做比較，並且在本實驗中試圖找出最佳的方法。第三章中則開始進行實驗的內容的介紹，從影像的切割，演員串列的建立，及影像前處理的方法比較，最後整合人臉以及身體的資訊來做分群。在文章最後會將所有實驗中不同方法比較以表格方式秀出，方便讀者了解各方法的優劣。



## 第二章 文獻探討

### 2.1 人臉偵測

在早期的人臉偵測，大部分是使用膚色作為判斷依據，Czirjek[1]提出的方法也是先以膚色偵測為基礎，再加上面積計算，旋轉等方法來修正。[2][3]都是屬於再影片中作人臉辨識分群的文章，因是接近近代的文章，在硬體及軟體都能跟得上的條件下，他們使用了需要大量計算的臉部器官偵測。一般而言是以眼睛、鼻子及嘴巴為主，但通常偵測臉部器官除了驗證此區域是否為人臉外，大多使用器官之間的距離與角度來做為校正臉部的依據。而近年來較為熱門的人臉偵測方式為 Haar-like[4]的特偵選取，透過計算人臉特殊的陰影來進行人臉偵測，雖然誤判的機會不小，但速度快，因此也廣泛被使用。

### 2.2 人臉辨識

人臉辨識長久以來就是一直被研究的議題，除了最常見的Eigenfaces[5]以外，[6][7][8][9]之中也使用FLD (Fisher's Linear Discriminant)、LDA (Linear Discriminant Analysis)、SVD (Singular Value Decomposition)，並針對各種2DPCA(2-Dimensional Principle Component Analysis)[10]變化進行實驗結果比較。而Ahonen[11]的LBP (Local Binary Patterns)除了在性別辨識上有很好的效果外，人臉辨識之中也常常使用此方法作為紋理的擷取。

### 2.3 人臉校正

在進行人臉辨識之前，有無對影像作校正會有非常明顯的差距，因此是否要對人臉的位置及角度進行正規化，已經有需多文章證明了校正的優異性。如[12]是一個比較標準的新聞主播辨識與分類，其中使用了計算最正接近正臉的人臉影像以提高辨識結果，顯示了正面的臉含有較多的資訊。在[13]中使用 affine invariant 來避免臉部姿勢角度對辨識結果的影響，但[14][15]更進一步利用偵測到的五官位置來模擬一個粗糙的 3D 模型，並且在進行人臉辨識比對時，以貼圖的方式將人臉圖像貼在模擬出來的，此兩篇所實驗結果所得到的數據非常優秀，但技術需求及計算量非常高。較為一般常見的方法是事先做人臉器官偵測，如[16][17][18][19][20]的實驗過程都需要事先偵測五官的位置，再利用這些資訊作臉部的校正工作。但利用臉部器官的方法都有同樣的問題，就是無法確保

五官偵測的穩定性是否對後續有重大的影響，並且也需要大量的計算。直到最近幾年來，有人提出了利用紋理擷取的方法來實現臉的方向判別，[21]之中提出 PES(Pose Eigen Space)的技巧，利用 Gabor Wavelet Transform 對人臉進行紋理擷取後，再投影至 3 維度空間，即可將不同角度的人臉依照左側到右側順序在此三度空間分布。[22]也是採用此方式來做為校正的依據。

## 2.4 人臉分群

人臉分群也就是使用人臉資訊作為分群的依據，但由於人臉的變化可能會因角度、光源或是時間而有所變化，因此往往在進行分群時會搭配額外的資訊，如在[23][24][25]中除了使用影像之外，還加上了不同的資訊來提高辨識結果，像是聲音或是文字等額外資訊，利用演員的嘴型是否有變化來判斷此時刻的聲音是屬於哪位演員，在進而利用這些相互關係來對人臉影像做分群，但由於牽涉過多技術，以至於效果有限，像是聲音變的分辨程度，影音是否同步，或是演員嘴型變化是否有正確被偵測出都是環節中問題的一部份。

[26]則是使用了身體的色彩資訊，除了使用不同的人臉辨識的技術外，希望能夠整合影片中演員的身上衣物特徵，通常是使用顏色做為比較一句，利用不同的權重比例來整合臉部與身體的顏色特徵。此方法效率很高，因一般影片中，演員不會在短時間內多次變換身上的衣物，因此身體的色彩資訊若是比重調整得宜，便可以得到大幅度的正確率提升，反之，此參數也必須視影像的類型來決定比重參數，故若是需要時做全自動的系統，則必須有最佳化參數的方法。而在[18]之中使用了[21]所提出的姿勢分辨方式來將人臉分為不同角度種類，進而在將群結合時使用同一類型的臉做為比較依據，並且在結合兩個人臉的群組時，加上群與群之間的額外限制，會令兩非常不相像的子人臉群不因為連結其他人臉的群而互相結合為一群。其中主要的方法也是利用 PES 來將人臉大致上分為三個主要角度，左側臉、正面臉及右側臉。在上一節提到有使用人臉器官偵測的文章中，也有部分文章僅利用這些臉部器官的位址將人臉的角度分類，而不是直接將圖像做位移、旋轉角度或是變形，此類方式的好處是不破壞原圖的幾何形狀，利用影片中包含大量不同角度人臉的特性來達到提高分群效率的目的。

### 第三章 實驗方法

本章節首先會將整個實驗流程以圖表方式秀出，其中會將各階段的方法及細節在小節中逐一介紹。圖 3-1 即為本實驗流程圖，從一開始的原始影片輸入，到最後輸出的標記位置。

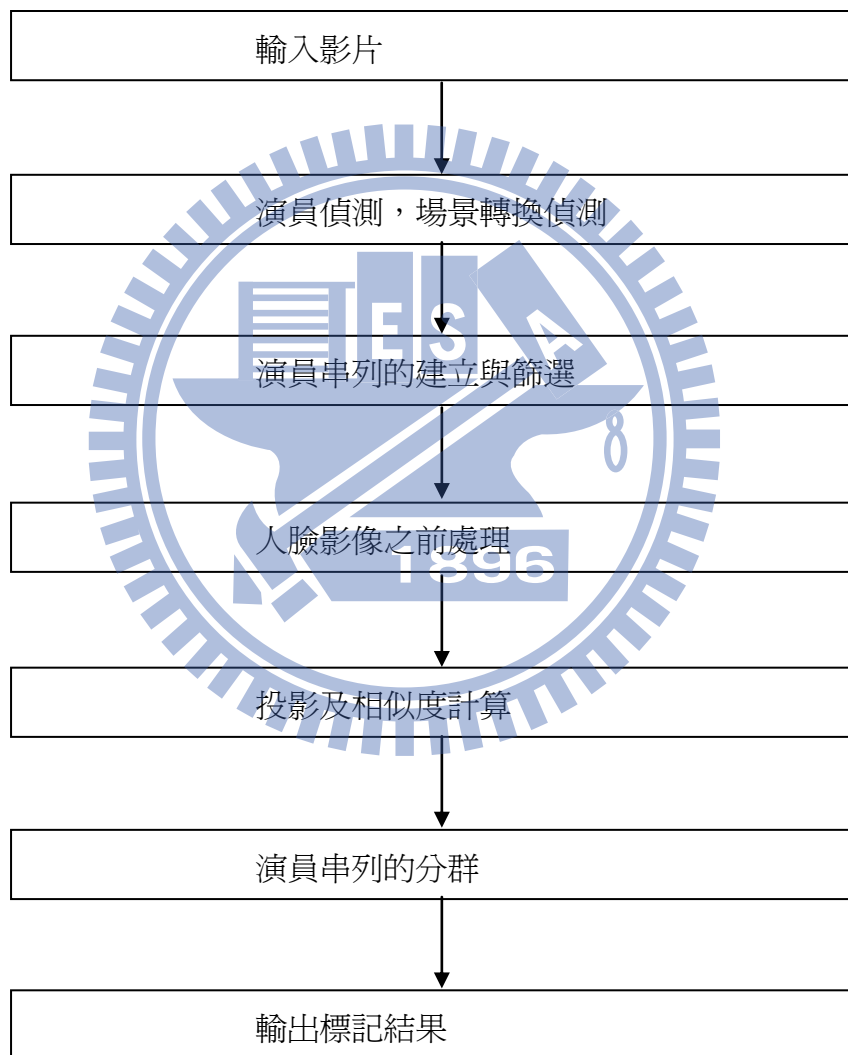


圖 3-1：實驗流程步驟流程圖。



### 3.1 前置作業

在開始進行一連串的實驗步驟之前，需要先做四個前處理步驟，從一開始的將影片剪裁成數張影像，接著進行人臉偵測，其中在抓取到可能的人臉區域時，會直接進行膚色偵測，將膚色區域過少的候選區塊篩選掉。而分鏡偵測是透過分鏡劇烈的顏色變化來找到可能發生的分鏡切換，在下個小節的演員串列建立即會用來切割正確的串列。

#### 3.1.1 畫面擷取

這裡我們利用 DVDVideoSoft.com 所提供的免費軟體“Free Video To JPG Converter” [27]來使切割影片更為便捷。也因為在進行各種實驗中需要不斷重複存取影像片段，因此在實驗過程中我們事先將所有切割後的影像存檔起來，以加速實驗流程。在實驗中，我們採取每秒擷取 5 張影像的頻率來進行影像切割，此數值相較於一般影片的每秒 30 張，可達到減少資料量，但不值於影響整體影片的流暢，經測試後此數據在本實驗可容許範圍之內。

#### 3.1.2 人臉偵測

在前面已提過的各種人臉偵測，而近年來較受學者們青睞的一個方法是“Harr-like feature”也就是本實驗中所使用的方法。Intel Corporation [28] 所開發之 OpenCV(Open Source Computer Vision) Library 人臉偵測所用的方法即是 Harr-like 的方法。本實驗中所著重的地方並非在人臉偵測，因此使用此函式庫來節省實驗的時間，不再自行收集和分析人臉資料。

#### 3.1.3 膚色偵測

在使用 OpenCV 時，由於函式中所使用的人臉特徵並不包含顏色，進行人臉偵測時往往會有些誤判，將不屬於人臉的區塊判定為人臉。為了篩選掉這些不屬於人臉的錯誤偵測區塊，我使用了[29]所用的膚色偵測。

此方法中所提出的公式中忽略了亮度的 Y 值，僅使用 Cb 及 Cr 作為判斷的依據。公式(1)為此膚色偵測的判定公式， $\bar{x}$  向量代表受測像素的 Cb 以及 Cr 數值，而 Z 及  $\sum$  為將人類膚色統計資料分析後所得的固定參數，此參數主要是針對不同人種膚

色的分佈範圍作些許調整。而下列公式  $P$  值為計算後之膚色之數據，此數值越高代表越接近膚色，介於 0 到 1 之間。此處的  $Z$  與  $\Sigma$  參數是以白種人作為預設值，若需要不同的人種參數可參閱[29]所提供的數據。圖 3-2 為膚色偵測的部位，以藍色表示之。

$$P = \exp\left(-\frac{1}{2}(\bar{x} - Z)^T \Sigma^{-1} (\bar{x} - Z)\right) \quad (1)$$

$$Z = \begin{bmatrix} 113.17 \\ 149.03 \end{bmatrix}; \Sigma = \begin{bmatrix} 44.98 & -31.01 \\ -31.01 & 46.60 \end{bmatrix}$$

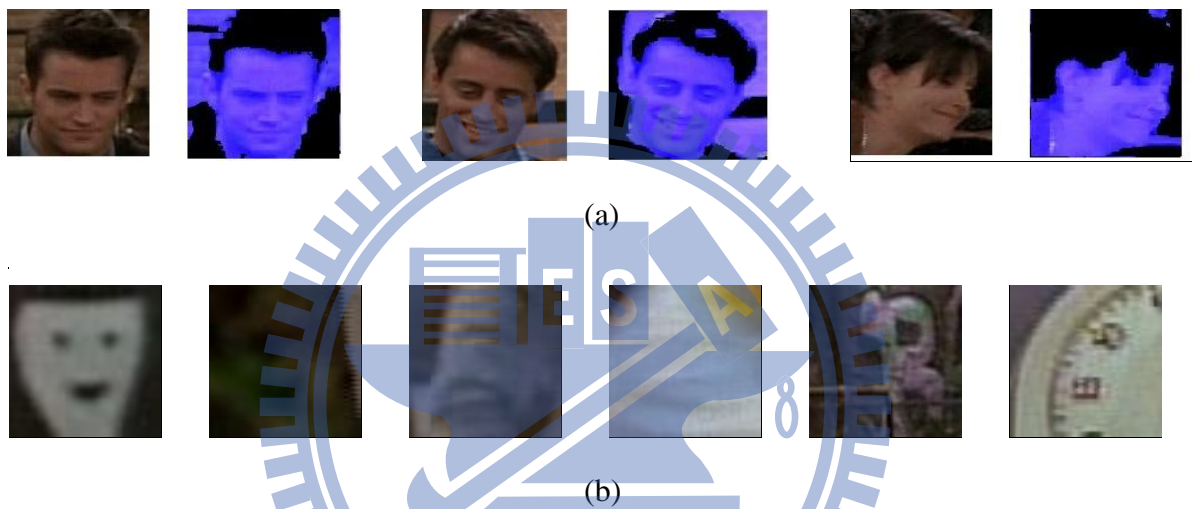


圖 3-2:用膚色偵測來篩選非人臉的區塊，(a) 在正常人臉情況下的膚色偵測，可以順利的通過檢測。淺藍色表示膚色偵測所抓取的區域。(b) 透過膚色偵測正確的將非人臉的誤判刪除。

### 3.1.4 影像前處理

在做影像處理之前，由於影像有許多不定因素如不同光線強度、不同角度光源、鏡頭對焦距離、抑或是攝影器材的差異都會造成影像處理時莫大的影響。因此不論對於何種實驗，對於資料的前處理都希望能將資料的條件一致化。在我實驗中，亦會將所有圖片做前置處理，之後在計算基底的維度、相似值時便可減少這些不定因子對實驗結果的影響。圖 3-3 展示了流程的分區部分，顯示了無論在做何種特徵選取或是投影，都會將影像前處理擺在第一步驟，然而針對我實驗內容會細分為兩部分，一是人臉辨識，另外一部分則是姿勢的相似度辨識。在以下的章節會進一步介紹。

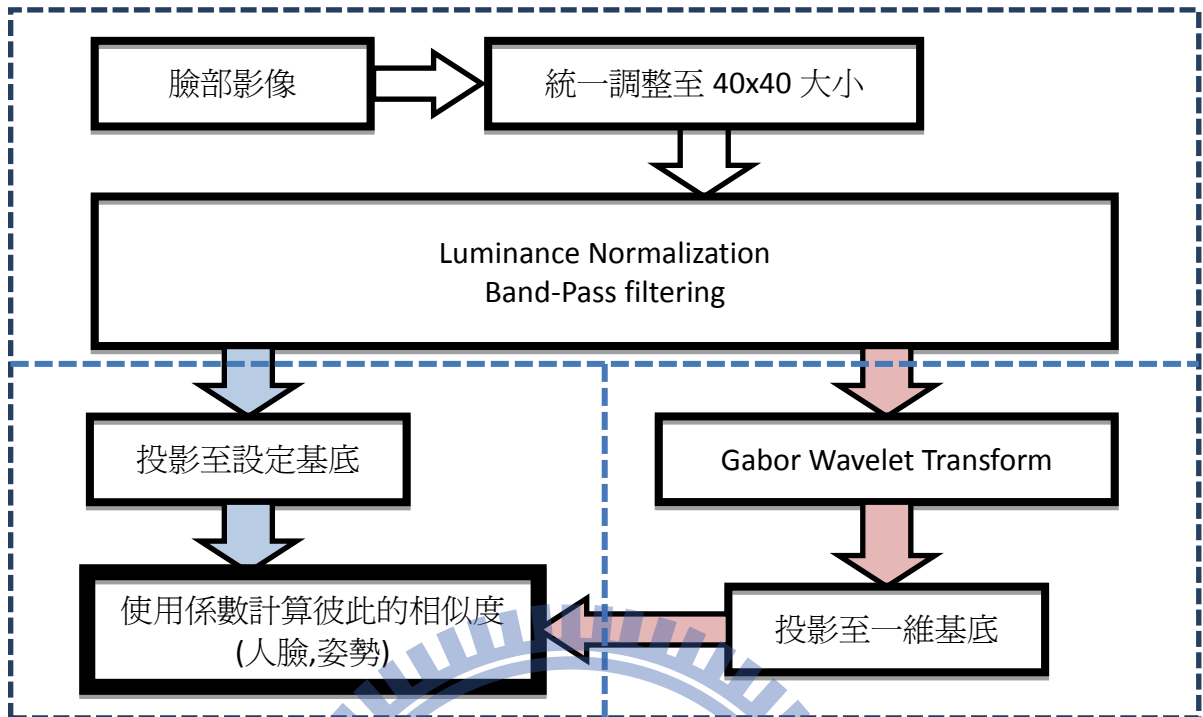


圖 3-3 前處理與辨識的運作流程圖。

在前處理的部分，可以簡單分為兩階段，一是光影的平衡，二是雜訊的過濾。一般而言會先處理光線的問題，再將處理完的影像作雜訊過濾的處理。在此我也照此順序來介紹。

光影平衡 (Luminance Normalization)，其實有許多不同的做法，如白平衡 (White Balance) 也是一種對光源變化的校正。在本實驗中所使用的方法與[32]相同，分別對於影像的三個圖層進行同樣的處理動作，首先計算此圖層的平均值與標準差，在使用公式(2)代入參數，將所有影像的標準差調整至相同。目的是將影像三個圖層的亮度調整相同。其中公式(2)中  $\delta_0$  及  $\mu_0$  分別代表所有影像欲調整的目標，而  $\delta$  及  $\mu$  分別為被調整影像的標準差以及平均值。

$$x \rightarrow (x - \mu) \cdot \frac{\delta_0}{\delta} + \mu_0 \quad (2)$$

調整完成後接下來利用 Band-Pass Filter 頻率濾波器來過濾掉影像中低頻及高頻的部分，低頻的部分包含不變的背景或是過於平滑的表面，而高頻的部分包含了雜訊以及影像交界處。實際操作的方法是利用兩組反向的二維高斯函式組合成頻帶濾波器 (Band-Pass Filter)。本實驗嘗試不同的高斯函式之標準差，來實驗不同頻帶影響影像



的結果。圖 3-4 顯示了幾組不同的參數，在經過觀察後，影響並不大，因此本實驗採取了  $d_0=10$  ( 反向 ) ，  $d_1=20$  ( 正向 ) 的參數。

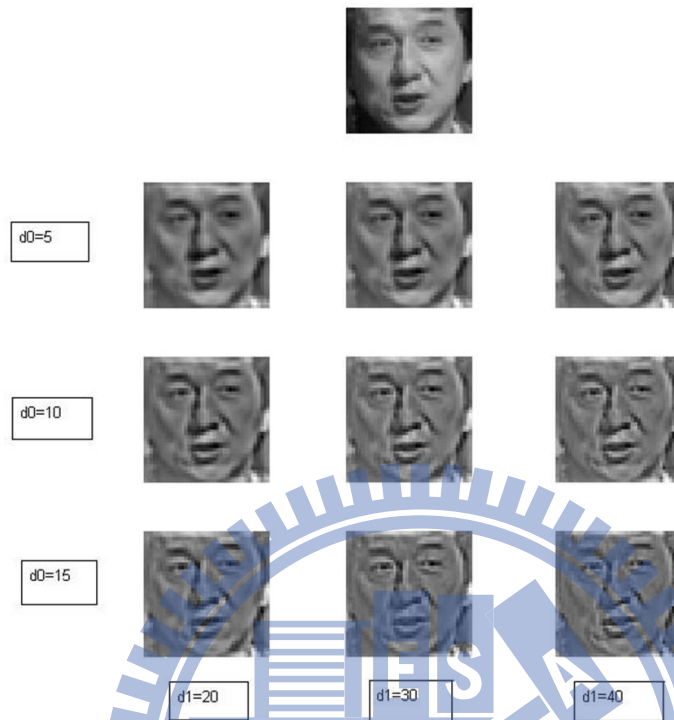


圖 3-4 同一張臉在不同的  $d_0$  與  $d_1$  之間作頻率域濾波的結果。

在作頻帶濾波時，不同的邊界影響程度的大小在允許範圍內並不會有太大落差， $d_0$  越大表示所要過濾的低頻範圍越大， $d_1$  越小表示要過濾的高頻範圍越大，可從圖中看到細節。由於若頻寬帶選取過小，可能會遺失過多的資訊，因此根據圖像中選取較不失真但可將雜訊或無意義資訊過濾的選擇。

### 3.2 演員串列建立

影片與影像最大的不同點就在與時間的連續性，在影片中，時間是非常重要的資訊之一，若能善用分鏡切換偵測，即能將每一幕分鏡內角色視為一個單位，也就是演員串列。製作演員串列的優點很多，除了能透過時間與座標偏移量來建立非常精確的演員串列外，得到的演員串列中可能包含了不同的背景、光線、姿勢等等。這些包含在演員串列中的資訊都是在做分群比對時重要的資訊，也因此在做影像分析處理之前，會先建立

所有的演員串列。

### 3.2.1 分鏡偵測

在較早的一些研究論文，研究者們是以較簡單的戲劇影片做為研究用的影像資料，如[30]是它會以分鏡變化來判斷視角，但僅限於戲劇的影像。[31]中 Zhong 等人提出的色彩直方圖，透過計算兩張圖像之間的直方圖差異來判斷此兩張圖象是否屬於相同分鏡。此方法簡單但有效率，利用此方法可以簡單的找出每張分鏡是否有變化，這對作影像的人臉辨識非常有幫助，可以事先裁斷不同分鏡的連續出現人臉。再進行影片中人臉分群時，切割影片以建立影片段落的人臉串列是非常有效的方法。因此偵測分鏡變換可以用來切割演員的連續影格。

在作影像中人臉偵測時，有許多額外資訊可以用來幫助我們進行更正確的辨識處理，如在進行人臉串列建立時，人臉的串列最終會在分鏡切換時被切割。意思是，任何一個演員串列都不會跨越分鏡的切換。因此在建立演員串列之前，我們會先偵測分鏡的切換時機點，以便於用來將不同的人臉因座標相近誤判的情況給正確切割。

分鏡偵測中最常使用也最簡單的方法即是計算彩色直方圖的差距(color histogram)，然後設定一個直方圖數據改變量的閾值[31]，若是兩影格的分鏡相異程度超過此數值，則此兩影格即判斷有發生分鏡。這是一種直覺的方法，目的是希望能從分鏡轉換時大範圍的景象改變中獲取分鏡改變的資訊。優點是簡單且快速，缺點是沒辦法正確的偵測出在相似背景下的分鏡轉換。但在我們的實驗中，此方法已足夠使用。

而彩色直方圖的求取方式為：設定欲將此影像像素切割成多少個區間，落在同一區間的數值都將被視為同樣的等級，而在個別的影像層中都做相同的動作。最後再將三個影像層(此處以 RGB 作範例)所統計出各等級個數合併成彩色直方圖，即成為簡易彩色直方圖。

在我的實驗中，由於希望能盡量將所有的分鏡轉換都找出來，除了實驗了不同的色彩空間，及不同大小的區間分段。最後選取了數值偏小的閾值以達到能正確切割後續演員串列的目的。然而取較小閾值的缺點是後續的演員串列將會被切割成較短的片段，經過後續的刪減處理後可能會將此類過短的人臉片段刪除。在實際測試後，雖然不可避免的會有不少人臉串列被刪除，但對於演員在影片時間軸上的分佈並沒有太大的影響，意思即主要演員所涵蓋的範圍並不會因為刪除片段而有太大影響。

圖 3-5 為簡易彩色直方圖的示意圖，表 3-1 為以 Jaccard 直方圖相似比(介於 0 到 1，1 表示完全相同)作為製作演員串列時所用的分鏡切換偵測。Jaccard 相似度是將兩個影像之直方圖的距離放在分子，而取較大的放在分母所計算出的簡略相似值。此表格所使用的影片為“firends”，而演員串列建立的演算法在後面章節會詳加介紹，在演員串列建立後，最後會將影格數目小於等於三個的串列刪除，因此類串列大多數是在人臉偵測時留下的誤判物體，並非正確人臉，而本實驗依據不同色彩空間以及分區塊大小作了 ROC Curve，並且決定在色彩空間 YCbCr 中，可以較低 False negative 情況下取得較大的 True positive。而區塊大小對於 32 和 16 之前並沒有太大的差異，因此在後續實驗我採用區塊切割大小為 16 作為固定參數，而在分鏡偵測所用的色彩空間則一致使用 YCbCr。

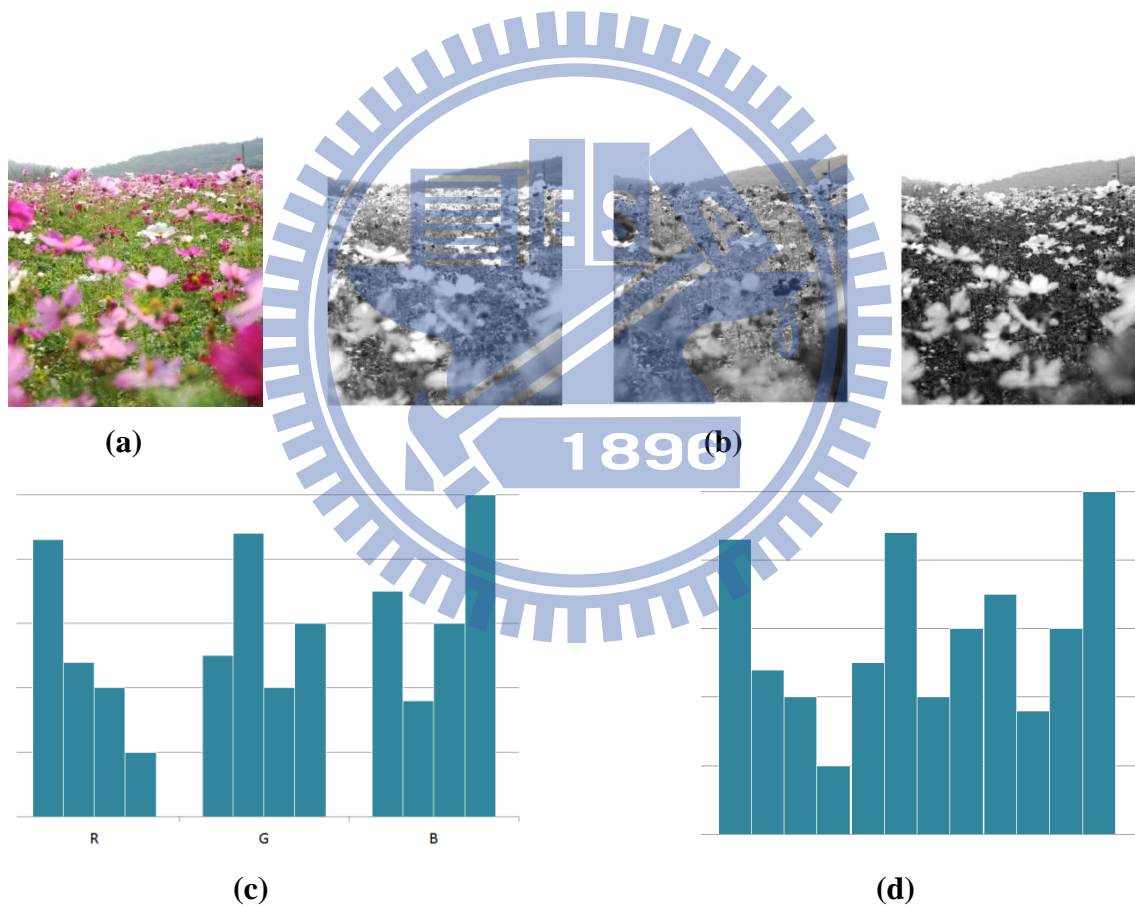


圖 3-5：簡易直方圖之示意圖(block size=4)

(a) 原圖，(b)為 RGB 三個個別圖層

(c)為三個個別圖層的直方圖

(d)將三組直方圖組合成彩色直方圖

### 3.2.2 人臉追蹤

經過 OpenCV 的人臉偵測，我們得到每個影格中的各個人臉區域位置及大小，但若是直接使用這些資料，除了資料量太過於龐大外，並沒有善加利用影片中的時間軸資訊，也就是影片中演員連續動作之間，人臉的大小以及移動位置並不會有太大的改變。依據論文[32]所提及的限制，建立演員串列時須滿足條件：

- ◎ 同一個人臉串列不會跨越任何一個分鏡。
- ◎ 人臉的位置的偏移量必須小於此臉區域的大小，也就是以此臉大小作偏移容忍比例。
- ◎ 兩張臉區域大小的比例必須在 0.67 ~ 1.5。
- ◎ 人臉個數必須在 4 個以上。

在我的串列建立方法中，希望能透過影片的特性將同一分鏡中某個演員的臉事先串成一個串列。也就是說，在做人臉辨識之前，將每幕分鏡的角色依照抓取到的位置及大小先串成一串，以利於在做人臉辨識及分群時的降低龐大資料量。在建立同一分鏡之間的演員串列時，除了考慮座標及大小，也會依據兩軀體的顏色差異(直方圖)來判斷是否屬於同一位演員。

此處計算直方圖距離時，定義的軀幹範圍是指人臉影像正下方，大小為人臉影像長寬兩倍大小，以此作為軀幹衣物的範圍。在連續的影象中，此區域包含了軀體以及背景，若是沒有分鏡切換，此區域的變化對於同一個演員很有限，是個很重要的參考依據。並且會將軀幹大小固定調整至 80x80，並且使用 Jaccard 來計算相似值。Jaccard 的計算公式如下。

$$J = \frac{\sum \min(x_i, y_i)}{\sum \max(x_i, y_i)} \quad (3)$$

數值 J 即為 Jaccard 的值，數值介於 0 到 1 之間， $x_i$  以及  $y_i$  分別為兩圖的直方圖向量之第  $i$  個數值，即是以最大的數值作為分母，而另一項作為分子當作相似度的關聯性。表 3-2 指出了在演員串列建立時所使用的不同色彩空間以及直方圖區塊大小所造成的不同結果。表中四個數據在表中都有所介紹，a 為在作串列刪減前演員串列數量，b

是刪減後串列的數量，c 是經過刪減後依舊無法刪減的混雜串列，意思是此串列中可能包含不同的演員，或是含有非人臉的區塊。d 是刪減後串列所有串列所包含的人臉個數(含 False positive)。下表中 J 即是 Jaccard 的數值，bin 則是使用之色彩直方圖所設定的區間數。可以看到在 HSV 色彩空間下，所造成的混雜串列較少，且保留的人臉數目也較完整，因此在建立演員串列時加入一項條件：

◎ 兩張臉下方的軀幹直方圖的 Jaccard 相似值必須在 0.65 以上 (HSV 色彩空間)。

表 3-1：演員串列建立時所用的分鏡偵測參數

四數字表示：刪減前演員串列數量，刪減後串列數量，無法刪減的混雜串列，刪減後串列所含的人臉個數(含 False positive)													
Jaccard - bin Color space		RGB				HSV				YCbCr			
		J	bin	a	b	c	d	a	b	c	d	a	b
0.8	16	4571	364	0	2397	5389	248	0	1489	3332	497	0	3795
0.75	16	3864	443	0	3182	4293	413	0	2773	3000	558	4	4511
0.7	16	3309	507	0	3839	3571	483	0	3581	3000	576	7	4989
0.65	16	3000	548	1	4379	3010	529	0	4222	3000	599	9	5332
0.6	16	3000	578	3	4782	3000	565	2	4726	3000	605	16	5503
0.75	32	4577	362	0	2411	5365	255	0	1524	3593	476	0	3527
0.7	32	3914	439	0	3141	4293	411	0	2772	3056	519	2	4098
0.65	32	3356	497	0	3807	3548	481	0	3614	2614	553	4	4638
0.55	32	2558	567	4	4704	2530	565	2	4770	2130	601	6	5314
0.5	32	2269	596	6	5119	2255	587	4	5119	2007	599	7	5440

計算軀幹顏色的直方圖是希望能以演員當下所穿的衣服來判斷兩張臉是否為同一人。軀幹的選取方式雖然並非精確的軀幹部位，但根據影像漸進式的改變，若是同一角色則此區域的顏色變化不會有太大改變。因此我們加入此判定的參考資訊，確實提升了初步演員串列建立的正確度。一般而言，在此處所設定的閾值都會偏低，理由是，寧可將演員串列切細，也不願建立出混雜不同角色的串列，在最後分群辨識時造成的影響更加龐大。





圖 3-6: 軀幹區域示意圖。

然而所得到的演員串列還是稍嫌過多，於是加入了二次機會的演算法。其主要目的是，根據 OpenCV 函式所抓取的結果來看，偶而會有些人臉因為某些角度而無法抓取，造成串列的斷裂。有一種情況是某些正確的人臉在膚色篩選時因光線而被刪除了，這些情況都會造成串列斷裂成更細的片段，也因此我加入了二次機會的演算法，主要流程為，每張臉在串接時，如果遇到中斷，可以往下一個影格尋找是否有符合上列後四項條件的人臉，若是符合，則繼續往下串接。根據實驗結果，不僅可以再度濃縮串列個數，並且可以將被部分切細的串列連接起來，增加整體的人臉個數。圖 3-7 即為判斷流程圖。如果下一張沒有符合的臉，則在下一個影格沒有被串接的餘留候選人中尋找符合條件的臉。

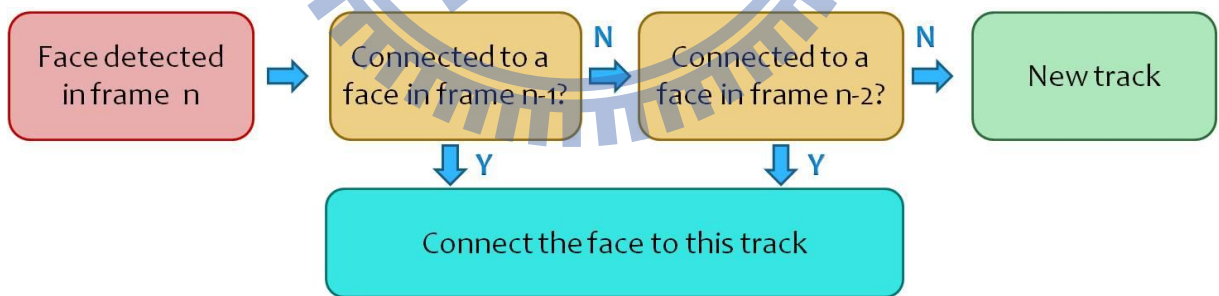


圖 3-7：二次機會串列銜接流程



圖 3-8：對演員串列進行二次機會串接，紅框中的人臉在並未正確被偵測出來，本來是左右兩組串列，經過二次機會串接可以將此兩串列成功再次組合。

### 3.3 結合身體與臉部相似資訊

#### 3.3.1 臉部相似度

在作影像投影之前，先談談人臉的大小對於實驗的影響。一般而言，影像解析度越精細，則影像中所包含的資訊也就越多。但資訊含量多與辨識的正確率並非成完全正比。因為解析度不統一的資料庫反而會降低辨識的準確度，也因為若是選取了太大的臉部大小，反而增加額外的計算了，這不是我們所樂見的。因此在本實驗中，我根據影集“Friends”中所有臉的平均，大略選取了 40x40 作為調整的固定大小。而並非所有影片的人臉範圍都會相近，因此 40 這數字對於整體已經是偏小的，目的是希望能以較低解析度的人臉來作到較一般化的辨識流程。而在第四章也會介紹實驗中以不同大小的臉區域調整作比較。

由於 40x40 的影像大小有 1600 個像素需要作比較，因此 EigenFaces 是最常用來降低維度的工具，而一般的 EigenFaces 是將整張圖當作是一串向量去處理，並沒有利用到整張臉二維的特性，因此近年來比較盛行的是用 2D-PCA，也就是以二維的方式去取得 EigenFaces 的投影基底，並且以二維的方式計算每兩兩的距離。

而 EigenFaces 的方法主要就是利用 PCA ( Principle Component Analysis ) 來降低維度，首先將所有像素排列成一個向量  $u$ ，接著把所有人臉影像都排列成一個向量組合成一個矩陣  $[u_1^T, u_2^T, u_3^T, \dots]$ ，計算此矩陣的特徵值 ( eigenvalue ) 與特徵向量 ( eigenvector )，視需要降至的  $n$  維度，選取前  $n$  大特徵值的特徵向量組合為投影矩陣，以此投影矩陣對原始的圖形向量作投影之矩陣相乘計算即可將原始的  $1 \times 1600$  向量降維度至  $1 \times n$  維度。在[32]的文章中已經先做過不同基底的實驗了，因此在本論文中只探討不同投影基底、不同臉大小及不同投影方式作實驗比較，而詳細的表格數據則在第四章中會再次說明。

2D-PCA 的理論基礎與一般 EigenFaces 投影相似，主要是以列或行的形式去計算投影基底，以及計算距離。首先將圖片以列的形式列出  $\{X_1, X_2, \dots, X_N\}$ ，然後計算出

$G_T$ ，再用此值來計算特徵向量與特徵值。如同 PCA 的概念，選出前  $N$  大個特徵向量作為投影基底。

$$G_T = \sum_{i=1}^N (X_i - \mu_X)(X_i - \mu_X)^T \quad (4)$$

$$Y_i = W^T(X_i - \mu_X), Y_i \in R^{k \times 1} \quad (5)$$

而  $\mu_X$  是所有  $X_i$  的平均值，再利用公式(5)作投影。而  $W$  就是我們希望投影到的低維度空間之投影矩陣，而此陣列是以前  $N$  大的特徵值對應的特徵向量所組成。最後透過公式(6)來計算兩張臉之間的距離。

$$d(Y, Y') = \sum_{i=1}^N \sqrt{(Y_i - Y'_i)(Y_i - Y'_i)^T} \quad (6)$$

此距離就如同我們在做 PCA 時所得到的歐式距離，在此是將每一組投影後的列向量取歐式距離並將之加總，用此距離作為相異值，之後會再進一步計算我們需要的相似值。在第四章有為投影至 PCA 以及 2DPCA 的結果比較。

在人臉辨識這領域，最常被廣泛使用的方法有 EigenFaces 以及 FisherFaces。而在我的論文中所需要的方法必須是不需要訓練資料的，也因此我採用 EigenFaces 相關的方法。在[32]的文章中僅對 PCA 做了不同維度投影的實驗，延續此文章的方式，我引入了 2D-PCA 的方法[15]，而 2D-PCA 又有許多延伸如：以行為基準的 2D-PCA，或是以列為基準的 2D-PCA 等等。但在這些新方法中，並沒有其中一樣是特別突出的，因此我在此僅使用了以列為基準的 2D-PCA。

臉部相似度的計算流程首先將輸入的人臉影像進行正規化，然後將色彩空間轉換為灰階來使用。接著將在人臉偵測時所得到的所有人臉用來計算出所需要的投影矩陣，在第四章會有針對 PCA 與 2D-PCA 的比較結果介紹。最後將所有人臉影像乘上投影矩陣，將所有人臉影像都投影至預設的空間維度上，我的實驗中取 10 個列空間當作投影基底。投影完成之後，便可以透過 2D-PCA 的距離計算公式來計算兩張臉之間的距離。得到了距離也就是相異度，我們便可以如法炮製，將距離帶入指數函式得到介於 0 到 1 的相似度。公式(7)即是計算臉相似度的公式， $S_{face}$  即是臉的相似值， $D_{face}$  則是在 2D-PCA 投影空間中計算的兩張臉距離，也就是公式(6)得到的數值。

$$S_{face} = e^{-D_{face}^2/2\delta^2}, \delta = \text{std}(D_{face}); \quad (7)$$

### 3.3.2 姿勢相似值計算



本論文的重點即是如何提升人臉辨識的準確度，而其中所使用的資訊就是姿勢相似值的計算。根據第二章內容所提到的，若能事先將資料庫的影像以固定座標角度來校正，能夠明顯的提升辨識的正確率。而現今就有許多學者們致力於這個領域，但所用的方法計算量越來越大，而對於座標（人臉特徵）的偵測卻沒有非常高的可信賴度。最常見的方式如眼睛嘴巴的偵測，此方法的優點是能精確的校正臉的位置以及角度，但最大的問題即是若發生偵測錯誤，對於後續辨識造成的反效果讓人難以接受。

在本實驗中所用的方法為姿勢的判定，根據[25]的實驗結果，我們可以透過所謂的 Gabor Wavelet Transform Filter 來對臉部影像作遮罩，目的是取的此臉部的紋理，但在從他的實驗中可以知道，透過不同方向及頻率的 Gabor Wavelet Transform 作遮罩，再將取得的資訊作 PCA(Principle Component Analysis) 降低維度，投影至三個維度後，可以看見不同旋轉角度的人臉均勻且分散的座落在此三維空間中。圖 3-9 是對同一張臉使用不同角度及頻率的遮罩結果，可以看見取得紋理的強度是根據頻率的大小。

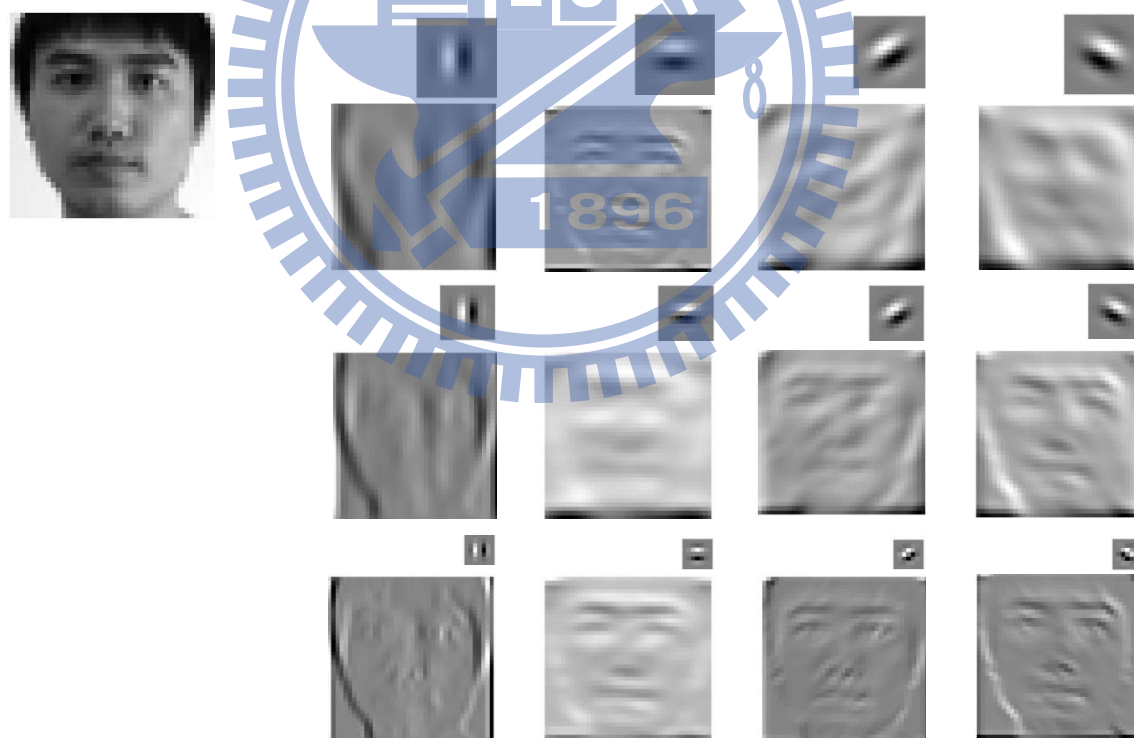


圖 3-9 共 12 種不同方向及頻率大小的 Gabor Transform 以及其結果。

在左上方的為原始圖型，每一行都代表正弦波的一種方向，不同列則是相異的頻率變化，由上至下頻率越來越大，在此限定高斯只包含兩個完整的正弦波波型。

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x'}{\lambda} + \psi\right) \quad (8)$$

$$x' = x \cos \theta + y \sin \theta$$

$$y' = -x \sin \theta + y \cos \theta$$

公式(8)為 Gabor Wavelet Transform 的計算公式，而此公式只包含實數部分，因在計算數值時，我們將虛數忽略不處理。其中  $\lambda$  代表此遮罩中的正弦波之係數，本實驗以高斯所涵蓋的區域限制兩個正弦波大小。而  $\theta$  代表此 Gabor Wavelet Transform 中波的行進方向，本實驗使了 0、45、90、135 共四種角度， $\psi$  是指正弦波的偏移量，一般而言將原點為波峰或是設在零點都可以，在此設為  $\theta/2$ 。 $\sigma$  是高斯函數的變異量係數，用來控制其分部寬度。而  $\gamma$  是此遮罩的寬高比例參數，在本實驗中因只使用圓形遮罩，故將此值設為 1 即可。

在[25]中投影的基底大小為 3，但在經過簡單的實驗測試後，發現若是要得到如此偏文章中如此完美的數據，勢必是在資料庫中作了非常優良的一般化，意思即是這些人臉影像可能在拍攝時都是使用同一台攝影器材，同樣的背景，同樣的光源。在圖 3-10 中的彩色符號為我自行收集的人臉資料下所呈現之分布圖。經過觀察，不難發現這些數據點的前進方向與人臉的轉動方向一致，也就是除了在其他文章中提到可以使用此方法分辨出最接近正面的人臉，我利用這特性將不同角度的臉以數字來呈現相似程度。本實驗利用自行收集的人臉來測試，發現實際在複雜背景及增加臉部配件的情況下，三維空間的分布會呈現兩種進行方向。此現象發生的可能原因之二即是前面提到的情況，因此在我們實驗後，發現若將投影空間維度降至一維，所得到的空間分布會比三維還要穩定。其優點是不同角度的人臉能在一維空間穩定的分布，但不像三維空間中可以有額外資訊來分辨左側臉、正臉及右側臉。由於本實驗目的是取得兩張臉之間的相異度(相似度)，故採取一維空間的投影，再用歐式距離來計算兩張臉的相似度。



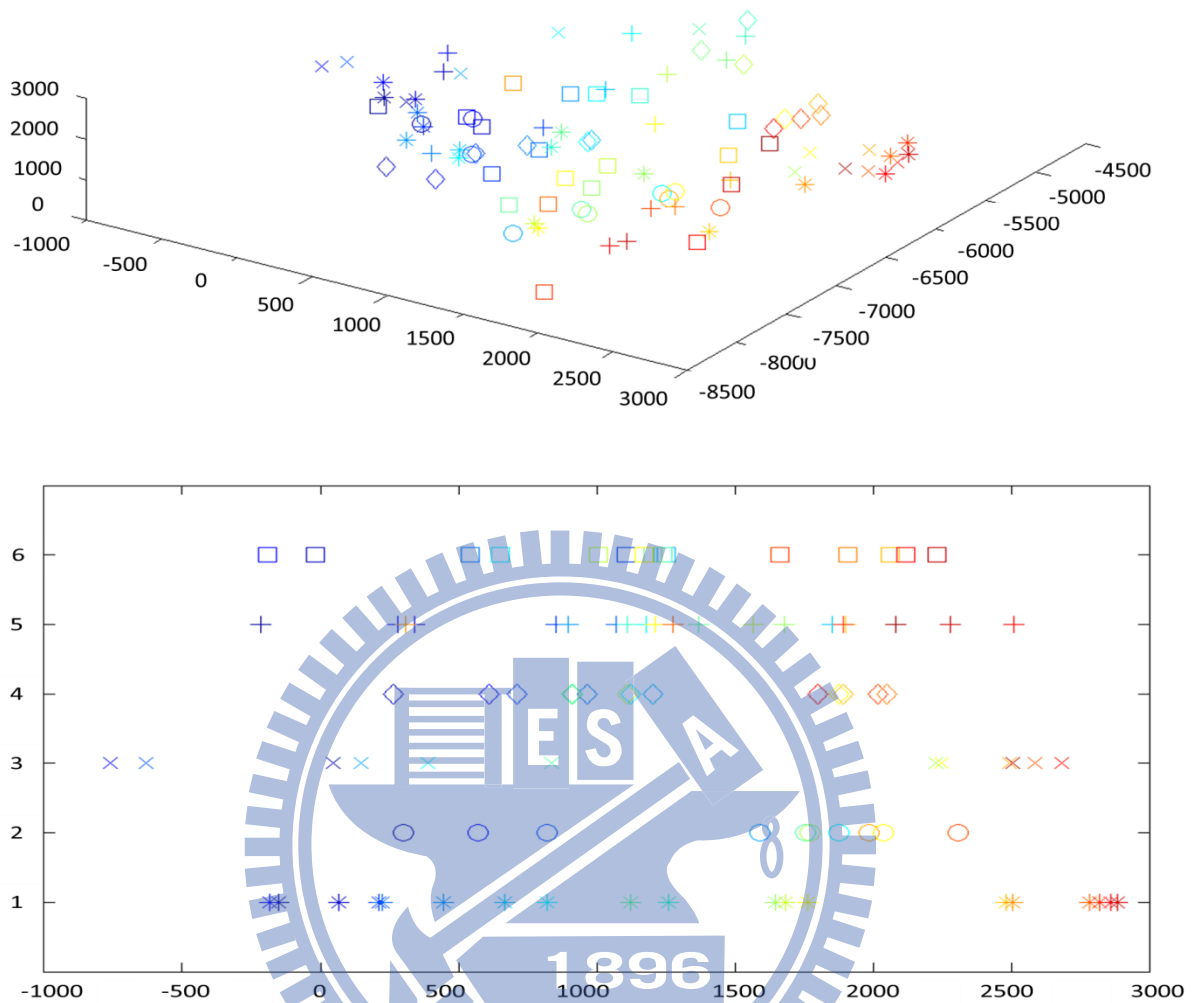


圖 3-10 最上方為人臉旋轉角度與顏色變化的示意圖。  
 中間的座標圖為投影至三維空間，而可看出相對距離較為混雜。  
 下圖為投影至一維空間，縱坐標是群組代號，橫坐標是一維空間的數值。

經過以上的實驗，於是我利用此一維空間的特性，使用兩張臉之間的距離來計算相似度。根據指數函式，我們可以輕易的將相異程度透過公式轉成相似值。

$$S_{pose} = e^{-D_{pose}^2/2\delta^2}, \delta = \text{std}(D_{pose}); \quad (9)$$

公式(9)中的  $S_{pose}$  代表兩張臉之間姿勢的相似度。 $D_{pose}$  表示歐式距離，在此因為我最後採納的是投影到一維度空間的方式計算，所得到的單一數值直接相減後取絕

對值即可。利用指數的公式將相似度限制在 0 到 1 間，而 1 代表完全相同。經過以上所有步驟，我們透過了 Gabor Wavelet Transform 的紋理提取，經過 PCA(Principle Component Analysis) 將空間維度降至 1 維，再透過指數函數將距離轉換為相似度後，對於每兩張臉之間都可以計算出彼此間姿勢的相似程度，在後面章節會以此為附加資訊來整合臉部辨識。

### 3.3.3 身體相似度

某些時候，要在影片中分辨兩個人，用衣服顏色來判斷會比用臉來的更明確。而且這常常發生在年紀相仿或是長相相似的演員們身上。在[5]的文章中就針對了各種可能的資訊實驗，透過不斷更新比重參數，最後得到了這些比重的最佳比例：頭髮、身體和臉。而其中比重最高的即是臉與身體，但他的實驗結果顯示，如果影像的間隔越長，身體資訊的比重就逐漸趨近於零。在本小節指介紹如何計算兩身體的相似度，而權重比例的參數設定留待章節 3.4.4 再介紹。

所謂身體的相似度，其實是指演員身上所穿的衣服，也因此要判斷一位演員的衣服是否不同，用顏色的改變量是非常直覺的方式。所以在此我使用與分鏡偵測相同的方法一直方圖。在計算之前，將欲計算的身體影像統一調整大小至臉部區域的兩倍長寬，也就是 80x80，如此一來像素的數量才會統一。而直方圖的區間大小經測試後，32 或是 16 之間的差別不大，因此在此採用大小為 16 的區間，也就是每張圖層都有  $\left(\frac{256}{16}\right)$  個數值，三張圖層合併後共的向量有 48 個數值。要計算相似度之前先要計算相異距離，這裡使用的是歐式距離。得到兩身體的距離後，代入與公式(9)相同的公式(10)，即取得兩身體之間的相似值。其中  $D_{torso}$  是指軀體彩色直方圖之間的距離，而  $\text{std}(D_{torso})$  在本實驗中為了減少計算量而使用了亂數選取的方式來計算標準差。

$$S_{torso} = e^{-D_{torso}^2/2\delta^2}, \delta = \text{std}(D_{torso}); \quad (10)$$

然而這數值將會在章節 3.4.4 與臉部資訊整合，其中權重參數的設計與實驗會在下面章節介紹。

### 3.3.4 身體與臉部資訊整合

在此小節裡面，會分別介紹如何計算臉部或身體的相似度，而在臉的這部分會與姿勢的相似度整合，其主要目的是改良在進行分群的時候，由於姿勢不相同而降低人臉辨識率的影響程度。簡單的說，即是在將兩張臉比較時另外計算可信賴度，在這邊指的是前一節中介紹的姿勢相似度。在姿勢不相同的情況下，往往不同的臉在相同姿勢下的相似程度會高於相同一張臉在不同姿勢下的相似度。這問題大大影響了人臉辨識的結果，尤其在影片當中，人們的臉部表情與姿勢非常豐富，與在拍照時的固定對準鏡頭有所區別。在實驗結果的部分，分群結果之數據也顯示此方式大幅提升了分群的準確度。

根據[5]文章中的實驗結果，我們可以了解在臉、身體、頭髮之中，以人臉以及身體的色彩最為之重要。因此在我的實驗中，也加入了身體的直方圖來修正人臉之間的相似程度。如同[32]所使用的方式，軀體的部分會先以直方圖計算此身體的特徵向量，以區間大小 16 而言，可以分為 16 個區間(0~255)，也就是得到一個具有 16 個數值的向量。計算每一對的軀體的三個色彩空間(16 x 3)，可以得到軀體之間色彩直方圖的距離，最後再利用公式 (10)得到軀體之間色彩直方圖相似值。

在本實驗中使用與[32]相同的結合方式來將臉部與軀體的相似值做結合。如公式(11)(12)。

$$S = \alpha \times S_{\text{torso}} + (1 - \alpha) \times S_{\text{face}} \quad (11)$$

$$\alpha = \alpha_h \times \exp\left(-\frac{FN}{2\sigma \times \alpha_t}\right); \sigma = \text{std}(FN) \quad ; \quad (12)$$

其中  $\alpha$  代表使用軀體的比重參數，反之  $(1 - \alpha)$  則是代表臉部資訊的比重。而  $\alpha$  是以公式(12)來計算臉與身體比重的比例，其中  $FN$  的值代表此兩張臉之間的隔了多少影格數量， $\sigma$  為此值的標準差，在此假設所有不同距離出現機率相同，使用固定比例的方式計算之，以降低計算複雜度。參數  $\alpha_t$  用來修正權重  $\alpha$  下降的速度，此值越大，

$FN$  的影響力也就越小，權重  $\alpha$  因  $FN$  增加而下降的影響也就越小。若是在影片之中角色的穿著並不會有太大的改變，則此參數  $\alpha_t$  設高一點可以得到較佳的結果。反之若角色在影片中多次改變穿著，或是不同角色穿著類似的衣物，則此參數設成較低的值可以得到較佳的分群結果。參數  $\alpha_h$  是公式(12)的整個權重比例，將原先介於 0 到 1 之間的後半部參數乘上一個百分比係數，意義是用來控制身體權重的影響程度。若



影片中的角色臉部外觀非常相似，而身穿不同的衣物，則此時將參數  $\alpha_h$  設高，使其在計算整合相似值時能根據較多的軀體權重來計算。所根據的觀念是，一段影片的同一个人不會在短時間內更換衣物，於是身體的部分即是很有意義的特徵，但在使用這特徵時也需要注意上面所提及的影片特性，於是除了利用間隔越遠身體資訊影響越小的方法外，還必須根據影片的特性來選取  $\alpha_h$  及  $\alpha_t$  參數。在第四章節中會使用兩段不同特色的同一系列影片做此兩參數的比對，可以用來驗證上面所提及的說明。

### 3.3.5 臉部與姿勢資訊整合

有了臉與姿勢的相似值  $S_{face}$  與  $S_{pose}$ ，在這裡會利用姿勢的資訊來修正人臉辨識的誤差。參閱在臉部辨識領域的文章，可以得到利用臉部校正來提升辨識率的結果。因此我希望能以相似姿勢的臉來辨識比較，以得到較為精確的結果。

在本實驗中所使用的判斷依據非常簡單，透過前幾章節所到的臉部相似值以及姿勢相似值，我們根據要做比較的臉之間的姿勢相似值來判斷此辨識結果是否有足夠的價值，也就是所謂的信心值 (Confidence)。若是兩臉間的信心值 (姿勢相似值) 低於我們自行設定的閾值 (Threshold)，則此兩臉的相似值就視為無效的值。在實際的程式碼中，我們僅加入了一段判斷式。

```
if (  $S_{pose} < \text{Threshold}$  )  
    將此  $S_{face}$  設為無效的值。
```

本實驗事先建立演員串列，在做分群合併時，以串列為合併之最小單位，也因此本公式主要使用在進行分群法時，針對串列與串列之間再計算相似值時所用，在串列與串列之間的相似值，僅使用姿勢在我們設定的合理範圍內之數值。在此章節僅介紹如何使用姿勢來篩選臉部的相似值，而在 4.3.4 會進一步介紹不同大小群合併時所取樣數的不同，以及此閾值的改變來提升整體的辨識率。在第四章的實驗結果中也顯示了使用姿勢當作閾值與否的結果數據，可以在第四章的表格中看到使用姿勢的結果得到了很大的提升。

### 3.4 演員串列分群

在此小節中所要做的就是利用計算出來的相似值來將影片中的人物做分群。由於為了使用時間軸的關聯性，我們根據之前人臉追蹤所製造的演員串列來加以分群。分群時將每一個演員串列視為已經分群完成的一小群，接著做階層式演算法（Hierarchical Clustering Algorithm）時，就是每兩群之間再計算新合併群與其他群的相似值。

#### 3.4.1 分群方法

有了臉與姿勢的相似值  $S_{face}$  與  $S_{pose}$ ，在這裡會利用姿勢的資訊來修正人臉辨識的誤差。本實驗使用的分群法為階層式演算法（Hierarchical Clustering Algorithm）[33]，而內部的演算法也有多種選擇性，在此我選擇凝聚法（agglomerative）演算法來做為主要的方法。

在群與群合併時，階層式演算法的凝聚法即是找距離最小（相似度最高）的兩群做合併。而群與群之間的距離要如何計算又可以分為下列四種：

Single-link :

$$d(C_q, C_s) = \min\{d(C_i, C_s), d(C_j, C_s)\} \quad (13)$$

Complete-link :

$$d(C_q, C_s) = \max\{d(C_i, C_s), d(C_j, C_s)\} \quad (14)$$

Average-link :

$$d(C_q, C_s) = \frac{n_i}{n_i+n_j}d(C_i, C_s) + \frac{n_j}{n_i+n_j}d(C_j, C_s) \quad (15)$$

$n_i$  及  $n_j$  分別為群  $C_i$  及  $C_j$  內元素個數

其中群組  $C_q$  是群組  $C_j$  與  $C_i$  的合併，而  $C_s$  表示其餘的群。

而  $d(C_q, C_s)$  表示  $C_q$  與  $C_s$  的距離。

本實驗採用公式(15)，在廣泛的實驗數據中，一般是使用平均值可以獲得最穩定的

數據，而本實驗針對前三中方法也有做數據的比較。而姿勢相似度與臉部相似度的整合演算法如下。

計算群  $C_j$  與  $C_i$  之間的相似度

計算群中所有臉之間的相似度  $S_{face}$  與姿勢相似度  $S_{pose}$ 。

if (  $S_{pose} < \text{Threshold}$  )

將此  $S_{face}$  設為無效的值。

用剩餘能用的數值來計算群之間的相似度。

輸出距離  $d(C_q, C_s)$  或是相似度  $S(C_q, C_s)$ 。

姿勢的主要作用是用來篩選刪除姿勢相差過大的配對，因此在使用階層式演算法計算群之間相似度時，就能只用信賴度高的配對來計算。至於 Threshold 的值以“Friend”影片數據而言，將其設為 0.4 是最佳的設定。

另外在階層式演算法中群之間距離的計算我也嘗試了一些不同的方法，在計算 Average-link 時，原始的方式是將全部的數值平均，其意義是降低特例點對於整體數值的影響，但加入了閾值來過濾信賴度低的配對時，也將大部分可能發生特例的情況一起刪除，也因為這個理由，我修改了 Average-link 的定義。改變的地方在於，在計算平均值時，如果所擁有可用的數據個數大於 N 時，則只使用 N 大的相似值(用距離則為最小)來計算平均值。而決定 N 的方法我實驗了三種不同的方式。

$$(1).N = \max (size(G_i), size(G_j))$$

$$(2).N = \min (size(G_i), size(G_j))$$

$$(3).N = \text{可用數值的個數(取全部)}$$

對於這三種 N 值的設定方式，在第四章的表格會將三種方式在處理同一段影片得到的數據列出，不難看出，雖然以第三種方式較為穩定，但很明顯第二種方式得到了最佳的結果。其推測的理由是，在經過姿勢的判定後，通過閾值的臉部相似度的可信度很高，



也就是說若 N 使用了全部的數值去降低特例的影響，反倒使得正確的臉部相似值被混濁了，反而是僅僅選取少量配對的方法(1)得了最佳的結果。

另外我也針對閾值 (Threshold) 做了些修正，此數值會依據合併時兩個群的大小作提高的修正。理由是在計算信心值 (姿勢相似值) 時，如果有更多數量的臉可以使用時，希望能提高閾值 (Threshold) 來篩選出較正確的姿勢配對。公式如下。

$$\text{Threshold} = 0.4 + (1 - \exp\left(-\frac{\epsilon}{2\sigma}\right)) \times 0.6 \quad (16)$$

符號  $\epsilon$  表示兩群之間的配對數量，以兩群數量 N 及 M 而言，之間的 edge 數量為  $N \times M$ 。而  $\sigma$  為以影片中全部臉的數目之一定比例計算的常數值。此公式代表 Threshold 會根據群的增大從 0.4 往 1 的方向增加。此方法在部分實驗情況下可以得到不錯數據，但是並不是非常穩定，故此參數的設定可以再深加探討。

### 3.4.2 分群限制

分群限制，指的是在做階層式演算法 (Hierarchical Clustering Algorithm) 的凝聚法 (agglomerative) 時，一般是選取最接近的兩個群，在此指的是相似值最大的兩群來做合併。而根據相關研究的文章中，大部分都會使用此限制來提高分群效率與準確度，即在同一張影格中出現的兩張臉不可能為同一個人。此假設非常直觀也非常有效，在[32]的實驗中即展示了此方法確實可提升分群結果的正確性。於是在本實驗中，在階層式演算法之凝聚法中，選取兩群時會判斷此兩群出現的時間軸之間是否有重疊，若有重疊則將此相似值直接設為 0，在計算新合併群的相似值時也會將此值傳遞下去。也就是說，在作凝聚法時，會利用傳遞的方式來判斷是否有重疊，可省去每次合併時都需要重新計算是否有所重疊。

## 第四章 實驗結果

### 4.1 評估工具

#### 4.1.1 Adjusted RAND index (ARI)

本實驗其中一種評估方式是 ARI[34]，是由 Hubert 等人提出的公式。而計算 ARI 所需要分別計算四種情況的不同結果，下面分別對四種情況做解釋，在此假設有  $n$  個物件，而正確分群結果為  $U$ ，實際分群結果為  $V$ 。 $n_{ij}$  代表在  $U$  中屬於群組  $u_i$  且在  $V$  中屬於  $v_j$  之元素個數， $n_i$  代表屬於群組  $u_i$  之元素數量， $n_j$  代表屬於群組  $v_j$  之元素數量。

- $a = \sum_{i,j} \binom{n_{ij}}{2}$

代表在分群  $U$  中在同一群分割中，且在  $V$  中也在同一群的物件對 (pairs) 數目。

- $b = \sum_i \binom{n_i}{2}$

代表在分群  $U$  中分為同一群的物件對 (pairs) 數目。

- $c = \sum_j \binom{n_j}{2}$

代表在分群  $V$  中分為同一群的物件對 (pairs) 數目。

計算完成此四個數值後即可計算 ARI，公式如下：

$$ARI(U, V) = \frac{a - (b * c) / \binom{n}{2}}{\frac{(b+c)}{2} - (b * c) / \binom{n}{2}} \quad (17)$$

ARI 的計算含意可以從四種情況中窺探出，此數值對於每一對物件對的分群結果是否保持著相同狀態較為敏感，而狀態可分為同一群和不同群兩種，此種評估工具會因最後分群的群組數而有巨大的影響。如實際上為 5 個群，而我們將它分為 3 群，則會有許多不同群的被分在同一群裡。因此，此評估工具在本實驗中以未知群組數量的方法下會比 CVC 這種取純度的評估方法還要有評量比較的意義。ARI 的數值若是 1 代表完全分群正確，數值越低代表分群結果較差，而某些極端的測試資料也會造成負數，如每一群只有

一個物件的測試資料。

#### 4.1.2 Classification via Clustering (CVC)

另外有一種較為常見的評估方法是 CVC，但評估的數值並不是太有意義，因此數值所計算的是群的純度(purity)。計算的方法即是將每一個群中，挑選占比例最大的物件種類作為正確分群種類，然後將所有群組中最大的種類個數加總放在分子，分母則是所有物件個數，此數值即為 CVC。

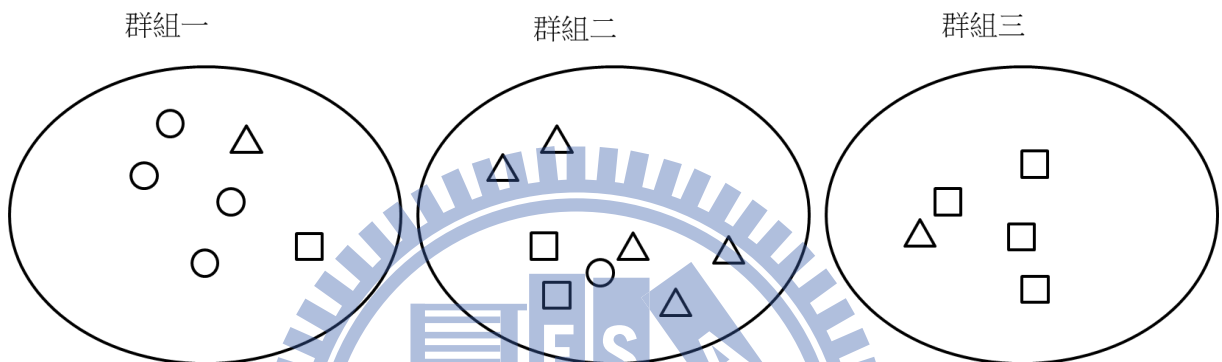


圖 4-1 : CVC 分群評估範例。

群組一之中最多數的是○，故將群組一當作是○的群，將○的總數當作群組中被正確分類的物件數目，而群組二以及群組三分別為△和□的群組，因此這三組群組的正確分群數目是  $4+5+4=13$ ，再除以總數( $6+8+5=19$ )就得到 CVC 的值。而 CVC 的數值就介於 0 與 1 之間，且最後的群組數量越小，CVC 值也會越小，因每兩群合併必定會使正確被分群的物件數變少，最好情況也只是相等而已，因此也可用來判斷演算法是否有出錯。

在本實驗中使用這兩種評估工具來做為判斷依據，但大部分情況下，此兩種數據不會同時上升或下降。所以我是以 ARI 作為最後分群結果的主要依據，而 CVC 可以用來觀察分群演算法是否有錯誤以及固定分群數目時的簡單評估依據。另外本實驗擷取演員串列作為分群的最小單位，因此在做評估時，會針對每一組串列去乘上它所包含的人臉數量，以便精準的算出以臉為單位的辨識率。

## 4.2 測試資料介紹

本實驗中共使用了三段影片作為測試資料，美國影集“Friends”及“Everybody Loves

Raymond”。第一份測試資料為“Friends”影集，經過影片切割後總共有 6520 張影格。經過人臉偵測和追蹤處理之後得到了 529 個演員串列，而所有串列包含了 4222 張偵測到的人臉(包含非人臉之錯誤偵測)。測試資料二也是使用“Friends”影集的其中一段，總共包含 5295 張影格，人臉追蹤處理後得到了 436 個演員串列，包含 4102 張偵測到的人臉。測試資料三使用的是“Everybody Loves Raymond”影集，共有 6456 張影格，463 個演員串列，3765 個人臉影像。

測試資料一的影像中，演員在整段影像中切換次數較少，演員身上的衣物也較少更換，而分鏡變換也僅有幾次，較接近一般家庭式錄像的影片，但其中有一段打美式足球的片段，大大影響人臉偵測的準確度，使得在演員串列建立時無法正確將人臉影像串接起來而被後續演算法刪除。測試資料二的影像與第一份測試資料不同的地方在於演員的服裝穿著，在影片中演員多次更換身上衣物，因此在分群時會將身體權重變數設為較低的值。且此影片中分鏡出現在公共場合，整段影片中被偵測出的演員共有 20 位，且其中包含了 3 位小朋友，也使得使用此測試資料得到結果較的一種測試資料差。測試資料三則是較為靜態的影片，演員身上的穿著並無太大變化，但影片中演員之間的年齡相差較大，其中有一對雙胞胎小嬰兒及一對老夫婦和一對中年夫妻。因此會將演員串列中的身體權重變數設為較高的值，但目前除了使用先前提的分群限制外，尚未處理雙胞胎的問題。圖 4-2 展示了完成人臉追蹤後的演員串列以及其身體的區塊。

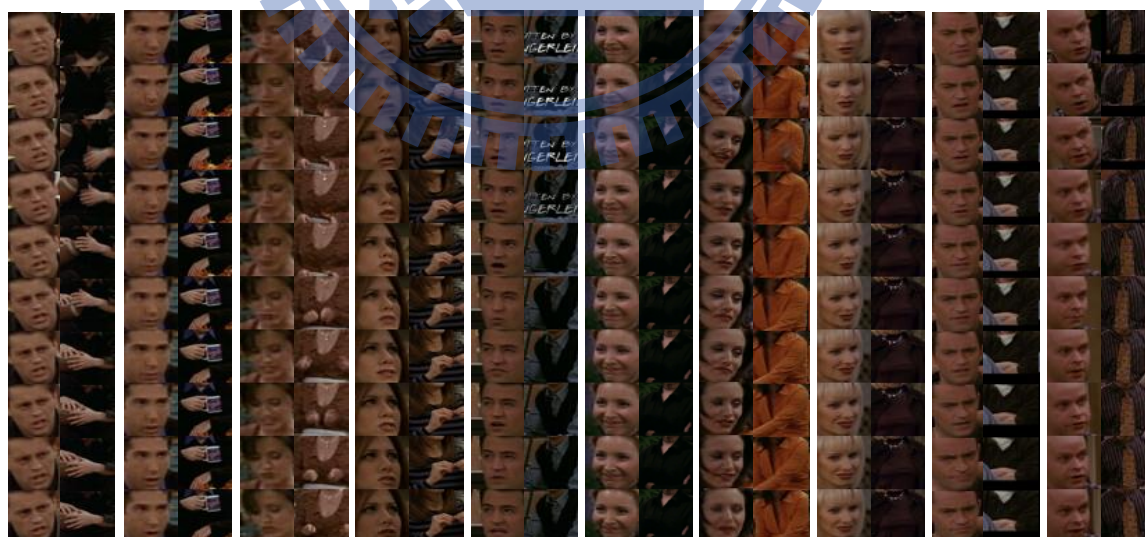


圖 4-2：演員串列以及身體區塊展示。

#### 4.3 不同條件下的實驗比較

在此章節中，我針對不同環境下的參數做分群結果的比較，如人臉投影的方法、人



臉尺寸大小的設定、使用姿勢資訊與法等比較結果。最後再列出本實驗的最終版本，並比較分析其各比重參數的意義。

#### 4.3.1 各種環境變數設計

根據[32]的文章，由於此文章中僅使用 PCA 作為投影的方式，故有針對投影到之較低維度的空間大小有作比較與測試，而本實驗用來作為對照組的 PCA 投影基底之維度就是參考此實驗結果。此處引用它的實驗結果數據，以四種不同基底而言，維度 70、130，160、190 所投影後的實驗結果相差甚小，且其中沒有絕對優勢的基底，因此在他的實驗中是使用 70 作為欲投影的空間維度大小，以減少計算量。

本實驗中將使用 2DPCA 作為投影的方法，在此與前面提及 PCA 作人臉辨識分群的結果比較。PCA 的投影基底大小使用 70 維度以及 400 維度用來和 2DPCA 比較，而 2DPCA 的投影基底大小我使用 10 個列的大小，故維度大小為 40x10。在本章節中希望以分群結果的正確率來判斷不同投影方法的優劣。但由於本實驗的整個流程包含了姿勢的參數變數以及分群合併時的取樣數變數等變因，因此在比較此兩種投影方法時，會將所有不同的參數設定跑出的結果一起做比較。表 4-1 是在做比較時的所有參數設定種類。K 代表最終分群的數量。Data 右方依序是不同的投影方法以及所使用臉部區域的尺寸大小。而加上“Pose”代表使用了姿勢的相似值作為閾值來修正分群結果，另外“Pose up”代表除了使用姿勢相似值，在合併兩個群時會依群的大小來提高閾值的作法。Clustering 右方代表的是在合併兩群時，計算新距離所使用範例的數量，在此都是使用 average-linked 來計算新的距離，而“min”表示 3.5.1 中所提到的取樣方式，僅取前 N 個最相近的，“all”則是一般的 average-linked，使用全部的資料來計算新的距離。

表 4-1 : PCA 與 2DPCA 的分群結果

		K=7	K=10	K=20	K=30
ARI	PCA(1x70)	0.130	0.127	0.122	0.117
	PCA(1x400)	0.105	0.145	0.153	0.135
	2DPCA(10x40)	<b>0.182</b>	0.188	0.165	0.164
CVC	PCA(1x70)	0.372	0.382	0.469	0.519
	PCA(1x400)	0.357	0.413	0.460	0.494
	2DPCA(10x40)	<b>0.462</b>	0.477	0.530	0.547

從表 4-1 中可以從左邊往右邊慢慢比較，最左邊兩個分別是針對 PCA 以及 2DPCA 的比較，其中所使用臉的尺寸大小都是  $70 \times 70$ 。可以從(1)和(2)看出除了在 K 較高的 CVC 數據，2DPCA 的實驗結果都較 PCA 優異，且在 ARI 更是明顯，無論在進行分群合併時使用了何種方式來選取樣本，此配對的實驗結果都顯示在進行人臉辨識時，相較於 PCA 而言 2DPCA 確實是更好的選擇。

表 4-2：使用姿勢相似只與不使用的分群結果

Agglomerate into K clusters.(2DPCA 40x40)					
Pose Threshold=0.4, didn't set $\alpha_t$ .					
		K=7	K=10	K=20	K=30
ARI	No Pose	0.182	0.188	0.165	0.164
	Use Pose	<b>0.358</b>	0.361	0.367	0.375
CVC	No Pose	0.462	0.477	0.530	0.547
	Use Pose	<b>0.607</b>	0.612	0.626	0.641

從表 4-2 中的數據可以看到在增加了姿勢相似度的限制後，得到的數據明顯較沒有使用姿勢資訊的優秀，且改良的幅度也非常多，因此透過資式修正我們可以得到更好的分群結果。

表 4-3: 在 3.4.1 中三種不同方法的實驗數據，

資料為使用 2D-PCA 投影，無加入身體資訊，左方為使用姿勢，右方為不使用。

含姿勢資訊		方法(1)	方法(2)	方法(3)	無姿勢資訊		方法(1)	方法(2)	方法(3)
Tracks Number =529					Tracks Number =529				
K=30	CVC	0.517	<b>0.734</b>	0.541	K=30	CVC	0.564	<b>0.641</b>	0.556
	ARI	0.202	<b>0.375</b>	0.197		ARI	0.341	<b>0.375</b>	0.226
K=20	CVC	0.505	<b>0.717</b>	0.505	K=20	CVC	0.549	<b>0.626</b>	0.527
	ARI	0.198	<b>0.386</b>	0.191		ARI	0.333	<b>0.367</b>	0.221
K=10	CVC	0.483	<b>0.67</b>	0.492	K=10	CVC	0.527	<b>0.612</b>	0.510
	ARI	0.193	<b>0.377</b>	0.187		ARI	0.327	<b>0.361</b>	0.219
K=8	CVC	0.482	<b>0.642</b>	0.486	K=8	CVC	0.549	<b>0.608</b>	無法合併
	ARI	0.208	<b>0.341</b>	0.183		ARI	0.333	<b>0.359</b>	無法合併
K=7	CVC	無法合併	無法合併	0.482	K=7	CVC	0.523	<b>0.607</b>	無法合併
	ARI	無法合併	無法合併	0.183		ARI	0.325	<b>0.358</b>	無法合併

表格中的粗體字代表每一列最大的數值，K 表示最後分成多少群。從表格中清楚看見無論是 CVC 或是 ARI 何種評估工具，對於這三種 N 的選取方式都顯示了第二種的優勢。然而在有無加入姿勢資訊方面，也都是第二種的結果較佳，但在第一及第三種方法中，加入姿勢反而降低的原因是此姿勢的參數並非已最佳化，僅是使用固定數字來做為比較用。

表 4-4：各種不同的參數設計比較。

Data		(1). PCA 70		(2). 2DPCA 70		(3). 2DPCA 40		(4). 2DPCA 40 Pose		(5). 2DPCA 40 Pose up	
Clustering		min	all	min	all	min	all	min	all	min	all
CVC	K=30	0.519	0.676	0.520	0.536	0.547	0.568	0.641	0.556	0.701	0.5092
	K=20	0.469	0.612	0.501	0.517	0.530	0.513	0.626	0.527	0.621	0.485
	K=10	0.382	0.361	0.451	0.476	0.477	0.489	0.612	0.510	0.430	0.4519
	K=7	0.372	0.274	0.438	0.435	0.462	0.459	0.607	0.348	0.351	0.3372
ARI	K=30	0.117	0.179	0.142	0.198	0.164	0.249	0.375	0.226	0.193	0.2053
	K=20	0.122	0.187	0.140	0.195	0.165	0.212	0.367	0.221	0.184	0.1998
	K=10	0.127	0.116	0.128	0.187	0.188	0.206	0.361	0.219	0.103	0.1933
	K=7	0.130	0.116	0.126	0.185	0.182	0.191	0.358	0.146	0.113	0.1448

接下來我做了不同臉部尺寸大小的比較，由於本實驗的目的是針對影像進行處理，在一般800×600的影片中，人臉的尺寸範圍可能從20~100都有可能出現，但若是取過大的尺寸做取樣，不但會使計算複雜度增加，且會發生多重解析度的問題。也就是說，在尺寸選取方面希望選取一個較平均大小還小一些的尺寸即可，目的將大部分臉部影像的解析度統一調整到統一的大小。在本實驗中依據了測試資料一的影片，選擇了大小 40 作為欲重新調整的大小。從表 4-4 的(2)和(3)可以看出兩者之差異。值得一提的是，(2)和(3)的資料參數中僅有臉部大小尺寸的不同，其他條件情況完全相同，但是實驗結果數據卻是解析度較小的得到了較佳的結果，這與我們一般直觀的想法正好相反。此處我們只能猜測在此實驗數據中雜訊的影響可能在較低解析度被降低了，也可能是解析度差異過大令人臉辨識的結果變差了。在本論文中暫不討論多重解析度的解決辦法，僅以較低解析度的取樣方式來解決，此議題也可以留做日後改良的一個部分。

在姿勢資訊的處理，也可以觀察(3)、(4)和(5)的數據結果。首先可以透過(3)和(4)來觀察使用姿勢資訊與否的實驗結果。這兩者差異只在於(4)使用了姿勢的相似度來篩選可

信的的臉部辨識結果，透過較信任的臉部辨識結果來計算新的相似度或距離。此處的重點在於 average-linked 在計算新距離時所使用的取樣數量，可以從表格中看到，若是一般的 average-linked 將全部的可用數據一起計算新的距離，(3)和(4)的結果差異不大，無法判斷何種比較優秀。但在我的實驗中，透過限制 average-linked 的取樣數量，我們選取最接近的 N 個臉部相似值重新計算新距離。透過表格數據，我們看到了令人驚訝的結果，經過上述方式處理後，不但在 CVC 有了明顯的提升，在 ARI 甚至提升了一倍之多，從 15% 提升到 35% 左右。這也是為什麼我會採用 3.5.1 所提出的方法。

最後我希望能加上新條件來提升分群的正确性。我提出了公式(16)將每次在分群法做合併時將姿勢相似度的閾值依群的大小作提升。但在表格 4-4 之中顯示的數據僅有在 CVC 且分群數量較高時有不錯的結果，而其他的數據相較於無使用此方法的結果卻是更差，因此這方法的可行性還有待未來在深加探討，我相信透過適當的設定此姿勢相似度閾值的可以提升一定程度的分群結果。

#### 4.3.2 臉部與身體資訊結合參數比較

接下來就是將上一階段得到的人臉相似值與身體相似值整合。如在 3.3.4 提到的方法，透過公式(11)和(12)來計算整合的數值。根據公式(11)，計算此權重值必須要設定兩個參數，分別是  $\alpha_h$  和  $\alpha_t$ 。分別代表身體資訊的權重比例與身體權重參數下降的速度。在我的實驗中設定  $\alpha_h$  從 0.1 到 0.9 間隔 0.1 做測試，而  $\alpha_t$  從 0.1 到 10 之間做測試。由於實驗結果數據過於繁雜，在此僅列出最佳參數選取理由的數據曲線。表 4-4 的測試資料為測試資料一。



表 4-5：測試資料一，合併身體與臉部資訊時的權重變數設定。

$\alpha_h=0.3$	$\alpha_t$	1.2	1.3	1.4	1.5	1.6	1.7
CVC	K=30	0.8129	0.8119	0.8119	0.8129	0.81	0.81
	K=20	0.7866	0.7856	0.7856	0.7883	0.8065	0.8065
	K=10	0.7513	0.7504	0.7414	0.7402	0.7215	0.7215
	K=8	0.6902	0.6892	0.6892	0.6206	0.7215	0.7215
	K=7	無法合併	無法合併	無法合併	無法合併	0.6485	0.6485
ARI	K=30	0.465	0.4631	0.4638	0.4709	0.4711	0.471
	K=20	0.4703	0.4684	0.4684	0.4714	0.4811	0.481
	K=10	0.4766	0.4748	0.4449	0.4591	0.4272	0.4275
	K=8	0.4319	0.4303	0.4389	0.3833	0.445	0.4468
	K=7	無法合併	無法合併	無法合併	無法合併	0.4106	0.4123

$\alpha_h=0.4$	$\alpha_t$	1.2	1.3	1.4	1.5	1.6	1.7
CVC	K=30	0.8096	0.8243	0.8243	0.8214	0.8228	0.8228
	K=20	0.8032	0.7968	0.7968	0.7954	0.8115	0.8103
	K=10	0.7421	0.7373	0.7373	0.7387	0.7565	0.7565
	K=8	0.6791	0.675	0.675	0.7387	0.7565	0.7565
	K=7	無法合併	0.6021	0.6021	0.6658	0.6836	0.6836
ARI	K=30	0.4505	0.4485	0.4486	0.423	0.424	0.424
	K=20	0.4585	0.4571	0.4572	0.4404	0.4526	0.4517
	K=10	0.4135	0.4193	0.4201	0.4993	0.514	0.5138
	K=8	0.3829	0.4021	0.4028	0.5065	0.5204	0.5198
	K=7	無法合併	0.3705	0.3711	0.4712	0.4843	0.4836

$\alpha_h=0.5$	$\alpha_t$	1.2	1.3	1.4	1.5	1.6	1.7
CVC	K=30	0.8212	0.8212	0.8238	0.8055	0.8055	0.8065
	K=20	0.801	0.8103	0.8138	0.7963	0.7975	0.7975
	K=10	0.7729	0.7729	0.7541	0.7688	0.7688	0.7688
	K=8	0.6805	0.6805	0.6253	0.6506	0.6506	0.6506
	K=7	無法合併	無法合併	無法合併	0.6506	0.6506	0.6506
ARI	K=30	0.3918	0.3926	0.3682	0.3735	0.3735	0.4005
	K=20	0.4062	0.4105	0.4146	0.4183	0.4193	0.4193
	K=10	0.4769	0.4769	0.4491	0.4259	0.4259	0.4259
	K=8	0.4113	0.4113	0.3852	0.3696	0.3696	0.3696
	K=7	無法合併	無法合併	無法合併	0.4096	0.4096	0.4096

以此實驗數據表格繪製出圖表會更容易看出最佳的變數設定應該在哪。圖 4-2 上下兩張

折線圖分別代表 CVC(上方)與 ARI(下方)在參數改變時的變化。

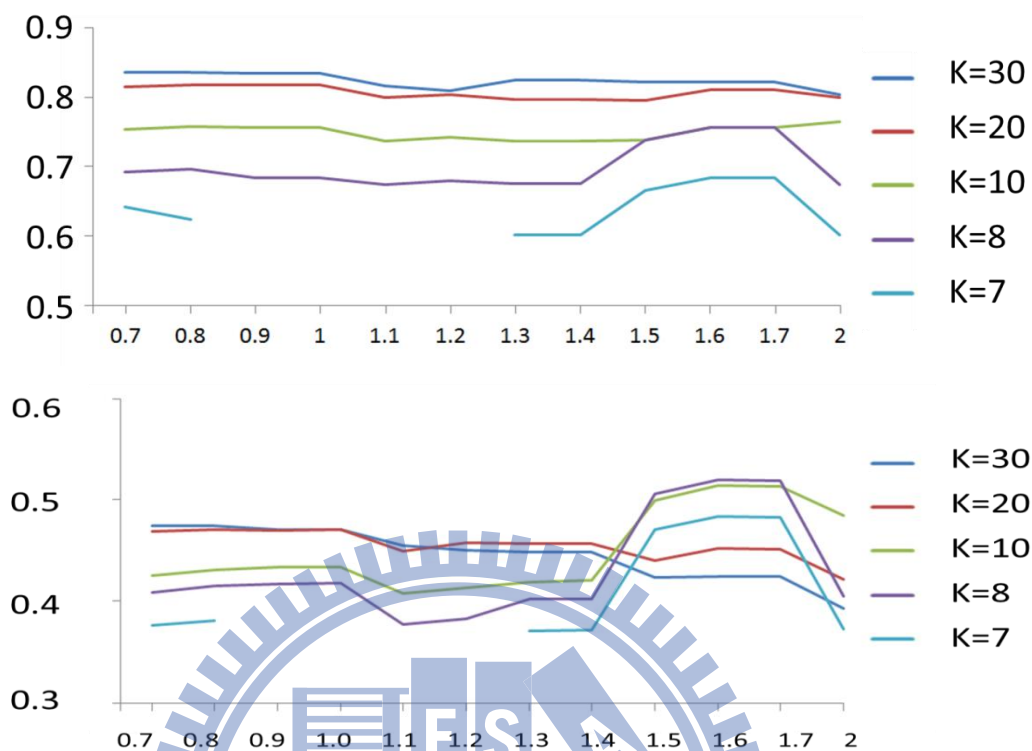


圖 4-3：測試資料一，縱座標為 CVC(上圖)與 ARI(下圖)在各種參數環境下的折線圖。 $\alpha_h$ 在此處為 0.4 而橫坐標為  $\alpha_t$  的數值。

雖然透過圖 4-3 可以輕易地找到最佳的設定值，但  $\alpha_h$  與  $\alpha_t$  必須事先設定好，因此針對不同類型的影片所要設定的最佳參數就不同。如同前幾章提及的，若影片中演員的衣物穿著多次變換，則身體資訊的權重設低才可以得到較佳結果，反之若是人臉之間太過於相似，則身體權重設高則能得到較好的結果。在經過一些測試影片的實驗後，此兩參數並沒有絕對優異的值，非常容易因影片的內涵而有大幅度的跳動。因此根據我們的三種測試資料，也僅能大概猜測某範圍的數值，而是否有更好的技術來解決此問題，也是個未來研究的方向。

表 4-6：測試資料二，合併身體與臉部資訊時的權重變數設定。

$\alpha_h=0.4$	$\alpha_t$	2	2.1	2.2	2.3	2.4	2.5
CVC	K=30	0.7445	0.7445	0.7762	0.7774	0.7784	0.7784
	K=20	0.7031	0.7031	0.7145	0.7145	0.7155	0.7155
	K=10	0.5965	0.5965	0.6029	0.6029	0.5636	0.6026
	K=8	0.5563	0.5563	0.567	0.567	0.5093	0.5436
	K=7	0.5032	0.5032	無法合併	無法合併	0.4734	無法合併
ARI	K=30	0.2986	0.2986	0.3023	0.3033	0.304	0.3062
	K=20	0.2912	0.2914	0.2865	0.2865	0.2885	0.2912
	K=10	0.3197	0.3195	0.2838	0.2838	0.2596	0.3182
	K=8	0.3044	0.3043	0.3043	0.3043	0.2224	0.2911
	K=7	0.2764	0.2765	無法合併	無法合併	0.2081	無法合併

$\alpha_h=0.5$	$\alpha_t$	2	2.1	2.2	2.3	2.4	2.5
CVC	K=30	0.7543	0.753	0.754	0.7633	0.7633	0.7696
	K=20	0.6892	0.6892	0.7009	0.7004	0.7216	0.7126
	K=10	0.59	0.579	0.6029	0.6107	0.6555	0.6058
	K=8	0.5341	0.5502	0.5446	0.5819	0.6197	0.5685
	K=7	無法合併	0.4861	無法合併	0.5361	無法合併	0.5327
ARI	K=30	0.2892	0.2884	0.2903	0.287	0.2922	0.2998
	K=20	0.2752	0.2754	0.2845	0.2843	0.3044	0.3033
	K=10	0.2618	0.2896	0.3462	0.3294	0.3678	0.3048
	K=8	0.2747	0.3063	0.3067	0.4114	0.4431	0.2892
	K=7	無法合併	0.2714	無法合併	0.3847	無法合併	0.2694

$\alpha_h=0.6$	$\alpha_t$	2	2.1	2.2	2.3	2.4	2.5
CVC	K=30	0.7448	0.7411	0.7413	0.747	0.7457	0.7457
	K=20	0.6755	0.6743	0.6743	0.686	0.6848	0.6848
	K=10	0.54	0.5468	0.5258	0.5553	0.5541	0.5551
	K=8	無法合併	0.5468	0.4783	0.5327	0.5314	0.5212
	K=7	無法合併	無法合併	無法合併	無法合併	無法合併	無法合併
ARI	K=30	0.282	0.2786	0.2776	0.2831	0.2824	0.2821
	K=20	0.2627	0.2609	0.2609	0.2731	0.2723	0.272
	K=10	0.2507	0.252	0.2469	0.2498	0.2489	0.2664
	K=8	無法合併	0.2799	0.1995	0.2261	0.2253	0.249
	K=7	無法合併	無法合併	無法合併	無法合併	無法合併	無法合併

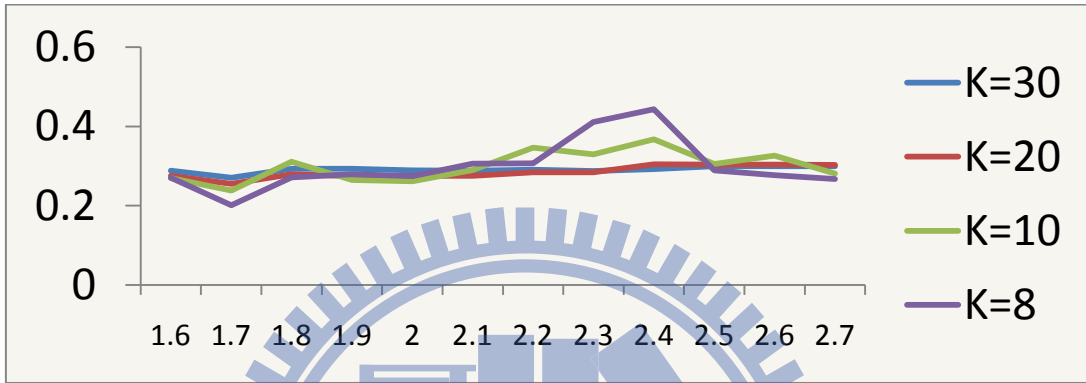
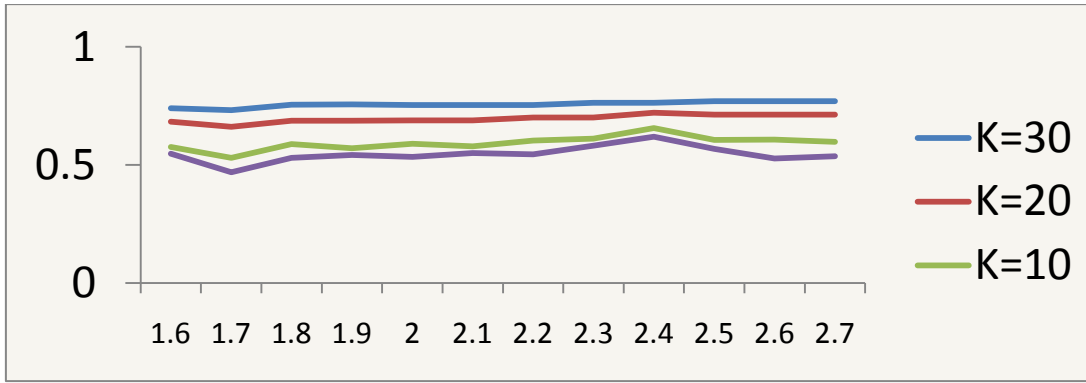


圖 4-4：測試資料二，縱座標為 CVC 與 ARI 在各種參數環境下的折線圖。 $\alpha_h$ 在此處為 0.5 而橫坐標為  $\alpha_t$  的數值。

表 4-7：測試資料三，合併身體與臉部資訊時的權重變數設定。

$\alpha_h=0.4$	$\alpha_t$	2	2.5	3	3.5	4	4.5
CVC	K=30	0.8712	0.8701	0.8789	0.8861	0.8744	0.8717
	K=20	0.8444	0.8274	0.826	0.8462	0.8398	0.8313
	K=10	0.7389	0.7402	0.7355	0.7963	0.7044	0.7819
	K=8	0.7124	0.7333	0.6183	0.6818	0.643	0.6911
	K=7	0.7124	無法合併	0.5772	0.6335	0.6005	0.6486
ARI	K=30	0.3448	0.3598	0.3649	0.3693	0.3912	0.3854
	K=20	0.3669	0.353	0.3556	0.3793	0.3825	0.3724
	K=10	0.3881	0.3406	0.365	0.4037	0.315	0.3737
	K=8	0.3546	0.3956	0.3122	0.3549	0.2982	0.3254
	K=7	0.3739	無法合併	0.2787	0.3279	0.2525	0.2783

$\alpha_h=0.5$	$\alpha_t$	2	2.5	3	3.5	4	4.5
CVC	K=30	0.8685	0.911	0.9004	0.9049	0.9015	0.9049
	K=20	0.8361	0.8624	0.8507	0.859	0.8707	0.8799
	K=10	0.7849	0.762	0.7721	0.7745	0.7745	0.7745
	K=8	0.6991	0.753	0.7161	0.7676	0.6871	0.7676
	K=7	0.6991	0.7057	無法合併	0.7116	0.6412	無法合併
ARI	K=30	0.3269	0.3496	0.3417	0.3597	0.3511	0.3538
	K=20	0.3383	0.3612	0.3505	0.3714	0.376	0.3931
	K=10	0.4585	0.375	0.3863	0.3974	0.3732	0.3677
	K=8	0.3977	0.4015	0.3886	0.4397	0.3896	0.4105
	K=7	0.407	0.3687	無法合併	0.3973	0.3492	無法合併

$\alpha_h=0.6$	$\alpha_t$	2	2.5	3	3.5	4	4.5
CVC	K=30	0.8614	0.8627	0.8916	0.8887	0.8969	0.8969
	K=20	0.8167	0.8141	0.8122	0.8205	0.8037	0.817
	K=10	0.7219	0.7384	0.7378	0.7062	0.6869	0.7001
	K=8	0.6659	0.6436	0.6837	0.6507	0.6396	0.6234
	K=7	0.6441	0.6436	0.6579	0.6507	0.6367	0.6151
ARI	K=30	0.3218	0.3235	0.3246	0.328	0.3339	0.3338
	K=20	0.3106	0.3091	0.3057	0.31	0.3065	0.3151
	K=10	0.3679	0.3752	0.389	0.3418	0.3436	0.3465
	K=8	0.3443	0.3052	0.3484	0.3061	0.3221	0.278
	K=7	0.3419	0.326	0.3294	0.3399	0.3247	0.2918



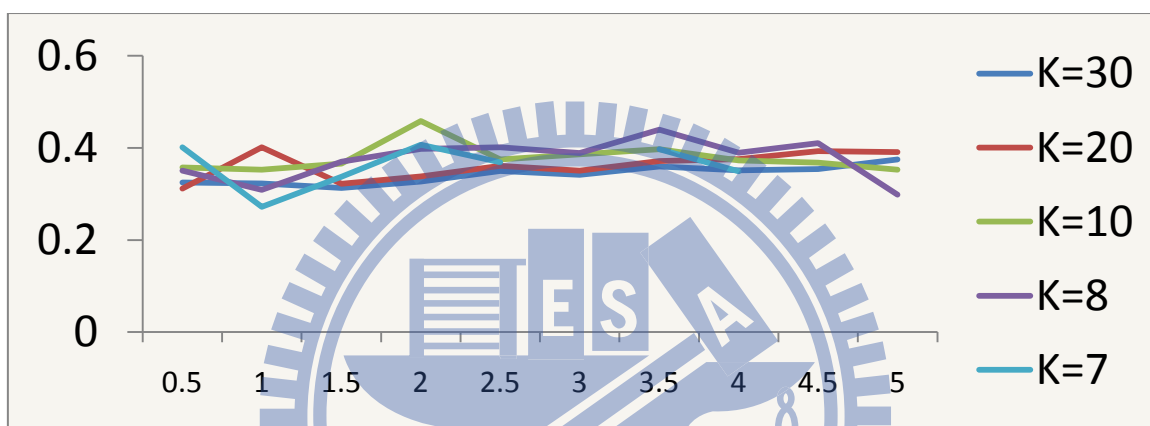
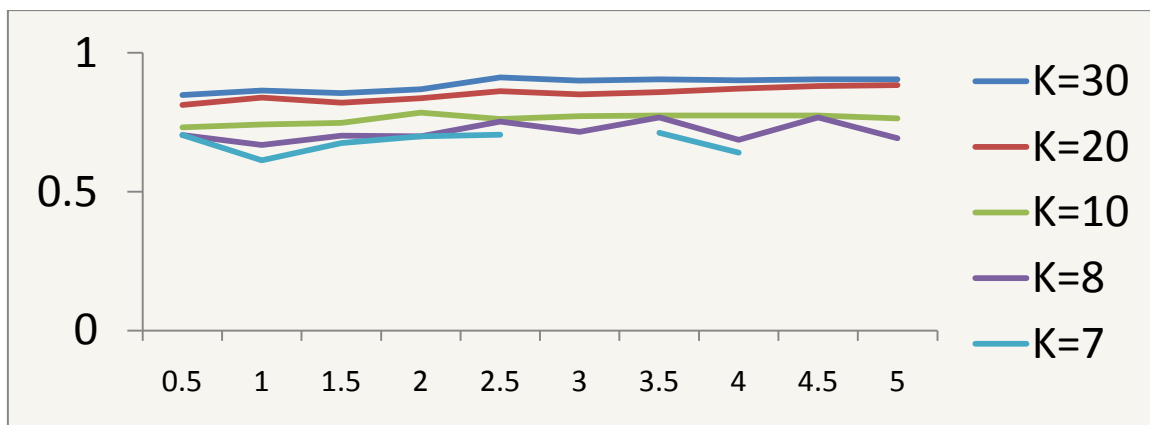


圖 4-5：測試資料三，縱座標為 CVC 與 ARI 在各種參數環境下的折線圖。 $\alpha_h$ 在此處為 0.5 而橫坐標為  $\alpha_t$  的數值。

雖然根據圖無法明確的找到最佳的權重參數，但確實在某範圍內中的參數設定可以達到較佳的結果。觀察這些數據可以發現，雖然在測試資料二中的演員多次更換衣物，但在此實驗結果得到的身體權重  $\alpha_h$  卻仍有 0.5，可以理解的是，這並不代表實驗中有所出錯，而是代表就算是身體變化較頻繁的影片，身體的相似度依然扮演著重要的腳色。而測試資料二的特性就反映在  $\alpha_t$  上，較高的  $\alpha_h$  表示身體的權重依時間軸下降的速度較快，非常符合實際的影片內容。另外在測試資料三，雖然  $\alpha_t$  差異不大，但  $\alpha_h$  卻高達了 3.5，主要原因就是在此影片中的許多演員身上衣物的顏色組成非常相似，因此在僅極短時間內身體的資訊較有價值，而連續分鏡中出現不同人物穿著相同衣物的特殊情況，目前沒有解決的辦法。

## 第五章 結論與未來展望

本論文不僅是提出了完整的影像人臉註記的流程，也加上了姿勢資訊來校正角度差異過大的臉。整個完整的流程包含了人臉偵測，以膚色偵測來過濾非人臉的區塊，追蹤影片中連續出現的同一張人臉，然後經過前置處理的光線平衡與高低通濾波將影像正規化，接著才進行 2DPCA 投影，其中加上了利用 Gabor Wavelet Transform 擷取紋理的方式來辨別姿勢的相似關係，最後還使用改良的分群計算方式做階層式分群法。每一階段的步驟都經過實驗測試，像是在膚色偵測也經過了兩種不同演算法的測試，而前處理也是有對無前處理的臉進行簡單的比較，在文章 3.3.2 也展示了 2DPCA 優於 PCA 的分群結果。其中本論文的重點，姿勢相似值，更是從紋理擷取、數據分析、測試，最後才加入實驗之中，並且在第四章有對此相似值大幅提升辨識準確加以說明，以 ARI 來比對的話，可以從 0.16 提升到 0.35。

但是在姿勢判斷上依舊有些瑕疵，如同一個角度的不同人臉，經過投影後的座標可能會很接近，但必定有些微的落差。若兩張臉投影座標前後落差過大，則可能對於判斷此張臉的相似值是否可用就還有些疑慮。因此在姿勢相似值作為閾值的實驗中，常常會有非常大的落差，這可能就是其中一個原因。而如剛剛提及的問題，我有嘗試著以動態的方式來調整此閾值，如 4.3.1 最後提到的方法，但目前僅提升了 CVC 的分群結果，因此希望將來能找到更有效率的方法來動態調整閾值。

實驗表格數據中，可以看出 ARI 的上升下降梯度與最終群的關係並不沒有絕對的正相關，其最大的可能即是 ARI，本身對群評估的依據。例如，有一群甲含有 A 和 B 兩位演員，而另一群乙含有 B 和 C 兩種演員串列，則若是此兩群因 A 之間非常相似而合併後，ARI 可能會提升，因為甲群及乙群中都有 A 演員，合併後此 A 演員從原先的不同群中合併到同一群之中，但另一方面，ARI 也會因此下降，因為甲乙群合併後，同一群中的 A 與 B 及 A 與 C 及 B 與 C 這些不同的演員串列也因此被合併在同一群，造成 ARI 值的下降。因此 ARI 是否提升與此階段群合併中正確合併的比例有關，若正確合併的比例比錯誤的合併比例高，則 ARI 值才會提升。

值得一提的是，在各種實驗階段中，研究者最好能嘗試所有組合可能性，因在姿勢相似值使用與否的實驗中，一開始並沒有得到比較好的實驗結果，但改變了在分群時計算新距離的取樣方式，卻可以大幅度的提高實驗數據，因此，一個不錯的想法雖然沒在

一開始就得到好的結果，並不一定代表此想法是錯的，或許轉個方向即可突破現狀。

另外，在同一類型的研究中都有著通病，權重的數據必須事先設定，且沒有固定的值是最佳解。在最後第四章中提及的問題，就是我們必須依據影片的特性來設定參數，這與我們自動化分群的目的大大相抵觸，但身體的額外資訊非常有使用的價值，不應該完全不使用，因此，如何解決此問題也很值得未來深加探討與研究。



## 參考文獻

- [1]. C. Czirjek, N. O'Connor, S. Marlow, and N. Murphy, "Face Detection and Clustering for Video Indexing Applications," *Proc. Advanced Concepts for Intelligent Vision Systems*, pp.2-5, 2003
- [2]. O. Arandjelović and A. Zisserman, "Automatic Face Recognition for Film character Retrieval in Feature Length Films," *Proc. IEEE Conference on Computer Vision Pattern Recognition*, vol. 1, pp. 860-867, 2005.
- [3]. Y. Gao, T. Wang, J. Li, Y. Du, W. Hu, Y. Zhang, and H. Ai, "Cast Indexing for Videos by NCuts and Page Ranking," *Proc. of the ACM International Conference on Image and Video Retrieval*, pp. 441-447, 2007.
- [4]. J. Barreto, P. Menezes, and J. Dias, "Human-robot Interaction Based on Haar-like Features and Eigenfaces," in *International Conference on Robotics and Automation*, New Orleans, pp. 1888-1893, 2004.
- [5]. M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991.
- [6]. S. Satoh, "Comparative Evaluation of Face Sequence Matching for Content-Based Video Access," *Proc. IEEE Conference on Automatic Face and Gesture Recognition*, pp. 163-168, 2000.
- [7]. S. Foucher and L. Gagnon, "Automatic Detection and Clustering of Actor Faces based on Spectral Clustering Techniques," *Proc. Computer and Robot Vision (CRV)*, pp.113-122, 2007.
- [8]. T.L. Berg, A.C. Berg, J. Edwards, M. Maire, R. White, Y.W. Teh, E.G. Learned-Miller, and D.A. Forsyth, "Names and Faces in the News", *Proc. CVPR* , vol. 2, pp.848-854, 2004.
- [9]. S. Foucher and L. Gagnon, "Automatic Detection and Clustering of Actor Faces based on Spectral Clustering Techniques," *Proc. Computer and Robot Vision (CRV)*, pp.113-122, 2007.
- [10]. J. Yang, D. Zhang, A.F. Frangi, and J. Yang, "Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition," presented at *IEEE Trans. Pattern Anal. Mach. Intell.*, pp.131-137, 2004.
- [11]. T. Ahonen, A. Hadid, and M. Pietikainen, "Face Recognition with Local Binary

- Patterns,” *Proc. European Conference on Computer Vision*, vol. 3021, pp. 469–481, 2004.
- [12].S. Satoh, Y. Nakamura, and T. Kanade, “Name-It: Naming and Detecting Faces in News Videos,” *IEEE Multimedia*, 6, pp. 22-35, 1999.
- [13].A. W. Fitzgibbon and A. Zisserman, “On Affine Invariant Clustering and Automatic Cast Listing in Movies,” *European Conference on Computer Vision (ECCV)*, vol. 3, pp. 304 – 320, Springer-Verlag, 2002.
- [14].M. Everingham and A. Zisserman, “Automated Person Identification in Video,” *Proc. CIVR*, pp.289-298, 2004.
- [15].M. Everingham and A. Zisserman, “Automated Visual Identification of Characters in Situation Comedies,” *Proc. ICPR*, vol. 4, pp.983-986, 2004.
- [16].O. Arandjelović and A. Zisserman, “Automatic Face Recognition for Film character Retrieval in Feature Length Films,” *Proc. IEEE Conference on Computer Vision Pattern Recognition*, pp. 581-588, 2005.
- [17].J. Sivic, M. Everingham, and A. Zisserman, “Person Spotting: Video Shot Retrieval for Face Sets,” *Proc. CIVR*, pp.226-236, 2005.
- [18].Y. Gao, T. Wang, J. Li, Y. Du, W. Hu, Y. Zhang, and H. Ai, “Cast indexing for videos by NCuts and page ranking,” *Proc. CIVR*, pp.441-447, 2007.
- [19].P. Huang, Y. Wang, and M. Shao, “A New Method for Multi-view Face Clustering in Video Sequence,” *Proc. IEEE International Conference on Data Mining Workshops (ICDMW)*, pp.869-873, 2008.
- [20].J. Tao and Y. P. Tan, “Efficient Clustering of Face Sequences with Application to Character-based Movie Browsing,” *Proc. IEEE International Conference on Image Processing*, pp. 1708-1711, 2008.
- [21].S. Gong, S. McKenna, and J. J. Collins, “An Investigation into Face Pose Distributions,” In *FG.*, pp. 265, 1996.
- [22].T. Ji and Y. P. Tan, “Face Clustering in Videos Using Constraint Propagation,” *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp.3246-3249, 2008.
- [23].Z. Liu and Y. Wang, “Major Cast Detection in Video using Both Audio and Visual Information,” In *ICASSP-2001*, pp. 1413-1416, 2001.
- [24].G. Iyengar, H.J. Nock, and C. Neti, “Semantic Indexing of Multimedia Content Using Audio, Text and Visual Cues,” *Proc. Multimedia Information Systems*, pp.134-134, 2002.



- [25].Z. Liu and Y. Wang, “Major Cast Detection in Video Using Both Speaker and Face Information,” *IEEE. Trans. Multimedia*, vol. 9, no.1, pp. 89-101, Jan. 2007.
- [26].M. Everingham, J. Sivic, and A. Zisserman, “Taking the bite out of automated naming of characters in TV video,” presented at *Image Vision Comput.*, pp.545-559, 2009.
- [27].<http://dvdvideosoft.com/download/FreeVideoToJPGConverter.exe>
- [28].<http://opencv.willowgarage.com/wiki/Welcome>
- [29].洪詩祐,“使用臉部訊息輔助自動化膚色偵測,” 交通大學多媒體工程研究所碩士論文, 2009
- [30].S. Satoh, “Towards Actor/Actress Identification in Drama Videos,” *Proc. ACM Multimedia*,pp. 75-78, 1999.
- [31].D. Zhong, H. Zhang, and S. Chang, “Clustering Methods for Video Browsing and Annotation,” *Proc. Storage and Retrieval for Image and Video Databases (SPIE)*, pp.239-246, 1996.
- [32].蘇偉志,“以人臉為依據建立視訊影片中人物出現時間之索引,” 交通大學多媒體工程研究所碩士論文, 2010
- [33].S. Theodoridis and K. Koutroumbas, *Pattern Recognition (4th Edition)*, Academic Press, 2008
- [34].L. Hubert and P. Arabic, “Comparing Partitions,” *Journal of Classification*, vol. 2, no. 1, pp.193-218, 1985