國立交通大學

多媒體工程研究所

碩士論文

用於人臉資訊分析的視訊資料集和視訊相似度之分析

The analysis of video datasets and similarity measures for face information analysis

研 究 生：廖向德

指導教授：王才沛　教授

中 華 民 國　一 百 零 二 年 八 月

用於人臉資訊分析的視訊資料集及視訊相似度之分析

The analysis of video datasets and similarity measures for face information analysis

研 究 生：廖向德　　Student： Xiang-de Liao

指導教授：王才沛　　Advisor：Tsai-pei Wang

國 立 交 通 大 學

多 媒 體 工 程 研 究 所

碩 士 論 文

A Thesis
Submitted to Institute of Multimedia Engineering
College of Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in
Computer Science
Aug 2013
Hsinchu, Taiwan, Republic of China

中華民國一百零二年八月

用於人臉資訊分析的視訊資料集及視訊相似度之分析

學生：廖向德　　指導教授：王才沛

國立交通大學多媒體工程所

摘要

人臉辨識一直是多年來許多人們研究的課題。早期從心理學方面開始研究人類如何辨別不同的人臉，到現在我們試圖找出一個可靠的方法來讓電腦辨識人臉，一直是一個很大的挑戰，且至今為止也沒有一個完美的方法被提出來。大部分人臉辨識的演算法都是以影像為基礎的，但是在很多情況下，我們卻需要應用在一段影片上而不是單一的影像。

比起單一影像，一段影片能夠提供更多的資訊，有利於提升人臉辨識的可靠性。因此本文主要以人臉影像串列之間的相似度為主要研究方向。本文在四種資料集中搭配一些前處理以及不同環境下比較了幾種人臉影像串列相似度計算方法，說明各方法的優劣和幾種可能會影響效能的因素，並且分析各種資料集的特性。

The analysis of video datasets and similarity measures for face information analysis

Student： Xiang-de Liao Advisor：Tsai-pei Wang

Institute of Multimedia Engineering
College of Computer Science
National Chiao Tung University

Abstract

Face recognition has been studied for many years, but it has stayed a challenging problem as no one perfect method has been proposed. Most face recognition algorithms are image-based. However, in many cases, it is useful and beneficial to apply face recognition algorithms to video data rather than single images.

Compared to a single image, a video can provide more information, thus improving the reliability of face recognition. This thesis focuses on facial image sequence similarity as the main research topic. We compare four different datasets under several different environments to analyze algorithm for computing face image sequence similarities. We illustrate the pros and cons of each method and also discuss several factors that may affect the performance. In addition, we also analyze the characteristics of these data sets.

誌謝

本論文的完成，首先我要感謝我的指導教授王才沛老師。感謝老師這兩年來的指導，在我遇到問題與困難時會耐心的教導與幫助，對於我的錯誤也會不厭其煩一步一步的引導，在研究方面給予我一些方向與想法。在此非常感謝老師的指導與栽培。

另外還要感謝實驗室的同學廖耿德、李育任、黃翰賢，尤其是廖耿德，每每在我遇到問題的時候都願意花時間幫助我解決疑難或與我討論。跟你們一起度過這多采多姿的兩年，一起打報告、寫作業、聊天、吃大餐，一起參加由田的比賽。大家遇到問題的時候也能夠互相討論、打氣，讓我的兩年研究所生活中不會感覺到孤單，豐富我的生活。

最後我要感謝我的父母，在這兩年的支持與關心，以及提供經濟上的援助，讓我可以毫無後顧之憂地完成我的學業。

目錄

圖例

表格

第一章 簡介

1.1 研究動機

人臉辨識一直是多年來許多人們研究的課題。早期從心理學方面開始研究人類如何辨別不同的人臉，到現在我們試圖找出一個可靠的方法來讓電腦辨識人臉，但這一直是個很大的挑戰，至今為止也沒有一個完美的方法被提出來。隨著科技日益進步的今日，到處充斥著可錄影的設備如手機、攝影機、街頭巷尾的監視器等，加上網路的發達，使得影片深深地融入我們日常生活中。而人臉辨識則可以應用在這些地方，例如搜尋、索引等，是非常重要的一個領域。

大部分人臉辨識的演算法都是以影像為基礎，但是在很多情況下，我們卻需要應用在一段影片上而不是單一的影像。而且比起單一影像，一段影片能夠提供更多的資訊，例如影片中就對於同一人可能包含人更多角度的影像，有利於提升人臉辨識的效能與可靠性。

兩組影像串列可能個別來自一段影片的片段，或者是某人的照片集等。人臉影像串列的相似度可以應用在許多地方，例如從監視系統所錄下的影片或從即時畫面中辨識裡面出現的人物身分，如此可以有效協助警政單位辦案；或是給定某一特定人物的影像集合(或影像串列)，在大量的影片中尋找該特定人物出現的片段，讓我們可以快速的瀏覽影集或電影中有興趣的部分。

在本文中要研究幾個人臉辨識的核心問題，例如前處理、人臉偵測準確度、視訊畫面的取樣數等對辨識效能的影響，還有什麼樣的資料集特性對現有辨識演算法會是很大的挑戰。本文中在不同類型的四組資料集上用幾種現有的演算法進行實驗，比較並分析獲得的實驗結果。另外本文介紹了一個自行收集並經過人工

檢查和校正人臉範圍資訊 ground truths 的資料集，並詳細描述建立此資料集的過程。

1.2 論文架構

在接下來的章節中，第二章介紹一些人臉偵測、校正、辨識、分群的相關文獻，以及介紹 4 個公開的人臉影片資料集。在第三章中介紹本文的實驗所使用的資料集，其中特別詳細敘述一個由我們自行收集的資料集的過程和特性。在第四章中介紹實驗所使用的方法，包括前處理、數種人臉影像集相似度的計算方法以及人臉分群的方法。第五章是各種實驗的結果與討論。最後第六章則是總結以及未來可以努力的方向。

第二章 文獻探討

2.1 人臉偵測(Face Detection)

人臉偵測是所有臉部分析與處理演算法的基石。人臉偵測的目標是從任意給定的一張影像中，找出是否有一個或是多個人臉存在於其中，並回傳人臉的位置以及範圍。人臉偵測的工作對人類來說不過是一件非常簡單而直覺的事情，但對電腦卻是一件困難的任務與挑戰。近年來較為熱門且常用的人臉偵測方法為 Viola & Jones 所提出的 Viola-Jones face detector[1]。此方法的特點在於速度快，可以實時運行(run in real time)，因此被廣泛的使用。

2.2 人臉校正(Face Alignment)

在進行人臉辨識或是其他臉部分析處理的演算法之前，若有對影像作校正結果會有非常明顯的差距，例如對人臉的朝向、姿勢等進行校正。要對人臉朝向和姿勢進行校正，首先需要有五官甚至更細微的臉部資訊，例如[2]就是簡單的偵測出人臉五官的位置，而[3]除了偵測出五官位置外，還偵測出整個臉部和五官的輪廓等，相對地也需耗費更多的計算時間。利用臉部五官甚至輪廓等方法都有一個問題，就是無法保證偵測出來的五官等資訊的正確性。後來有人提出用紋理擷取的方法來判別臉部的方向，如[4]利用 GWT(Gabor Wavelet Transform)擷取人臉的紋理資訊，再用 PCA(principal components analysis)投影，即可將不同角度的人臉分別出來。[5, 6]中也都有利用 GWT 來加強辨識成功率。

2.3 人臉辨識(Face Recognition)

根據 Zhao 等人在[7]中的統整，人臉辨識的工作可以分成兩大類：第一類是針對靜止影像(still images)，第二則是針對影片(video)或影像集(image set/sequence)。

- 以靜止影像為基礎的人臉辨識(Image-based Face Recognition)

早期人臉辨識大多都是針對靜止影像(still images)的研究。而其方法大致上可以分成以下幾類：

1. 整體匹配方法(Holistic matching methods)

此類方法是直接將整個人臉影像輸入辨識系統進行辨識。有很多人臉辨識方法都利用 PCA(principal-component analysis)發展而來，例如由 Turk 等人在[8]提出著名且常見的方法─Eigenfaces；由 Belhumeur 等人提出利用的 Fisherfaces[9]；FLD(Fisher's Linear Discriminant)/LDA(Linear Discriminant Analysis)[10]；及對區域特性能有較佳表現的 2D-PCA[11]等。

2. 基於特徵的匹配方法(Feature-based (structural) matching methods)

此類方法則是將一些區域性的特徵如眼睛、嘴巴、鼻子等資訊取出，再利用這些資訊進行辨識。例如 HMM(Hidden Markov Model)[12]。

3. 混合方法(Hybrid methods)

此類方法則是融合前兩種方法，就像是人類在辨識時會同時對整個臉部的範圍以及區域的特徵進行比對。

除了以上幾種方法之外，較著名的還有 Gabor Wavelet[5, 6]。以及 Ahonen 等人[13]利用 LBP(Local Binary Patterns)作為描述臉部影像的工具。

- 以影片為基礎的人臉辨識(Video-based Face Recognition)

此類型的方法是針對一群人臉影像(face image set)或是從影片中取出連續的臉部影像串列(face image sequence)進行辨識的工作。從影片中擷取連續的人臉影像牽涉到人臉追蹤(face tracking)，如 Kim 等人就在[14]利用人臉追蹤的方法來提高辨識率。此外有人提出將聲音的訊息也當作辨識的依據之一，如 Bigun 等人提出的 Multi-modal method[15]。

以影片為基礎的人臉辨識方法大致上可以分成以下幾類：

1. 直接擴展靜止影像的方法(direct extension of still-image-based recognition)

從人臉影像集(face image set)或人臉影像串列(face image sequence)中隨機或某些數學方法選出一張到數張代表臉(Representative Face)，然後再用以靜止影像為基礎的方法對代表臉進行人臉辨識，甚至搭配多數決等方法。

2. 建立三維人臉模型(3D Face model)

利用人臉影像集(face image set)或人臉影像串列(face image sequence)建立 3D 模型，進行比對時再以貼圖的方式將人臉影像貼在模擬出來的 3D 模型上。如[16-18]。

3. 影片對影片的相似度(video to video similarity)

這類型較著名的有 MSM(Mutual Subspace Method)[19, 20]、MMD(Manifold-manifold Distance)[21]、由 Cevikalp 等人在[22]中提出的 AHISD(Affine Hull based Image Set Distance)和 CHISD(Convex Hull based Image Set Distance)、SANP(Sparse approximated nearest points)[23]、Dictionary-Based method[24]等。

MSM 是由主成分分析法計算其線性子空間之間的夾角來決定相似性。AHISD 和 CHISD 則是用一個 affine hull 和 convex hull 來代表一個影像集。SANP 則是在兩個影像集各自的影像所形成的點中算出最短距離。

2.4 人臉分群(Face Clustering)

人臉分群主要是利用人臉資訊作為分群的依據。人臉分群和人臉辨識的差別在於，人臉辨識是將某一測試人臉影像或人臉影像集(testing face image or face image set)經過辨識系統後將其認定為某一已知的身分，而人臉分群則是將一群相似的人臉影像或人臉影像集分成同一群。利用分群的資訊，可以將影片中各個人物出現的片段標示出來，方便使用者可以針對想看的人物進行選擇性的瀏覽。

實際進行人臉分群的時候，時常會因為人臉角度、光源等因素而造成人臉影像差異過大，進而造成分群效果不佳。因此有人提出可利用一些額外的資訊來增加分群效能，如在[25]中就利用了影片中的聲音做為額外的資訊；在[26]中則利用演員嘴型的變化來判斷聲音是屬於哪位演員，進而利用這些關係來做分群；在[27, 28]中則使用了身體或衣服的色彩資訊作為輔助。在[29]中則使用姿勢(pose→ out-of-plane rotation)分辨方式將人臉依據不同角度分類，並以同類角度的人臉作為群和群結合的依據，這麼做的原因在於：「相同膚色、不同姿勢」比「不同膚色、相同姿勢」更為相似，所以用相同姿勢的人臉來分群會有較佳的結果。除了人物本身的資訊之外，在[30]中則使用了場景資訊，藉由人臉所在場景的特性以提升分群之效能。

2.5 資料集(Dataset)

一個公開的資料集是非常重要的，因為這可以提供在這領域的眾多研究者一個可以比較研究成果的基準。資料集主要分為兩大類，以影像(image)為基礎和以影片(video)為基礎。由於本文主要著重在影片上，因此以下只介紹幾個較為著名且被廣為使用的影片資料集(video data set)：

2.5.1 Honda/UCSD

這是目前被廣為使用的影像資料集(video dataset)，由 Kuang-Chih Lee 等人[31]所提供。此資料集會在後面的章節 3.1 做詳細說明。

2.5.2 CMU Mobo(motion and body)

此資料集是由 Ralph 等人[32]所提供。這個資料集中共含有 96 段影片，其中有 24 個不同的人在跑步機上進行一些不同的運動。這個資料集原本是為了人類動作姿勢辨識(human pose recognition)而建造的，但是也有許多人臉辨識的相關文獻利用此資料集進行實驗。

2.5.3 YouTube Celebrities Face Tracking and Recognition Dataset

此資料集是由 Kim 等人[14]所提供。這個資料集中共有 1910 段影片片段(video sequences)，其中共有 47 個不同的人。每個人各有 3 段原始的影片，每個影片再進一步的切割成數段片段(video sequences)。這些影片是從 YouTube 下載所得，影片的解析度從240×180到320×240都有。這個資料集一開始是為了臉部追蹤(Face Tracking)而建立的，其中有附上每一段影片片段的起始畫面(frame)

的人臉資訊(包含位置和範圍)，用來當作臉部追蹤的種子。這個資料集的特色在於非常低解析度以及很高的壓縮比(compression ratios)，難度也比前面兩個資料集高出許多。

2.5.4 YouTube Faces Database

此資料集是由 Lior 等人[33]所提供。這個資料集包含了 3425 個從 YouTube 下載而來的影片，其中有 1595 個不同的人。這個資料集和它的基準(benchmarks)是依照 LFW(Labeled Faces in the Wild)資料集[34]的模式來設計和建造的。這個資料集的主要任務是判斷任意的一對影片(a pair of videos)中的人臉是否屬於同一個人。

第三章 實驗用資料集

在此章節中，將介紹本文中的實驗所使用的資料集(Dataset)。本文中總共使用了四個資料集，分別為 Honda/UCSD[31]、Sinica Face Video Dataset、Friends、Raymond。第一個資料集是一個公開且常見的資料集，剩下三個則為自行蒐集。Honda/UCSD 的資料量是四個資料集中資料量最少的一個，適合用來在較短時間內測試演算法的正確性。而 Sinica Face Video Dataset 的資料量則是四個資料集中資料量最多的一個，而 Friends 和 Raymond 的資料量則差不多。這四組資料集中只有 Honda/UCSD 的背景較單純且不會有光影變化，其餘皆有複雜的背景與光影變化。

由於本文主要著重於影片間相似性的演算法，因此這四組資料集皆為影片，而非較常見的以影像為基礎的資料集。Honda/UCSD 和 Sinica Face Video Dataset 的資料型態較為接近，資料集都是由數個各別獨立的影片所組成，以一個影片當作一筆資料，也就是以影片當作單位，每一個影片中都只含有單獨的一人，且資料集中的總人數較多，比較適合用在人臉辨識上。而 Friends 和 Raymond 這兩組資料集，則都只有單一的影片，影片中含有不多的主要人物，然後以一串在時間上連續的人臉串列作為單位，因此整個資料集中有非常大量的資料筆數(很多人臉串列)，但是總人數卻不多，因此比較適合用在人臉分群上。本章將分成兩節，第一節中介紹適合人臉辨識的 Honda/UCSD 和 Sinica Face Video Dataset，第二節將介紹適合人臉分群的 Friends 和 Raymond。

3.1 人臉辨識用資料集

3.1.1 Honda/UCSD

這個資料集是由 Kuang-Chih Lee 等人[31]所提供。此資料集中共有 59 個影片(video)，其中包含了 20 個不同的人。影片的解析度為 640×480。每一個影片(video)中都只有單一的人物，並且包含了一些臉部表情(facial expression)、朝向(orientation: in-plane rotation)和姿勢(pose: out-of-plane rotation)的變化。此資料集的拍攝環境是在背景單純且幾乎相同的室內，沒有光照的變化。在原始的標準協議(standard protocol)中用 20 個影片(每個人一個)做為訓練，剩餘 39 個影片則為測試用。圖 3-1 為 Honda/UCSD 的數張範例截圖。

圖 3-1: Honda/UCSD 的範例截圖。

3.1.2 Sinica Face Video Dataset

這個資料集中的影片都是從網路上下載而來，大部分是來自 YouTube 和 Flickr，由[35]所蒐集。此資料集的每一部影片中都只有一人，共 100 個人，包含

40 個小孩，30 個成年男人和 30 個成年女人，每個人各有 3 部影片，共 300 個影片。在本文的實驗中除了年齡辨識外，皆只有使用成人的影片。此種分布的特性使得這個資料集可以用在性別和年齡辨識(辨識為成人或小孩，可應用在嬰幼兒戒護系統上)。影片的解析度為320 × 240，畫面更新率為每秒 30 張影像(frame rate = 30 fps)。這個資料集裡面有複雜的背景、光影變化，不同的臉部表情、姿勢和朝向等，有時臉部還可能被衣物或手等物體遮住。圖 3-2 為 Sinica Face Video Dataset 的數張範例截圖。

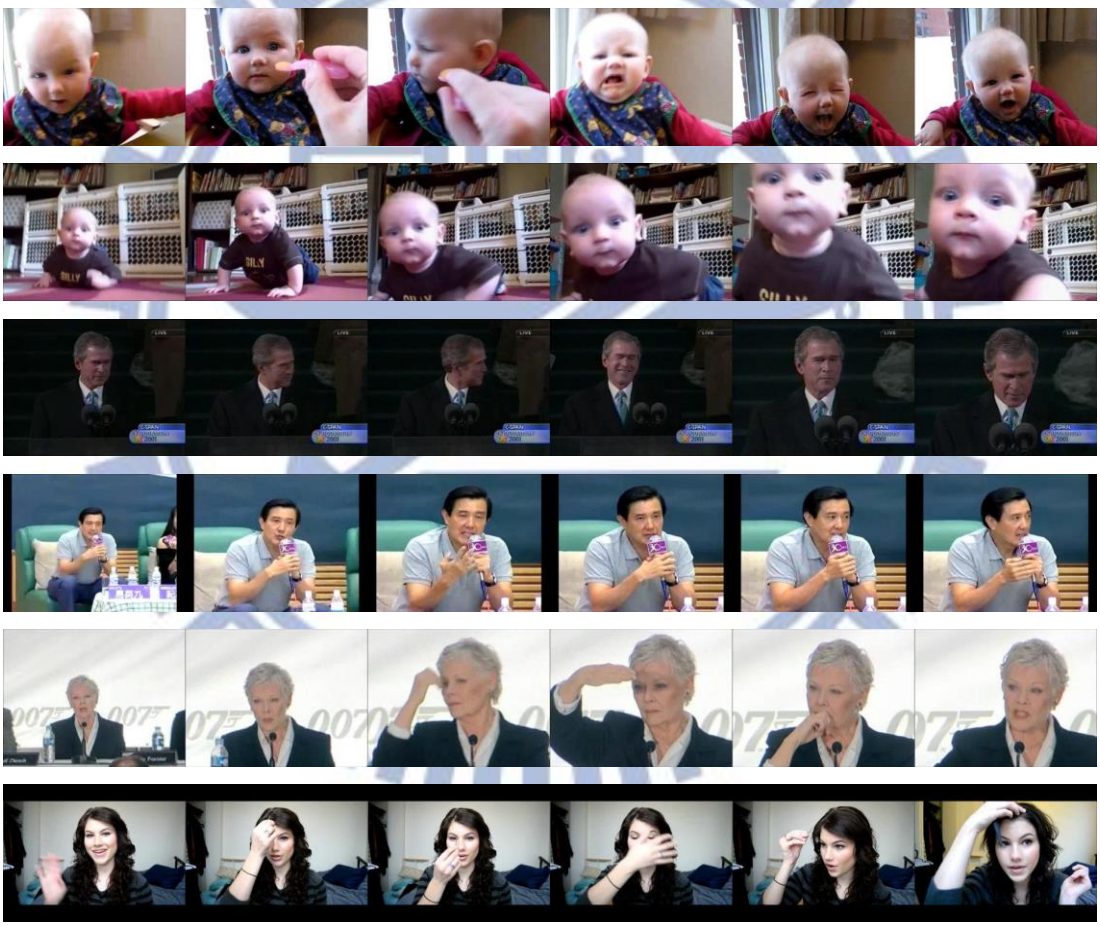

圖 3-2: Sinica Face Video Dataset 的範例截圖。

有些原始影片中的某些片段可能完全沒有人物出現或是含有除了目標以外的人物。為了移除這些問題，我們用人工手動的方式從每部影片中選出 3 至 5 個片段(segment)，使得每一個片段中都只會包含目標人物。每一個選出的片段

(segment)的長度約在 10 到 30 秒之間。整個資料集中共有 1236 個影片片段(video segment)。

人臉偵測和追蹤是更進一步的人臉辨識和其他人臉分類任務的必要先決條件。然而沒有一個現存的公開影片資料集(video dataset)中有提供完整的人臉資訊。而本資料集提供了所有視訊畫面(frame)的人臉偵測的 ground truths，也就是人臉在視訊畫面中的位置和範圍資訊。因此我們的資料集使得以下的研究成為可能：

- 使人臉辨識或其他類型的高階處理可以獨立於人臉偵測和追蹤的結果來進行最佳化。
- 我們可以比較用自動偵測和手動校正的人臉集合(face set)的辨識效能，並得知辨識演算法在不正確或不準確的人臉偵測下是否依然強健(robust)。
- 可以使用 ground truths 來評估以影片為基礎(video-based)的人臉偵測和追蹤演算法是否準確。

接下來將描述如何產生人臉偵測的 ground truths 的過程。首先，我們使用 Viola-Jones face detector[1]取出大部分的人臉，每一個視訊畫面至多一個臉。這邊要注意的是，雖然每一個視訊畫面都一定包含一個人物，但是當有過大的姿勢、角度改變或遮擋(occlusion)時，有可能會偵測不到人臉。

在資料集中的影片包含了成千上萬的視訊畫面，為了加速人工檢查和校正自動偵測的人臉的過程，我們開發了一個標記用程式來解決這個問題，圖 3-3 為標記用程式的銀幕截圖。這個標記用程式讓使用者可以在不同的在不同的畫面更新率(frame rates)下檢查影片片段(video segment)的視訊畫面。如果使用者覺得有必要校正一些自動化偵測的人臉結果，可以依照以下程序進行：

- 首先使用者選擇一個視訊畫面，在此稱為關鍵視訊畫面(key frame)，並且在此視訊畫面中手動調整人臉範圍。
- 接著使用者可以在關鍵視訊畫面(key frame)之前或之後的任意視訊畫面中選擇另一個視訊畫面，此視訊畫面以及關鍵視訊畫面(key frame)之間的所有視訊畫面為使用者所要更正的範圍。
- 由關鍵視訊畫面(key frame)開始，在選取的範圍內使用 Template-match based face tracking 來更正人臉範圍。這邊程式只使用了簡單的 tracker，在 5%縮放範圍內尋找差距最小的部分作為新的人臉範圍，差異是由前後兩張影像的像素值(pixel values)的歐氏距離(Euclidean norm)來決定。
- 最後使用者決定是否接受更正後的結果。

上述的方法可以讓數十甚至數百個人臉偵測結果在短短幾步驟內就更正完成。以上的程序可以反覆執行直到使用者滿意影片片段中校正後的人臉範圍。

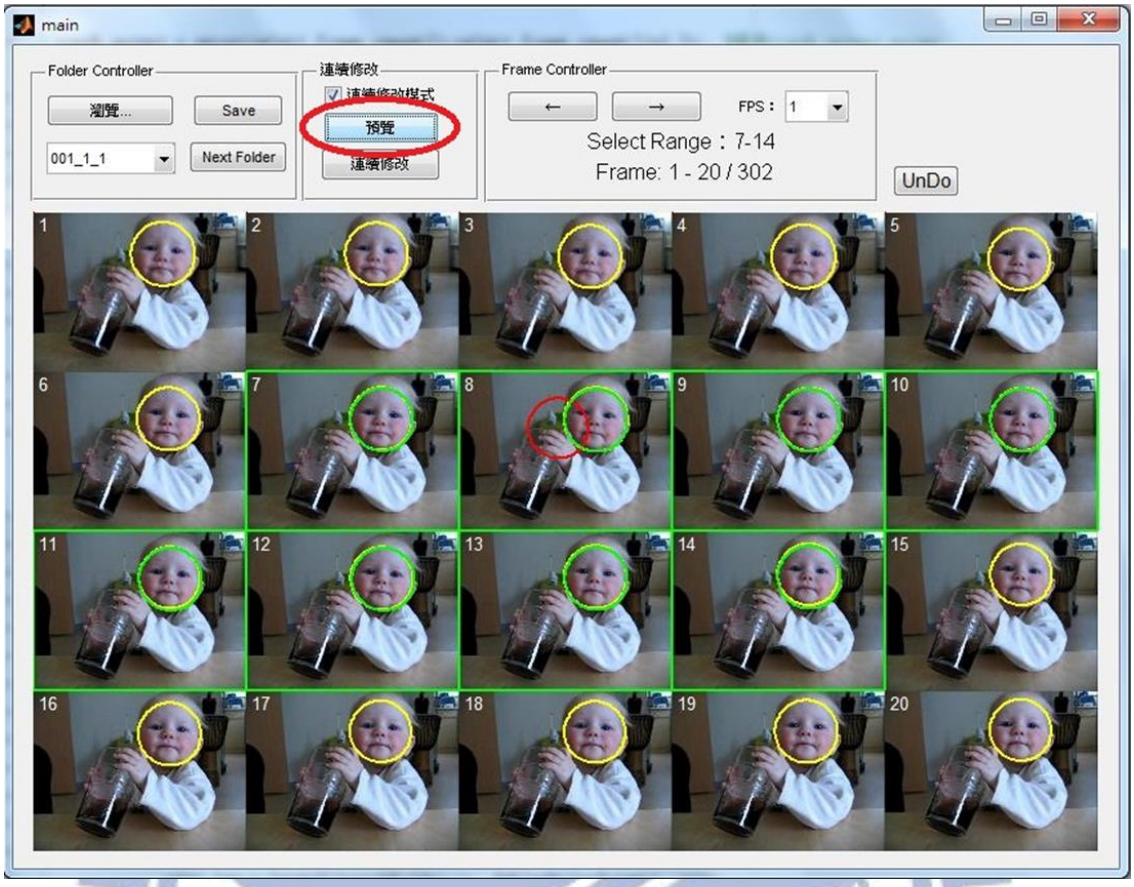

圖 3-3: 標記用程式的銀幕截圖。黃色的圓圈代表自動化偵測的人臉範圍，綠色的方框代表要進行校正的範圍，綠色圓圈代表校正後的估計人臉區域。紅色圓圈為人工選取或是經過追蹤校正後的人臉範圍。

3.2 人臉分群用資料集

3.2.1 Friends

此資料集只有單一的一個影片，這個影片是美國影集“Friends”中的某一集。主要腳色有 3 男 3 女共六人。影片的解析度為720 × 480，畫面更新率為每秒 5 張影像(frame rate = 5 fps)。經過人臉偵測和追蹤處理程序後共得到 529 個演員串列(actors sequence) (擷取演員串列的方法見[36])，串列中含有數個到數十個不等的人臉影像，每個人臉影像大小為40 × 40。由於是利用程式自動化產生演員串

列，因此有些串列中的人臉影像是錯誤的。圖 3-4 為 Friends 的數張範例截圖，圖 3-5 為 Friends 的數個演員串列範例截圖。

圖 3-4: Friends 的範例截圖。

圖 3-5: Friends 的演員串列範例截圖。

3.2.2 Raymond

此資料集和 Friends 一樣只有單一的一個影片，是美國影集“Everybody Loves Raymond”中的某一集。主要腳色有三個小孩、一對中年夫婦、一對老年夫婦、還有一位警官和一位朋友。三個小孩中有兩個是非常相似的雙胞胎，因此在進行人臉辨識或分群的實驗時我們將其視為同一人。影片的解析度為 640 × 480，畫面更新率為每秒 5 張影像(frame rate = 5 fps)。經過人臉偵測和追蹤處理程序(方法同資料集「Friends」)後共得到 463 個演員串列(actors sequence)，串列中含有數個到數十個不等的人臉影像，每個人臉影像大小為40 × 40。由於是利用程式自動化產生演員串列，因此有些串列中的人臉影像是錯誤的。圖 3-6 為 Raymond 的數張範例截圖，圖 3-7 為 Raymond 的演員串列範例截圖。

圖 3-6: Raymond 的範例截圖。

圖 3-7: Raymond 的演員串列範例截圖。

第四章 實驗方法

4.1 臉部影像前處理

本文中人臉辨識和分群的實驗測試了數種前處理對實驗結果的影響，在第五章的實驗中會個別註明使用哪種前處理。共通的前處理步驟依序如下：

1. 將彩色影像轉為灰階影像。
2. 將影像大小調整到40×40。

除了以上兩個固定最先執行的前處理步驟之外，以下步驟則是視不同實驗來搭配執行：

- Histogram Equalization

  此步驟是為了對影像的亮度進行校正，讓所有的影像有一個統一的標準。本論文中大多使用這種方式，主要原因是為了與[22]文中的實驗設定相同

- Gaussian band pass filtering ($\sigma_{low} = 10,\ \sigma_{high} = 20$)

  利用 Gaussian band pass filter 頻率濾波器來過濾掉影像中低頻及高頻的部分，低頻的部分可能包含了背景或平滑的表面，而高頻的部分則可能包含了雜訊。實作的方法是利用兩組反向的二維 Gaussian 函式(如式(1)、(2))組合成頻帶濾波器。

$$f_{low}(x,y) = exp\left(-\frac{(x-c_x)^2+(y-c_y)^2}{2{\sigma_{low}}^2}\right) \quad (1)$$

$$f_{high}(x,y) = 1 - exp\left(-\frac{(x-c_x)^2+(y-c_y)^2}{2{\sigma_{high}}^2}\right) \quad (2)$$

在執行頻帶濾波的時候，σ_{low}的值愈大，則被過濾掉的低頻範圍愈大；σ_{high}的值愈大，則表示被過濾掉的高頻範圍愈小。σ_{low}和σ_{high}的值是依據[36]中的設定。

4.2 人臉影像集之相似度計算與識別

在此節中，將依序介紹數種計算人臉影像集(或人臉影像串列)相似度的方法。相似度的評估在本文中是計算兩影像集的距離，距離越短代表相似度越高。而這些方法的評估方式有兩種：一種是計算人臉辨識的辨識率，另一種則是用 ROC curve。

人臉辨識就是進行分類的動作。首先要在一個資料集(data set or data base)中將某些人臉影像集分為訓練資料(training data)，剩下的則是當作測試資料(testing data)。每一個測試影像集(testing image set)都和所有訓練影像集(training image set)各別算出距離，並用 Nearest Neighbor Classifier 將該測試影像集歸類到相異程度最小(距離最短)的訓練影像集，換言之就是將該測試影像集的身分定為最相似的訓練影像集的身分。最後統計所有測試影像集的分類結果正確性來計算辨識率。

ROC curve 的部分，則是以人臉影像集(face image set)為單位，給定兩個人臉影像集形成一組配對(pair)，驗證這兩個影像集是否為同一人。同一人則是陽性(positive)，否則為陰性(negative)。設定某一距離為閥值(threshold)，若該組配對的距離低於閥值，則視該組配對為同一人，否則為不同人。每一個閥值可以算出一個 TPR(True Positive Rate)和 FPR(False Positive Rate)，計算方法如式(3、4)，每次以不同的距離當作閥值可以得到一組 TPR 和 FPR，最後用所有得到的 TPR 和 FPR 畫出 ROC curve(橫坐標是 FPR，縱座標是 TPR)。ROC curve 越靠近左上角代表效果越好。若是一條從左下原點至右上且斜率為 1 的直線則是隨機猜測的結果。

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{TP+FN} \tag{3}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \tag{4}$$

TP：程式結果認定是相同身分而實際上也是相同身分的配對數量
FN：程式結果認定是不同身分而實際上是相同身分的配對數量
FP：程式結果認定是相同身分而實際上是不同身分的配對數量
TN：程式結果認定是不同身分而實際上也是不同身分的配對數量

我們接下來在以下各小節中介紹實驗中用到的相似度計算與人臉辨識的方式。

4.2.1 LAHISD

LAHISD (Linear Affine Hull based Image Set Distance)用一個 affine hull 來代表一個影像集(image set)或影像串列(image sequence)，而 image set 中的影像(images)則是作為 affine hull 的特徵向量(feature vectors)。而影像集之間的相異程度(dissimilarity)則是由 affine hull 之間的幾何距離來決定。本文中的實驗是直接使用由[22]的作者所提供的程式碼。

4.2.2 MSM

MSM(Mutual Subspace Method)是由 Fukui 等人[19, 20]所提出。兩個圖像集之間的相似性是由主成分分析法計算其線性子空間之間的夾角。本文中的實驗是直接使用由[22]的作者所提供的程式碼。

4.2.3 PCA+voting

1. 將人臉影像存成行向量(column vector) → 人臉向量。
2. 利用 PCA(principal-component analysis)將所有人臉向量轉換低維度的特徵向量(feature vector)。在第五章的實驗中將用「d_{PCA}」代表降維的目標維度。

3. 用 k-means clustering 對個別 training video(or image set)中的所有人臉向量做分群，並以群中心當作該代表臉向量(representative face vector)。因此每個人臉影像集各有 k 個代表臉向量。在第五章的實驗中將用「n_{key}」來表示代表臉向量的數量。此目的主要是為了選出較有代表性的人臉來代表此人物並且降低後續步驟的運算量。
4. 對某一 testing video(or image set)中的所有人臉向量，個別找出最相似(即歐氏距離最小)的 training video 中的代表臉向量，並以多數決方式決定此 testing video 的身分。

   p.s. 此方法只適用於人臉辨識，人臉分群不適用。因為人臉分群的時候需要有明確的兩群之間的距離，而此方法只是用多數決決定身分，並沒有一個明確的距離值，需要用在分群上時需改用 4.2.4 的方法。

4.2.4 PCA+AvgDist

1. 利用 PCA(principal-component analysis)將所有人臉向量轉換成維度較低的特徵向量(feature vector)。
2. 算出兩影像集(image set)所有特徵向量(feature vector)之間的歐氏距離。若影像集 A 和 B 各有 a 和 b 張人臉影像，則共有$(a \times b)$組距離。
3. 將步驟 2 的所有距離的平均值，即為兩兩影像集(image set)之間的相異程度(dissimilarity)。

4.2.5 2DPCA+AvgDist

利用 2DPCA(2 dimensional principal-component analysis)將所有人臉向量轉換成維度較低的特徵向量(feature vector)，本文中實驗取 10 個 basic component。

1. 算出兩影像集(image set)所有特徵向量(feature vector)之間的歐氏距離。若影像集 A 和 B 各有 a 和 b 張人臉影像，則共有$(a \times b)$組距離。
2. 將步驟 2 的所有距離的平均值，即為兩兩影像集(image set)之間的相異程度(dissimilarity)。

4.2.6 SANP

SANP (Sparse Approximated Nearest Points distance)計算兩個影像集相異程度(dissimilarity)的方法是根據兩個影像集中最接近點的距離決定，而最近點(nearest points)是從個別影像集合的影像樣本(image samples)稀疏近似(sparsely approximated)而得。本文中的實驗是直接使用由[37]的作者提供的 SANP 程式碼。

4.3 性別與年齡辨識

4.3.1 性別辨識

這裡的性別辨識方法和[38]中的相同。用 LBP 來提取人臉特徵，再用 Adaboost 分類器進行性別辨識。從網路上另外獨立蒐集的 800 張人臉影像(男女各 400)作為訓練資料。在本文實驗中，使用「Viola-Jones Face Detector」進行人臉偵測。前處理則是將人臉影像縮放到 80x80 的大小並且轉成灰階影像，影像強

度則用固定的平均值和標準差進行正規化。接著是進行人臉校準，最後在裁減取得中間 50x50 部分來排除背景。人臉校準的步驟是經由旋轉人臉影像使得雙眼的眼角四個標記點(landmarks)所形成的直線呈現水平。臉部特徵標記的方法是使用[2]中的方法。

4.3.2 年齡辨識

年齡辨識的研究和應用在近年來獲得越來越多的關注。然而，從網路上下載的影片中難以獲得準確的年齡 ground truths。因此這裡專注在更有限的問題上，也就是分別成人和小孩，而這依然有相當多的應用。

本文用 2DPCA[11]特徵來代表影像(取 10 個 basic component)，並且用 k-nearest-neighbor 分類器進行年齡辨識(k=3)。這裡使用一組獨立從網路上蒐集的 500 張人臉影像集作為訓練資料，其中有 200 個小孩和 300 個成人。這個小孩/成人的比例和「Sinica Face Video Dataset」中 40 個小孩和 60 個小孩相同。人臉影像大小為50×50。使用的前處理和性別辨識實驗相同。

4.4 人臉分群

人臉分群和人臉辨識最大不同之處在於人臉分群不會將資料集(data set or data base)分為訓練用和測試用。首先使用章節 4.2 中的任意方法計算出資料及中包含的所有影像集之間的距離，然後再以這些距離用來分群。

本文中採用的分群法是 agglomerative hierarchical clustering。此方法在一開始將每一筆資料視為一群，在每一回合中，找出最相似(也就是距離最小)的兩群C_i及C_j進行合併，反覆進行合併的動作直到群的數量降到所要求的數目為止。另外每

次兩群合併後要將新群與剩下的全部群之間的距離做更新，而更新的方式有以下幾種：

- Single-link:

$$\mathrm{d}(C_q, C_s) = min\{d(C_i, C_s), d(C_j, C_s)\} \quad (5)$$

- Complete-link:

$$\mathrm{d}(C_q, C_s) = max\{d(C_i, C_s), d(C_j, C_s)\} \quad (6)$$

- Average-link:

$$\mathrm{d}(C_q, C_s) = \frac{n_i}{n_i+n_j} d(C_i, C_s) + \frac{n_i}{n_i+n_j} d(C_j, C_s) \quad (7)$$

n_i 及 n_j 分別為群 C_i 及 C_j 內元素個數，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。

- Fusion:

合併後的新群，將其中的所有影像合起來視為一個新的影像集，再用人臉影像集相似度計算方法與其它群重新算距離。

第五章 實驗結果

5.1 人臉影像集相似度計算方法比較

在此小節的實驗中，主要是測試並比較各種人臉串列相似度計算方法的效能優劣，將依序在四個資料集「Honda/UCSD」、「Sinica Face Video Dataset」(只取成人部分)、「Friends」、「Raymond」上分別進行實驗，在這四個實驗中人臉偵測的方法都是使用 Viola-Jones Face Detector。對人臉影像的前處理方法是使用 Histogram Equalization。由於「Honda/UCSD」和「Sinica Face Video Dataset」的每一個影像集的影像數量較多，為了節省計算時間，對每一個影像集各隨機取樣 50 張人臉影像。

在本小節的實驗中都沒有使用「PCA+voting」這個方法，因為畫 ROC 曲線的時候需要有明確的兩影像集之間距離，而此方法只是用多數決決定身分，並沒有一個明確的距離值，因此改用「PCA+AvgDist」。另外在這四組資料集中只有「Honda/UCSD」裡面有使用「SANP」這個人臉影像集相似度的計算方法，因為在實驗的過程中發現「SANP」在計算影像集之間的距離時難以達到收斂，幾乎都是達到限制次數之後才停止，因此造成時間消耗過久，是其他方法的數百甚至數千倍的運算時間，所以只有在「Honda/UCSD」這個最小的資料集中進行試驗。

從圖 5-1 中會發現在「Honda/UCSD」這個資料集中每種方法皆有非常好的表現，尤其以「SANP」的辨識效果最好，但是和「LAHISD」的效果比起來並沒有明顯好太多。在資料集「Sinica Face Video Dataset」中「LAHISD」的效果是最好的，但是四種方法的效能差異並不明顯，在資料集「Friends」中也是同樣

的情況。在資料集「Raymond」中是「LAHISD」和「PCA+AvgDist」的效果比較好「MSM」的效果則明顯落後許多。整體來看，排除「SANP」不談，「LAHISD」在四個資料集上皆有較佳的表現，是所有方法中效果最好的，而剩餘的三個方法則互有勝負，但是「MSM」的效果似乎較為落後一點。

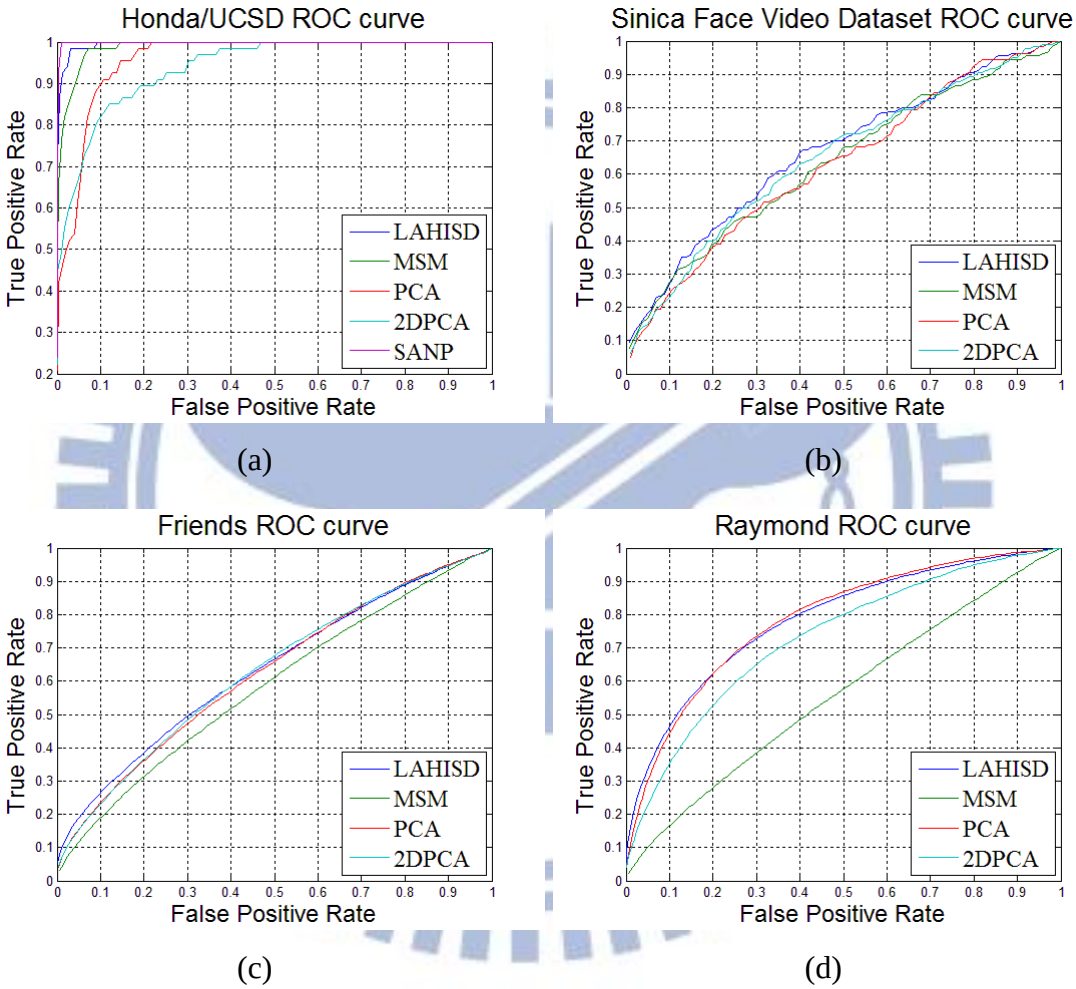

圖 5-1: 數種人臉影像集相似度方法的 ROC 曲線。(a)資料集「Honda/UCSD」的 ROC 曲線。(b) 資料集「Sinica Face Video Dataset」的 ROC 曲線。(c) 資料集「Friends」的 ROC 曲線。(d) 資料集「Raymond」的 ROC 曲線。

綜合圖 5-1 的這些實驗可以發現資料集的難度由簡單到困難依序是「Honda/UCSD」→「Raymond」→「Friends」→「Sinica Face Video Dataset」。在「Honda/UCSD」中每種方法都有趨近於完美的結果，因此在此資料集中不容

易分別出方法的好壞。在「Sinica Face Video Dataset」中則是難度過高，所有方法的效果都不理想，在「Friends」中也有類似的情形。從「Raymond」這個資料集的結果來看，各種方法之間有比較明顯的差距，因此這個資料集比較適合用來比較不同方法的優劣。

接下來的實驗要比較使用時間資訊是否會對效能有所影響。時間資訊的使用方式為，當兩個人臉影像集(或人臉影像串列)之間在時間上有重疊的部分，則將這兩個人臉影像集的距離設為無限遠(即相似度為 0)。實驗結果見圖 5-2。

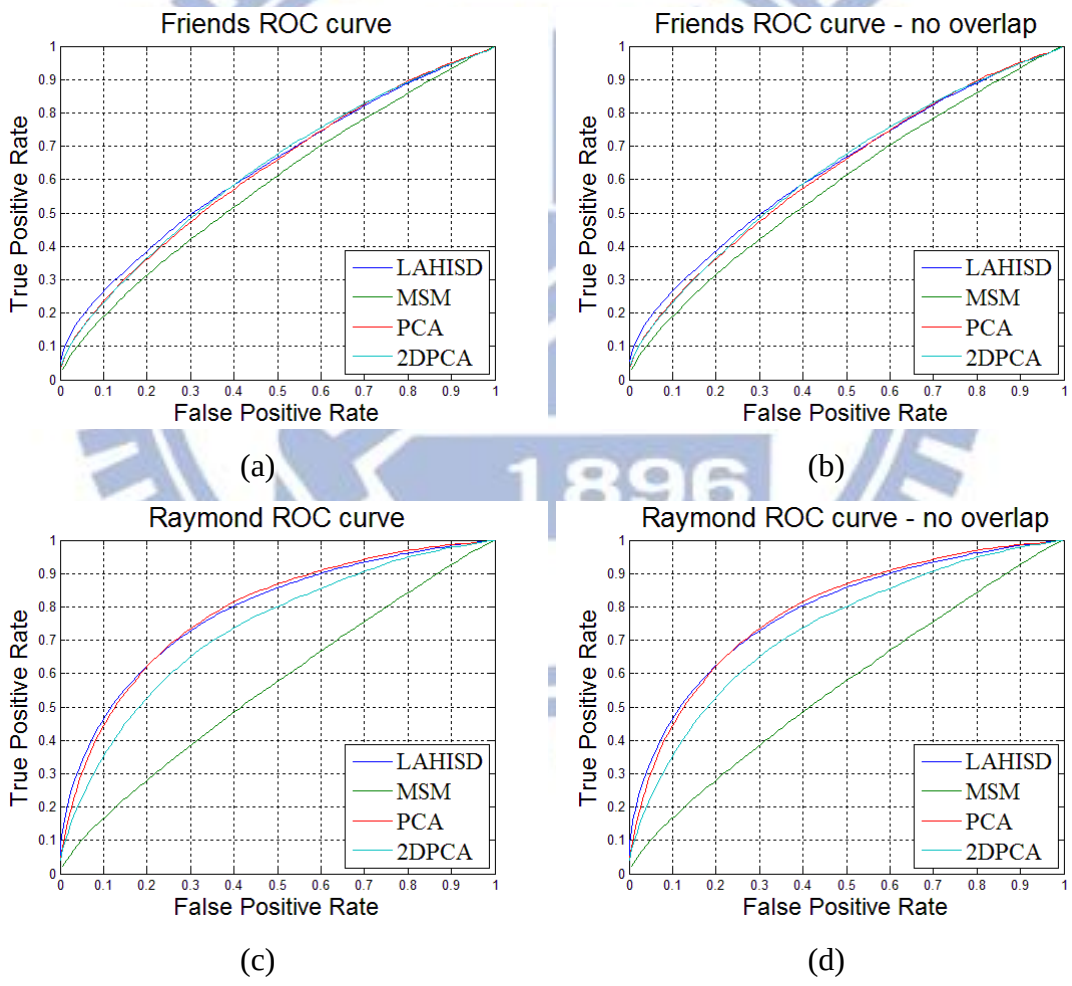

圖 5-2: 使用時間資訊的 ROC 曲線比較圖。(a)資料集「Friends」無使用時間資訊。(b)資料集「Friends」有使用時間資訊排除重疊。(c)資料集「Raymond」無使用時間資訊。(d)資料集「Raymond」有使用時間資訊排除重疊。

比較圖 5-2 的(a)和(b)小圖，可以發現不論有無使用時間資訊，在資料集「Friends」上看不出明顯的差別。比較圖 5-2 的(c)和(d)小圖，可以發現在資料集「Raymond」上也有相同的狀況。推測會產生這種結果是因為影片中有重疊的人臉串列所佔有的比例相當低，所以影響不大。雖然對辨識效果幾乎沒有貢獻，但是利用時間資訊仍有一些好處，第一就是確保不是同一人的影像串列不會被誤認，第二則是可以節省計算時間，因為直接將這些重疊的部分排除，就不需要再花時間計算重疊的人臉影像串列之間的距離。

5.2 人臉辨識

5.2.1 人臉影像集相似度計算方法比較

此實驗的目的是要比較不同的人臉影像集之相似度計算方法的效能。分別在兩個資料集「Honda/UCSD」和「Sinica Face Video Dataset」(只取成人部分)上進行測試。此實驗的人臉偵測使用「Viola-Jones Face Detector」來取得人臉影像，再使用 Histogram Equalization 做前處理。在「Sinica Face Video Dataset」上的實驗為了節省計算時間，對影像集進行隨機取樣，每一個影像集各取樣 100 張人臉影像。

資料集「Honda/UCSD」原本就有區分訓練用資料和測試用資料，因此直接使用。資料集「Sinica Face Video Dataset」則沒有進行區分，因此這裡使用 three-fold cross-validation 進行辨識。本實驗分割「Sinica Face Video Dataset」的資料作為訓練用和測試用有兩種方式：Separated Video 和 Mixed Video。Separated Video 的作法是，從每個人的三部影片中選擇其中兩部作為訓練用，剩下一個是測試用。Mixed Video 的做法是，訓練用資料包括了每個影片中三分之二的影像，剩下三

分之一則是測試用。Separated Video 和 Mixed Video 的取法都各有三種組合，每種組合皆實驗一次在平均其結果。

從表格 5-1 中可以發現在「Honda/UCSD」這個資料集下這四種方法的辨識率幾乎沒有實質上的差距。而在「Sinica Face Video Dataset」上用 Separated Video 分割法的效能差距也不是很明顯，辨識率大約在 20%。由好到壞大致上是「LAHISD」→「MSM」→「PCA + voting」→「SANP」，此結果與章節 5.1 的結果也大致上符合。而「Sinica Face Video Dataset」上用 Mixed Video 分割法的效能全部都是 100%，看不出差異。

表格 5-1: 數種人臉影像集之相似度計算方法的辨識率。

Image set distance method	Honda/UCSD	Sinica Face Video Dataset	
		Separated Video	Mixed Video
LAHISD	0.9744	0.2006	1.0
MSM	1.0	0.1944	1.0
PCA + voting (d_{PCA}=50, n_{key}=5)	1.0	0.1889	1.0
SANP	1.0	0.1611	1.0

在「Sinica Face Video Dataset」上的兩種分割方法的辨識率差距非常大，這種差距說明了在進行影片的人臉辨識效能評估時，訓練用和測試用資料不該同時包含來自相同影片的片段或是非常類似的原始影片，這也可能是在資料集「Honda/UCSD」上的結果都很完美的原因。做為比較，由[14, 37]在資料集「Youtube Celebrities Face」上所進行的實驗和這裡的 Mixed Video 分割法非常類似，他們的實驗所得到的辨識率大約在 50%到 60%附近，但是他們並沒有考慮到這兩種不同分割方法的影響，他們若改成使用 Separated Video 的方法，他們的實驗數據應該會大幅度的降低。

另外在使用「SANP」進行實驗的過程中發現，幾乎每次計算距離時都難以收斂，總是達到運算次數的上限後才停止，因此耗費相當多的計算時間。雖然四種方法的辨識率差距不大，但是在使用的時間方面，「SANP」很明顯比其他方法要耗費多非常多的時間(「SANP」需要上萬秒，其餘在百秒內)，因此「SANP」並非一個理想的方法而且不適合應用在 real time 的系統上。

5.2.2 隨機取樣的影響

在一個影片中，通常都是由許多的視訊畫面所構成，但是通常在相近時間的視訊畫面並不會有很大的不同，為了降低程式運算時間，只取其中一部分的視訊畫面進行運算是非常必要的動作。而本實驗的目的是要對資料集中的每一組人臉影像集各別隨機取樣(random sampling)數張影像(在取樣過程中不破壞人臉影像的順序)，觀察不同的取樣數量對辨識率的影響，以找出對辨識率的影響在合理範圍內的最低取樣數量，換句話說就是最少需要多少張人臉影像才有足夠的代表性來代表一個個體。

為了避免出現極端的狀況，每種取樣數量皆實驗 10 次並平均其結果。本實驗將分別在資料集「Honda/UCSD」和「Sinica Face Video Dataset」(只取成人部分)上進行測試，使用的人臉偵測方法皆為「Viola-Jones Face Detector」，並且用 Histogram Equalization 對人臉影像進行前處理。在「Sinica Face Video Dataset」的部分，使用 three-fold cross-validation，也就是每人各自從 3 段 video 中選一個當作 test data，剩餘兩者為 training data，每種取法各做一次在計算其平均辨識率和計算時間，如同章節 5.2.1 中的 Separated Video 方法。

表格 5-2 中的取樣數從 72 開始是因為在資料集「Honda/UCSD」中，擁有最少人臉影像的影像集其影像數量是 72。從表格 5-2 的數據可以發現當取樣數低於 10 的時候，辨識率才會有比較明顯的降低，這個趨勢也證明了使用影片進行辨識比單獨使用影像會有更好的準確度。

表格 5-2: 隨機取樣在「Honda/UCSD」上對辨識率的影響。

Image set distance method	All	72	50	20	10	5	3
LAHISD	0.9744	1.0	1.0	0.9974	0.9846	0.90	0.8436
MSM	1.0	1.0	0.9974	0.9974	0.9795	0.9231	0.7846
SANP	1.0	1.0	1.0	0.9974	0.9795	0.9205	0.7436

表格 5-3: 隨機取樣在「Sinica Face Video Dataset」上對辨識率的影響。

Image set distance method	取樣數	All	100	50	20	10	5	3
L AHISD	辨識率	0.094	0.201	0.229	0.235	0.224	0.196	0.191
	Time(sec)	5157	133	62	17	6	2	1
MSM	辨識率	0.117	0.194	0.221	0.223	0.207	0.212	0.186
	Time(sec)	3692	3	1	<1	<1	<1	<1
PCA+voting (d_{PCA}=50, n_{key} =20)	辨識率		0.220	0.216	0.213			
	Time(sec)		74	56	46			
SANP	辨識率		0.161					
	Time(sec)		11768					

在表格 5-3 中的空白之處代表沒有進行該實驗。從表格 5-3 的實驗數據可以發現「LAHSID」和「MSM」在隨機取樣數從全取到取 20 這之間都是當取樣數量下降時，辨識率反而提升，尤其是全取和取樣數 100 之間的辨識率差距非常大，全取耗費的計算時間也多出非常多。「LAHSID」和「MSM」在隨機取樣數低於 20 辨識率才開始下降。「PCA+voting」的三種取樣數量在辨識率方面則比較沒有

明顯變化。「SANP」在辨識率方面沒有特別突出的表現，而且耗費的計算時間過長。在表格 5-3 的實驗中有一個比較特別的地方，就是沒有進行隨機取樣的辨識率反而低於有進行隨機取樣的辨識率，這個現象目前還沒發現是什麼因素造成。

比較表格 5-2 和 5-3 可以發現兩者的辨識率差距相當大，這表示「Sinica Face Video Dataset」的難度遠遠高於「Honda/UCSD」。

5.2.3 前處理的影響

此實驗的目的主要是測試三種不同的人臉影像前處理對辨識率的影響。使用的資料集是「Sinica Face Video Dataset」(只取成人部分)，人臉影像的取得是使用人工手動取得。三種前處理都包含了固定會執行的部分：將彩色影像轉為灰階影像，並且將影像大小調整到 40×40。然後三種前處理分別依序是：

Preprocess A: 「Histogram Equalization」

Preprocess B: 「Gaussian band pass filtering ($\sigma_{low} = 10,\ \sigma_{high} = 20$)」

Preprocess C: 「Histogram Equalization」

「Gaussian band pass filtering ($\sigma_{low} = 10,\ \sigma_{high} = 20$)」

此實驗使用 three-fold cross-validation。也就是每人各自從 3 段 video 中選一個當作 test data，剩餘兩者為 training data，每種取法各做一次在計算其平均結果，如同章節 5.2.1 實驗的 Separated Video 方法。此實驗中皆對每一個 image set 隨機取樣 50 張人臉。

從表格 5-4 中可以看出來當隨機取樣數在 20 的時候，三種前處理方法差異不大，在隨機取樣數為 50 和 100 時，Gaussian band pass filtering 對辨識率的提升有較大的貢獻，而兩種方法的混合和只用 Gaussian band pass filtering 的差異較不

明顯。本論文中大多使用 Preprocess A 這種方式，主要原因是為了與[22]文中的實驗設定相同。

表格 5-4: 不同前處理的辨識率。

Image set distance method	Sample Number	Preprocess A	Preprocess B	Preprocess C
LAHISD	20	0.2361	0.2289	0.2428
	50	0.1922	0.2433	0.2317
	100	0.1650	0.2172	0.2256
MSM	20	0.2267	0.2206	0.2422
	50	0.1917	0.2250	0.2289
	100	0.1717	0.2044	0.2150

5.2.4 人臉偵測準確度對人臉辨識的影響

在此小節的實驗中，將探討人臉偵測的準確度對人臉辨識的影響。在實驗中將比較「Viola-Jones Face Detector」和人工手動(Manual)所得到的人臉影像在資料集「Sinica Face Video Dataset」(只取成人部分)上的辨識率。兩者皆使用 Histogram Equalization 對人臉影像進行前處理。此實驗分割訓練用和測試用資料的方式如同章節 5.2.1 中的 Separated Video 方法。在表格 5-5 中列出了不同隨機取樣數量下的結果。

表格 5-5: 人臉偵測對人臉辨識率的影響。

Image set distance method	Sample Number	Viola-Jones Face Detector	Manual
LAHISD	20	0.2350	0.2361
	50	0.2289	0.1922
	100	0.2006	0.1650
MSM	20	0.2228	0.2267
	50	0.2206	0.1917
	100	0.1944	0.1717

從表格 5-5 的數據中可以發現，當隨機取樣數在 20 的時候，使用自動偵測和人工標定的人臉的辨識率相當接近。在隨機取樣數為 50 和 100 時，反而是「Viola-Jones Face Detector」的辨識率較高，而且兩種取得人臉方法的辨識率差距比在取樣數 20 的時候明顯高出許多。目前還不清楚造成這個結果的原因，猜測可能是進行人工校正時，將一些原本在「Viola-Jones Face Detector」中不會偵測到的側臉也標示為人臉有關。但是從這個實驗的結果至少可以知道一件事，就是人臉偵測的準確度並不是影響人臉辨識效能的主要因素，也就是在一般的人臉辨識應用中使用「Viola-Jones Face Detector」來偵測人臉已經足夠。

5.3 性別辨識

這個實驗的目的是要對影片中的人物進行性別辨識。本實驗在「Sinica Face Video Dataset」上進行測試，並且只使用 60 個成人的部分，因為小孩子的性別辨識較困難，就算由人類來進行此項工作也不容易。

這裡用兩種方法來從影像集中選擇測試用的人臉影像：方法 A 是根據人臉影像左右對稱性選擇較為為正面朝向的人臉影像作為測試用，對稱性與否是計算影像的左半部分和右半部分的 cosine similarity。方法 B 則是簡單的從每間隔六張影像中選擇一張人臉影像作為測試用。然後用多數決來決定每個人的分類結果。

表格 5-6: 性別辨識的辨識率。

Method	per segment	per video	per person
A	0.58	0.57	0.65
B	0.60	0.62	0.63

在表格 5-6 中有三個欄位"per segment"、"per video"和 "per person"，分別代表用於每一個分類的圖像集包括從“單一片段(segment)來的人臉影像”、“從一個影片的所有片段(segment)來的人臉影像”和“從一個人的所有影片來的人臉影像”。

從表格5-6的實驗數據來看，當使用越多的影像時分類結果有越準確的趨勢。在這個例子中就是"per person" > "per video" > "per segment"。另一方面，方法 A 和 B 的差別不大，此結果應該和訓練資料集裡面都是較為正面朝向的人臉有關(訓練資料集是另外蒐集的，並沒有使用「Sinica Face Video Dataset」中的人臉影像)，需要進一步研究以了解原因。

5.4 年齡辨識

此實驗的測試的人臉影像選擇方式和 5.3 節中的方法 B 相同。

辨識的結果列在表格 5-7。從實驗結果可以發現和 5.3 節性別辨識類似的狀況：當使用越多的影像來時有越高的分類準確度。

表格 5-7: 成人/小孩辨識的辨識率。

	per segment	per video	per person
Child	0.83	0.88	0.90
Adult	0.62	0.66	0.75
Overall	0.71	0.75	0.81

5.5 人臉分群

5.5.1 分群效能評估方法：CVC

CVC(Classification Via Clustering)[39]是計算群的「純度」的一種方法。計算的方法是將每一群中挑選出佔有最大比例的物件種類視為正確的分群種類。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。CVC 的數值範圍在 0 到 1 之間，數值越大代表分群效果越好。

圖 5-3(圖片取自[36])的「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「Δ」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。

圖 5-3: CVC 分群評估範例。

5.5.2 分群效能評估方法：ARI

ARI(Adjusted RAND Index)這個分群效能評估方法是由 Hubert 等人[40] 所提出的公式。給定一個含有 n 個元素的集合 S，和兩種分群結果$X = \{X_1, X_2, \ldots, X_r\}$和$Y = \{Y_1, Y_2, \ldots, Y_s\}$，在這邊可以假定其中一種是正確分群結果一種是要被評估的分群結果。接下來則是算出 contingency table 如下圖 5-4。圖 5-4 中n_{ij}代表同時屬於X_i和Y_j的元素數量，此圖取自[41]。

X\Y	Y_1	Y_2	$\cdots$	Y_s	Sums
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
Sums	b_1	b_2	$\cdots$	b_s	

圖 5-4: The contingency table。

計算 ARI 之前要先計算以下三個數值：

- 在 X 被分為同一群且在 Y 也被分為同一群的物件，其對(pair)數為 $c = \sum_{i,j}\binom{n_{ij}}{2}$
- 在 X 被分為同一群組的物件，其對(pair)數為 $d = \sum_i \binom{n_{i.}}{2} = \sum_i \binom{a_i}{2}$
- 在 Y 被分為同一群組的物件，其對(pair)數為 $e = \sum_j \binom{n_{.j}}{2} = \sum_j \binom{b_j}{2}$

$$\text{ARI} = \frac{c - (d \times e)/\binom{n}{2}}{\frac{(d+e)}{2} - (d \times e)/\binom{n}{2}} \tag{8}$$

ARI 的數值範圍在-1 到 1 之間，數值越大代表分群效果越好。

5.5.3 分群實驗

本實驗的目的是比較不同的更新群距離的方法以及不同的最終分群數目對分群效果的影響。本實驗分別在資料集「Raymond」和「Friends」上進行，使用的人臉偵測方法是「Viola-Jones Face Detector」，另外在開始分群之前，有先用人工的方式將「Viola-Jones Face Detector」偵測錯誤的人臉串列排除，以免影響分群的結果。本實驗僅列出用「LAHISD」作為計算人臉影像集相似度方法的結果。在計算人臉影像串列相似度時，使用時間資訊將在時間上有重疊的人臉影像串列的距離設為無限遠，讓這些串列在分群時不會被分在同一群。此實驗中的 ARI 計算方面分別列出以人臉影像串列為單位(ARI_video)和以人臉影像為單位(ARI_image)的實驗結果。

由前面章節 5.1 和 5.2 的眾多實驗中得知了幾個重要資訊：一、人臉影像串列相似度計算方法中「LAHISD」各方面的表現最佳，因此本實驗只列出此方法的實驗數據。二、在四個資料集中，「Honda/UCSD」過於簡單，不論用何種方法都會得到很好的效果；「Sinica Face Video Dataset」則是難度太高，而「Raymond」有較適中的難度且可以較好的分別出不同方法的效能。

從表格 5-8 中依序觀察最終分群數為 9 的幾種更新群中心的方法，可以發現到不論是 ARI_video、ARI_image 或是 CVC 都是 Average-link 的分群效果最好，Complete-link 次之，Single-link 最差。同樣觀察最終分群數為 10、20、30 的幾種更新群中心的方法也可以發現同樣的結果。另外比較表格 5-8 和 5-9 可以發現分群效果上「Raymond」比「Friends」的好，這也和之前章節 5.1 的結果符合。

表格 5-8: 「Raymond」上 agglomerative 分群結果。

最終群數		ARI_video	ARI_image	CVC
9	Single-link	0.0232	0.0172	0.4269
	Complete-link	0.2102	0.2741	0.5616
	Average-link	0.4882	0.4860	0.6804
	Fusion	0.0036	0.0044	0.3995
10	Single-link	0.0232	0.0173	0.4292
	Complete-link	0.2124	0.2768	0.5731
	Average-link	0.4935	0.4896	0.6826
	Fusion	0.0036	0.0044	0.3995
20	Single-link	0.0211	0.0190	0.4521
	Complete-link	0.2694	0.3389	0.7306
	Average-link	0.4115	0.4486	0.7146
	Fusion	0.0074	0.0089	0.4132
30	Single-link	0.0223	0.0214	0.4635
	Complete-link	0.2381	0.3180	0.7603
	Average-link	0.4291	0.4864	0.7534
	Fusion	0.0099	0.0109	0.4269

表格 5-9: 「Friends」上 agglomerative 分群結果。

最終群數		ARI_video	ARI_image	CVC
7	Single-link	0.0001	0.0000	0.2255
	Complete-link	0.0234	0.0426	0.3134
	Average-link	0.0362	0.0362	0.3114
	Fusion	0.0006	0.0011	0.2275
10	Single-link	-0.0002	0.0007	0.2275
	Complete-link	0.0686	0.1045	0.3633
	Average-link	0.0381	0.0374	0.3154
	Fusion	-0.0002	0.0007	0.2315
20	Single-link	0.0061	0.0071	0.2535
	Complete-link	0.0736	0.1164	0.4351
	Average-link	0.0428	0.0406	0.3473
	Fusion	0.0024	0.0022	0.2455
30	Single-link	0.0054	0.0069	0.2754
	Complete-link	0.0795	0.1298	0.4870
	Average-link	0.0956	0.1115	0.4192
	Fusion	0.0034	0.0035	0.2615

圖 5-5 是用「LAHISD」搭配「Average-link」在資料集「Raymond」的分群效果趨勢圖，從中可以看到不同的最終分群數對分群效果的趨勢，當分群數越多時可以發現分群效果有逐漸降低的趨勢。但是最佳的分群效果的群數(約 12 群)卻不等於此資料集的人數(9 人)。

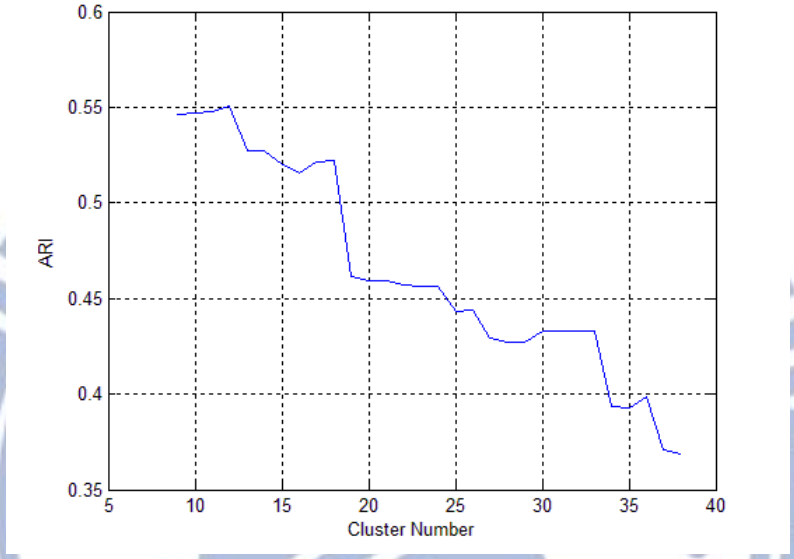

圖 5-5: 不同最終分群數的分群效果趨勢圖。

第六章 結論與未來展望

本文中分別測試了幾種計算人臉影像集距離的方法，並且在 4 個不同的資料集上進行實驗。這幾種方法中「LAHISD」在辨識率和計算時間上有較佳的表現。這個方法是在 2010 年所提出，是近年來較新的計算影像集相似度的方法之一，雖然在簡單的環境如資料集「Honda/UCSD」可以有非常好的表現，但是在資料集「Sinica Face Video Dataset」的表現卻非常不理想，顯示出在人臉辨識這領域裡計算影像集相似度的演算法還有許多進步的空間。

從本文中的數個實驗結果得知在進行人臉辨識任務時的數項影響因子：對辨識率的影響在合理範圍內的隨機取樣數大約為每組影像集各取 10~20 張人臉影像；證明了使用影片進行辨識比單獨使用影像辨識會有更好的辨識率；訓練和測試資料之間不該有重疊或是來自相似甚至相同影片的片段；少數偵測錯誤的人臉對整體辨識率影響不大。另外在人臉偵測準確度的影響實驗中，在隨機取樣數 50 和 100 時，用自動化偵測的人臉進行辨識比反而比用人工標註的人臉進行辨識有更好的辨識率，這部分可能可以透過只從人工標註的人臉取出自動化偵測的到的人臉進行實驗，或許可以排除一些人工標記出來較不正向的臉對辨識率的影響。

在本文中介紹了一個新的影片資料集「Sinica Face Video Dataset」的蒐集和標記過程。為了進一步的人臉辨識任務，此資料集讓性別分類和成人/小孩分類的研究可以在富有挑戰性的真實生活中的影片上進行。從本文所展現的實驗結果，可以證明「Sinica Face Video Dataset」這個資料集的挑戰性，並且在今後相關的研究上，這些實驗結果也可以做為其他相關方法的比較基準。這個資料集的特色(人工標註的人臉區域 ground truths)適合用來研究及評估在真實世界中影片的臉

部追蹤演算法。此外手動標記的程式可用於其他類型的影片資料集，以建立人物追蹤的 ground truths。

新的資料集帶來了一些新的挑戰和相關研究。在本文中已經用人工方式標記了人臉的區域範圍 ground truths，未來可以更進一步的將臉部特徵例如五官輪廓等也用人工方式標記出 ground truths。並使真實生活中的影片臉部特徵標記與追蹤演算法將能夠像人臉偵測與追蹤演算法一樣進行評估與研究，還可以研究臉部影像校準(facial image alignment)對辨識效能的影響。

參考文獻

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, 2001, pp. I-511-I-518 vol.1.

[2] Mark Everingham, Josef Sivic, and A. Zisserman, "Hello! My name is... Buffy -- automatic naming of characters in TV video," *BMVC 2006, 4-7 September 2006, Edinburgh, UK*, 2006.

[3] A. Katsamanis, G. Papandreou, and P. Maragos, "Audiovisual-to-articulatory speech inversion using Active Appearance Models for the face and Hidden Markov Models for the dynamics," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 2237-2240.

[4] G. Shaogang, S. McKenna, and J. J. Collins, "An investigation into face pose distributions," in *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, 1996, pp. 265-270.

[5] L. Shen, L. Bai, and M. Fairhurst, "Gabor wavelets and General Discriminant Analysis for face identification and verification," *Image and Vision Computing*, vol. 25, pp. 553-563, 5/1/ 2007.

[6] R. Thiyagarajan, S. Arulselvi, and G. Sainarayanan, "Gabor feature based classification using statistical models for face recognition," *Procedia Computer Science*, vol. 2, pp. 83-93, // 2010.

[7] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *Acm Computing Surveys (CSUR)*, vol. 35, pp. 399-458, 2003.

[8] M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 71-86, 1991/01/01 1991.

[9] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, pp. 711-720, 1997.

[10] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *JOSA A*, vol. 14, pp. 1724-1733, 1997.

[11] Y. Jian, D. Zhang, A. F. Frangi, and Y. Jing-Yu, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, pp. 131-137, 2004.

[12] A. V. Nefian and M. H. Hayes III, "Hidden Markov models for face recognition," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, 1998, pp. 2721-2724.

[13] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face Recognition with Local Binary Patterns," in *Computer Vision - ECCV 2004*. vol. 3021, T. Pajdla and J. Matas, Eds., ed: Springer Berlin Heidelberg, 2004, pp. 469-481.

[14] K. Minyoung, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1-8.

[15] J. Bigün, B. Duc, F. Smeraldi, S. Fischer, and A. Makarov, "Multi-modal person authentication," in *Face Recognition*, ed: Springer, 1998, pp. 26-50.
[16] M. Everingham and A. Zisserman, "Automated person identification in video," in *Image and Video Retrieval*, ed: Springer, 2004, pp. 289-298.
[17] M. Everingham and A. Zisserman, "Automated visual identification of characters in situation comedies," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, pp. 983-986 Vol.4.
[18] U. Park and A. K. Jain, "3D model-based face recognition in video," in *Advances in Biometrics*, ed: Springer, 2007, pp. 1085-1094.
[19] K. Fukui and O. Yamaguchi, "Face Recognition Using Multi-viewpoint Patterns for Robot Vision," in *Robotics Research*. vol. 15, P. Dario and R. Chatila, Eds., ed: Springer Berlin Heidelberg, 2005, pp. 192-201.
[20] O. Yamaguchi, K. Fukui, and K. i. Maeda, "Face recognition using temporal image sequence," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, 1998, pp. 318-323.
[21] W. Ruiping, S. Shiguang, C. Xilin, and G. Wen, "Manifold-Manifold Distance with application to face recognition based on image set," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1-8.
[22] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 2567-2573.
[23] H. Yiqun, A. S. Mian, and R. Owens, "Sparse approximated nearest points for image set classification," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 121-128.
[24] Y.-C. Chen, V. Patel, P. J. Phillips, and R. Chellappa, "Dictionary-Based Face Recognition from Video," in *Computer Vision – ECCV 2012*. vol. 7577, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., ed: Springer Berlin Heidelberg, 2012, pp. 766-779.
[25] Z. Liu and Y. Wang, "Major cast detection in video using both audio and visual information," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, 2001, pp. 1413-1416.
[26] Z. Liu and Y. Wang, "Major cast detection in video using both speaker and face information," *Multimedia, IEEE Transactions on*, vol. 9, pp. 89-101, 2007.
[27] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automated naming of characters in TV video," *Image and Vision Computing*, vol. 27, pp. 545-559, 2009.
[28] E. El Khoury, C. Senac, and P. Joly, "Face-and-clothing based people clustering in video content," in *Proceedings of the international conference on Multimedia information retrieval*, 2010, pp. 295-304.
[29] Y. Gao, T. Wang, J. Li, Y. Du, W. Hu, Y. Zhang, *et al.*, "Cast indexing for videos by ncuts and page ranking," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, 2007, pp. 441-447.
[30] K. YAMAMOTO, O. YAMAGUCHI, and H. AOKI, "Fast face clustering based on shot similarity for browsing video," 2010.
[31] L. Kuang-chih, J. Ho, Y. Ming-Hsuan, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Computer Vision and*

Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, 2003, pp. I-313-I-320 vol.1.
[32] R. Gross and J. Shi, "The cmu motion of body (mobo) database," 2001.
[33] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 529-534.
[34] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.
[35] 林維昭, "private communication," ed, 2010.
[36] 蘇裕傑, "根據姿勢與外貌整合的影像人臉註記," *交通大學多媒體工程研究所學位論文*, pp. 1-43, 2011.
[37] Y. Hu, A. S. Mian, and R. Owens, "Face Recognition Using Sparse Approximated Nearest Points between Image Sets," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, pp. 1992-2004, 2012.
[38] S. Y, L. Y, C. J, and W. L, "辨識率 5 成~鳩咪," 2010.
[39] A. Strehl and J. Ghosh, "Cluster ensembles---a knowledge reuse framework for combining multiple partitions," *The Journal of Machine Learning Research*, vol. 3, pp. 583-617, 2003.
[40] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193-218, 1985/12/01 1985.
[41] *Rand index - Wikipedia*. Available: http://en.wikipedia.org/wiki/Rand_index