國立交通大學

電機與控制工程學系

碩士論文

結合影像及語音
之雙模情緒辨識系統

Bimodal Emotion Recognition System Using Image and Speech Information

研 究 生：許晉懷

指導教授：宋開泰 博士

中 華 民 國 九十五 年 七 月

結合影像及語音
之雙模情緒辨識系統

Bimodal Emotion Recognition System
Using Facial Image and Speech Information

研 究 生：許晉懷　　Student: Jing-Huai Hsu

指導教授：宋開泰 博士　　Advisor: Dr. Kai-Tai Song

國 立 交 通 大 學

電 機 與 控 制 工 程 學 系

碩 士 論 文

A Thesis

Submitted to Department of Electrical and Control Engineering

College of Electrical and Computer Engineering

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of Master

in

Electrical and Control Engineering

July 2006

Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 五 年 七 月

結合影像及語音之雙模情緒辨識系統

學生:許晉懷　　指導教授:宋開泰 博士

國立交通大學電機與控制工程學系

摘要

本論文研究結合人臉影像及語音之雙模情緒辨識系統。文中提出一種新的雙模辨識策略：透過兩種資訊辨識情緒，對辨識的可靠度設定不同的權重，用以決定要該採用何種資訊。雙模資訊權重的數值，是透過 SVM 理論中測試資料距超平面的距離，以及訓練資料之標準差，接著再經由訓練資料距超平面的平均距離正規化後決定，此權重係數即可以判斷分類的可靠度。由於需要辨識多種情緒，在辨識某兩類的情緒時，採用權重較高資訊的辨識結果，可以修正其它資訊錯誤的分類，而原本被修正的資訊在下一步辨識另兩類情緒時，可能其權重較高，亦可修正另一方的資訊辨識結果，如此可以提高原有單一資訊辨識的準確性。在以人臉表情辨識方面，基於以 SVM 兩兩表情辨識之設計，本論文提出針對不同表情進行關鍵特徵辨識。整套系統已在實驗室開發的嵌入式數位訊號處理器(DSP)平台予以實現，擷取特徵後予以分類辨識，完成高興、生氣、普通、傷心與驚訝五種情緒即時辨識系統，雙模辨識率可達 86.85%，比只用人臉影像辨識高 5.1%。

Bimodal Emotion Recognition System Using Image and Speech Information

Student: Jing-Huai Hsu Advisor: Dr. Kai-Tai Song

Department of Electrical and Control Engineering
National Chiao Tung University

ABSTRACT

A bimodal emotion recognition system has designed and realized by combining image and speech information. This thesis proposes a new bimodal recognition strategy by setting different weights to each mode based on their recognition reliability. The weights are determined by the distance between test data and hyperplane and the standard deviation of training data. The weights are normalized by the mean distance between training data and hyperplane, representing the classification reliability of different information. Five emotions are to be recognized, we adopt recognition result with higher weight, which corrects the other information. At the next step, the corrected information may correct wrong classification result of the other mode of information. It can raise the accuracy of single modal information. Further, regarding facial expression recognition, we propose to use key features for the design of facial expressions classification in SVM theorem. The whole procedures have been implemented on an embedded image system. After extracting the features, classification is applied to recognize five emotion expressions: happiness, anger, neutral, sadness, and surprise. The experimental results show that bimodal emotion recognition rate is 86.85%, an increase of 5% compared with using single modal image emotion recognition.

誌謝

謹向我的指導教授宋開泰博士致上感謝之意，感謝他兩年來在專業上的指導，以他豐富的學識與經驗，配合理論的應用，使得本論文得以順利完成。感謝口試委員胡竹生教授、楊谷洋教授與范欽雄教授的指導與意見，讓本論文能夠更加嚴謹。

感謝與我共同奮鬥的同學宏宜、忠憲及鎮謙的相互鼓勵及提攜，以及學弟富聖、濬尉、振暘、俊瑋及志昇在生活上帶來的樂趣，同時感謝學長孟儒、嘉豪、奇諡及崇民在實作與理論上的指點。

感謝我的朋友們的勉勵與支持，並在研究上相互討論，讓我獲益良多。特別感謝女友宜珊對我的體諒以及生活上的照顧，並在我最無助及失意的時候給予我意見及鼓勵。最後，感謝我的父母與家人，由於他們的辛苦栽培，在生活上給予我細心地關懷與照料，使得我才得以順利完成此論文，在此我願以此論文獻給我最感激的父母親。

目錄

圖例

表格

第一章 緒論

1.1 研究動機

近年來在人口老化的衝擊下，藉由機器人提供的照護與娛樂功能逐漸受到重視[1]。其中娛樂型機器人在這幾年最為人所津津樂道，像是 Sony 開發的 AIBO 寵物型機器狗，具有獨立行動能力，能夠識別主人的名字，聲音和面貌，還可以自動充電；和人們的互動中，它可以表現情緒，藉此達到娛樂甚至安慰人們的功能[2]。Sony 另外開發的 Qrio 雙足機器人[3]，其設計的用意是在提供家庭娛樂，它會辨別主人的樣子與聲音，並且可以與人簡單對話。日本產業技術綜合研究所(AIST)開發的海豹型機器人”Paro”[4]，它的臉部表情豐富，還能做出伸展身體、眨眼睛等小動作，更重要的是，它對人動作的反應相當靈敏，其主要的特點是可以對人的觸摸產生交互的反應，達到安慰人的目的，進而具有治療的作用[4-5]。由上述的例子看來，寵物機器人朝向智慧化與多元化邁進，將與人們共同生活，提供家庭娛樂且撫慰人心。

機器人要能感測從外界得來的資訊，以便可以和人有所互動且自主決定其行為，其中最重要的功能之一是需要一個可靠的人機介面，能從外界擷取重要的訊息，讓機器人知道下一步的行為為何。而要使得機器人和人之間能夠更自然地互動，可以藉由情緒辨識讓機器人偵測到人類的情緒狀態，並依照人類不同反應有其對應的自主式行為，使得機器人不再是冷冰冰的機器，讓人可以對機器人能夠感興趣，經由進一步的互動後進而產生感情。而寵物機器人可以辨識人類的情緒反應，就像是一個真實的寵物，陪在人類身旁能夠提供娛樂甚至是撫慰人心的功能，使得人與機器人間的互動更為自然。

人類表達情緒包含多種形式，包含肢體語言、聲音和臉部表情，而主要還是透過臉部表情和聲音的語調表達情緒，因此本論文將設計一個結合人臉影像和語音辨識人類情緒的方法，提高單一人臉影像或單一語音的辨識率，使得人與機器人之間的溝通更為有效且更加自然。

1.2 相關研究回顧

目前已有相當多數的語音與人臉影像情緒辨識研究，幾乎都是單一模組辨識，單純只靠語音特徵[17-22]或臉部影像特徵[7-16]來辨識情緒，雙模辨識方法，即結合語音及人臉影像的特徵，在辨識架構上能達到資訊互補，提高辨識率的效果，然而這方面的研究報告並不多[23-29]。。

在單一模式人臉影像表情辨識上，最早是 Ekman 和 Friesen[6]發展了 Facial Action Coding System(FACS)，根據人在表達不同表情時會牽動不同臉部肌肉的原理，定義出 44 種不同的 Action Unit (AU)，如圖 1-1 所示，圖(a)是上半臉 Action Unit 和一些 Action Unit 的組合，圖(b)是下半臉的 Action Unit 種類，根據特定 Action Unit 的組合，人臉的表情即可以被描述出來。多數的研究都採用 FACS 的理論來判斷臉部表情， Song *et al.* [7]提出了一個基於小波轉換的 Multi-resolution 性質設計，可以將雜訊從二值化的影像中移除，二值化後的邊緣影像可以用來辨識多個不同的 Action Unit。Cohn *et al.* [8] and Seyedarabi *et al.* [9] 使用 Optical flow 追蹤多個特徵點的移動，這些特徵點是在影像序列中的第一張影像中手動設置。在第一張與最後一張影像中，每個特徵點位移組合而成的特徵向量則用來辨識不同的 Action Unit 及臉部表情。

Tian *et al.* [10]和 Donato *et al.* [11]也都採用 FACS 理論，不同的是在臉部特徵擷取與辨識分類方法的選擇上。Tian *et al.*將臉部的特徵分為常在與瞬間特徵，常在特徵是眼睛與嘴巴的變化，瞬間特徵是眉間、雙頰與嘴角上端的皺紋變化，先手動設置好這些特徵點的初始位置，再追蹤將這些點的變化，用類神經網路作辨識。雖然辨識的效果很好，但以 FACS 為基礎的完全自動表情辨識方法尚未被提出。

在特徵擷取上，人臉的特徵大致上可以分為 Appearance based 與 Feature based 兩種方法，Appearance based 的特徵擷取，常見的為用主成份分析法(Principal Component Analysis, PCA)、獨立成份分析法(Independent Component Analysis, ICA)對整個臉部影像作特徵擷取。Wilhelm *et al.*[12]就用 Appearance

NEUTRAL	AU 1	AU 2	AU 4	AU 5
				
Eyes, brow, and cheek are relaxed.	Inner portion of the brows is raised.	Outer portion of the brows is raised.	Brows lowered and drawn together	Upper eyelids are raised.
AU 6	AU 7	AU 1+2	AU 1+4	AU 4+5
				
Cheeks are raised.	Lower eyelids are raised.	Inner and outer portions of the brows are raised.	Medial portion of the brows is raised and pulled together.	Brows lowered and drawn together and upper eyelids are raised.

(a)上半臉的 Action Units 和 Action Unit 的組合

NEUTRAL	AU 9	AU 10	AU 12	AU 20
				
Lips relaxed and closed.	The infraorbital triangle and center of the upper lip are pulled upwards. Nasal root wrinkling is present.	The infraorbital triangle is pushed upwards. Upper lip is raised. Causes angular bend in shape of upper lip. Nasal root wrinkle is absent.	Lip corners are pulled obliquely.	The lips and the lower portion of the nasolabial furrow are pulled pulled back laterally. The mouth is elongated.
AU15	AU 17	AU 25	AU 26	AU 27
				
The corners of the lips are pulled down.	The chin boss is pushed upwards.	Lips are relaxed and parted.	Lips are relaxed and parted; mandible is lowered.	Mouth stretched open and the mandible pulled downwards.

(b)下半臉的 Action Units

圖 1-1 臉部的 Action Units[10]

based 方法作特徵擷取，採用 PCA 與 ICA 擷取特徵，減少輸入分類器的資料量，並計算投影向量的合適性，選取具有分辨性的特徵向量作分類，以提高辨識的準確性，再分別採用 Nearest Neighbor(NN)分類器、Multi Layer Perceptron(MLP)和 Radial Basis Function(RBF)網路當作表情的分類器。使用 Appearance based 的方

法作人臉資料庫的建立與表情辨識，可以減少載眼鏡與人臉鬍鬚的影響。Buciu[13]等人採用 ICA 以及 Gabor 小波轉換擷取人臉影像的特徵，實驗結果證實採用 Gabor 小波轉換結合 support vector machine (SVM)可以大幅提高辨識的效果。

利用 Appearance based 方法辨識表情，可能會因為臉部在影像中位置的改變而影響到辨識的結果，於是 Y. Zhang[14]等人利用整張臉部影像，計算 PCA、ICA 與 Linear Discriminant Analysis(LDA)得到臉部的表情特徵，接著算出臉部資料庫與測影像的差異性權重矩陣，在變化較大的地方給予較小的權重；反之給予較大的權重，用 Nearest Neighbor 辨識 4 種表情，實驗結果證實 LDA 結合權重矩陣效果比單純用 eigenface 辨識好。除了 Appearance based 作特徵擷取的方法，還有 Feature based[15-16]方法，擷取人臉特徵點，像是眼睛、眉毛和嘴巴各設置一些特徵點，再將各特徵點之間的距離或是特徵點在影像中的位移當作分辨不同表情的特徵。

在語音的情緒辨識上，主要是選擇帶有情緒資訊的特徵，像是聲韻(Prosody)特性和能量相關的特徵，最常見的特徵是音高(Pitch)和能量(Energy)[17-20]，少數文獻採用共振峰(Formant)[19]，主要是根據這些特徵的統計值當作不同情緒分類的特徵，像是平均值、標準差、最大值、最小值、梯度變化等。Schuller *et al.*[20]即根據線性鑑別分析對不同統計方式的特徵進行排名，用音高相關的特徵明顯比能量相關的特徵更能分辨出不同的情緒種類。採用梅爾倒頻譜係數（Mel-scale Frequency Cepstral Coefficients, MFCC）或線性預估參數(Linear predictor coefficient, LPC)[21-22]也都可以找到關於情緒的訊息，透過分類方法可以辨識出語音中帶有的情緒，辨識方法上包括類神網路[20]、隱藏式馬可夫模型(Hidden Markov Model, HMM)[17][18][22]、高斯混合模型(Gaussian Mixture Model, GMM)[20-21]、線性鑑別分析[22]等。

結合語音與臉部影像的情緒辨識，其架構和單模組的語音或臉部影像辨識相同，首先經過特徵擷取，接下來為特徵的分析與分類，最後為情緒的決定，只是

在特徵的擷取上需要語音和人臉影像的輸入，以及在情緒的決策上較單一模式複雜。

De Silva *et al.*[23]分別利用兩種語言測試在表達情緒時，若是影像與語音兩種資訊辨識出的結果不同時，觀察是臉部影像還是語音具有支配性(Dominant)的地位，實驗結果發現傷心和害怕兩種情緒是聽覺支配(Auditory dominant)，表示當語音辨識情緒為傷心或害怕時，而臉部影像卻辨識為其他不同的情緒時，則主要採信語音的辨識結果。而其他情緒可能是影像支配(Visually dominant)或是沒有固定的支配地位，因此可以運用這些要素給予臉部影像和語音在不同表情上有著不同的權重值，設計一個給各個表情及模式的權重矩陣，屬於 Rule-based 的分類方法。

Chen *et al.*[24]也是承接著 De Silva *et al.*的 Rule-based 方法繼續研究，他們透過語音與人臉影像之間的互補特性作辨識，例如音高最大值座落在某一範圍內，可能會是某兩種情緒，再透過人臉影像資料對這兩種表情作辨識。但可能是文化、語言或是個人問題，這一套結合方法並不適用於每一個人。

後來 De Silva[25]發現在負面情緒上，像是生氣、害怕與傷心的臉部表情很接近，辨識分類的效果不可靠，於是即根據上一篇的實驗結果，將負面情緒：生氣、憎惡、害怕與傷心設為聽覺支配，高興和驚訝設為影像支配，結合語音與臉部影像作辨識。語音部分用 Hidden Markov Model(HMM)辨識語音訊號的特徵，臉部影像用類神經網路辨識臉部特徵點的位移量，在辨識結果決策上，首先若語音辨識為負面情緒中的一種，但影像辨識結果不同，即完全採用語音部分的結果；反之，影像辨識為正面情緒，而語音辨識結果不同，則完全採用影像部分結果，如此大幅提高負面情緒的辨識可靠度，也提高了所有表情的辨識率。

Go *et al.*[26]採用類神經網路，直接將分別輸入的語音特徵與人臉影像特徵結合辨識，得到一個辨識的輸出結果。不同的是語音的特徵擷取上，先將語音信號用小波轉換分成不同頻段，針對不同頻段擷取 MFCC、能量和過零率的特徵；而人臉影像也是用小波轉換，將人臉低頻成份用 LDA 擷取特徵向量。但未比較

單一模組和雙模組的差異，因此未能知道雙模組辨識是否有其必要性，可以提升辨識率。

Song *et al.*[27]提出用 Tripled HMM (THMM)辨識結合語音及影像的情緒，語音特徵採用音高和能量的相關特徵，影像特徵則是臉部特徵點的位移特徵，THMM 將語音及影像分開平行處理，最後再根據特徵的信號雜訊比(SNR)設定權重，計算辨識的結果。用 THMM 的優點是語音及影像在辨識的狀態過程中允許非同步輸入，仍保持信號間的相關性，且兩種不同的特徵訊號都用同一套辨識系統，減少辨識系統的複雜度。

Wang *et al.*[28]將語音與臉部影像的特徵串接起來，語音特徵採用 Pitch 和 MFCC 的相關特徵，臉部影像採用 Gabor 小波轉換係數，根據歐幾里德距離，選出顯著的特徵，對每一種表情採用一對多(One-against-all；OAA)的 LDA 辨識，算出每一種表情的機率值。不同的是在辨識的決定上，設定了幾個大規則，第一：若是只有一類表情的機率大於 50%，則結果為此類表情，第二：若是沒有任一表情超過 50%，則會再採用一個全部表情分類器；若是有 2 種以上的表情機率都大於 50%，則會針對這幾種表情再作分類，找出機率最大的表情。

吳鑑峰[29]在語音特徵上採用音高、共振峰、能量和過零率，並計算這四種特徵的趨勢與統計特徵，影像特徵是用臉部特徵點之間的距離當作特徵，用主成份分析法選出具代表性的特徵，而語音與臉部影像兩類特徵分別用 Continuous Density Support Vector Machine(CDSVM)作辨識，根據測試資料與訓練資料距超平面的距離以及訓練資料的正確率，計算測試資料為各種情緒的機率。最後再計算語音與臉部影像分類後的機率值的幾和平均，得到總情緒機率值，辨識效果比單一模組的語音和人臉影像要好。

1.3 問題描述

由於機器人技術逐步成熟，正步入一般民眾的家庭之中，機器人提供的娛樂

功能成為在應用上的發展目標。而要使得機器人和人之間能夠互動地更加自然，希望讓機器人可以自然地偵測到人類的情緒狀態，因此它可以有對應的自主性行為娛樂人類甚至達到安撫的功能。目前情緒辨識的研究多是採用單一模式，即只靠語音或臉部影像來辨識不同的情緒，而這兩類資料都可以辨識不同情緒，表示這兩類資訊具有一定的相關性，在辨識不同情緒應該會有互補的作用。

所以本論文的目標為建構一個雙模式情緒辨識系統，結合這兩類不同的資訊作情緒辨識，提出辨識策略以提高辨識準確性。我們採用 SVM 演算法辨識不同情緒，SVM 可以根據訓練資料將兩種不同情緒用訓練出的 hyperplane 分割開來，再利用資料在空間中距 hyperplane 距離和分佈狀態這兩種訊息，提供我們該類資料分類的可靠度，如此在人臉影像和語音情緒分類不同時，即可知該信任語音或是人臉影像的辨識結果，以修正某一組錯誤的辨識結果。

在人臉影像特徵的擷取上，以 Appearance-based 作特徵擷取的方法，其計算量相當大，且人臉表情的影像必須要一致，但在實際應用中，人臉的位置與大小常常是未知的，所以本論文採用 Feature-based 方法擷取人臉特徵點，將各特徵點之間的位移向量，當作分辨不同表情的特徵，即使人臉在影像中的大小不一，擷取出的人臉特徵依然可以有效代表不同情緒。

1.4 系統架構與章節說明

系統架構圖如圖 1-2 所示，主要分為三個部分，分別是語音的特徵擷取、人臉影像的特徵擷取以及將兩類特徵結合的辨識決策部分。結合語音與人臉影像情緒辨識中特徵擷取的方法，目的在擷取出每種表情具有代表性和獨特性的語音和人臉影像特徵，使得在辨識階段能夠區分出彼此間的差別，辨識出正確的情緒。

在語音特徵的擷取上，經由麥克風將情緒語音輸入後，將類比信號取樣轉為類比信號，由於不是整段信號都是有效的資料，因為這段信號可能包含真正的語音資料和剩餘的靜音或雜訊部分，所以要先計算出語音資料的起點與終點，

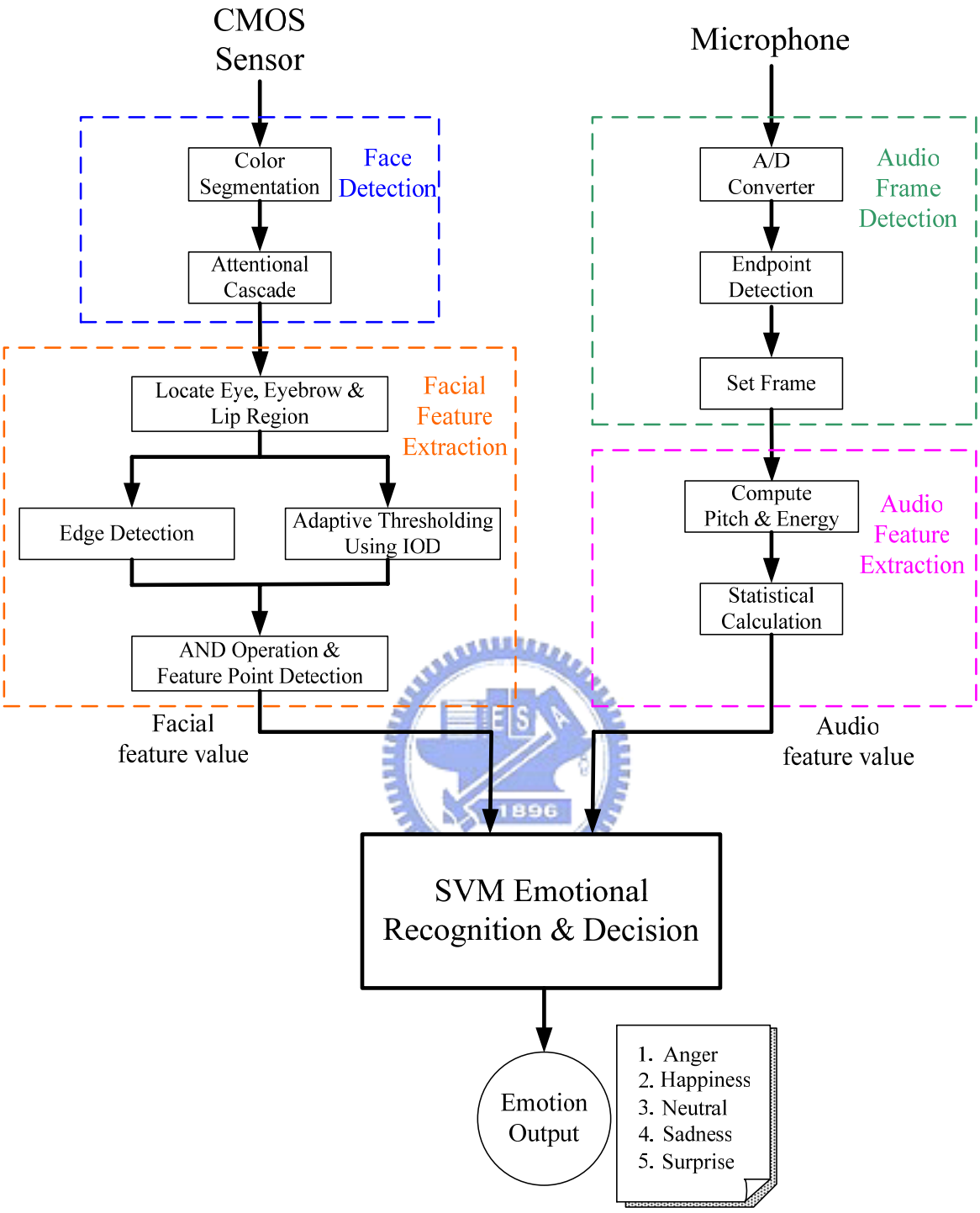


圖 1-2 系統架構圖

對這有效的語音信號設置音框(frame)。再以音框為單位，計算每一個音框內具有情緒資訊的特徵：音高(pitch)與能量(energy)，計算出基於音高和能量的統計特徵。人臉特徵的擷取上，經由膚色和人臉規則資訊作為人臉偵測的依據，找到人臉位置後，根據瞳孔在正面上半臉影像灰階值強度最低的資訊，找出的瞳孔的確

切位置，再根據瞳距和瞳孔位置定出眼睛和眉毛可能區域，利用影像灰階值強度和邊緣偵測找出眼睛與眉毛特徵點的位置。而嘴巴的特徵點是利用它在人臉的幾何位置和影像強度找出，再將各特徵點之間的特徵向量作為分辨不同表情的依據。

得到這兩類的特徵後，便可進入情緒辨識決策階段，本論文採用 SVM 演算法，分別得到兩類特徵的辨識結果，若是兩類資料分類結果不同，即根據分類情形決定該採用何種辨識結果，得到受測者表現的情緒狀態。

本論文共分六章，第一章為介紹相關研究背景及研究動機。第二章介紹語音的情緒特徵擷取。第三章介紹人臉影像特徵的擷取。第四章介紹雙模式情緒辨識的演算法和辨識決策，根據SVM理論中資料距hyperplane的距離和在空間中的分佈狀態做決定。第五章為實驗的方法和結果，包括介紹硬體平台、資料庫的建立、用SVM的關鍵性特徵的表情辨識，以及用柔性邊界SVM的bimodal情緒辨識配合設計的辨識決策。第六章為結論與未來展望。

第二章 語音特徵擷取

語音特徵擷取目的在於取出每種情緒具有代表性和獨特性的特徵，使得在辨識階段能夠區分出彼此間的差別，辨識出正確的情緒。首先經由麥克風將情緒語音輸入後，將類比信號取樣轉為類比信號，由於整段信號可能包含真正的語音資料和剩餘的靜音或雜訊部分，所以要先由端點偵測計算出語音資料的起點與終點，利用短時距能量(Short time energy)和越零率(Zero crossing rate)作為判斷的標準。接著對這有效的語音信號設置音框，以音框為單位，計算每一個音框內具有情緒資訊的特徵：音高(pitch)與能量(energy)，計算出基於音高和能量的統計特徵，作為之後辨識語音情緒的依據。

2.1 語音信號前置處理

語音信號之前置處理是語音信號在作特徵擷取之前，以語音處理之方法將所需之語音段落擷取出來，便於之後的特徵擷取，其中最主要的步驟是端點偵測，在此我們參考實驗室顏坤銘學長[30]論文所使用的方法。

2.1.1 語音信號取樣

語音信號藉由麥克風介面輸入，為準位在數十毫伏特之類比電壓信號，之後再經過放大。取得之語音類比訊號在經過取樣(Sampling)的動作之後，可以得到離散化的數位語音訊號。由於一般人說話的頻譜多集中在 4KHz 以下，根據取樣定理，取樣頻率(Sampling frequency)的設定至少要設定在信號頻寬的兩倍以上，才不會造成失真的現象，所以設定為 8KHz。

2.1.2 端點偵測(Endpoint Detection)

得到一段原始的數位語音訊號後，並不是整段訊號都是有效的資料，因為這段訊號可能包括真正的語音資料和剩餘的靜音或雜訊部分，所以需要先知道語音資料的起點和終點位於何處，一方面可以找出有效的語音段，另一方面可

以減少不必要的資料。因此，端點偵測可以說是最重要的前置過程，若是不正確地抓取語音段的端點，將會影響後續計算的特徵值，影響辨識的成功率。端點偵測主要依靠的是短時距能量(Short time energy)及越零率(Zero crossing rate)兩項資訊作為偵測的依據標準，兩者的說明如下所述。

短時距能量是把每一小段(Frame)的語音信號切出來。每一Frame之短時距能量 $E(k)$ 可以定義為第 k 個音框內之 N 個信號樣本 $x(n)$ 之能量取平方後相加，如式(2-1)所示。但是取平方後之數值有可能會太大。因此採用另一種定義，即能量值先取絕對值後再作相加，如式(2-2)所示。若 $E(k)$ 超過了預先設定好之能量的臨界值，則認為此音框是真正含有語音資訊的，因為通常真正有說話的段落，其能量值會明顯高於其他靜音部份。

$$E(k)=\sum_{m=0}^{N-1}x(n+m)^2 \tag{2-1}$$

$$E(k)=\sum_{m=0}^{N-1}\left|x(n+m)\right| \tag{2-2}$$

越零率 $Z(k)$ 的定義是信號通過原點的次數，即相鄰信號樣本之振幅值若有一正一負的變化則越零率累計值加一，如式(2-3)、(2-4)所示，$Z(k)$ 是第 k 個音框之越零率值，$S(n)$ 是音框中之取樣點，音框長度為 N 。而從越零率的次數大致上可以反映出信號之頻率範圍。越零率的偵測是為了輔助短時距能量偵測在判斷上的不足之處，例如說話時之摩擦音、鼻音、子音，因為在能量的表現上並不足夠超越短時距能量偵測的臨界值，所以會產生端點誤判之情形。

$$Z(k)=\frac{1}{2}\sum_{m=0}^{N-1}\left|\mathrm{sgn}(S(n+m))-\mathrm{sgn}(S(n+m-1))\right| \tag{2-3}$$

$$\mathrm{sgn}\left[S(n)\right]=\begin{cases}1 & if\ S(n)\geq 0\\ -1 & if\ S(n)<0\end{cases} \tag{2-4}$$

基本上，越零率偵測可以分出有聲語音跟無聲語音的分別，無聲語音如摩擦音能量多集中在3KHz以上，所以越零率值會比較高，反之，若越零率值低則為有聲語音，因此，如圖2-1所示，先利用短時距能量偵測大致判斷出有聲語音的開始與結尾處，再利用越零率偵測找出語音段之真正的開頭跟結尾處。找尋

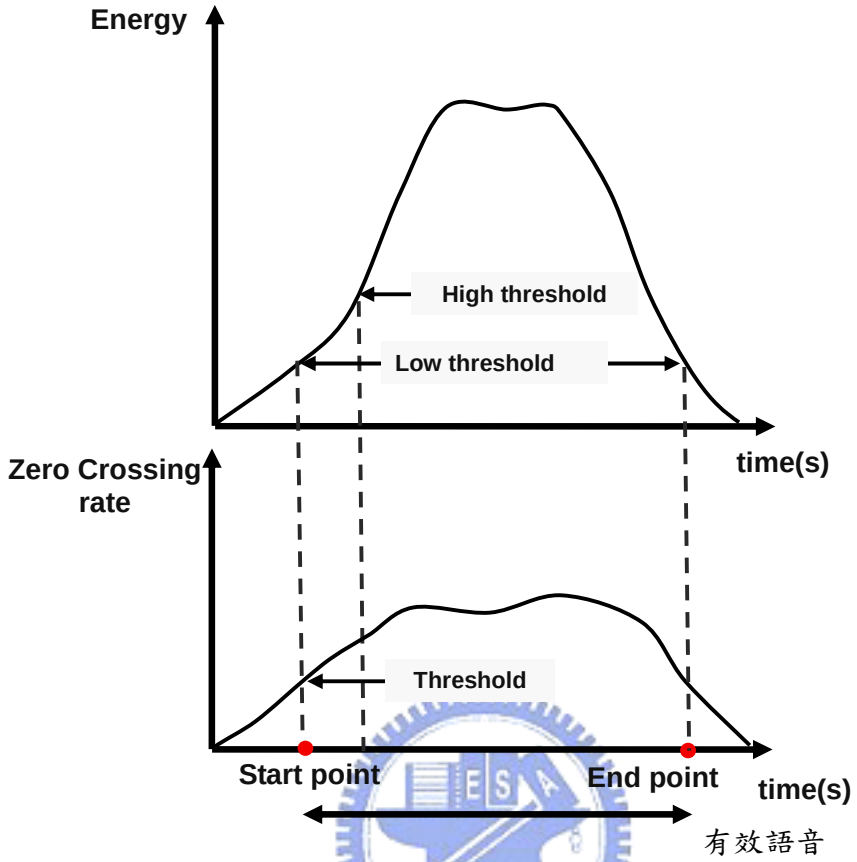


圖 2-1 端點偵測

之規則列出如下：

(1)若 $E(k)$ 小於短時距能量之低臨界值(Energy low threshold)，則認定是非語音段的部份。

(2)若 $E(k)$ 大於短時距能量之低臨界值(Energy low threshold)，且也高於短時距能量高臨界值(Energy high threshold)，則認定為語音段之起點。

(3)若 $E(k)$ 小於短時距能量之高臨界值(Energy high threshold)，則必須要再加上越零率臨界值(Zero crossing rate threshold)來加以輔助，$Z(k)$ 大於越零率臨界值(Zero crossing rate threshold)可以判定為語音段之起點。

(4)要找尋語音段終點時，則是從抓取信號之尾端開始作反向搜尋，若 $E(k')$ 大於短時距能量之低臨界值(Energy low threshold)，則認定是語音段之終點部份。

利用這兩個參數來偵測端點，首先必須要分別找出兩者適合的臨界值

(Threshold)，臨界值的取得必須經由反覆測試才能得到各個系統之適用值，且臨界值會受到系統採用之麥克風種類等等因素影響而有所不同，因此，不同的系統所使用的臨界值並不一定相同。為使臨界值能夠隨著擷取到的語音信號作出調整，本論文之做法是利用擷取到的最初一段語音信號的偵測值作為判斷的基準。利用這一小段時間之能量作為短時距能量偵測的低臨界值，而高臨界值則是低臨界值乘上一個倍數，如式(2-5)、式(2-6) ，N_1為取基準所設定之Frame大小。越零率臨界值的決定則是同樣的方式，以得到之基準為越零率臨界值如式(2-7)、式(2-8)。至於乘上之倍數(Factor)是經由反覆測試而決定。依據上述步驟，圖2-2為實際端點偵測後之結果，圖中兩側之粗黑線為鎖定之語音段起點及終點，如圖中所示一般可以正確地標註出語音段之兩側端點。

$$Energy\ low\ threshold = \sum_{m=0}^{N_1-1} |S(n+m)| \tag{2-5}$$

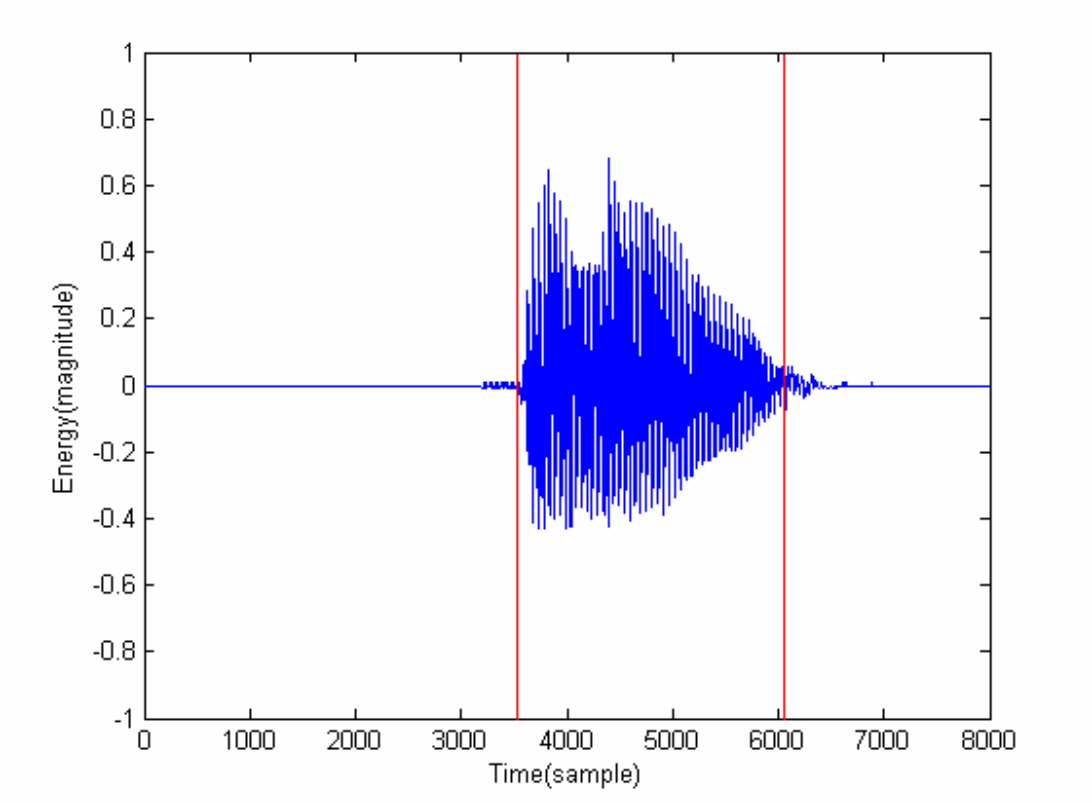


圖2-2 端點偵測之結果

$$Energy\ high\ threshold = Energy\ low\ threshold \times factor \tag{2-6}$$

$$\text{Zero crossing rate threshold} = \frac{1}{2}\sum_{m=0}^{N_1-1}\left|\text{sgn}(S(n+m)) - \text{sgn}(S(n+m-1))\right| \tag{2-7}$$

$$\text{sgn}[S(n)] = \begin{cases} 1 & if\ S(n) \geq 0 \\ -1 & if\ S(n) < 0 \end{cases} \tag{2-8}$$

2.1.3 取音框(Frame)

在得到端點資訊之後，由於語音是一時變(Time variant)的訊號，藉由觀察可以知道語音訊號在短時間內的變化是很緩慢的。所以一般都會採取語音信號是在短時間內維持穩定的假設，即以固定的取樣點數來形成一個音框(Frame)，作為後續處理、分析的基本單位。為了使得一個個音框之間的移動不至於產生太突兀的變化，並保持連貫性，所以音框之間會有部分的重疊，一般會採用二分之一或是三分之一的重疊。

2.2 語音特徵參數計算

語音信號經過前置處理後，將語音段落擷出來並設成一段一段的音框，接著就進行特徵擷取，計算每段音框的音高(Pitch)和能量(Energy)，統計能表示每段音框的特徵值，得到一組可以在辨識時區分出不同情緒的特徵值。

2.2.1 音高(Pitch)

音高（Pitch）代表者聲音頻率的高低，而此頻率指的是聲腔在發出聲音時的共鳴率，也稱為「基本頻率」（Fundamental Frequency）。對整段語音訊號進行抓取音高的過程，通常稱為「音高追蹤」（Pitch Tracking）[31]，音高追蹤的基本流程如下：首先將整段聲音訊號切成多個音框，相鄰音框之間可以重疊，接著算出每個音框所對應的音高，排除不穩定的音高值，得到整段的音高值。

音高追蹤的方法可以分為時域和頻域兩大類，由於時域方法中的自相關函數(Autocorrelation function; ACF)運算量較少，在實作上也比較容易，所以本論文採用

自相關函數計算整段音訊的音高值。自相關函數是把一段序列的訊號與訊號本身的延遲作運算，我們先將整段音訊切成一段一段的音框 $x[n]$，一段音框總共有 N 個信號，再針對每段音框計算自相關係數，公式如式(2-9)所示，$x(n)$ 是一段音框信號，k 是信號位移的大小，音框內信號乘上它的移位信號再全部相加，算出每次移位後的自相關係數 $R(k)$。

$$R(k) = \sum_{k=0}^{N-1-k} x(n) \cdot x(n+k) \tag{2-9}$$

音高追蹤用自相關函數運算的意義在於找尋當訊號移位多少時，訊號會最接近於原本的訊號，也就是尋找訊號的週期，如圖所示自相關函數的計算結果，可由圖 2-3 中看出，除了原點之外，最高點出現在第 41 點，因此這個音框的音高為 fs/(49-1)= 8000/52= 153.85Hz。

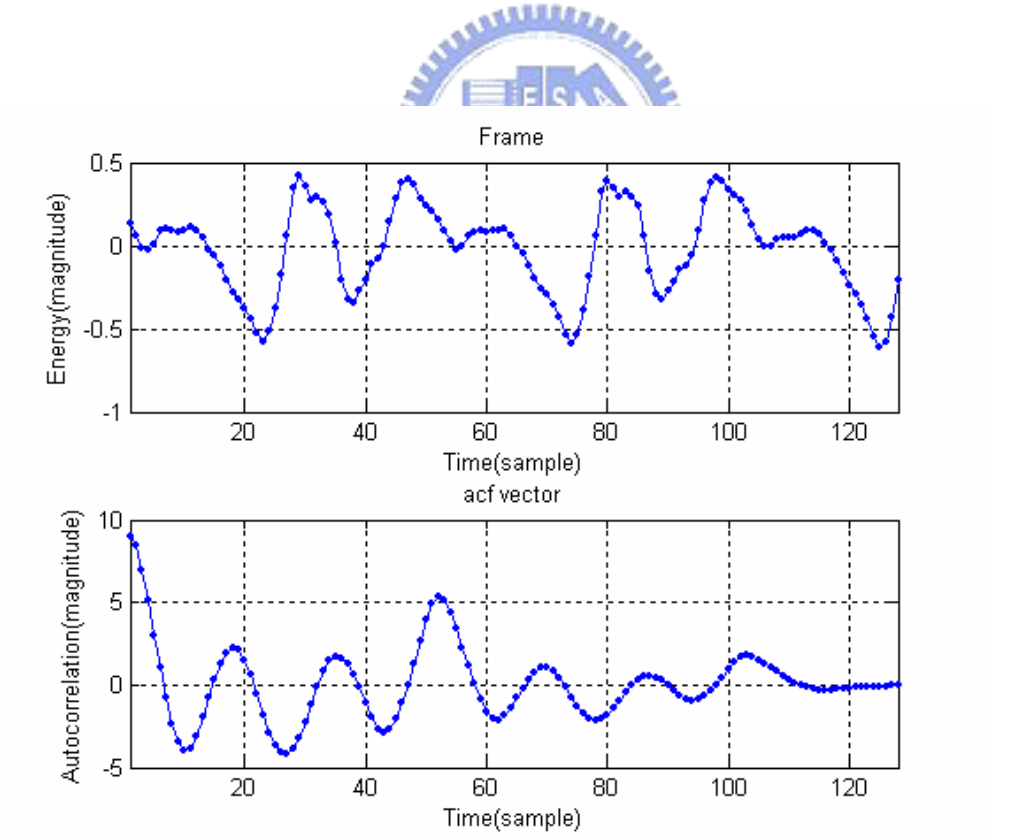


圖 2-3 自相關函數計算結果

2.2.2 能量(Energy)

這邊指的能量就是端點偵測時的短時距能量，先把每一個音框切割出來，對一個音框內的能量取絕對值後再作相加，如式(2-10)所示，算出整段有效語音內每一音框的能量 $E(k)$，k 為音框之編號，音框內信號 $x(n)$ 共有 N 個信號樣本。

$$E(k)=\sum_{m=0}^{N-1}\left|x(n+m)\right| \tag{2-10}$$

2.2.3 語音特徵值計算

計算出每個音框的音高和能量，由於一段語音訊號內有許多音框，所以根據每個音框的音高和能量，統計出整段語音信號的特徵值，如表 2-1 所示，包括平均值、標準差、最大值、最小值以及梯度變化[18-20]。可以藉這些特徵看出音高和能量的分佈與變化情況，從梯度變化的特徵值可以看出特徵變化的趨勢。像是類似語調上揚的語音，如驚訝時發出的驚嘆聲，其音高變化為逐漸往上提高，如圖 2-4(a)所示；反之，當生氣時發出的情緒聲，為語調下挫的語音，

表 2-1 語音的 12 個特徵

Pitch	1. Pave: Average pitch
	2. Pstd: Standard deviation of pitch
	3. Pmax: Maximum pitch
	4. Pmin: Minimum pitch
	5. PDave: Average of pitch derivation
	6. PDstd: Standard deviation of pitch derivation
	7. PDmax: Maximum of pitch derivation
Energy	8. Eave: Average energy
	9. Estd: Standard deviation of energy
	10. Emax: Maximum energy
	11. EDave: Average of energy derivation
	12. EDstd: Standard deviation of energy derivation

其音高變化為漸漸變低，如圖 2-4(b)所示；而當沒有情緒時，說話呈現語調平穩，音高大致上是一致的，如圖 2-4(c)所示。而從其他的統計特徵，像是平均值、標準差、最大值與最小值，可以知道音高和能量的分佈情形，像是聲量小時，能量相關的特徵值較小；反之，能量相關特徵值相對上會比較大。

2.3 本章總結

本章節主要完成的部分為語音特徵擷取，先是利用越零率和短時距能量作端點偵測，找出有效語音的起點和終點，對這段有效語音設置音框，以音框為單位計算音高和能量，對整段有效語音內每個音框的音高和能量統計出 12 個語音特徵，以供後續辨識決策階段辨識不同的情緒。

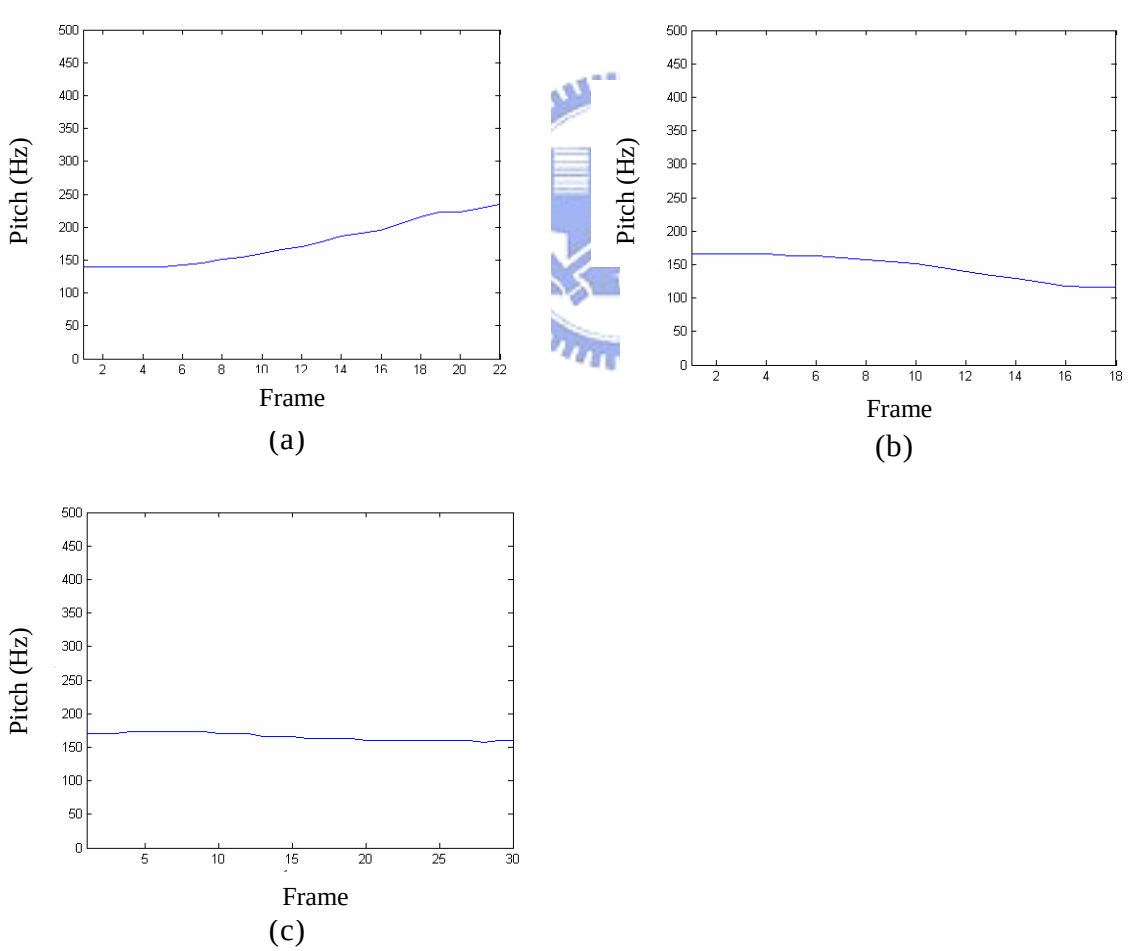


圖 2-4 音高測試結果

第三章 人臉影像特徵擷取

人臉影像特徵擷取，目的在於取出每種表情具有代表性和獨特性的特徵，使得在辨識階段能夠區分出彼此間的差別，辨識出正確的情緒。人臉特徵的擷取上，經由影像的前置處理，藉由膚色和人臉規則資訊作為人臉偵測的依據，找到人臉位置。接著根據瞳孔在正面上半臉影像灰階值強度最低的資訊，找出的瞳孔的確切位置，再根據瞳距和瞳孔位置定出眼睛和眉毛可能區域，利用影像灰階值強度和邊緣偵測找出眼睛與眉毛特徵點的位置。而嘴巴的特徵點是利用位於人臉的幾何位置和影像強度找出，找出所有特徵點之後，定出 12 個特徵點之間的距離當作影像特徵，作為分辨不同表情的依據。

3.1 人臉偵測

當 CMOS sensor 取得臉部影像後，首先需要人臉偵測，找出人臉的位置以便作接下來的人臉特徵點擷取，在此我們採用實驗室周崇民學長論文[32]設計的人臉偵測方法，流程如圖 3-1 所示。為了使系統能夠快速的搜尋到人臉位置，在進行人臉偵測之前，我們會先將輸入的 320×240 彩色影像縮小為 160×120 彩色影像，以減少人臉偵測所需的時間。在人臉偵測的過程當中，系統會先將輸入的影像進行膚色的

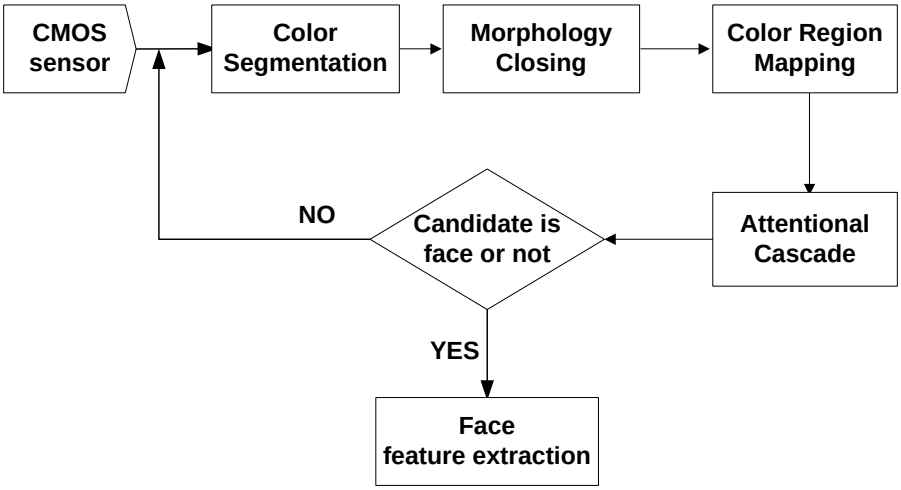


圖 3-1 人臉偵測系統流程圖

色彩分割，並且透過形態學的閉合運算（Closing）填滿膚色區域內空洞及不連續之處。接著我們透過膚色區域投影將屬於膚色的區域框選出來，並且利用投影的長寬比先排除可能不屬於人臉的區域，再將可能屬於人臉的區域交由專注式串聯法（Attentional Cascade）[33]來判別此區域是否為正面的人臉。以下各章節將對圖 3-1 的各元件分別加以說明。

3.1.1 膚色搜尋及選取

色彩分割就是找出膚色收斂的範圍並且設定膚色之色彩分佈上下閥值，若設定的範圍太大，則會將不屬於膚色的背景判定為膚色像素。因為膚色為人臉偵測的第一個步驟，必須先找到膚色分佈才可以進行色彩分割，所以我們根據在不同的環境及光源的膚色色彩分佈的情況，調整其膚色分佈閥值。圖 3-2(a)為膚色分割測試圖，圖 3-2(b)將膚色分割出來。接下來透過形態學的閉合運算（Closing）填滿膚色區域內空洞及不連續之處，如圖 3-2(c)所示。

接著我們希望從這些二值化的影像中找出可能為人臉的區域，以縮小搜尋的範圍，即利用先前所求得的色彩分割後的資訊 I_{skin} 進行可能人臉區域的判別。我們利用色彩分割後的資訊作水平軸與垂直軸上的投影以找出符合要求的區域。前面我們已求得的特定色彩資訊 I_{skin} 為一 M×N 的矩陣，其中 M 為影像高，N 為影像寬，$I_{skin}(x,y)=1$ 表示影像中(x,y)座標上的像素屬於膚色的色彩分佈，反之，則為非膚色。我們將 I_{skin} 在 Y 軸上進行投影求得 I_{skin} 在 Y 軸上的投影量 H(y)

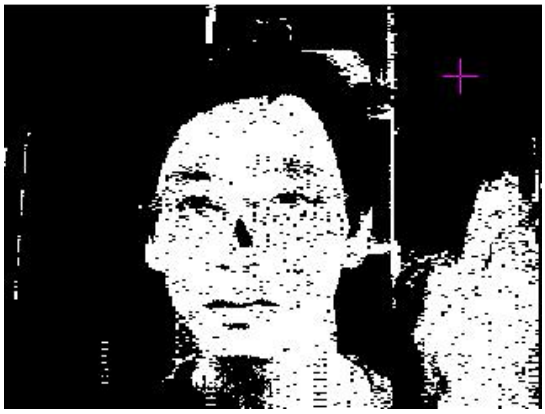
$$H(y)=\sum_{x=0}^{M} I_{skin}(x,y) \tag{3-1}$$

從 H(y)中根據設定最小的人臉長，選出 K 個 Y 方向的區間，接著我們再對這 K 個 Y 區間內的膚色資訊利用式(3-2)各別進行 X 軸上的投影，其投影量各別為 $W_i(x)$，i 為 1 到 K，YS 為起始投影 Y 值，YE 為最終投影 Y 值。

$$W_i(x)=\sum_{y=YS_i}^{YE_i} I_{skin}(x,y) \text{ ，i=1~K} \tag{3-2}$$

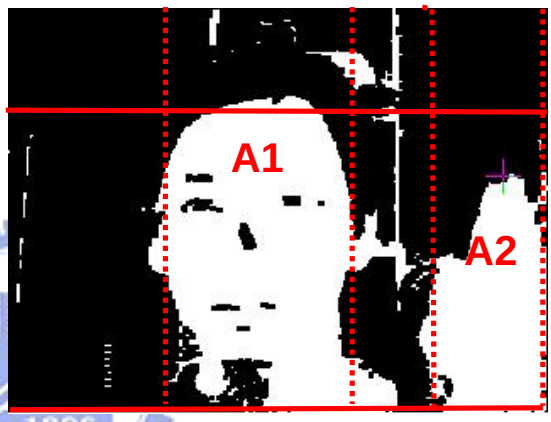
從 $W_i(x)$根據設定的最小的人臉寬進行分割，分割出數個可能為人臉的區間，如圖

(a)膚色分割測試圖

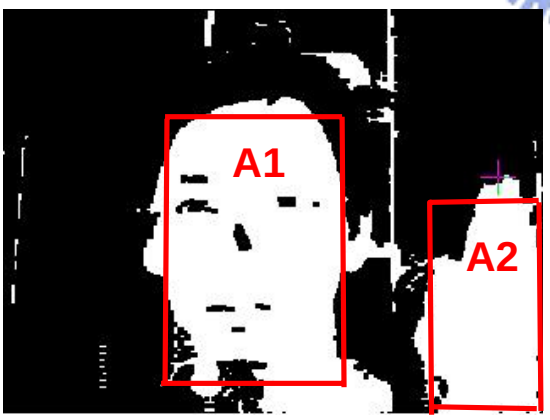


(b)膚色分割結果

(c)閉合運算



(d)第一次投影



(e)第二次投影

圖 3-2 膚色分割及投影

3-2(d)所示，經由第一次的投影，框選出 A1 及 A2 兩個區域，然後再對這兩個區域進行第二次的投影，如圖 3-2(e)所示，用意在於移除不是膚色的區域。

3.1.2 專注式串聯法

透過上一節將屬於膚色的區域框選出來，並且利用投影的長寬比先排除可能不屬於人臉的區域，再將可能屬於人臉的區域交由專注式串聯法（Attentional Cascade）來判別此區域是否為正面的人臉。專注式串聯法可解決過多的運算量浪費在判斷非人臉影像視窗的問題。在此人臉偵測器主要是使用數個簡單的臉部特徵，稱為簡易分類器（Simple Classifier），串聯成為一個複雜分類器（Complex Classifier）。愈前面的分類器規則愈容易，所需運算量也愈少，但是具有快速判斷影像視窗為人臉或者是背景之效果；愈後面特徵則是愈複雜人臉規則，所以運算量也隨之提高，不過當層數愈多時，人臉偵測的準確性也會提高。以圖 3-3 來說明，在判斷的過程中以 T 來表示通過某一層特徵的檢測，若為 F 則表示偵測失敗，也就是此影像視窗不是屬於人臉。

透過色彩投影及臉部長寬比初步的篩選之後，我們將可能為人臉區域送入串聯法判斷此區域是否為正面的人臉。我們用以下幾點規則做為判斷條件：

1) 人臉投影之長寬比為 1 到 2 倍。
2) 人臉上半部兩眼區域與兩眉之間之灰階值總和會小於下半部灰階值的總和。
3) 眼睛區域的灰階值總和會小於眉毛中間灰階值的總和。
4) 在兩側臉頰相鄰的上下區域其灰階值總和會小於某閥值。

利用串聯的方法來判斷輸入的影像視窗是否為正面的人臉，因為結合了膚色的條件，所以在人臉偵測時不必進行全域的掃描，如此可以減少運算量以達到即時偵測人臉的效果。

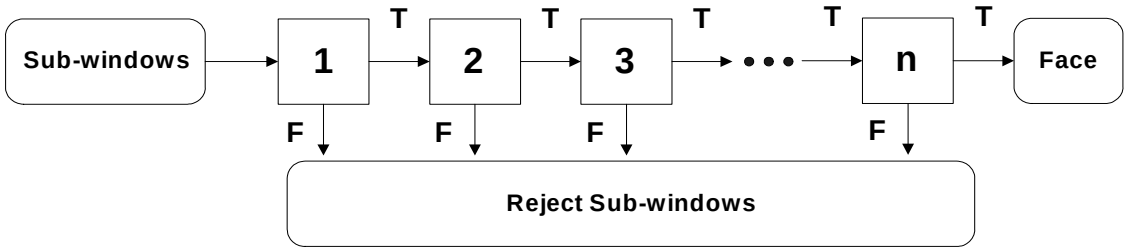


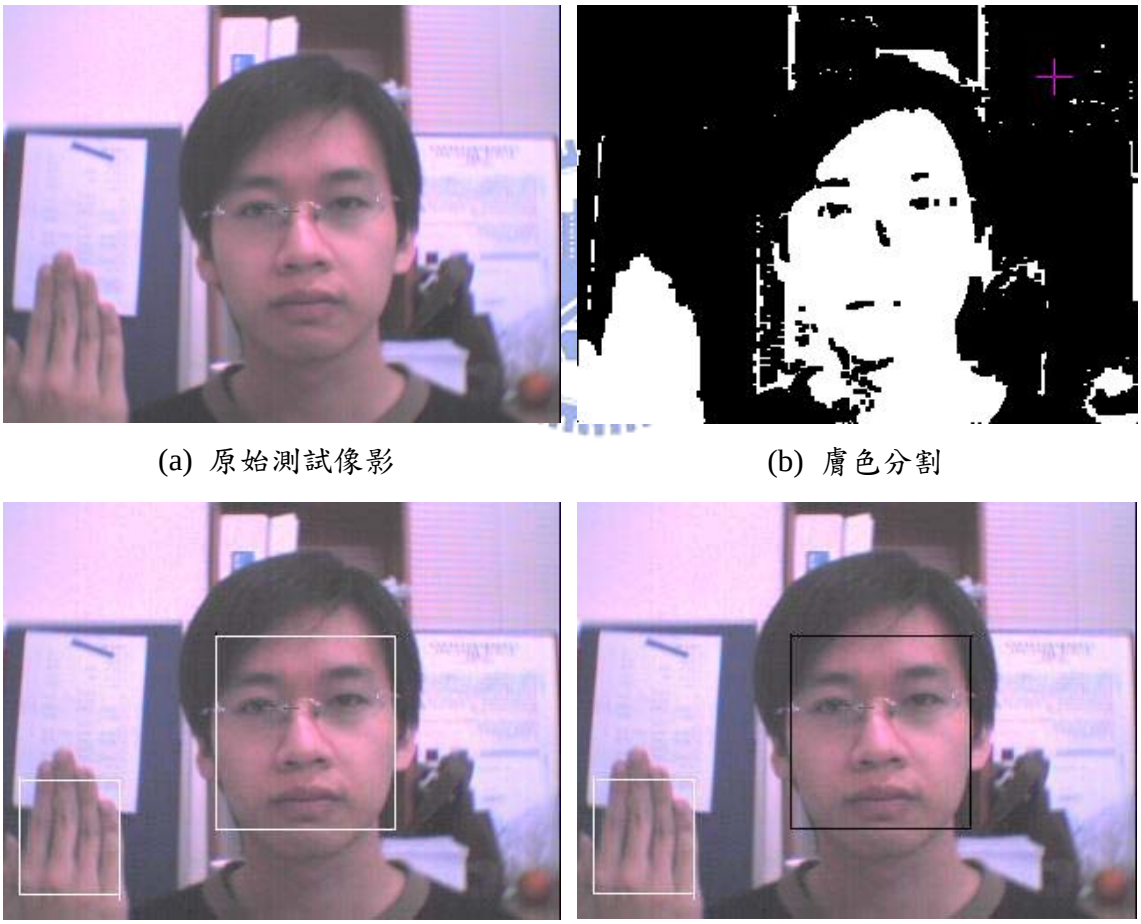
圖 3-3 人臉偵測專注式串聯法

3.1.3 人臉偵測結果

圖 3-4 為正面人臉偵測之結果。圖 3-4(a)為原始之測試影像，利用色彩分割將屬於膚色的範圍分割出來且作閉合運算，圖 3-4 (b)即為運算之後的結果。圖 3-4 (c)為利用膚色投影之後所框選之區域，圖 3-4 (d)為加入人臉特徵判斷之後得到的結果，黑色框框表示找到的人臉位置。

3.2 人臉特徵點擷取

經過人臉偵測後，可以找出人臉的位置和大小，若是將整個人臉影像當作表情辨識特徵作判斷，有過多冗餘的資料，計算量會過大；另外，我們也不能確定人臉



(a) 原始測試像影　(b) 膚色分割

(c) 膚色投影　(d) 加入特徵判斷之處理結果

圖 3-4 正面人臉偵測實驗結果

在影像中的大小，所以必須找出幾個足以代表人臉表情的特徵點，利用這些特徵點位置的變化當作特徵，除去不必要的資訊，減少運算量。

圖 3-5 為人臉特徵點的選取[34]，利用人眼及眉毛會在上半臉出現，嘴巴會在下半臉出現，找尋各特徵點的位置，共有 14 個特徵點，左右眼睛各 3 個，左右眉毛各 2 個，嘴巴有 4 個特徵點，找到這些特徵點的位置後，利用它們之間的距離當作特徵，辨識臉部表情。

3.2.1 眼睛特徵點擷取

眼睛特徵點擷取方法如圖 3-6 所示，經由人臉偵測找到人臉後，首先要找出瞳孔的確切位置，由於瞳孔和上半臉附近區域比較起來，會有較黑的顏色，灰階值是最低的部分，因此針對上半臉區域眼睛所在的可能區域，我們是設定在臉部寬$\frac{1}{8}$到$\frac{7}{8}$，長為由上$\frac{1}{12}$到$\frac{6}{12}$的區域內，如圖 3-7(a)所示，計算出灰階值分佈的直方圖

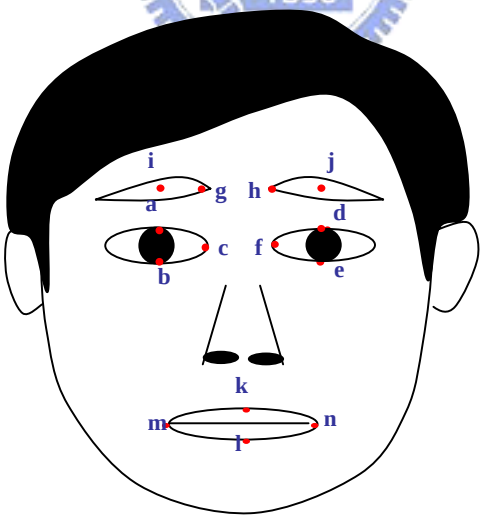


圖 3-5 人臉特徵點的選取

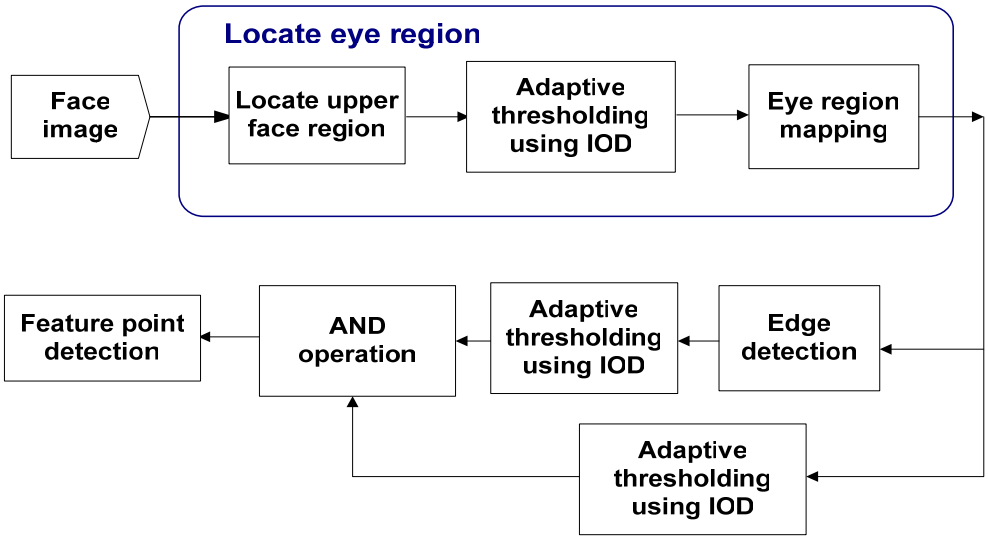
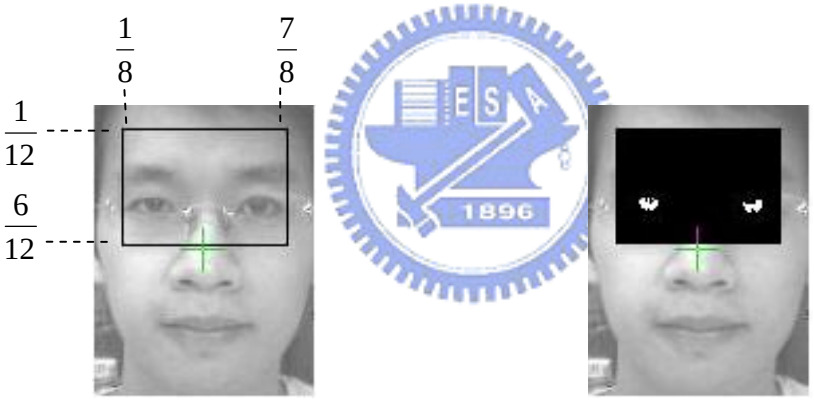
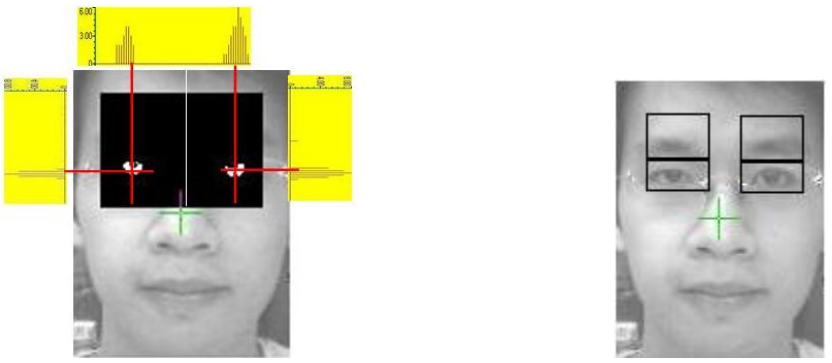


圖 3-6 眼睛特徵點偵測方塊圖



(a) 人眼可能位置　　(b) IOD 動態閥值二值化



(c) 人眼位置投影　　(d) 框選出眼睛範圍

圖 3-7 眼睛區域定位

(histogram)，決定臨界值取出最黑的像素。但環境的亮度會有所變動，要取出最黑的像素點即需要動態地設置臨界值，我們採用 IOD(Integral Optical Density)[35]，D 為眼睛區域，$Y_k(u,v)$ 為以 k 為閥值之眼睛區域 $f(u,v)$ 的二值化影像：

$$Y_k(u,v)=\begin{cases}0 & if \quad f(u,v)>k\\ 1 & if \quad f(u,v)\le k\end{cases} \tag{3-3}$$

由直方圖中灰階值最低的部份開始算起，累加到此區域像素個數總和的 3% ($IOD=0.03$)，將此像素的灰階值 k 設為臨界值

$$IOD=\frac{\iint_D Y_k(u,v)dudv}{\iint_D dudv} \tag{3-4}$$

最後將小於臨界值則將影像灰階值二值化設為 1，否則設為 0，如圖 3-7(b)所示。接著對二值化後的眼睛區域影像兩端分別作水平投影 H(y)與垂直投影 V(x)，M 為眼睛區域的影像高，N 為眼睛區域的影像二分之一寬

$$H(y)=\sum_{x=0}^{M} I_{eye}(x,y) \tag{3-5}$$

$$V(x)=\sum_{y=1}^{N} I_{eye}(x,y) \tag{3-6}$$

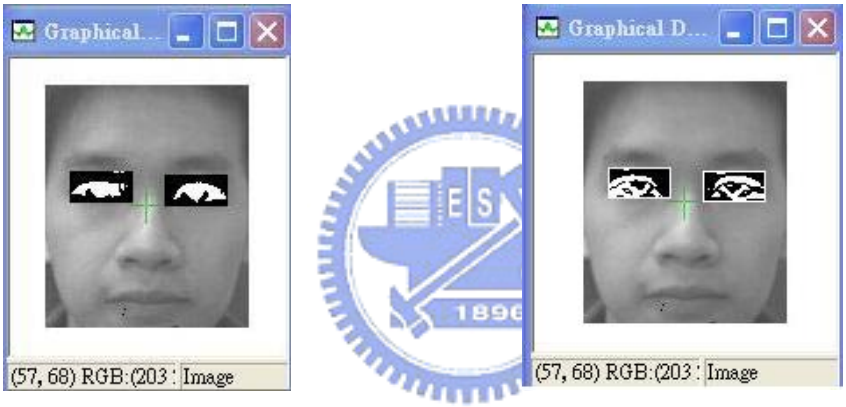
分別找出投影量最大值，即為找到瞳孔的確切位置，如圖 3-7(c)所示，並依照瞳孔距離的大小比例，框選出眼睛和眉毛的範圍，眼睛區域的高度為三分之一瞳距，寬度為三分之二瞳距，而找到眼睛的區域後，也一道將眉毛的區域框選出來，眉毛區域的高度為二分之一瞳距，寬度同樣為三分之二瞳距，如圖 3-7 (d)所示。

找到眼睛所在區域後，為了要能穩健地找出眼睛的上下與內側特徵點，先對眼睛區域內找出亮度最深的部分，用 IOD 動態地設置臨界值 Th1 找出眼睛，如式(3-7)所示，這裡我們設眼睛區域像素達像素總和的 25%為臨界值，結果如圖 3-8(a)所示。

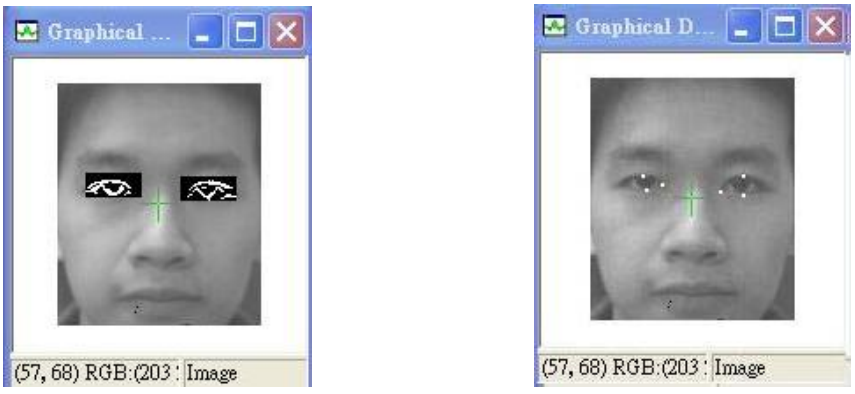
$$I_{eye}=\begin{cases}1 & if \quad I(x,y)<Th1\\ 0 & ,otherwise\end{cases} \tag{3-7}$$

另一方面，對眼睛區域作邊緣偵測且用 IOD 作二值化找出眼睛的輪廓，臨界值設為達像素總和的 50%。邊緣偵測是應用影像中連續像素之間的灰階值達到某種程度的變化，也就是在某一像素上的梯度變化大於某個閥值，此像素可視為邊緣影像的一部分。影像的梯度變化是有 x 方向和 y 方向所組成，梯度向量可由式(3-8)表示，梯度大小可由式(3-9)表示。

$$\nabla I(x,y)=\begin{bmatrix}\dfrac{\partial I(x,y)}{\partial x}\\ \dfrac{\partial I(x,y)}{\partial y}\end{bmatrix}=\begin{bmatrix}G_x\\ G_y\end{bmatrix} \tag{3-8}$$



(a) IOD 設閥值二值化　　(b) 邊緣偵測與 IOD 設閥值二值化



(c) 圖(a)與圖(b)作 AND 運算　　(d) 找出雙眼的上下與內側特徵點

圖 3-8 雙眼 6 個特徵點

$$\nabla f = |G_x| + |G_y| \tag{3-9}$$

梯度的運算元有兩個，分別是水平邊緣檢測運算元G_x，和垂直邊緣檢測運算元G_y，我們是採用 Sobel 運算元，將原始灰階影像 I 分別和水平與垂直運算元作摺積運算得到 I_x 及 I_y。

$$I_x = I * G_x \tag{3-10}$$

$$I_y = I * G_y \tag{3-11}$$

$$G_x = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, \quad G_y = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{3-12}$$

邊緣影像 I_{edge} 為 I_x 與 I_y 的絕對值之和，再用 IOD 決定的閥值 Th2 濾掉較小的梯度值，如圖 3-8(b)所示。

$$I_{edge}(x,y) = \begin{cases} 1 & if \quad |I_x(x,y)| + |I_y(x,y)| > Th2 \\ 0 & , \quad otherwise \end{cases} \tag{3-13}$$

最後再將兩者作 AND 運算，如圖 3-8(c)所示，找出眼睛的確切位置，找出上、下與內側的特徵點，如圖 3-8(d)所示。

$$I_{eye\_contour}(x,y) = I_{eye}(x,y) \,\&\&\, I_{edge}(x,y) \tag{3-14}$$

3.2.2 眉毛特徵點擷取

找尋眉毛特徵點的方法大致上和眼睛相同，偵測流程圖如圖 3-9 所示，由於眉毛會位在眼睛上方的區域，所以先根據前一節所述找到眼睛的可能區域後，接著在眼睛的上方定出眉毛的可能區域，高度為二分之一瞳距，寬度與眼睛同樣為三分之二瞳距。接著用 IOD 動態設置臨界值並二值化眉毛區域影像，臨界值設為達像素總和 30%的灰階值，找出眉毛的大致位置。為了要穩健地找出眉毛位置，同樣與加上

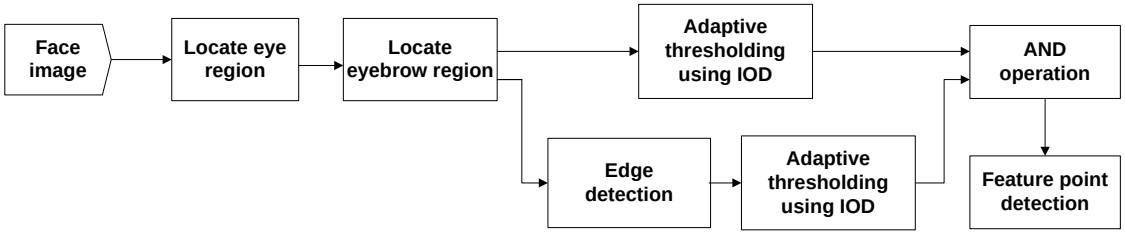
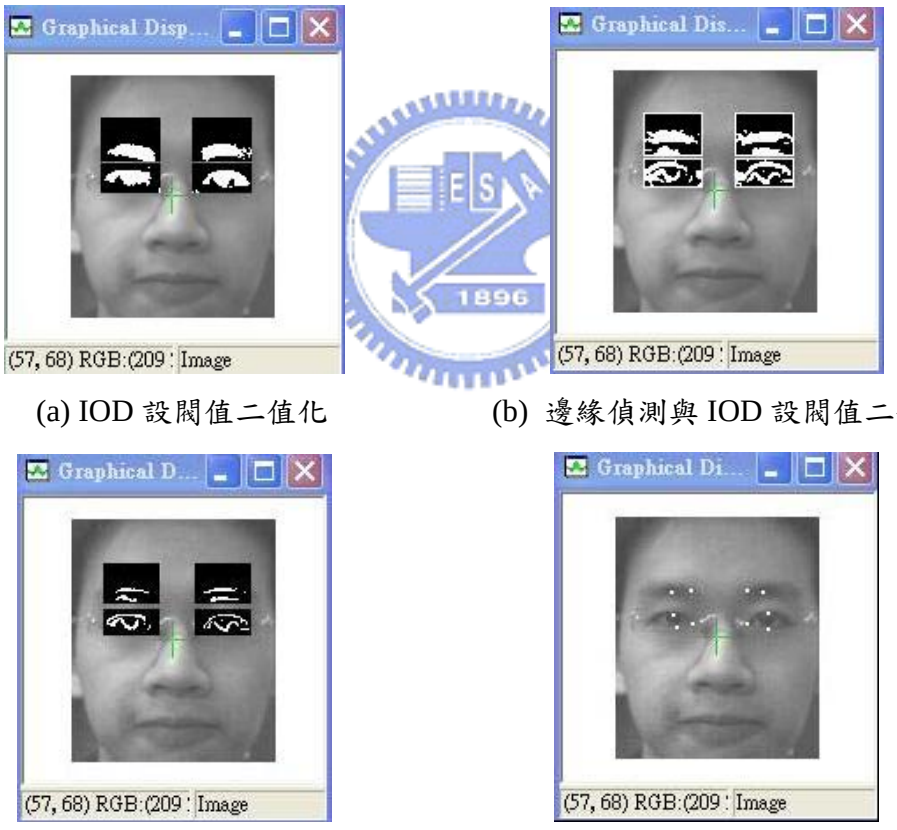


圖 3-9 眉毛特徵點偵測方塊圖

邊緣偵測作二值化後的影像作 AND 運算，邊緣偵測的臨界值設為像素總和 40%的灰階值，最後找出雙眉的中心與內側特徵點，過程如圖 3-10 所示。



(a) IOD 設閥值二值化　　(b) 邊緣偵測與 IOD 設閥值二值化

(c) 圖(a)與圖(b)作 AND 運算　　(d) 找出眉毛的內側與中心特徵點

圖 3-10 眉毛與雙眼 10 個特徵點

3.2.3 嘴唇特徵點擷取

在下半臉的可能區域裡找尋嘴唇特徵點，由於 CMOS sensor 取得的影像在色彩的表現上不是很鮮豔，造成原本要用色彩資訊找尋嘴唇位置的困難，於是只用影像灰階值強度找尋嘴唇，偵測方塊圖如圖 3-11 所示。首先根據人臉的比例定出一個區域，我們是設定在臉部寬$\frac{3}{16}$到$\frac{13}{16}$，長為由上$\frac{6}{10}$到底部的區域內，如圖 3-12(a)所示。因嘴唇的灰階值大致比臉頰和下巴暗，故採用 IOD 動態設置臨界值作二值化，臨界值設為達像素總和 25%的灰階值。接著對這個區域作水平與垂直投影，找出嘴唇的上下左右的特徵點位置，如圖 3-12(b)(c)所示。

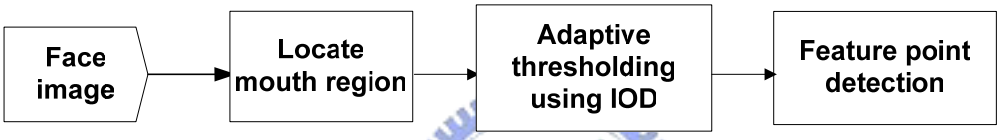
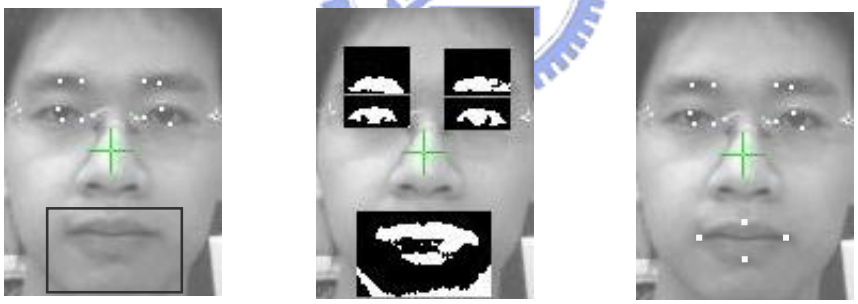


圖3-11 嘴唇偵測流程圖



(a) (b) (c)

圖 3-12 嘴唇區域作 IOD、二值化，找出特徵點

3.2.4 臉部影像特徵值選取

經由前兩節所述的方法擷取出人臉特徵點後，需要決定可以表達人臉表情變化的特徵值，由於表情的變化多是靠著眼睛、眉毛與嘴巴的大小及位置變化，所以在特徵值的決定上就根據這個原則定出 12 個特徵值，為眉毛到眼睛的距離、眼睛與嘴

巴大小和眼睛到嘴巴距離，如圖 3-13 所示，描述如表 3-1 所述。但由於人臉與 CMOS sensor 之間距離的不同，會造成影像中人臉大小的不同，影響特徵值的大小，所以特徵值需要針對距離作正規化(normalized)，減少因距離的不同造成的影響。由於雙眼內眼角的距離是固定不變的，這個不變的距離即當作基準，每個特徵除以這個距離，得到正規化後的特徵值。

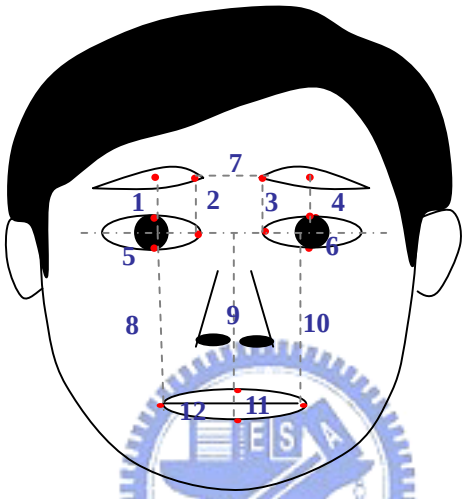


圖 3-13 臉部 12 個特徵值

編號	特徵	描述
1	E_1	右眉中心與右眼中心距離
2	E_2	右眉內側與雙眼平行線距離
3	E_3	左眉內側與雙眼平行線距離
4	E_4	左眉中心與左眼中心距離
5	E_5	右眼上下距離
6	E_6	左眼上下距離
7	E_7	雙眉內側距離
8	E_8	右嘴角與雙眼平行線距離
9	E_9	上嘴角與雙眼平行線距離
10	E_{10}	左嘴角與雙眼平行線距離
11	E_{11}	上下嘴角距離
12	E_{12}	左右嘴角距離

表 3-1 人臉的 12 個特徵

第四章 雙模情緒辨識系統演算法

本章說明雙模式情緒辨識系統設計概念，圖 4-1 為辨識系統架構圖。對 CMOS sensor 取得的影像作人臉偵測後，並接著計算出人臉表情的 12 個特徵值，同樣地經由麥克風取得帶有情緒的聲音，經由前置處理後，以音框為單位來計算並統計出語音的 12 個特徵值。這兩類特徵先經由 SVM 分類，計算出每筆特徵距 hyperplane 的距離，並與資料庫中的訓練資料的距離比較並正規化，比較語音和人臉影像的權重大小，即選擇權重較大的分類結果，即可決定出較可靠的分類結果，得到受測者表現的情緒狀態。

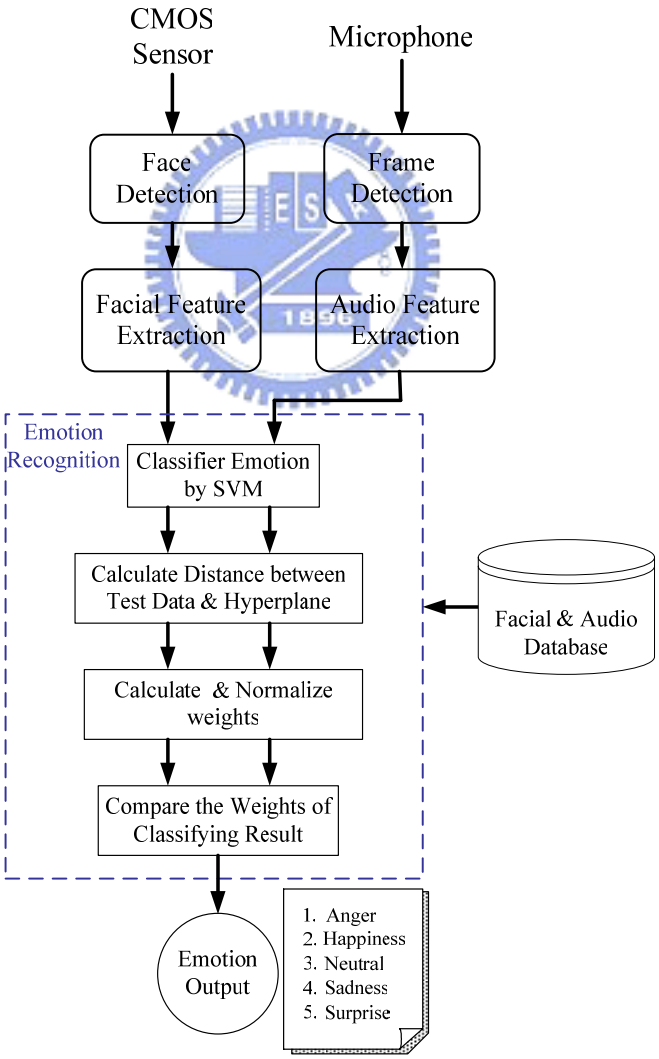


圖 4-1 情緒辨識系統架構圖

4.1 SVM 分類—線性可分割問題

取得人臉影像和語音資料後，擷取代表這筆資料的特徵，接著就需要作特徵的比對分類，所以提供特徵分類的方法是一個重要的問題。在分類方法上，我們採用SVM(Support Vector Machines)[36-37]的分類方法，是一種以統計學習理論(Statistical Learning Theory)為基礎，而發展出來的機器學習系統，設計概念是用於處理兩類的分類問題，優點是具有清楚的理論與完整的架構，並且實作分類效果良好。SVM首先要經由訓練的過程，用訓練的資料算出分開兩類資料的分割平面(hyperplane)，在辨識過程即用這訓練出的hyperplane分類。

在圖 4-2 中，表示在這個空間中有多筆資料 $x_i (i = 1 \sim l)$，經由一個線性函數定義 hyperplane 的位置，將所有的輸入資料 x_i 區分為兩類，標記為 $y_i = \{+1, -1\}$，hyperplane 的函數定義為 $w \cdot x + b = 0$，w 為 hyperplane 之法向量(Normal Vector)。距離 hyperplane 最近的幾筆資料就是所謂的 support vector，代入 hyperplane 的函數即等於+1 或–1，就是圖中的兩條虛線。在處理可區分為二類的資料時，SVM 會找尋一個具有最大邊界距離的分隔平面，符合下列兩個限制式[37]：

$$w \cdot x_i + b \geq +1 \quad \text{for} \quad y_i = +1 \tag{4-1}$$

$$w \cdot x_i + b \leq -1 \quad \text{for} \quad y_i = -1 \tag{4-2}$$

可將(4-1)與(4-2)二式結合為以下不等式：

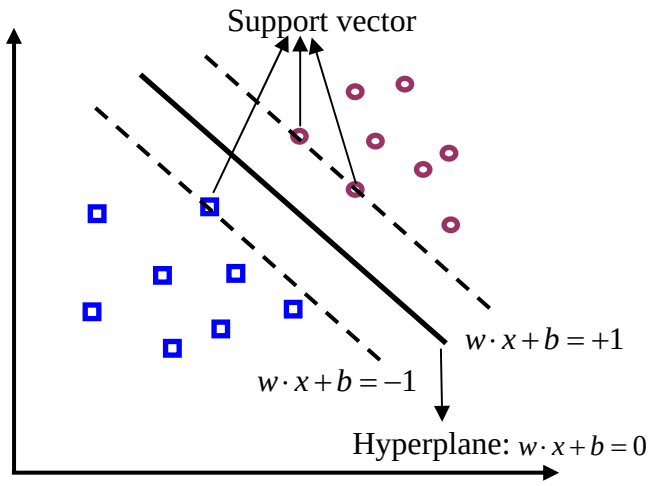


圖4-2 SVM示意圖

$$y_i(w\cdot x_i+b)-1\geq 0 \qquad \forall i \tag{4-3}$$

Support vector 與 hyperplane 的距離為 $\frac{1}{\|w\|}$，因此可以得到不只一個 hyperplane 可將兩類資料分開，為求得具有最大邊界的 hyperplane，邊界距離為 $\frac{2}{\|w\|}$，在符合限制式(4-3)的條件下，因此要求得 $\frac{\|w\|^2}{2}$ 的最小值。在線性不等式的限制下求最佳化問題，根據 Karush-Kuhn-Tucker 條件，把原本最佳化問題轉為其對應的對偶(dual)問題。它的拉格蘭吉(Lagrange)為

$$L(w,b,\alpha)\equiv\frac{1}{2}\|w\|^2-\sum_{i=1}^{l}\alpha_i[y_i(w\cdot x_i+b)-1] \tag{4-4}$$

其中拉格蘭吉係數(Lagrange Multipliers) $\alpha_i, i=1,...,l$ ，且 $\alpha_i\geq 0$ 。且要滿足

$$\frac{\partial L(w,b,\alpha)}{\partial w}=0 \quad ， \quad 得到 \quad w=\sum_{i=1}^{l}\alpha_i y_i x_i \tag{4-5}$$

$$\frac{\partial L(w,b,\alpha)}{\partial b}=0 \quad ， \quad 得到 \quad \sum_{i=1}^{l}\alpha_i y_i=0 \tag{4-6}$$

將(4-5)(4-6)代入(4-4)式後，得到

$$L(w,b,\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j x_i\cdot x_j \tag{4-7}$$

原本求 $L(w,b,\alpha)$ 的最小值問題，其對偶問題變為求最大值，限制式為(4-5)(4-6)和 $\alpha_i\geq 0$ 。

在求對偶問題的最佳解時，每一個拉格蘭吉係數 α_i 都有對應每一筆訓練資料；如果 $\alpha_i>0$ 表示該資料是此問題的支持向量，會落在分隔平面的邊界上，將 α_i 代入(4-5)式可求得 w 。為了求得 b ，可以利用 Fletcher 提出的 Karush-Kuhn-Tucker complementarity conditions：

$$\alpha_i(y_i(w\cdot x_i+b)-1)=0 \qquad \forall i \tag{4-8}$$

最後得到一個可以處理分類問題的函數：

$$f(x)=\text{sgn}\left(\sum_{i=1}^{l}y_i\alpha_i\cdot(x\cdot x_i)+b\right) \tag{4-9}$$

當 $f(x)>0$ 時，表示該資料與標註為"+1"的資料屬於同一類；反之則屬於另一類。

4.2 關鍵性特徵辨識人臉表情

我們是採用 SVM 當分類器，是由多個兩個不同的表情的分類器作分類，可以看出每組比對的兩種表情在選定的特徵上的不同變化程度。另外，由於 SVM 要分類資料的類別已經決定，屬於監督式學習(Supervised learning)，常會有如圖 4-3 的情況發生，使用分類的資訊複雜度愈高，像是特徵的維度愈高，分類器在分類資料時會過於精確，完全適合於訓練資料分佈情形的分類，使得訓練資料分類錯誤的情形會愈少；然而，對於測試資料卻會造成較高的誤差[38]。因此在分類上選定 Model 的複雜度愈高，並非測試資料的錯誤率就愈低。另一方面，將分類的問題從機率的角度來看，通常使用貝氏分類器(Bayes classifier)來做分類決定，其決策方式如下：

If $p(C_1 \mid x) > p(C_2 \mid x)$ then 判定為第一類 (4-10)

If $p(C_2 \mid x) > p(C_1 \mid x)$ then 判定為第二類 (4-11)

其中 $p(\cdot)$ 表示機率函式，C_1 與 C_2 分別表示第一、二類，x 表示已知的特徵。以圖 4-4 為例，根據貝氏分類器可以得到分類的臨界點 Thr_a 與 Thr_b，然後靠判斷 x 與 Thr_a 或 Thr_b 間的大小關係來決定所屬類別。在圖(a)中，由於兩類之機率分佈重疊較大，因此其判斷所屬類別之誤差將較大；反之，在圖(b)中，由於分佈重疊較小，因此其判斷所屬類別之誤差將較小。這結果說明了在分類時，若能選定兩類的機率函式中是屬於較分開者，那將會使得分類後的誤差較小。

基於以上兩個理由，根據要比較的兩類表情，挑選出機率函式重疊部份較小者，也就是肉眼可觀察到變化較明顯的部份作為具代表性的關鍵性特徵(Key feature)，取代原有的 12 個全特徵(All feature)。如此設計一方面使得兩類機率函式之間的重疊部分減少提高辨識效果；另一方面亦可以提高辨識的執行速度。

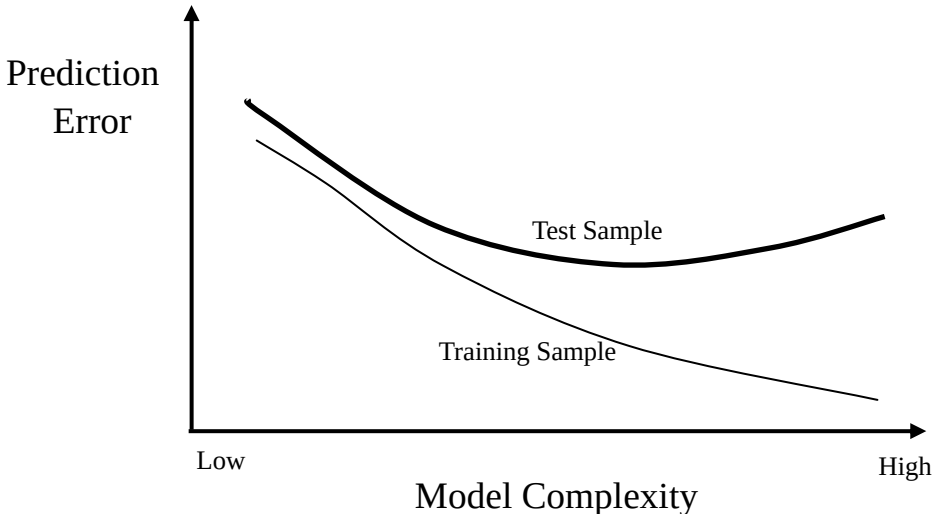


圖 4-3 訓練和測試資料複雜度與辨識誤差的關係

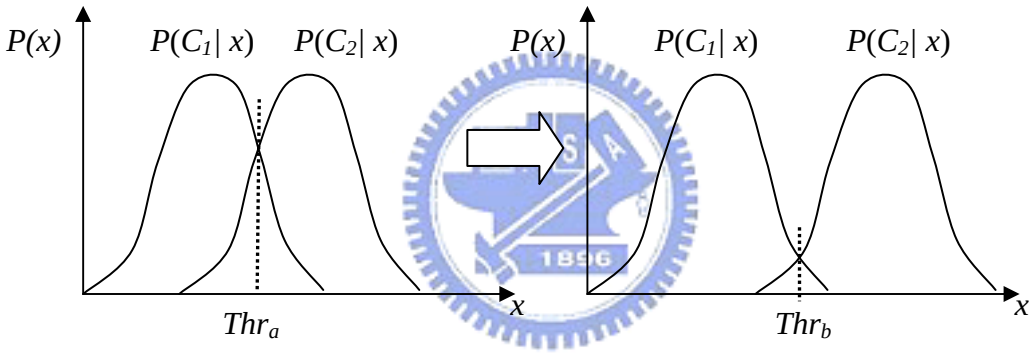


圖 4-4 資料重疊部分變少

選用表 4-1 的 8 個特徵當作下列 4 組表情比對用的關鍵性特徵，主要是在於這幾組表情的眉毛眼睛之間的距離變化、眼睛大小與嘴巴的高度變化較明顯。

表 4-1 第一組關鍵性特徵

1	右眉中心與右眼中心距離 E_1	5	右眼上下距離 E_5
2	右眉內側與雙眼平行線距離 E_2	6	左眼上下距離 E_6
3	左眉內側與雙眼平行線距離 E_3	7	雙眉內側距離 E_7
4	左眉中心與左眼中心距離 E_4	8	上下嘴角距離 E_{11}

Surprise vs. Sadness

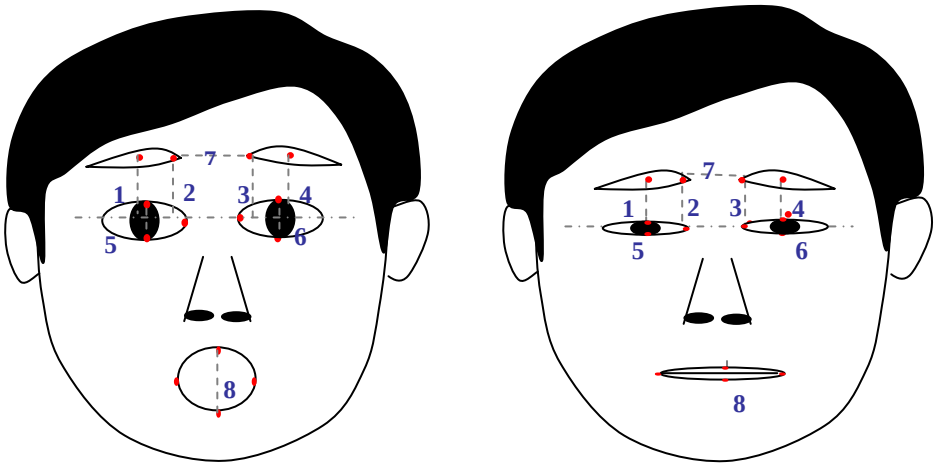


圖 4-5 驚訝與傷心表情的比較

Sadness vs. Anger

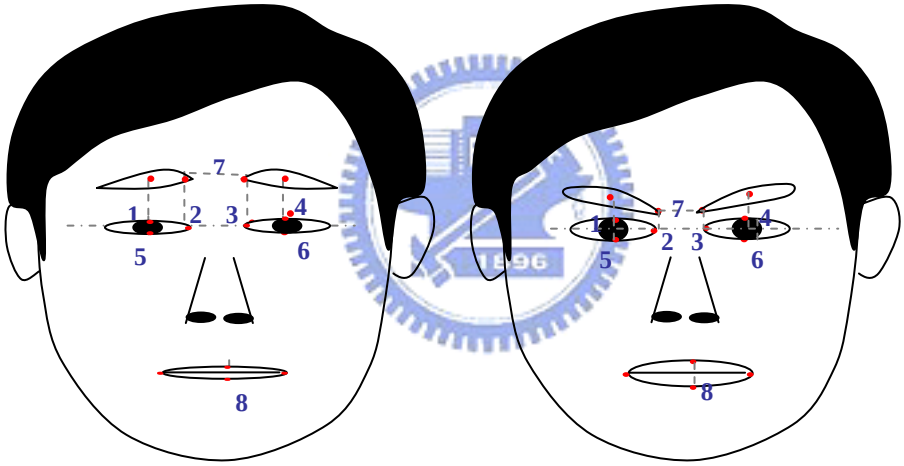


圖 4-6 傷心與生氣表情的比較

Neutral vs. Happiness

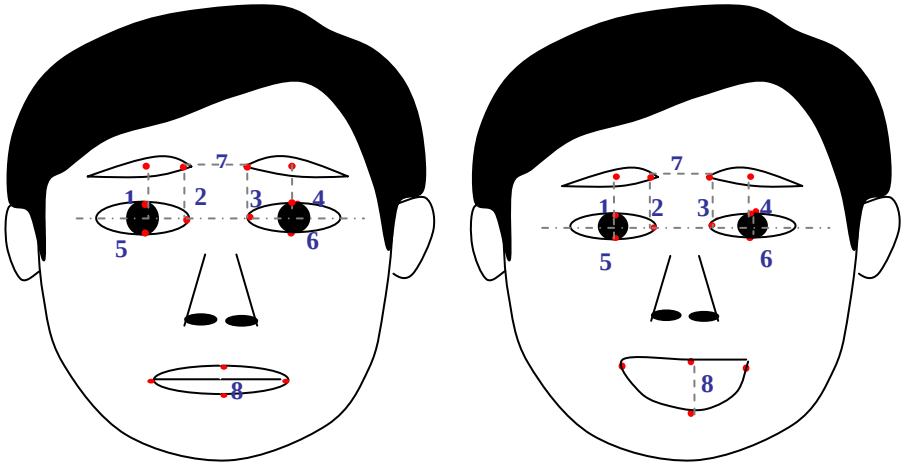


圖 4-7 普通與高興表情的比較

Anger vs. Happiness

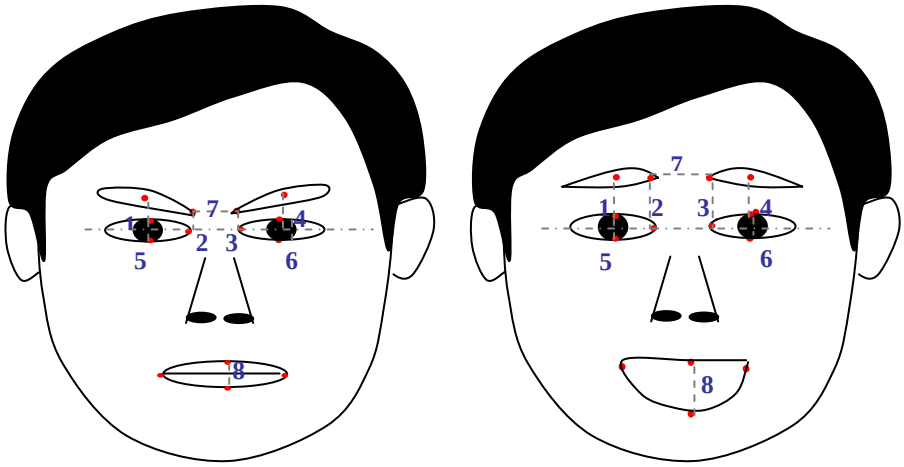


圖 4-8 生氣與高興表情的比較

選用表 4-2 的 8 個特徵當作下列 1 組表情比對用的關鍵性特徵，除了嘴巴的變化由高度變化改為左右大小變化，特徵的選取和第一組是相同的，主要在於強調驚訝和快樂時的嘴形左右寬度不同。

表 4-2 第二組關鍵性特徵

1	右眉中心與右眼中心距離 E_1	5	右眼上下距離 E_5
2	右眉內側與雙眼平行線距離 E_2	6	左眼上下距離 E_6
3	左眉內側與雙眼平行線距離 E_3	7	雙眉內側距離 E_7
4	左眉中心與左眼中心距離 E_4	8	左右嘴角距離 E_{12}

Surprise vs. Happiness

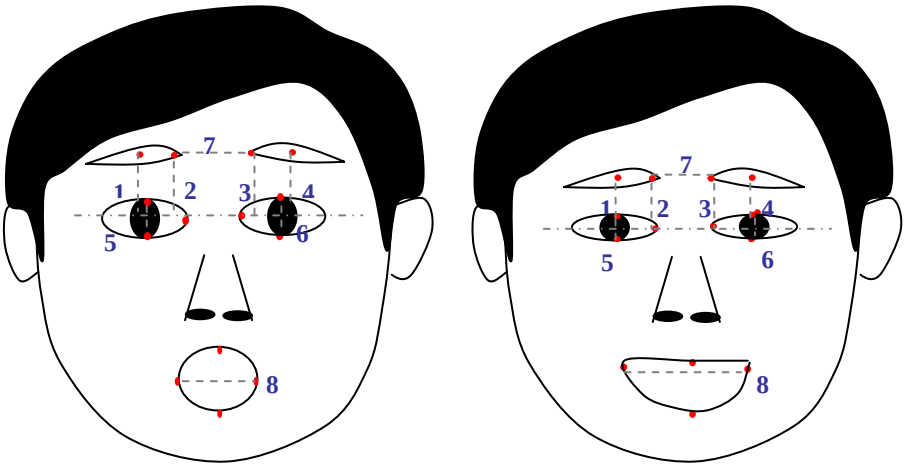


圖 4-9 驚訝與高興表情的比較

選用表 4-3 的 6 個特徵當作下列 2 組表情比對用的關鍵性特徵，主要是在於眉毛眼睛之間的距離變化與嘴巴的高度變化，像是驚訝通常都會眉毛上揚，相對於普通情緒狀態來說，這部分就會很明顯；而生氣通常會皺眉頭，這部分和驚訝的眉毛上揚就有很明顯的差別。

表 4-3 第三組關鍵性特徵

1	右眉中心與右眼中心距離 E_1	4	左眉中心與左眼中心距離 E_4
2	右眉內側與雙眼平行線距離 E_2	5	雙眉內側距離 E_7
3	左眉內側與雙眼平行線距離 E_3	6	上下嘴角距離 E_{11}

Surprise vs. Neutral

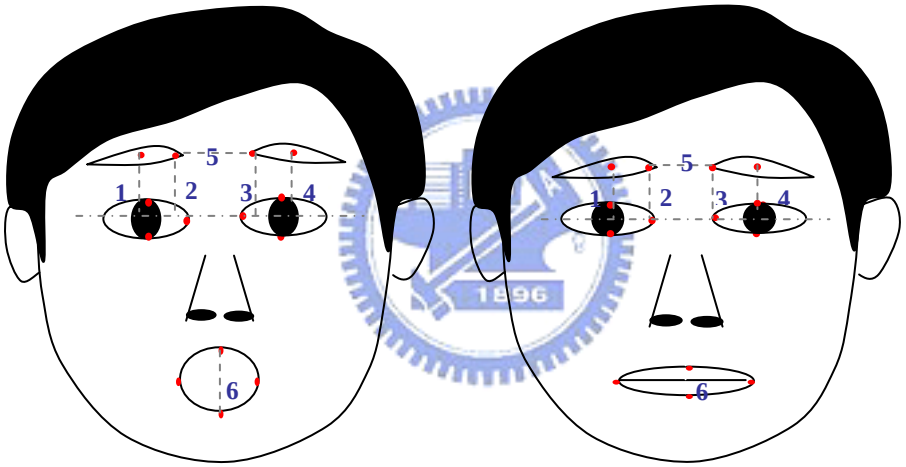


圖 4-10 驚訝與普通表情的比較

Surprise vs. Anger

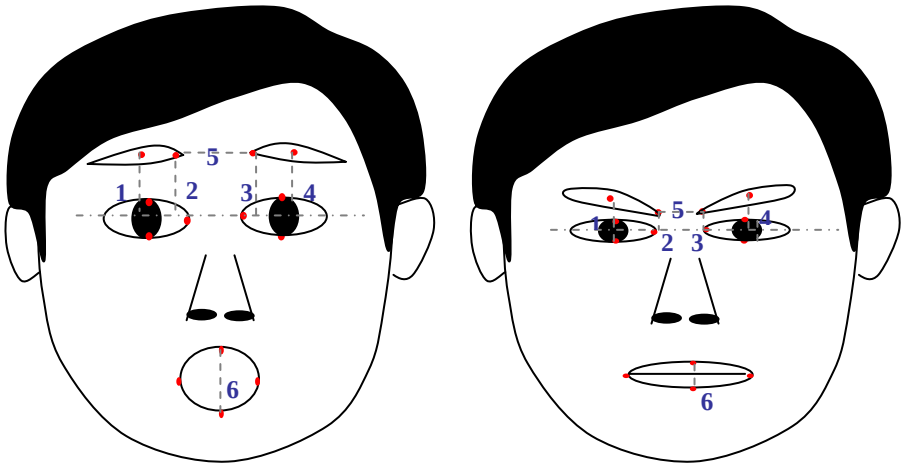


圖 4-11 驚訝與生氣表情的比較

選用表 4-4 的 7 個特徵當作下列 2 組表情比對用的關鍵性特徵，主要是在於眉毛眼睛之間的距離變化、眼睛大小與嘴巴的高度變化，像是傷心會使得眼睛下看，造成眼睛相對上較小，嘴巴也會內縮，和其他情緒狀態相比，這部分就會很明顯。

表 4-4 第四組關鍵性特徵

1	右眉中心與右眼中心距離 E_1	5	右眼上下距離 E_5
2	右眉內側與雙眼平行線距離 E_2	6	左眼上下距離 E_6
3	左眉內側與雙眼平行線距離 E_3	7	上下嘴角距離 E_{11}
4	左眉中心與左眼中心距離 E_4		

Sadness vs. Neutral

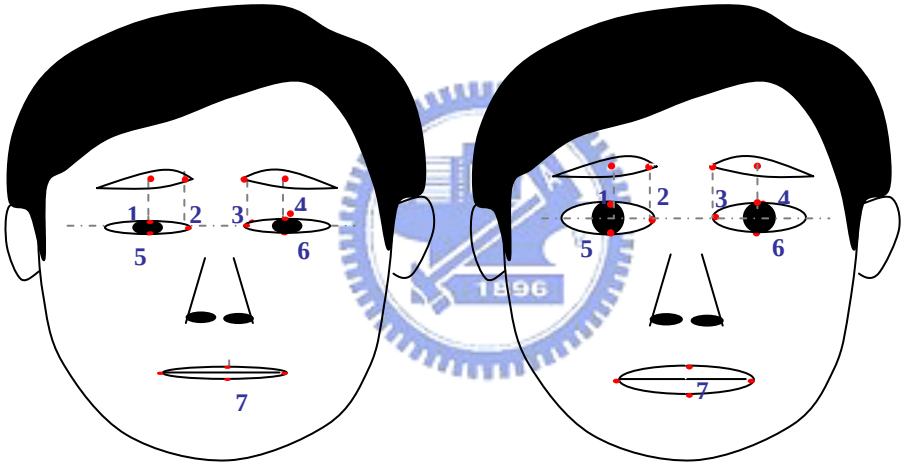


圖 4-12 傷心與普通表情的比較

Sadness vs. Happiness

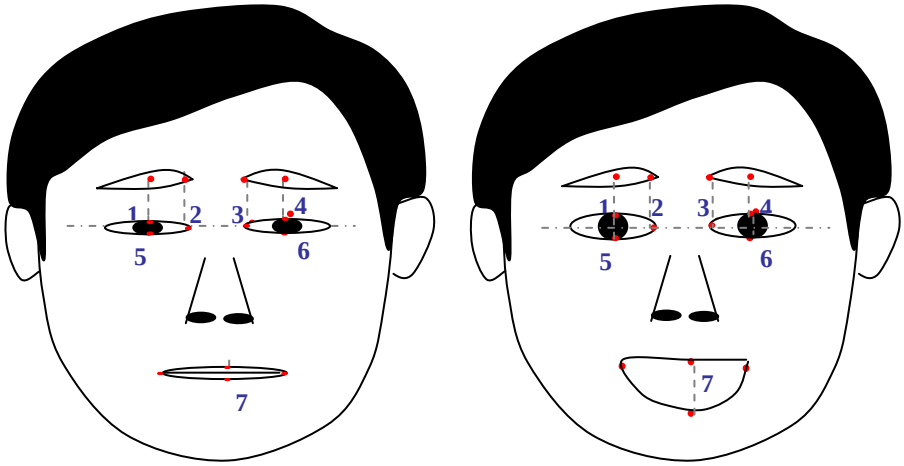


圖 4-13 傷心與高興表情的比較

選用表 4-5 的 7 個特徵當作下列 1 組表情比對用的關鍵性特徵，主要是在於眉毛眼睛之間的距離變化與眼睛大小，像是生氣會皺眉，和普通狀態相比，這部分就很明顯，而嘴巴的改變不明顯就不採用。

表 4-5 第五組關鍵性特徵

1	右眉中心與右眼中心距離 E_1	5	右眼上下距離 E_5
2	右眉內側與雙眼平行線距離 E_2	6	左眼上下距離 E_6
3	左眉內側與雙眼平行線距離 E_3	7	雙眉內側距離 E_7
4	左眉中心與左眼中心距離 E_4		

Neutral vs. Anger

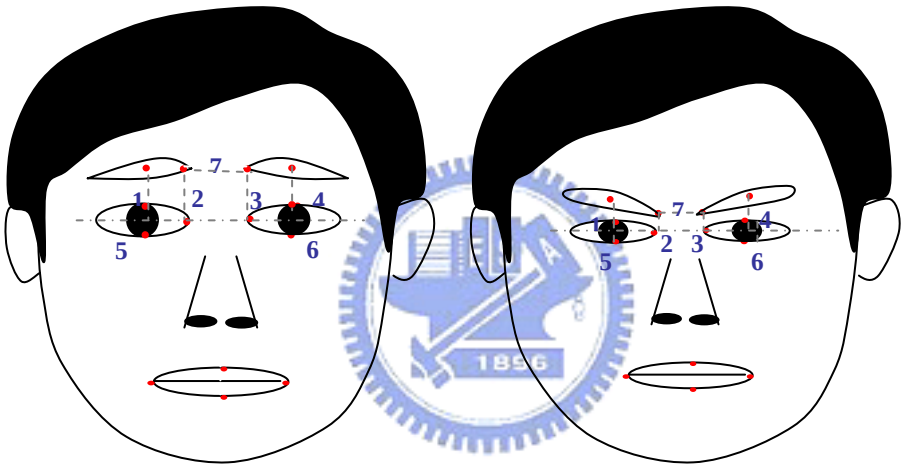


圖 4-14 普通與生氣表情的比較

4.3 SVM 分類—線性不可分割問題

前一節敘述的方法，是在兩類資料在可線性分割的情況下作分類，但資料若是線性不可分割的兩類資料(Nonseparable classes)，前一節的分類方法就不能有效分類，因此在原來的 hyperplane 限制式中加上鬆弛變數(Slack variable：$\xi_i \geq 0$)[37]， 原本的最大邊界的限制式變為式(4-12)與(4-13)，稱為柔性邊界(soft margin)。

$$w \cdot x_i + b \geq +1 - \xi_i \qquad \text{for} \quad y_i = +1 \tag{4-12}$$

$$w \cdot x_i + b \leq -1 + \xi_i \qquad \text{for} \quad y_i = -1 \tag{4-13}$$

可將(4-12)和(4-13)合併為

$$y_i(w\cdot x_i+b)\geq 1-\xi_i \qquad \forall i \tag{4-14}$$

可由上式看出，當訓練資料在分類發生錯誤時，ξ_i 會大於零，使得所有資料能落在 $w\cdot x_i+b=\pm 1$ 之間的間距之外，如圖 4-15 所示。因此在計算分割平面時，$\sum_i \xi_i$ 是希望愈小愈好，所以原本要求得 $\frac{\|w\|^2}{2}$ 的最小值，變為求 $\frac{\|w\|^2}{2}+C\sum_{i=1}^{l}\xi_i$ 的最小值，且要符合式(4-14)的限制式。利用上一節的概念，最佳化的問題轉換為最大化式(4-15)

$$L(w,b,\xi,\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j+\frac{1}{C}\delta_{ij}) \tag{4-15}$$

$$\text{where}\quad \delta_{ij}=\begin{cases}1 & if \quad i=j\\ 0 & others\end{cases}$$

限制式為

$$w=\sum_{i=1}^{l}\alpha_i y_i x_i \tag{4-16}$$

$$\sum_{i=1}^{l}\alpha_i y_i=0 \tag{4-17}$$

剩餘計算 w 和 b 的方法和上一節都一樣，將 α_i 代入式(4-16)求得 w，再利用 Karush-Kuhn-Tucker complementarity conditions 求得 b：

$$\alpha_i\left(y_i\left(w\cdot x_i+b\right)-1+\xi_i\right)=0 \qquad \forall i \tag{4-18}$$

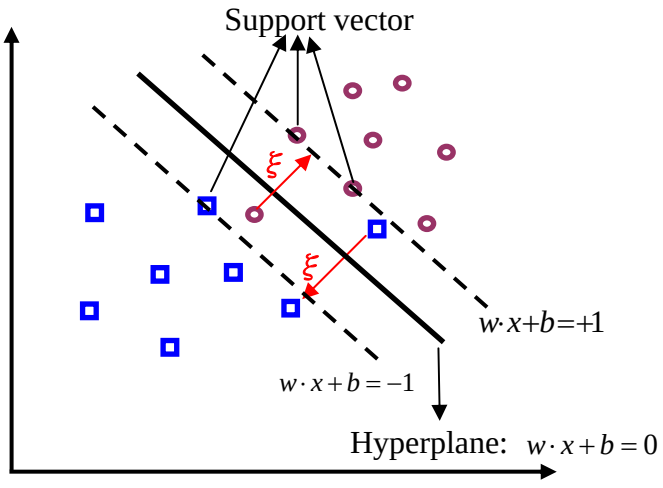


圖4-15 鬆弛變數導入SVM示意圖

4.4 結合影像及語音之雙模情緒辨識決策

經由特徵擷取後，得到語音和人臉影像這兩特徵，分別經由 SVM 分類，可以得到兩類的分類結果。當分類的結果不同時，該採用何者的分類結果，是要解決的問題。根據 SVM 的理論，受測資料在空間中距 hyperplane 的距離較遠，誤判的機會較小，分類結果的可靠度會較高；反之，離 hyperplane 距離較近，誤判的可能性較高，另外，若同類資料在空間中分佈分散，表示資料變異的程度很大，即使訓練資料距 hyperplane 有一段距離，但可能某筆資料很接近 hyperplane，誤判的可能性相對來說也比較大。圖 4-16 分別有人臉影像及語音兩類訓練資料訓練出來的 hyperplane，這兩類訓練資料計算出的平均距離 D_{Fave} 、D_{Aave} 都相同，但資料在空間中分佈的情形不同，人臉影像的資料分佈較集中，而語音資料分佈鬆散，有些資料甚至已相當接近 hyperplane，相較之下，即使兩類的平均距離相同，但還是要考慮到資料的分佈狀況。所以基於以上的概念來設計本論文中的辨識決策：

1. 分別統計出語音與人臉影像訓練資料距 hyperplane 的平均距離 D_{ave} 和標準差 σ，每筆資料計算出距 hyperplane 的距離。

	人臉影像特徵	語音特徵
訓練資料	D_{Fave}，σ_F	D_{Aave}，σ_A
測試資料	D_{Fi} for $i=1\sim N$	D_{Ai} for $i=1\sim N$

2. 計算並正規化人臉影像權重 Z_{Fi} 與語音資料權重 Z_{Ai}

$$Z_{Fi}=\frac{D_{Fi}-\sigma_F}{D_{Fave}-\sigma_F} \qquad for \quad i=1\sim N$$

$$Z_{Ai}=\frac{D_{Ai}-\sigma_A}{D_{Aave}-\sigma_A} \qquad for \quad i=1\sim N$$

3. 若是兩類資料分類結果不同，即比較兩類的權重大小，即可決定出較可靠的分類結果，得到受測者表現的情緒狀態：

   $Z_{Fi}\geq Z_{Ai}$ 則採用人臉影像特徵辨識結果

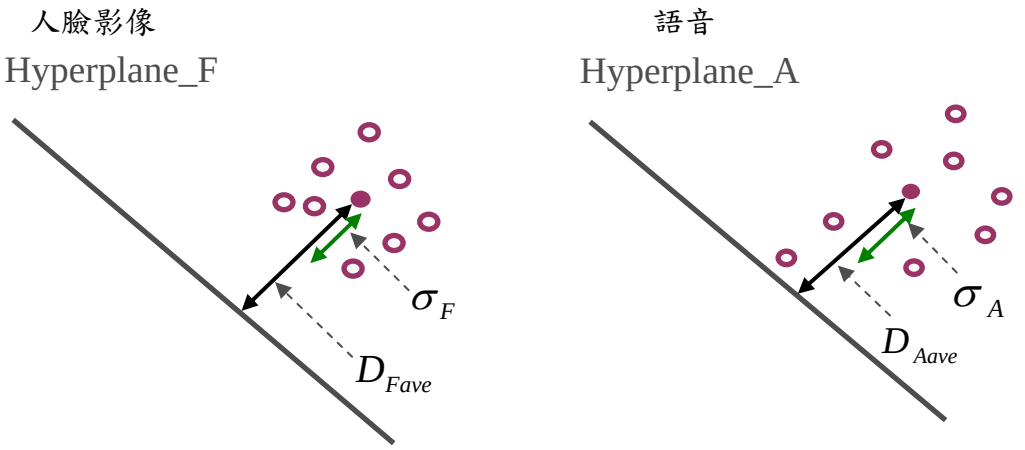


圖 4-16 資料在空間中的不同分佈狀況說明

$Z_{Fi} < Z_{Ai}$ 則採用語音特徵辨識結果

我們採用 SVM 當作情緒分類器，整個辨識流程如圖 4-17 所示，比對的順序和排法是根據 SVM 的辨識結果而定，將辨識率較高的一組表情優先比對，於 5-3 節的實驗結果作介紹。首先，輸入的未知情緒經由分類決定是 Happiness 抑或是 Sadness，同時也分類決定是 Surprise 或是 Neutral，假設辨識結果分別是 Sadness 和 Neutral，接著 Neutral 再和 Angry 作判斷，分類的結果最後再和 Sadness 作分類，決定這個未知情緒是屬於哪一類。由於分別有人臉影像和語音兩類資訊分別作分類，用提出的辨識策略結合兩類資訊辨識，若是在人臉影像在辨識某兩類情緒時分類錯誤，還可以透過語音資料修正錯誤的分類，繼續完成後續的分類，

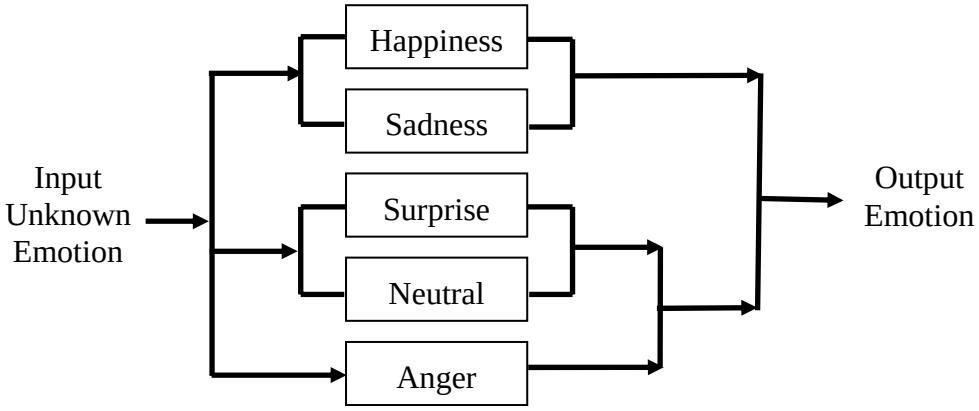


圖 4-17 SVM 辨識流程

而修正後的人臉影像分類結果，在後續的分類也可能修正語音資料的錯誤分類，提高辨識的正確率。

圖 4-18 為結合語音及人臉影像情緒辨識決策流程，將未知情緒的人臉影像特徵與語音特徵分別用各自訓練資料計算出的 hyperplane 作辨識，首先經由 Happiness 與 Sadness 以及 Surprise 與 Neutral 的 SVM 分類器，分類的結果都同樣是 Happiness 及 Surprise。接著 Surprise 與 Anger 進行 SVM 分類，結果人臉影像特徵分類結果為 Surprise，但語音特徵分類為 Anger，如圖 4-18(a)，此時就要比較兩類的特徵權重，人臉影像權重計算結果為 1.56 大於語音權重-0.289，表示這筆資料在人臉影像分類 Surprise 與 Anger 的空間裡距 hyperplane 的距離較遠，分類的可靠度較高，因此採用人臉影像的結果，Surprise 與 Anger 的分類為 Surprise，而語音特徵的 Surprise 與 Anger 分類結果要由 Anger 改為 Surprise。接著對辨識出的 Happiness 及 Surprise 進行 SVM 分類，結果分類的結果又不一樣，人臉影像特徵分類為 Surprise，但語音特徵分類為 Happiness，此時再比較兩類的特徵權重，人臉影像權重計算結果為-0.6685 小於語音權重 1.8215，因此採用語音特徵分類的結果 Happiness，所示這個輸入的未知情緒分類為 Happiness，如圖 4-18(c)所示。

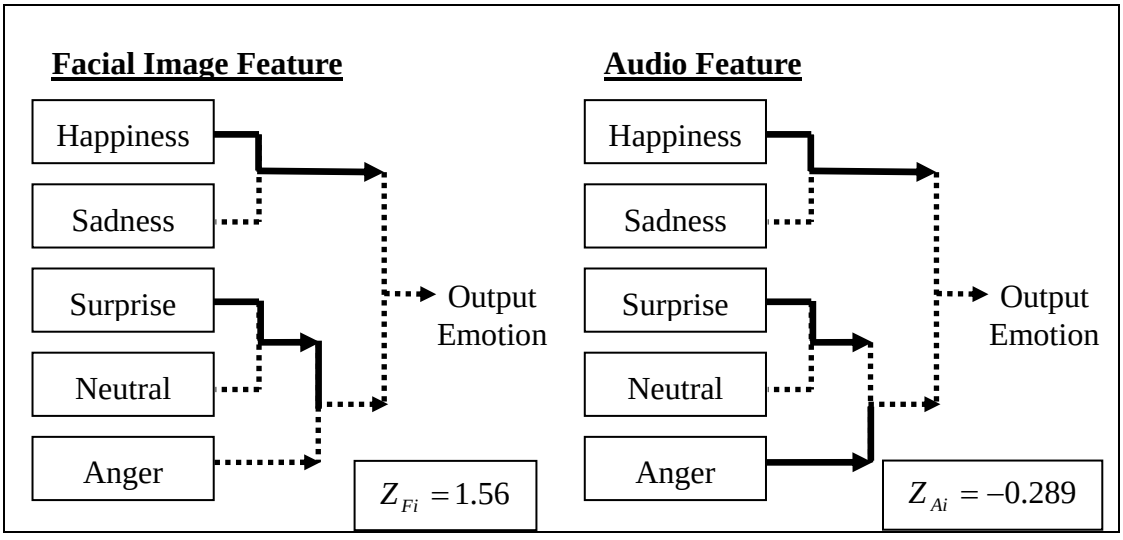


圖 4-18 Bimodal 情緒辨識流程(a)

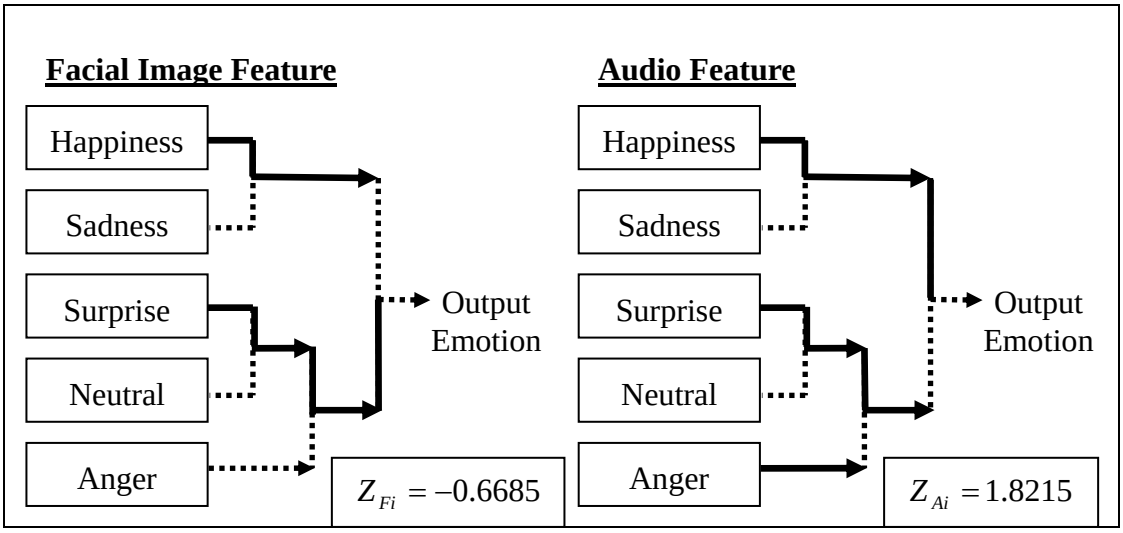


圖 4-18 Bimodal 情緒辨識流程(b)

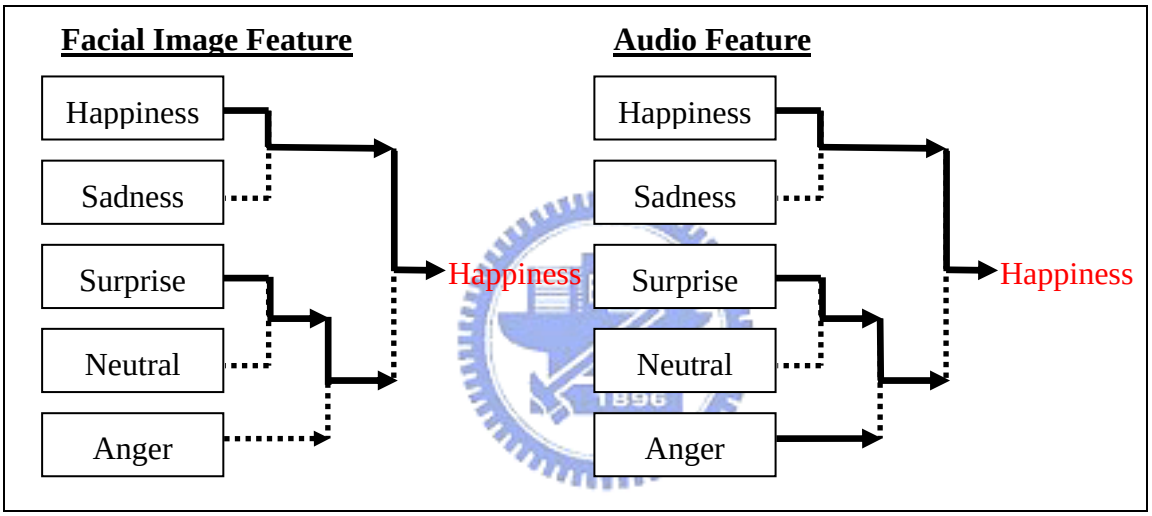


圖 4-18 Bimodal 情緒辨識流程(c)

第五章 實驗結果

為了驗證本論文提出的關鍵性特徵人臉表情辨識，及基於 SVM 理論結合影像語音之雙模情緒辨識決策設計的效能，我們將藉由以下幾個步驟來驗證：

1. 建立測試用的影像及語音資料庫。
2. 對人臉影像全部特徵以及影像關鍵性特徵，分別用 SVM 進行表情辨識，比較這兩種辨識結果。
3. 分別對影像、語音用柔性邊界 SVM 作辨識，再將兩個資訊結合使用本論文提出的辨識決策，修正錯誤的分類，比較最終的辨識結果。

藉由嵌入式影像平台的取像和收音，是希望利用影像平台即時取像和運算的特性，可以快速地擷取出影像及語音特徵，實行後端辨識的機制。另外，在機器人系統整合方面，包含了機器視覺系統(人臉偵測、追蹤及辨識)、聽覺系統(辨識及定位)、機器人移動系統等等，若每一部分都在 PC 伺服端上處理的話，將會佔據很大的運算量，不僅會使機器人的反應緩慢並且產生驗證困難等問題，若能將每個系統模組化的話，則在系統整合及驗證方面相對而言會較容易實現與解決，因此我們在嵌入式影像平台上實現我們的系統。

5.1 嵌入式影像平台

本論文所採用的影像擷取系統乃為本實驗室所開發之 DSP 嵌入式語音影像平台[32]，圖 5-1 為修改後的系統架構圖，主要包含了一個麥克風、CMOS 影像感測板、FPGA Board、DSK6416T 語音影像發展板和一顆緩衝記憶體(Frame Buffer)。CMOS sensor 部分為銳相(IC-media)公司所生產之型號 ICM205B CMOS 感測器，其最大有效像素為 640 x 480，最大更新率為 30 frames/sec，影像輸出格式包含了 8-bit raw data、8/16-bit YCrCb(本系統所選用的即為此)、16-bit RGB 和 24-bit RGB。此外，此 CMOS 感測板尚包含一些影像調整功能，包括了自動

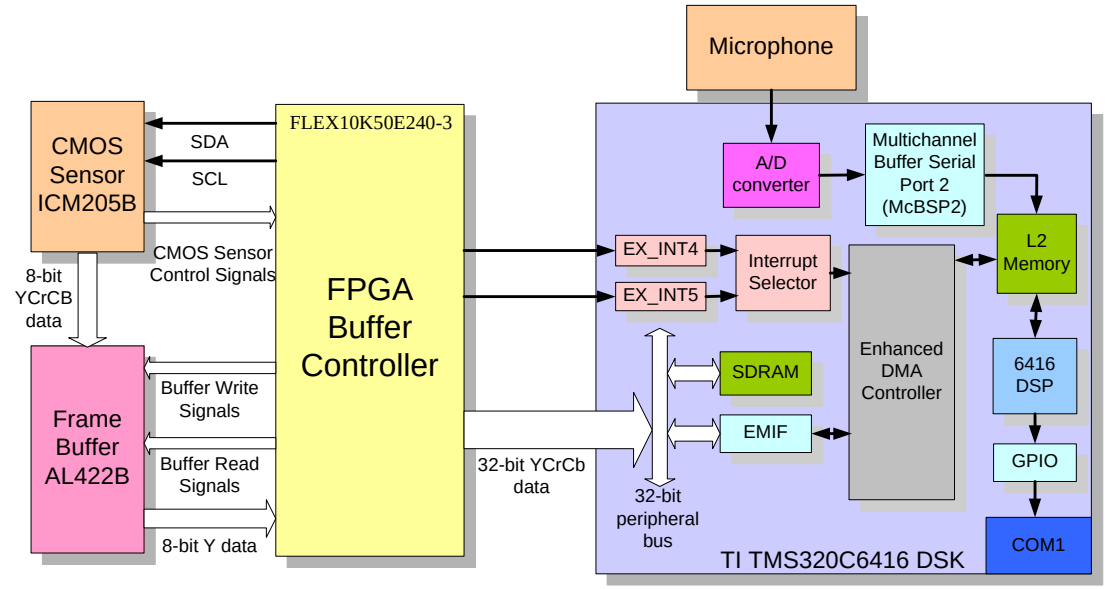


圖 5-1　DSP 影像平台系統架構圖

白平衡(Auto White Balancing)及 Gamma 參數調整可對整張影像的亮度及色彩飽和度進行調整等功能。在 FPGA 板方面所使用的為 ALTERA 公司生產，型號為 FLEX10K50E240-3，主要功能為處理緩衝記憶體之讀出與寫入的控制訊號、資料排序及產生讓 DSK6416T 讀取影像資料之觸發訊號，另外要透過 I2C 來做 CMOS 感測器的初始化過程，其初始化內容如下：

- 影像更新速率：每秒 30 張 Frames
- 影像輸出格式：8-bit 4:2:2 YCrCb QVGA (320 x 240 像素大小)
- 輻射校正：Gamma = 2.2
- 色彩飽和度：Saturation=1.5

在緩衝記憶體(AL422B)的部分主要做為影像資料儲存之用，其記憶體長度為 384Kbytes 可以完整儲存兩張 QVGA 大小之影像，且資料存取動作為 FIFO(First In First Out)，可以以較低的速度接收資料，而以非常快的速度放出(此緩衝記憶體之最快工作時脈為 50MHz)。在語音影像處理板方面選用的為德州儀器(TI)公司所出產之 TMS320C6416T DSK 發展板，主要做為影像資料搬運及即時運算的功能，其工作頻率為 1GHz。

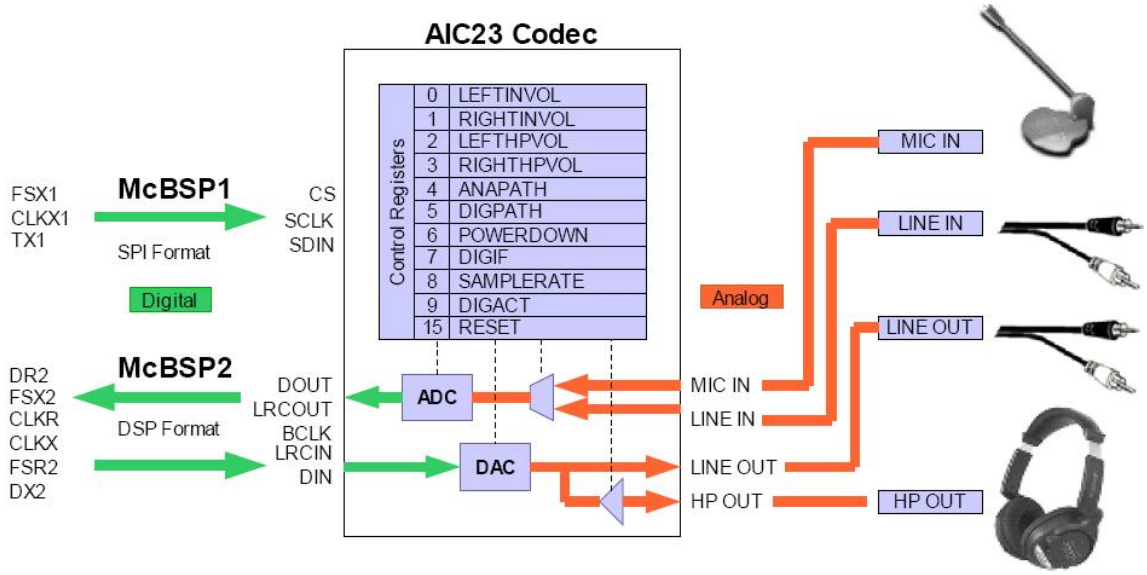


圖 5-2　DSK6416 Codec 介面[39]

DSK6416 提供了聲音輸入及輸出的功能，如圖 5-2 所示，主要是靠著 AIC23 這個立體聲 Codec，可以對類比的聲音信號取樣，輸入轉換為數位信號供後續的處理，當 DSP 處理完這個語音信號，可以再藉由 Codec 將它轉為類比信號輸出。它是利用 2 個多通道緩衝串列埠(Multichannel Buffered Serial Port；McBSP)控制 Codec 的內部設定以及接收發送聲音信號，McBSP1 是單向的控制通道，用來初始化和設定 A/D 輸入和 D/A 輸出的聲音信號格式，McBSP2 是雙向的資料通道，所有的聲音信號都是經由這個通道輸出入，所以讀取從麥克風的語音信號，即先初始化 McBSP1，再去讀取從 McBSP2 讀進來由類比轉成數位的語音信號。

5.2 建立資料庫

為了要驗證本論文提出的結合人臉影像與語音情緒辨識及其決策，需要建立一套合適之資料庫以供實驗使用。由於人臉影像和語音特徵都是由嵌入式語音影像平台所擷取，先是收取情緒語音，再擷取人臉影像，所以資料庫的建立基本上就根據這樣的原則進行，透過連接到 DSP 的麥克風和 CMOS sensor 分別取得語音和人臉影像信號，先由 DSP 上的麥克風輸入判斷是否有語音訊號，一旦有語音訊號則先計算語音特徵值，接著取一段時間的表情影像，然後再分別計算此兩

種特徵值，得到一組語音特徵值和多組影像特徵值。

圖 5-3 為資料庫建立情形，請協助建立資料庫者先發出情緒語音，接著在 CMOS sensor 前方約 40~50 公分處作出相對應的表情，擷取出具有相關性的語音和人臉影像特徵，因此我們可以用這兩種特徵資料作訓練與測試我們使用的演算法和提出的辨識決策。初步測試用的資料庫規格為 14 個人，建立的情緒包括生氣、高興、普通、傷心和驚訝五種情緒，每種情緒每人各作 10 次，因此可以用的語音和人臉影像每種表情各有 140 筆。圖 5-4 為資料庫中其中 5 人影像範例。

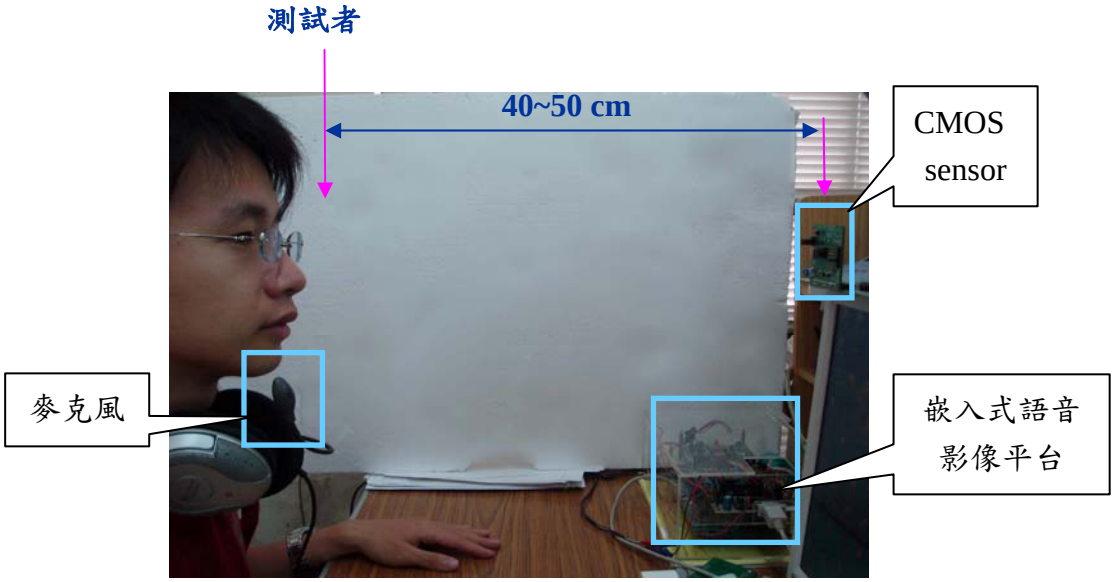


圖 5-3 資料庫建立情形

圖 5-4 資料庫影像範例

圖 5-4 資料庫影像範例(續)

5.3 SVM 搭配關鍵性特徵辨識結果

表 5-1 為不同表情比對下的全部特徵與關鍵性特徵比對的差異，由 5 位受測者作出 5 種表情各 10 次，每個表情各有 50 組的資料，24 組資料當作用來訓練 hyperplane 的資料，剩下的 26 組為測試資料，實驗的個人電腦為 Pentium IV、2.8GHz、1GB RAM 的配備，在訓練時是根據兩種表情之間訓練出一個 hyperplane，有五種情緒則有 10 種比對方式，每個 hyperplane 的訓練時間約為 0.047 秒，所以訓練時間約為 0.47 秒。可由圖 5-5 中看出全部特徵的辨識效果都不比關鍵性特徵的辨識率高，可看出選取少數關鍵性的特徵對於分類的準確率有一定的幫助。可能就是由於用 SVM 辨識，兩類的 12 個特徵資料在空間的分佈有大部分的交疊，所以辨識效果有限，改為少數關鍵性變化明顯的特徵作分類，除去一些對於表情兩兩比對上變化不大的特徵，使得特徵之間的交疊部分較小，使得分類效果較好，平均辨識率多了將近 14%。圖 5-6 是全部表情的辨識結果，同樣地用關鍵性特徵的辨識率比較高，5 種表情的平均辨識率由原本的 46%達到 77%，提升了 31%左右的效果。至於比對的順序是由表 5-1 的辨識結果中所挑選，辨識率高的表情優先比對，可看出 Surprise 和其它表情的比對辨識率都比較高，而 Neutral 和 Happiness 的比對辨識率很低，所以要讓這兩種表情的比對順序儘

表 5-1 全部特徵與關鍵性特徵辨識率結果

特徵選擇 / 比對的表情	全部特徵	關鍵性特徵
Surprise vs. Sadness	78.85%	92.31%
Surprise vs. Neutral	73.08%	92.31%
Surprise vs. Anger	84.61%	90.38%
Surprise vs. Happiness	63.46%	75.00%
Sadness vs. Neutral	57.69%	75.00%
Sadness vs. Anger	76.92%	80.76%
Sadness vs. Happiness	59.62%	78.85%
Neutral vs. Anger	63.46%	78.85%
Neutral vs. Happiness	63.46%	67.30%
Anger vs. Happiness	57.69%	84.61%
平均辨識率	67.89%	81.54%

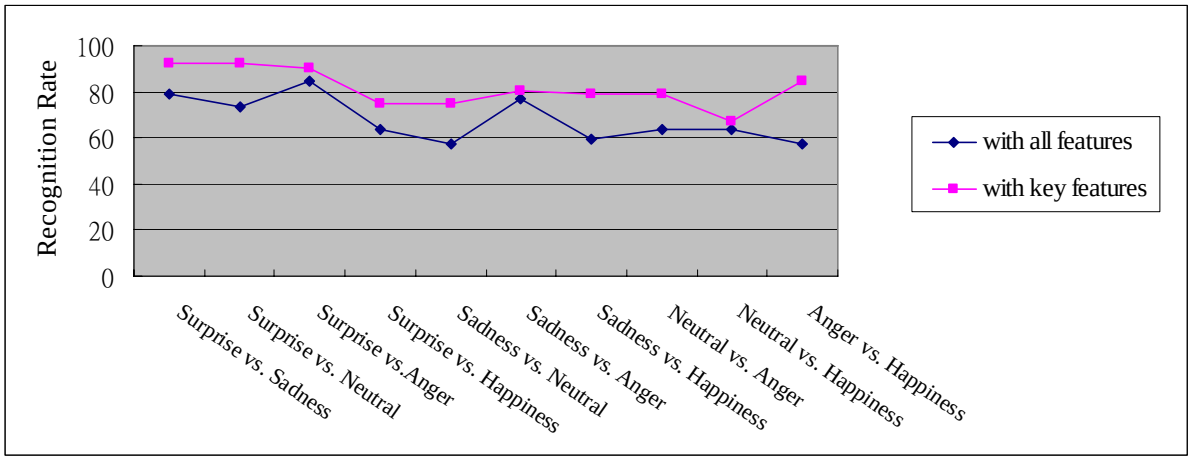


圖 5-5 全部特徵與關鍵性特徵兩兩分類的辨識結果比較

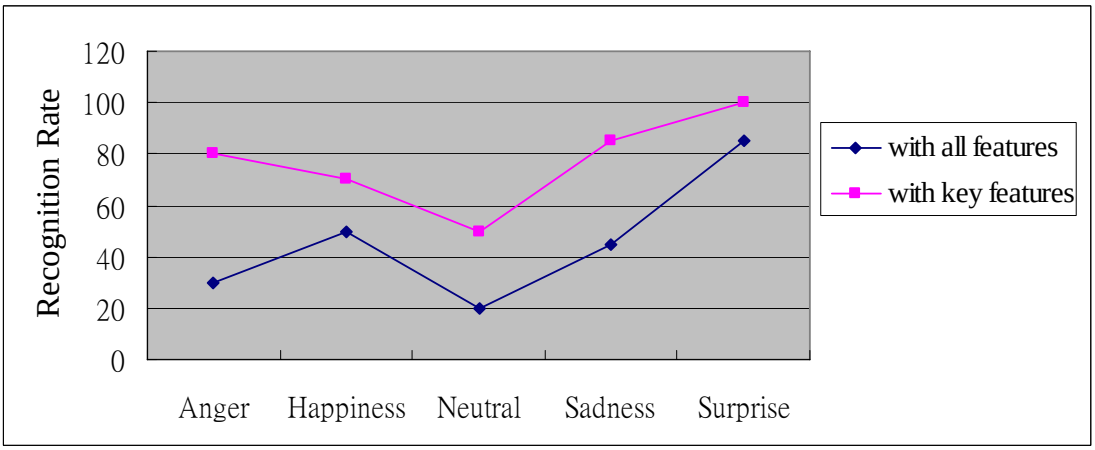


圖 5-6 全部特徵與關鍵性特徵的辨識結果比較

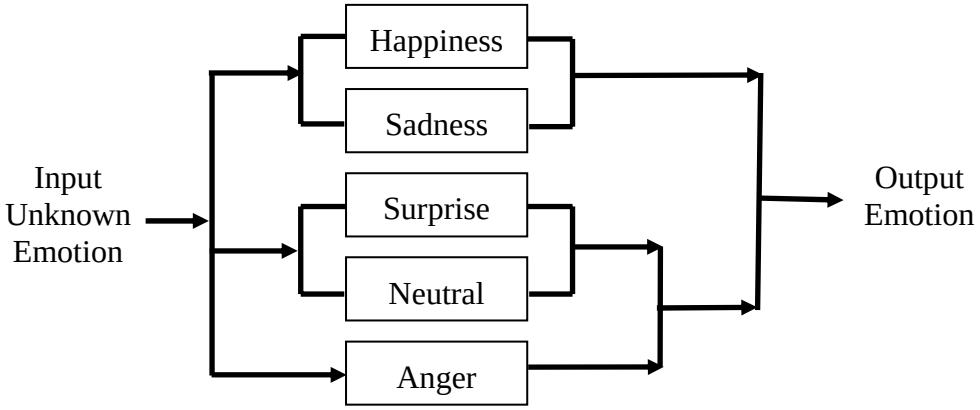


圖 5-7 SVM 辨識流程

量延後，所以先讓 Surprise 和 Neutral 比對，再和 Anger 比對。另一方面，讓辨識率不差的 Sadness 和 Happiness 也作比對，最後再將兩邊的辨識結果作比對，如圖 5-7 所示。

5.4 柔性邊界 SVM 用在結合語音及人臉表情的情緒辨識結果

由於資料是線性不可分割，所以採用柔性邊界 SVM 做為辨識與訓練的方法，表 5-2 與表 5-3 分別用語音特徵與用人臉影像全部特徵辨識的結果，辨識率比原來用 SVM 辨識要來得高。在訓練時是根據兩種情緒之間訓練出一個 hyperplane，有五種情緒則有 10 種比對方式，因此兩種模式各要訓練出 10 個分辨不同情緒的 hyperplane，這裡用的訓練由 5 位受測者作出 5 種表情各 10 次，每個表情各有 140 組的資料，70 組資料當作用來訓練 hyperplane 的資料，剩下的 70 組為測試資料，每個 hyperplane 的訓練時間約為 0.17 秒，所以全部 20 個 hyperpalne 的訓練時間約為 3.4 秒。圖 5-8 為語音、人臉影像與結合兩種特徵的辨識結果，語音 5 種情緒的平均辨識率為 73.71%，人臉影像為 81.71%，而結合以上兩種特徵用柔性邊界 SVM 辨識和本論文提出的辨識策略，可使得平均辨識率達到 86.85%，修正了部分錯誤的辨識結果，使得 5 種表情的平均辨識率提升了 5.1%左右，證實本論文提出的方法確實能提升辨識效果。

表 5-2 柔性邊界 SVM 用在語音特徵分類的結果

輸出 / 輸入	Anger	Happiness	Neutral	Sadness	Surprise	辨識率
Anger	**48**	12	5	3	2	68.57%
Happiness	8	**43**	6	10	3	61.43%
Neutral	5	9	**48**	5	3	68.57%
Sadness	2	6	3	**59**	0	82.85%
Surprise	0	1	7	1	**61**	87.14%

表 5-3 柔性邊界 SVM 用在人臉影像特徵分類的結果

輸出 / 輸入	Anger	Happiness	Neutral	Sadness	Surprise	辨識率
Anger	**53**	3	7	6	1	75.71%
Happiness	2	**57**	11	0	0	81.42%
Neutral	9	7	**48**	6	0	68.57%
Sadness	1	1	6	**62**	0	88.57%
Surprise	1	1	2	0	**66**	94.28%

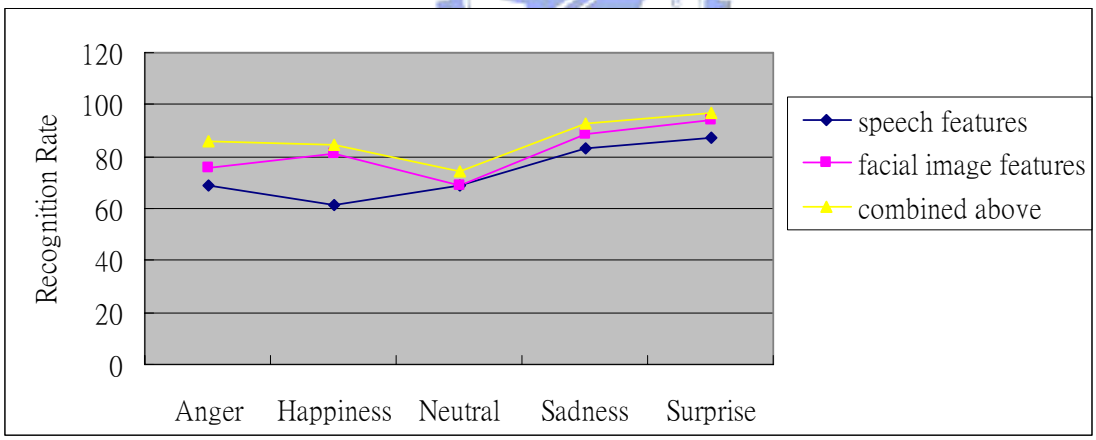


圖 5-8 柔性邊界 SVM 情緒辨識率

我們提出的雙模辨識策略主要是探討經由語音和影像資訊分別辨識出情緒後，最後該採信何者的辨識結果，因此與類似的方法架構做比較，其中 De Silva 採用 Rule-based 方法[25]，不同情緒在語音和影像上具有支配(Dominant)的地位，像是屬於負面情緒的傷心和生氣只採用語音的辨識結果，而快樂和驚訝只採用影

像的辨識結果，結合兩種資訊的辨識策略較簡單，但在實測中並不能預測受測者會表達何種情緒而決定採用何種模式。我們採用的辨識策略是根據資訊的可靠度決定要採用何種資訊的辨識結果，是依據資訊分類情況作調整，方法較具彈性，由實驗結果可知，每種情緒的辨識率都比只用單一模式要來的高，我們的雙模辨識策略在效能表現上較 De Silva 所提出的 Rule-based 方法為佳，表 5-4 為兩種方法的辨識率比較表。

表 5-4 兩種方法的辨識率比較

Emotion	**De Silva's method**	**Proposed method**
Anger	68.57%	85.71%
Happiness	81.42%	84.28%
Neutral	68.57%	74.29%
Sadness	82.85%	92.58%
Surprise	94.28%	97.14%
Average	79.14%	86.80%

5.5 即時情緒辨識結果

表 5-4 為 5 位受測者，如圖 5-9 所示，在 CMOS sensor 前作出 5 種表情，嵌入式影像平台立刻辨識出表情的測試結果，受測者每種表情交替變換各作 10 次，結果如表 5-5 所示，平均辨識率為 74.4%；表 5-6 為結合語音與臉部影像雙模情緒辨識結果，平均辨識率為 77.6%。由於 Surprise 和 Anger 和它表情的差異程度較大，可由表中看出這兩種表情辨識率較高，而 Sadness 有多次都辨識為 Neutral，Neutral 和 Happiness 也有部分是誤判成彼此的錯誤，可能是因為嘴巴特徵點的抓取錯誤造成表情誤判，嘴巴特徵點只用 IOD 作決定並不穩固，不適用於每個人。

由於表情辨識是一種漸進的過程，在實作中，和靜態影像相比，由於不知受測者的表情是何時開始變化，採取的作法是 CMOS sensor 連續取 50 張影像，且即時辨識出 50 種辨識結果，時間約為 2 秒左右，再統計出辨識結果最多的一

種表情，即為這段時間內的表情，如圖 5-10 所示。

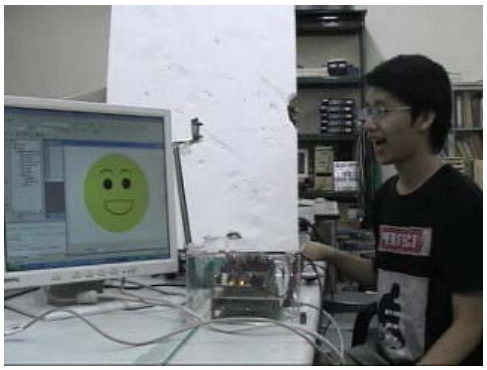
圖 5-9 五位受測者

表 5-5 即時表情辨識的結果

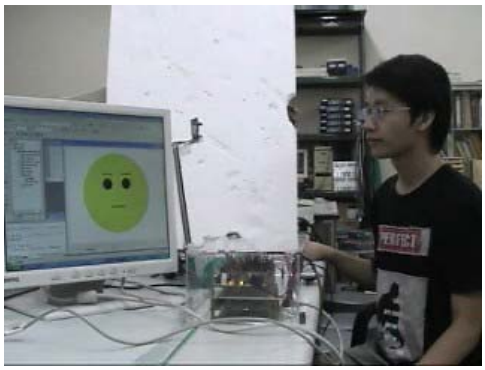
輸出 / 輸入	Anger	Happiness	Neutral	Sadness	Surprise	辨識率
Anger	**40**	1	8	0	1	80%
Happiness	3	**38**	9	0	0	76%
Neutral	1	10	**34**	4	1	68%
Sadness	1	4	14	**30**	1	60%
Surprise	0	2	3	1	**44**	88%

表 5-6 即時雙模情緒辨識的結果

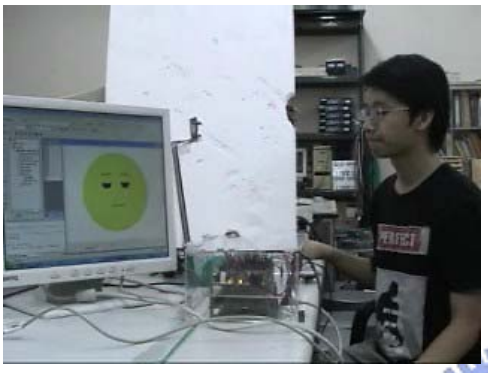
輸出 / 輸入	Anger	Happiness	Neutral	Sadness	Surprise	辨識率
Anger	**41**	1	7	0	1	82%
Happiness	1	**38**	9	0	2	76%
Neutral	2	10	**35**	2	1	70%
Sadness	1	1	14	**34**	0	68%
Surprise	0	2	2	0	**46**	92%



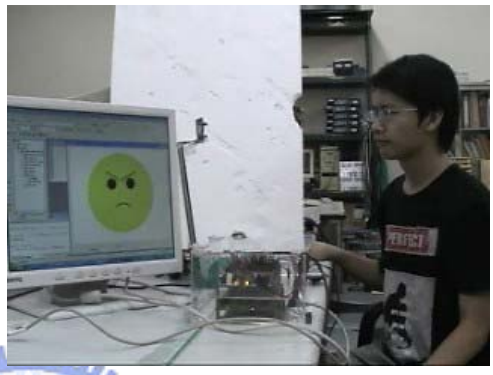
(a) Happiness



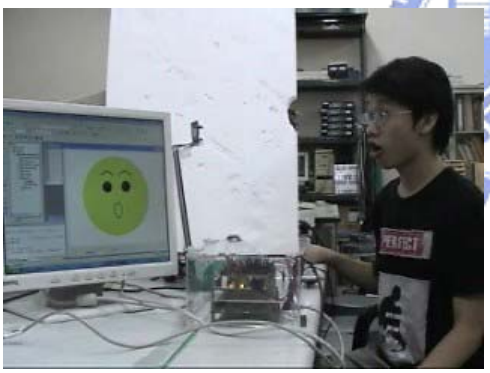
(b) Neutral



(c) Sadness



(d) Anger



(e) Surprise

圖 5-10 五種情緒即時測試情形

第六章 結論與未來展望

6.1 結論

本論文提出了結合人臉影像與語音的雙模式情緒辨識系統架構，利用實驗室發展的嵌入式語音影像平台擷取人臉影像和語音特徵，辨識出生氣、高興、普通、傷心與驚訝這五種情緒狀態。

在人臉影像特徵擷取方面，經由膚色和人臉正面特徵完成人臉偵測，得知人臉在影像當中的位置和大小，接著根據瞳孔在人臉中的幾何位置和灰階值強度低的特性，用影像灰階值強度找出其位置，再根據雙眼的瞳距定出眼睛和眉毛的可能位置，再利用影像灰階值強度和邊緣偵測找出眼睛和眉毛確切的範圍，定出特徵點。嘴巴部位特徵點同樣根據在人臉中的幾何位置和灰階值強度特性，找出特徵點位置。語音特徵擷取方面，計算出語音信號內多個音框的音高和能量值，再統計出情緒相關的特徵。

用 SVM 辨識這五種情緒狀態，由於 SVM 是二類的分類器，因此在比對的順序上，我們是針對表情差異性較大的兩種情緒先作辨識，辨識出的情緒再和其他情緒加以辨識，如此可以減少類似情緒在最初即誤判的可能性。在結合人臉影像與語音的情緒辨識上，根據 SVM 的理論，受測資料在空間中距 hyperplane 的距離較遠，分類結果的可靠度較高，並把訓練資料的分佈狀況一併考量，定出人臉影像和語音的權重參數，若是兩類資料分類不同時，即根據權重參數決定該採用何種分類結果，可以修正錯誤的資料分類，使得辨識率提升，經由初步的實驗結果，結合人臉影像與語音的情緒辨識到達 86.85%，比只使用人臉影像的辨識率高 5%，提高了辨識的準確性。

6.2 未來展望

目前本系統仍有許多可改善的空間：

1. 在特徵擷取方面

找尋眼睛和眉毛的位置是尋找上半臉影像的低灰階值，由於頭髮的灰階值大小和瞳孔幾乎一樣，若是受測者額頭散佈著過多頭髮，會造成瞳孔位置的誤判，影響以瞳孔之間距離為基準的後續判斷，使得眼睛與眉毛可能區域會找錯，找出的特徵點位置因此不是預期的位置，計算出的特徵值也非原先所定義；偵測嘴巴的特徵點只用灰階值強弱作判斷也不夠穩固，不適用於每個人。若要使特徵擷取更為強健的話，可以加入其他資訊，如用模型比對人臉器官的定位或是各器官之間位置的幾何比例。另外，可以增加其他特徵值，特徵值個數的增加或許可以提高辨識率，例如加入皺紋特徵。另外，在語音特徵的選擇上，可以增加更多的統計特徵，再由其中找尋具有鑑別性的情緒參數當作情緒辨識的特徵。

2. 在情緒辨識方面

辨識的情緒只能判斷出是屬於何種情緒，但卻難以判斷情緒表達的強度或是組合的情緒，可以在辨識部分加上模糊邏輯，增加一些可辨識的組合情緒，如驚訝加上高興的驚喜，或是判斷出情緒表達的強度。或是將 SVM 輸出改為機率輸出，計算受測情緒分別為 5 種情緒的機率，根據機率值的大小，也可辨識出組合的情緒與情緒表達的強度，應用在機器人技術上，使得機器人的更為智慧與人性化。

參考文獻

[1] http://cdnet.stpi.org.tw/techroom/policy/policy_05_013.htm

[2] M. Fujita, "On Activating Human Communications with Pet-type Robot AIBO," *Proc. of the IEEE*, vol. 92, no. 11, pp. 1804-1813, 2004.

[3] M. Fujita, Y. Kuroki, T. Ishida and T.T. Doi, "A Small Humanoid Robot SDR-4X for Entertainment Applications," *International Conference on Advanced Intelligent Mechatronics*, Kobe, Japan, 2003, pp. 938-943.

[4] http://www.robotdiy.com/article.php?sid=141

[5] http://www.ars-journal.com/ars/Free_Articles/IREX-2005.htm

[6] P.Ekman and W.V. Friesen, *The Facial Action Coding System: A Technique for The Measurement of Facial Movement*, San Francisco, Consulting Psychologists Press,1978.

[7] J. Song, Z. Chi, J. Liu and H. Fu, "Extraction of Face Image Edges with Application to Expression Analysis," *Proc. of 2004 8th International Conference on Control, Automation, Robotics and Vision*, Kunming, China, 2004, pp. 804-809.

[8] J. F. Cohn, A. J. Zlochower, J. J. Lien and T. Kanade, "Feature-Point Tracking by Optical Flow Discriminates Subtle Differences in Facial Expression," *Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998, pp. 396-401.

[9] H. Seyedarabi, A. Aghagolzadeh and S. Khanmohammadi, "Recognition of Six Basic Facial Expressions by Feature-Points Tracking using RBF Neural Network and Fuzzy Inference System," *Proc. of the 2004 IEEE International Conference on Multimedia and Expo*, Taipei, Taiwan, 2004, pp. 1219-1222.

[10] Ying-li Tian, T. Kanade and J.F. Cohn, "Recognizing Action Units for Facial Expression Analysis," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97-115, 2001.

[11] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman and T.J. Sejnowski, "Classifying Facial Actions," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974-988, 1999.

[12] T. Wilhelm, H.J. Bohme, H.M. Grofi and A. Backhaus, "Statistical and Neural

Methods for Vision-based Analysis of Facial Expressions and Gender," *IEEE International Conference on Systems, Man and Cybernetics*, Hague, Netherlands, 2004, vol. 3, pp. 2203-2208.

[13] I. Buciu, C. Kotropoulos and I. Pitas, "ICA and Gabor Representation for Facial Expression Recognition," *Proc. of 2003 International Conference on Image Processing*, Barcelona, Spain, 2003, pp. 855-858.

[14] Z. Zhang, M. Lyons, M. Schuster and S. Akamatsu, "Comparison Between Geometry-Based and Gabor-Wavelets-Based. Facial Expression Recognition Using Multi-Layer Perceptron," *Proc. of IEEE International Conference on Automatic Faceand Gesture Recognition*, Nara, Japan, 1998, pp 454-459.

[15] Z. Zhang, M. Lyons, M. Schuster and S. Akamatsu, "Comparison Between Geometry-Based and Gabor-Wavelets-Based. Facial Expression Recognition Using Multi-Layer Perceptron," *Proceedings of IEEE International Conference on Automatic Faceand Gesture Recognition*, Nara, Japan, 1998, pp 454-459.

[16] I.O. Stathopoulou and G.A. Tsihrintzis, "An Improved Neural-network-based Face Detection and Facial Expression Classification System," *Proc. of IEEE International Conference on System, Man and Cybernetics*, Hague, Netherlands, 2004, pp. 666-671.

[17] D.N. Jiang and L.H. Cai, "Speech Emotion Classification with the Combination of Statistic Features and Temporal Features", *IEEE International Conference on Multimedia and Expo* , Taipei, Taiwan, 2004, pp. 1967-1970.

[18] B. Schuller, G. Rigoll and M. Lang, "Hidden Markov Model-based Speech Emotion Recognition", *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, China, 2003, vol. 2, pp 1-4.

[19] D. Ververidis, C. Kotropoulos and I. Pitas ,"Automatic Emotional Speech Classification," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Quebec, Canada, 2004, vol. 1, pp 593-596.

[20] B. Schuller, G. Rigoll and M. Lang, "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Quebec, Canada, 2004, vol. 1, pp. 577-580.

[21] X.H. Le, G. Quénot and E. Castelli, "Recognizing Emotions for Audio-Visual

Document Indexing," *Proceedings of 9th Symposium on Computers and Communications*, Alexandria, Egypt, 2004, pp. 580-584.

[22]T.L. Pao, Y.T. Chen and J.H. Yeh, “Emotion Recognition from Madarin Speech Signals,” Proceedings of IEEE International Symposium on Chinese Spoken Language Processing, Hong Kong, pp. 301-304, 2004.

[23]L.C. De Silva, T. Miyasato and R. Nakatsu, ”Facial Emotion Recognition Using Multi-modal Information,” *Proceeding of IEEE International Conference on Information, Communications and Signal Processing*, Singapore, 1997, pp. 397-401.

[24]L.S Chen, T.S. Huang, T. Miyasato and R. Nakatsu, "Multimodal Human Emotion /Expression Recognition," *Proceeding of International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998, pp. 366-371.

[25]L.C. De Silva, “Audiovisual Emotion Recognition,” *IEEE International Conference on Systems, Man and Cybernetics*, Hague, Netherlands, 2004, pp. 649-654.

[26]H.J. Go, K.C. Kwak, D.J. Lee and M.G. Chun, "Emotion recognition from the facial image and speech signal," *SICE Annual Conference*, vol. 3, pp. 2890-2895, 2003.

[27]M. Song, J. Bu, C. Chen and N. Li, “Audio-Visual Based Emotion Recognition - A New Approach,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1020-1025, 2004.

[28]Y. Wang and L. Guan, “Recognizing Human Emotion from Audiovisual Information,” *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1125-1128, 2005.

[29]吳鑑峰，*應用語音及臉表情之雙模態情緒辨識*，碩士論文，國立成功大學資訊工程學系，2002.

[30]顏坤銘，*家用機器人之語音辨識系統*，碩士論文，國立交通大學電機與控制工程學系，2001.

[31] Jyh-Shing Roger Jang, “Audio Signal Processing and Recognition,” *(in Chinese) available at the links for on-line courses at the author's homepage at* *http://www.cs.nthu.edu.tw/~jang*.

[32]周崇民，*光源變化下之即時人臉追蹤*，碩士論文，國立交通大學電機與控制

工程學系, 2005.

[33] P. Viola and M. Jones, “Rapid Object Dectetion Using a Boosted Cascade of Simple Features,” *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[34]吳明衛, *自動化臉部表情分析系統*, 碩士論文, 國立成功大學資訊工程學系, 2003.

[35]J.H. Lai, P.C Yuen, W.S. Chen, S. Lao and M. Kawade, ”Robust Facial Feature Point Detection Under Nonlinear Illuminations,” *Proceedings of IEEE ICCV Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems*, Canada, 2001, pp.168-174.

[36]王景南, *多類支向機之研究*, 碩士論文, 元智大學資訊管理學系, 2003.

[37]N. Christianini and J.S. Taylor, *An Introduction to Support Vector Machines*, Cambridge, 2000.

[38]J. Friedman, T. Hastie and R. Tibshirani, *The Elements of Statistical Learning*, Springer, 2001.

[39]*TMS320C6416 DSK Technical Reference*, April 2003, Texas Instrument. Inc.

[40]TMS320C6000, Reference Guide, “TMSC6000 Peripherals Reference Guide, “ Literature Number: SPRU190D, December 2002, Texas Instrument. Inc.

[41]TMS320C6000, Reference Guide, “TMSC6000 Multichannel Buffered Serial Port (McBSP),” Literature Number: SPRU580D, September 2004, Texas Instrument Inc.