科技部補助

大專學生參與專題研究計畫研究成果報告

＊ ********* ** ＊
＊ 計 畫 ＊
＊ ： 結合人臉表情及語音之情緒辨識系統 ＊
＊ 名 稱 ＊
＊ ********* ** ＊

執行計畫學生： 賴柏村
學生計畫編號： NSC 102-2815-C-040-003-E
研 究 期 間 ： 102 年 07 月 01 日至 103 年 02 月 28 日止，計 8 個月
指 導 教 授 ： 徐麗蘋

處理方式： 本計畫涉及專利或其他智慧財產權,2 年後可公開查詢

執 行 單 位： 中山醫學大學醫學資訊學系

中華民國 103 年 03 月 25 日

(一) 摘要

近年來情感辨識已成為人們生活中特別的角色，人們藉由開發各式各樣的情感辨識方法，賦予電腦具人性化與人們之間的互動，且現在智慧型手機的廣泛和應用，更是影響深遠。所以在本計畫中，我們希望將情感辨識應用在智慧型手機上，讓使用者可以即時地、便捷的了解他人，本系統除了平時人們的臉部表情辨識之外，還加入了人與人間的言語表達溝通，來增加整體的辨識準確性。其中最主要的目的在於，人可能在判斷他人時，只用單一的資訊來辨別，卻沒有注意他人可能透露出的身體語言或其他不容易察覺的部分。因此我們希望情感運算能結合語音辨識及人臉表情辨識之雙核心，以及智慧型手機的便捷性，來加強對於人們情緒辨識的正確性及成功率，並討論情感運算在人機互動的過程當中所扮演的角色以及貢獻。運用情感運算除了了解人的情緒之外，還能加深其人性化互動的層次。在這裡，本計畫是以人的五官面相為基礎配合語音辨識來了解使用者的基本情緒，於臉部表情辨識中，我們會利用影像找出人臉，接著再區分出五官以 Action Units 進行特徵抽取；在語音辨識中，我們為了影像的一致性與方便性，需要將 1D 的語音訊號轉為 2D 的影像訊號，而 1D 轉 2D 的方法在這裡我們使用 Spectrogram 方法把聲音訊號以 2D 影像的形式呈現出來。接著將得到的 2D 影像訊號使用 Law's Mask 的方法來抽取聲音特徵，以便進行辨識語音情緒。最後，我們辨識情緒的方法是經由 iFuzzy LDA 來進行情緒聲音特徵的測試及訓練，以此來獲得辨識情緒結果。

(二) 研究動機與研究問題

人類在情緒的表達上是相當複雜的，像是透過各種表情的傳達、眼神的傳遞、各式各樣的肢體語言、生理現象或利用文字的表達來顯現當下的情感狀態。由於人臉的外形在不同觀察角度、以及透過臉部的變化產生表情，還有受光照條件(例如白天和夜晚，室內和室外等)、遮蓋物（例如口罩、墨鏡和頭髮等)、年齡等多方面因素的影響，使得人臉的視覺圖像也顯得不穩定。在語音方面，也受許多因素的影響，包括不同的說話人、說話方式、環境噪音、傳輸通道等，都會造成系統在不同的應用環境、條件下辨識，讓結果無法真正符合其本身正確性，而誤導我們對系統的功效。

因此，單純的使用單一方法來取得的情感或狀態都是不完整的，所以我們利用臉部表情辨識加上語音情緒辨識之雙模機制，試圖了解人們更多的情緒資訊。並希望藉由情感運算的參與，來加強情辨識結果的有效性、真實性及正確性。也由於近年來的文獻趨勢發展，逐漸從單一模組進步到雙模以上的系統模組來提高情緒辨識度，這也代表系統的開發方向逐漸往多樣化。所以，本計畫提出以雙核心系統為主的情緒辨識方法，來改善只用單一方法來得出真正的結果，而且也運用在智慧型手機上，在無須龐大的系統資源以有效的降低系統成本，而達到系統整合的目的。

(三) 文獻回顧與探討

近多年來，許多的研究工作主要集中在情緒或語音辨識。而在這方面有單模組與多模組辨識兩大類。在單模組多以採取語音特徵或臉部表情特徵來單獨進行情緒辨識[1-3]，而多模組辨識方面則是混合語音及臉部表情特徵的架構來進行[4-5]。在辨識分類的相關方法上，有採用如 Hidden Markov Model (HMM)[6], Gaussian Mixture Model (GMM)[7], Artificial Neural Network (ANN), Linear Discriminant Analysis (LDA), pattern recognition or Support Vector Machine (SVM)[8]，這使我們在進行情緒辨識上，能從許多的聲音特徵，such as prosody, pitch, energy or formant 中有效的分類出不同的情緒特徵，也讓我們更清楚地分辨不同種類的情緒。Björn Schuller et al. [9] 藉由 MPEG-4 的情緒標準，即快樂，憤怒，厭惡，恐懼，悲傷，驚訝和中立。在聲音特徵選擇和分類樣本使用了的 HMM 模組，判別情緒和音頻訊號，並使之能夠有效分類及識別人的情緒和語音。Yoshitomi et al. [10] 結合了 HMM 和神經網絡分別去訓練情緒語音和臉部表情，使最後得到的結果能進行整合分析，以提升辨識率。

人臉自動分析，如人臉檢測、人臉識別或人臉情緒辨識，儼然已成為計算機視覺領域的重要研究內容，對比於其它生物特徵的識別方法，人臉辨識具有方便採集特徵、紋理、非侵入性等優勢，使得在人機互動與人工智慧系統等領域有廣泛的應用。在計算機視覺研究中，能有效尋找人臉外觀所具有的特徵，是人臉分析中一個關鍵的議題，經過有關學者和研究人員的研究進展，各種人臉辨識的演算法也相繼的被提出，主要分為四類：(1)基於局部特徵的人臉辨識方法，如 LBP[11]；(2)基於人臉整體特徵的辨識方法，如 Principal Component Analysis (PCA)[12]；(3)基於局部和人臉整體特徵的辨識方法，如人臉與眼的特徵或鼻子與嘴唇的特徵等方法；(4)利用類神經網路進行識別的演算法。

雖然人臉辨識的研究得到了許多結果，但是在現實的生活環境中，由於我們身處的環境其干擾因子非常的多，如光照、姿態、表情或佩帶物等，使得我們要進行人臉辨識仍然存在要克服的困難。特別是當人在不同的光照之下，其辨識率會因光源照射的區域不平均或人臉皮膚反射光源時，這在辨識系統裡還是需要面對的問題[13-14]。因此，近幾年就有提出了許多關於光照的不同識別方法，如 Binary Face Edge Map (BFEM)[15] 、 wavelet-based illumination invariant algorithm[16]或使用視頻的人臉識別系統[17]等。

因此，基於 Law's Mask 是一種用來描述區域紋理變化的特徵計算方式，其中所提取的人臉特徵對人的姿態、表情或光照影響的變化具有較強的 robustness，且能夠更快的從人臉影像中提取出來，而為數也小很多。所以對於其運算簡單、快速，且不受陰影干擾，以及適合使用在真實的即時系統（real-time system）上，我們選擇了 Law's Mask 作為特徵抽取的方法，並且使用在語音情緒辨識中的語音特徵上。而在人臉情緒辨識方面，我們使用 Action Units(動作單元)來擷取臉部的特徵，以獲得到人臉影像的眼睛、嘴唇或相對位置的特徵$\{AU_1, AU_2, \cdots, AU_N\}$。我們希望利用雙核心的概念來加強系統的正確性和準確性與補足只有單一核心的辨識方法，以及利用 iFuzzy LDA 的分類器進行整個系統的辨識分析，藉此來完成我們的雙核心情緒辨識系統，以達到顯要的辨識率與可信度。

(四) 研究方法及步驟

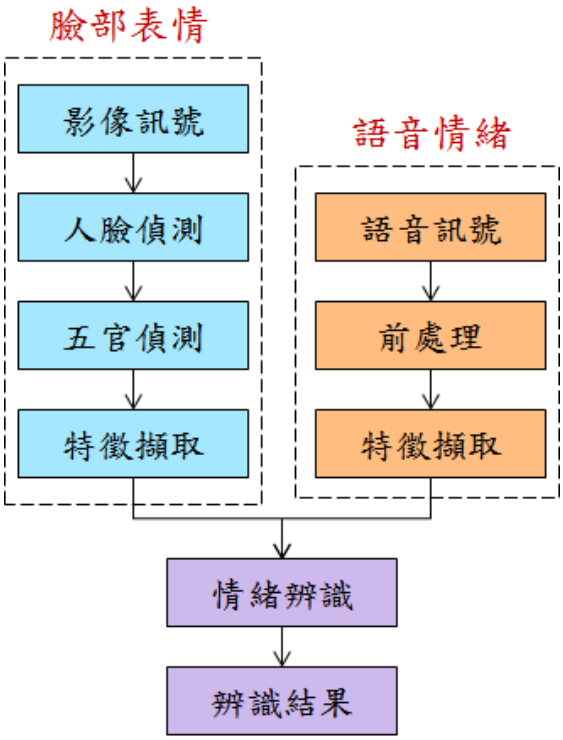


圖 1、系統流程圖

圖 1 為雙模情緒辨識系統的流程圖，它主要可分為三大部分，第一部份為本計畫所提出的臉部表情辨識，在這裡臉部表情的步驟分為擷取影像訊號、人臉偵測、五官偵測及特徵擷取；第二部份則是語音情緒辨識，在這裡語音情緒的步驟是為擷取語音訊號、語音訊號前處理和特徵擷取；第三部份則是利用前面兩大核心所擷取出的特徵進行情緒辨識，最後在將辨識後的整合結果呈現出來。接下來我們將會針對這三大部分作較詳細的介紹。

(1) 臉部表情

A. 擷取影像訊號

我們會透過手機的照相機功能，來擷取人臉的正面圖像，不管我們所擷取的影像品質如何，這些資料得出的結果，都將成為我們的參考依據。

B. 人臉偵測

首先，為了要對於臉部動作進行偵測，我們要先瞭解影像中人臉的位置在哪裡及範圍大小，在這裡我們將使用到膚色偵測及橢圓偵測，為避免亮度的影響，我們將彩色影像 RGB 分別轉成 HSV 與 YC_bC_r 的色彩空間，一開始我們利用方程式 1 將影像從 RGB 色彩空間轉換至 YC_bC_r 色彩空間：

$$\begin{cases} Y = 0.2989R + 0.5866G + 0.1145B \\ C_b = 0.5647(B - Y) \\ C_r = 0.7132(R - Y) \end{cases} \quad (1)$$

其中 Y 代表亮度，C_b 和 C_r 代表藍色的色度及紅色的色度，接著再利用 HSV 方法將影像從 RGB 色彩空間轉換至 HSV 色彩空間，其公式如下：

$$H_1 = \cos^{-1}\left(\frac{0.5[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right) \quad (2)$$

$$H = \begin{cases} H_1, & \text{if } B \le G \\ 360 - H, & \text{otherwise} \end{cases} \quad (3)$$

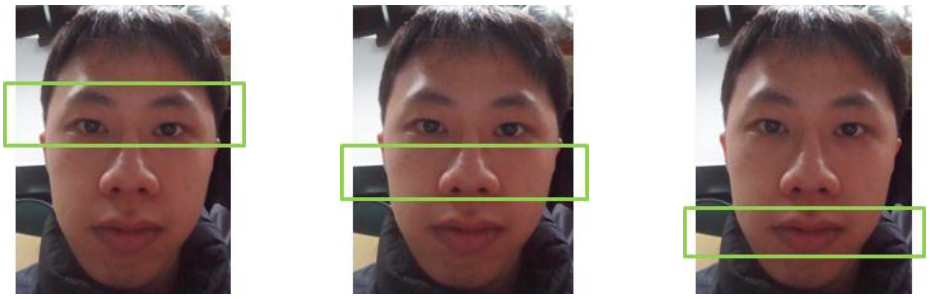
$$S = \frac{Max(R,G,B) - Min(R,G,B)}{Max(R,G,B)} \quad (4)$$

$$V = \frac{Max(R,G,B)}{255} \quad (5)$$

其中 H_1 值是 RGB 影像任一像素對應到 HSV 中 H 的像素值，H、S、V 各代表彩度、飽和度和色深度，而 Max 是為 R, G, B 三個值中最大的，相對的 Min 則是 R, G, B 三個值中最小的。經過 HSV 和 YC_bC_r 的色彩空間轉換，我們將依據各色彩空間的特性標訂出最適合擷取膚色的範圍，最後將三個色彩空間作交集來擷取膚色，但經過膚色偵測後，並不是只會偵測到人臉部分，所以我們藉由人臉是橢圓的這個特性，我們將進行橢圓偵測，以此擷取出人臉部位。

C. 五官偵測

由於眼睛及嘴唇還是會受到不同環境光線的影響，為了減少眼睛及嘴唇的錯誤偵測率，我們將在人臉影像中限制眼睛及嘴唇的搜尋區域。首先將人臉分為上、中及下三部份，眼存在於上部份，高度為人臉區域高度的 1/2；口則位在下部份，高度為人臉區域高度的 1/3；鼻位於中部份，其高度則為減去眼和口高度的剩下區域。因此，針對可能的人臉區域，以上、中、下三部份分別做 x 方向及 y 方向的邊緣偵測，藉由上述規則來找尋眼、鼻和口的位置，如圖 2。



(a) (b) (c)

圖 2、(a)眼於上部份(b)鼻於中部份(c)口則在下部份

在這裡我們是使用 Sobel 邊緣偵測的方法，找尋人臉五官位置。在 Sobel 邊緣偵測中，我們使用橫向與縱向之兩組 3×3 矩陣，以分別求出 G_X 和 G_Y，如公式 6 所示。接著再把 G_X 和 G_Y 帶入公式 7 以求出像素變換量，讓我們能找出各個位置，預期結果如圖 3，之後再擷取 Action Unit 特徵。

$$G_X = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * I \quad and \quad G_Y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * I \tag{6}$$

$$\nabla f(x, y) \cong |G_X| + |G_Y| \tag{7}$$

在這裡 I 是代表一張影像，∇f則是代表作完 Sobel 邊緣偵測後的影像。

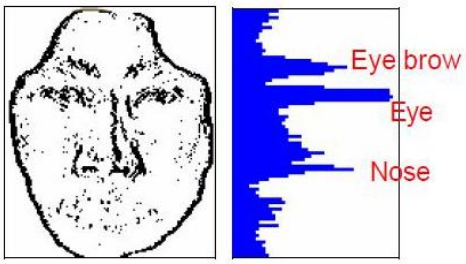


圖 3、Sobel 邊緣偵測預期結果

D. 特徵擷取

Action Units(基本動作單元)定義達 56 種之多，其中還有包括頭的角度，將可能包含的情緒分類的很詳細，但是若分類的過細時，在辨識各種表情可能區別的較不明確，容易造成系統設計上的困難，而且過多的 Action-Unit 會超出訓練與測試的負擔，不同類別之間差異性很小，對於實際應用上幫助不大。所以我們為了方便和快速的辨識，我們將人的臉部特徵區分成幾個較明顯的 Action Units，如此可以讓系統有較高的準確度，而且在收集上較固定、也較容易識別表情來做測試，如圖 4。因此，我們將對不同的表情找出最有代表性的 Action-Unit 做為訓練的 model。

從這裡我們得知人的表情是透過特定幾條肌肉控制而有不同變化，如眼睛閉上時，就有可能表示悲傷；眉毛上揚時，就有可能表示驚訝；嘴角往下時，就有可能表示悲傷或生氣。且每個人的五官特徵的寬高比不盡相同，所以我們在訓練之前會將所有類別物件進行相同寬高比正規化，接著再透過 iFuzzy LDA 來分不同的類別，以獲得到不同臉部表情的情緒特徵$\{AU_1, AU_2, \cdots, AU_N\}$。

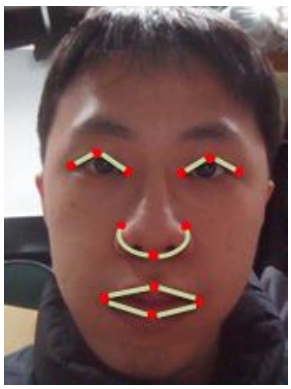


圖 4、人臉 Action Units 示意圖

(2) 語音情緒

A. 擷取語音訊號

當對談者在說話時，我們利用手持裝置內建的錄音功能，錄製一段對談者的聲音，如手機、錄音機、錄音筆或攝像機錄音等。不管我們所擷取的影像品質如何，這些資料都會丟入系統進行語音特徵擷取和情緒的判別，以增加我們的資料參考樣本。

B. 前處理

1. 去雜訊

首先，我們先輸入一個 1D signal，由於 signal 的解析度常常會受雜訊和 slew rate 的影響，因此我們對所需的訊號進行去雜訊工作，在這裡我們使用 1D median filter 把不必要的雜訊干擾排除，讓訊號能顯示出原音，並有效地去除因環境而產生的干擾，來大大的改善訊號素質。

2. 聲音訊號增強

去雜訊之後，接著我們將一段語音訊號切成 256 個樣本，再利用傅立葉轉換代替具有較高的複雜度和具有較難實現的高精密度的多通帶濾波器。為了避免音框之間變化過大，所以我們會讓音框之間有一段重疊區域，接著將每一個音框乘上漢明窗(Hamming Window)。而主要目的就是要加強音框的左端和右端的連續性，藉此將語音訊號增強。其公式如下所示：

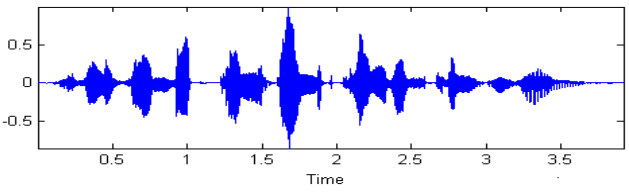
$$y(n) = x(n) \cdot w(n), 0 \le n \le N-1 \tag{8}$$

$$w(n) = \begin{cases} 0.52 - 0.46\cos(\frac{2n\pi}{N-1}), & 0 \le n \le N-1 \\ 0 \quad , & \text{otherwise} \end{cases} \tag{9}$$

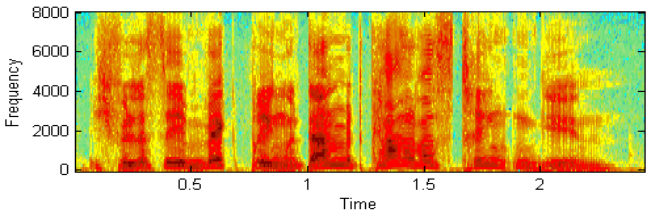
在這裡 N 表示寬度，其樣本是以一個離散時間對稱於窗函數w(n)，當 N 是奇數時，非平面的窗口會有一個訊號的最大值。當 N 是偶數時，他們有一個雙重的最大值。所以，我們才利用這方法來加強音框的連續性以增強訊號。而 n 是一個整數，其值範圍是$0 \le n \le N-1$。

3. 1D 轉 2D 的處理

經由去雜訊和訊號增強的處理，此時的語音訊號還是為 1D 形式。為了辨識語音的情緒，我們需要將如圖 5(a)的一維的聲音轉為如圖 5(b)二維的影像訊號。在這裡我們使用 Spectrogram 的方法把聲音訊號以二維影像的形式呈現出具有隨時間變化的頻譜圖和時域波形，然後由頻譜圖執行特徵擷取。



(a)



(b)

圖 5、(a)聲音一維訊號(b)二維訊號

C. 特徵擷取

經由 spectrogram 的方法後得到一個二維 Speech 的訊號後，接著我們就會使用 Law's Mask 抽取聲音特徵以便進行辨識語音情緒。此方法主要是一個應用於估測影像紋理能量變化的方法，其偵測的方式是根據紋理能量通過預設的遮罩(Mask)來計算，它包含 5 個 Masks，分別為邊(Edge)、水平(Level)、點(Spot)、漣漪(Ripple)和波浪紋(Wave)，來建構出 5 個 1 維的 Law's vectors。然後對設定好的 Edge、Level、Spot、Ripple 和 Wave 做 25 個 2 維的 Law's Masks，以取得大小為 5×5 的 25 個不同的特徵點，如表 1 所示。

在這個部分我們將訊號切成 N 等份之後，接著就以 Law's Mask 方法取得紋理資訊進行特徵值分析和比對，經處理過後，我們便會取得到到 N 個 2D 語音的特徵$\{V_1, V_2, V_3, \cdots, V_N\}$。最後，在由 iFuzzy LDA 進行分類與分析。

表 1、Law's Mask 相互組合的 2 維結果表

$L_5^TL_5$	$E_5^TL_5$	$S_5^TL_5$	$W_5^TL_5$	$R_5^TL_5$
$E_5^TL_5$	$E_5^TE_5$	$S_5^TE_5$	$W_5^TE_5$	$R_5^TE_5$
$S_5^TL_5$	$E_5^TS_5$	$S_5^TS_5$	$W_5^TS_5$	$R_5^TS_5$
$W_5^TL_5$	$E_5^TW_5$	$S_5^TW_5$	$W_5^TW_5$	$R_5^TW_5$
$R_5^TL_5$	$E_5^TR_5$	$S_5^TR_5$	$W_5^TR_5$	$R_5^TR_5$

接下來我們將相似的紋理能量影像結合，因此便能得出十四種不旋轉不變的紋理影像，接著將針對這些影像經由統計的方法分別求出 Mean, SD, Entropy，以取得紋理資訊進行特徵值分析和比對。

$$Mean = \frac{\sum_{i=0}^{M}\sum_{j=0}^{N}[TR_{ij}]}{M \times N} \tag{10}$$

$$SD = \sqrt{\frac{\sum_{i=0}^{M}\sum_{j=0}^{N}(TR_{ij} - Mean)^2}{M \times N}} \tag{11}$$

$$Entropy = \frac{\sum_{i=0}^{M}\sum_{j=0}^{N}(TR_{ij})^2}{M \times N} \tag{12}$$

利用此結果，我們使用三個方程式(10)-(12)與 TR 做紋理的判斷，此方程式分別為平均值(Mean)、標準差(SD)和熵(Entropy)，將這些使用 3 個特徵作為判斷聲音的依據，將其值儲存紀錄下來。之後，我們便會取得到到 N 個 2D 語音的特徵。最後在由 iFuzzy LDA 進行分類與分析。

(3) 情緒辨識

由於 Linear discriminant analysis (LDA)這個方法最主要是能將不同類別間的中心點(Centroid)拉開，並且能將屬於同一類別間的影像分散程度縮小，也明確的將不同類別間的距離拉開，所以可找出一條較明顯的決策邊界(Decision Boundary)區隔開兩個類別。但是 PCA 雖然有保持樣本總體離散度最大的一種降維分析方法，卻因為降維過程沒有引入分類信息，所以在使用最小距離方法進行識別時，其精度一般是小於最近鄰測量方法。反而 LDA 可得到有助於分類的最佳鑑別投影信息，所以在使用最小距離方法和最近鄰方法時精度會比較相近。因此我們選擇了以 LDA 為基底，並加入了模糊理論，因為模糊理論可用來表現某些無法明確定義的模糊性概論，尤其是在表現人類語言特有的模糊性現象方面有具體成效。所以針對模糊邏輯理論來將模糊不清的事件明確化、數據化。

在這裡我們使用了一模糊線性判斷分析(Fuzzy LDA)，其乃是一種根據 LDA 衍生而來的分類法，原先的 LDA 是一種依照已知的群組，選定判別標準來判別新樣本所歸屬的群體，但是在多類別的情況無法達到理想的分類效果，因此我們透過模糊理論的概念，來提高分類的效果。又由於先前研究以 LDA 為基礎的 Fuzzy 辨識工具，都是使用時間複雜度較高的 Fuzzy C-mean 和 Fuzzy K-NN，以及沒有修正 Fuzzy LDA 的 membership function 等缺點，所以我們使用了 iFuzzy LDA 來作為本計畫的分類器[20]。

由於我們認為若將語音情緒辨識與臉部表情辨識分開的話，有可能會得到結果是聲音在哭，但是使用者臉部卻在笑，這時我們便難以決策出使用者到底是何種情緒，所以我們將會把語音特徵及臉部特徵兩者結合，得到 N 個特徵數$\{AU_1, \mathrm{AU}_2, \cdots, AU_N\}\{V_1, V_2, \cdots, V_N\}$，接著我們將利用 iFuzzy LDA 進行聲音特徵測試、訓練和分析，其中 iFuzzy LDA 演算法如下：

Algorithm: iFuzzy LDA

1.Compute the membership function $U_{initial}$ to random numbers in the range [0, 1]
2. while(Threshold or Frequency is not equal to performance value)
3. Compute the membership function U_{LDA}

$$\text{Uphi} = U_0 \times \text{Fuzzy Exponent}$$

$$\text{Centroid}=\frac{\text{Uphi} \times \text{Data}}{\sum \text{Uphi}}$$

$$\text{Weight} = \text{inverse}(\text{Covariance}(\text{data}))$$

$$U_{LDA} = \frac{1}{1 + (\frac{\text{dist}(\text{Centroid}, \text{Weight})}{\max(\text{dist})})^{\frac{1}{\text{Uphi}-1}}}$$

4.calculate $FS_W = \sum_{i=1}^{c} \sum_{k=1}^{l_i} U_{LDA}(X_k - m_i)(X_k - m_i)^T$

$$FS_B = \sum_{i=1}^{c} \sum_{k=1}^{l_i} U_{LDA}(\text{Centroid} - m_i)(\text{Centroid} - m_i)^T$$

5. Calculate the eigenvector of $(FS_W)^{-1} * FS_b$
6.Compute the eigenvector V of step 5
7. Compute the project data using eigenvector V
8. performance value is correction rate which is equal to 0.95.
9. Updata membership function with $U_o(P+1) = U(P)$ multiplication
10. Endwhile

在此我們將取出的 N 筆特徵值輸入 iFuzzy LDA 分類器中，其學習過程通常以一次訓練一筆數據的方式進行，直到學習和完成所有筆數資料的訓練，即一個學習回合(Learning Epoch)。過程中會反覆學習，直到資料的學習達到收斂。訓練出各個情緒語音特徵向量，並對所得到的特徵向量做強化，以作為在不同情緒下之語音參考樣本，接著在計算出測試樣本與參考樣本間之距離，以期能達到情緒精確辨識之結果。之後，我們再跟情緒資料庫做分析，並從中先分類出兩類喜(Happiness)、怒(Anger)和哀(Sadness)、害怕(Fear)類型的情緒，接著再從喜、怒個別分出，同樣的再從哀、害怕個別分出。最後，我們便可以利用 iFuzzy LDA 演算法訓練後的結果來辨識出聲音目前是四種情緒類型中的哪一種，並將結果儲存於資料庫中以利後續分析及探討。

(五) 結果

本計畫主要開發在智慧型手機上，透過語音和人臉的結合來辨識人的情緒，並致力於提升即時辨識效能、便捷性及降低系統成本。在這裡本計畫會針對喜(Happiness)、怒(Anger)、哀(Sadness)和害怕(Fear)四種類型的情緒進行分析，在語音情緒方面，我們是使用 eNTERFACE 的 Database 來做為 training 的資料，且 testing 在本系統所搜集的語音上；而在人臉表情方面，我們是使用 JAFFE 和 TFEID 來做為 training 的資料，且 testing 在本系統所搜集的人臉影像上。

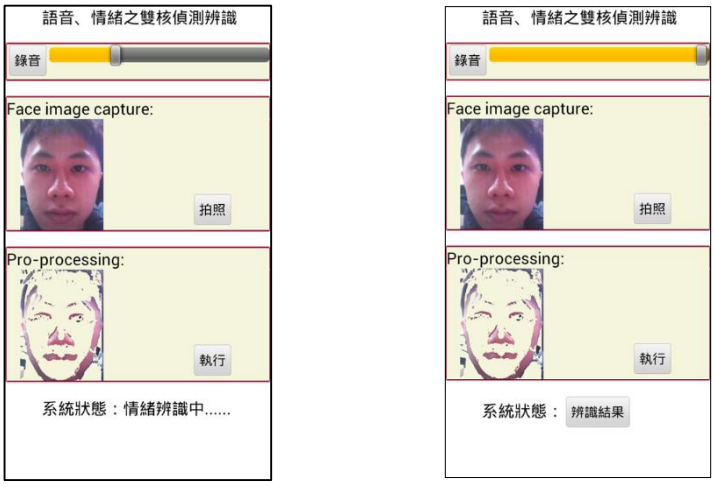
本系統整體用於 Eclipse 上開發使用，以 Java 程式撰寫，並以 HTC Android 智慧型手機來做為本系統使用的實作平台，平均情緒辨識處理時間為 10 秒。如

圖 6(a)為本系統之語音 Testing and Training 介面圖，一開始我們會將錄製的聲音儲存置手機中，接著會讀取每一筆資料與 eNTERFACE Database 進行測試訓練，等到每筆資料測試訓練後，最後系統就會依據辨識的數據差異得到平均門檻值，如圖 6(b)。而這裡我們就可以根據得到的門檻值來測試調整，以了解門檻值是不是能夠將 data 的情緒類別分類的清楚。其臉部表情方面也是和語音情緒一樣運用相同的方法測試訓練。

(a) (b)

圖 6(a) 語音錄製中畫面、(b) 語音測試、訓練和辨識畫面

當我們事先在手機上測試訓練完成後，我們就會透過如圖 7(a)的介面實際測試隨機的 data，本部分就會將語音及人臉影像的資訊一起進行辨識，而不是單獨分開辨識，最後系統辨識完成，使用者就可以按下「辨識結果」的按鈕來得知系統辨識情緒的結果是喜、怒、哀還是害怕，如圖 7(b)，其辨識結果畫面如圖 8(a)-(d)。



(a) (b)

圖 7(a) 實際測試中畫面、(b) 辨識完成後畫面

(a) (b) (c) (d)

圖 8(a)辨識結果為「喜」、(b) 辨識結果為「怒」、(c) 辨識結果為「哀」、(d) 辨識結果為「害怕」

表 1、情緒辨識結果比較

	情緒	辨識結果			
		Happiness	Fear	Sadness	Anger
情緒資料庫	Happiness	70.8	2.8	0	26.4
	Fear	2.2	81.5	16.3	0
	Sadness	2.3	11.7	85.5	0.5
	Anger	20.6	0	0.4	79

經過了實際測試之後，我們將喜、怒、哀和害怕的辨識情緒分析結果整理成表 1，在表中真正是 Happiness 被辨識成 Anger 的錯誤率為 26.4%，而真正是 Anger 被辨識成 Happiness 的錯誤率為 20.6%。上述 Happiness 和 Anger 的兩種情緒被誤判的錯誤率是相當高的。因此在最後的結果我們認為快樂和憤怒的錯誤率比率是如此高，使的讓系統容易誤判。所以這是我們需要解決的未來問題。但總體而言，本計畫提出的雙核心手持式情緒辨識整體平均 78.3%。

(六) 結論

由近年來的文獻趨勢發展，逐漸從單一模組進步到雙模以上的系統模組來提高情緒辨識度，這也代表系統的開發方向逐漸往多樣化。並且我們也用於智慧型手機中，因為現今智慧型手機的可攜性、方便性與區域網路不受距離限制等特性，讓廣大的人們可以透過我們的系統，來及時了解自己的情緒或他人的情緒。本計畫所提出的方法以雙核心的語音情緒和人臉情緒辨識應用在情緒分析上為主的系統，來改善單一訊號資訊的辨識缺陷，且無須龐大的系統資源以有效的降低系統成本，而達到系統整合的目的。

本計畫提出了一個手持式雙核情緒辨識系統，其特色是利用手持裝置，將人臉表情透過手機照相機功能，來擷取人臉的正面圖像，再將影像進行 RGB 轉換、五官偵測等處理，以得到 Action Units 特徵；以及將聲音利用手機裝置錄製聲音，

再將 1D 聲音訊號利用 spectrogram 轉成 2D 訊號影像,再經由 LBP 抽取特徵後。最後將臉部和語音特徵由 iFuzzy LDA 分類器上測試、訓練成功率，實驗結果成功率為 78.3%，證明本系統是可以有效的區分喜(Happiness)、怒(Anger)、哀(Sadness) 和害怕(Fear)四種情緒，但是在於快樂和憤怒的辨識是我們還需要再去加強的，由於兩個語音振幅相似，使得在判斷上容易有較高的錯誤率。因此，在這方面我們必須重新思考新方法，以其來改善高錯誤率。

本計畫未來希望可以改進在許多的干擾因子影響下，能透過多核心來進行多方位分析，以加強先前不足的正確性與準確性，我們還希望能讓這套系統建立於醫院中，利用雲端技術將醫院與一般使用者行動設備結合，讓彼此之間的資訊能互通有無，並建立線上情緒專家諮詢，讓我們的系統能有更大發揮的空間。

(七) 參考文獻

[1] Chao-Fa Chuang, FrankY. Shih, "Recognizing facial action units using independent component analysis and support vector machine," Pattern Recognition, vol. 39, pp. 1795 – 1798, 2006.

[2] Mahmoud Khademi, Mohammad Taghi Manzuri-Shalmani, Mohammad Hadi Kiapour, et al., "Recognizing Combinations of Facial Action Units with Different Intensity Using a Mixture of Hidden Markov Models and Neural Network," Lecture Notes in Computer Science, vol. 5997, pp. 304-313, 2010.

[3] Dimitrios Ververidis, Constantine Kotropoulos, "Emotional speech recognition: Resources, features, and methods," Speech Communication, vol.48, pp. 1162–1181, 2006.

[4] Ze-Jing Chuang, Chung-Hsien Wu, "Multi-Modal Emotion Recognition from Speech and Text," Computational Linguistics and Chinese Language Processing, vol. 9, pp. 45-62, 2004.

[5] Martin W¨ollmer, Angeliki Metallinou, Florian Eyben, et al., "Context-Sensitive Multimodal Emotion Recognition from Speech and Facial Expression using Bidirectional LSTM Modeling," INTERSPEECH Conference , pp. 2362–2365, 2010.

[6] Dan-Ning Jiang, Lian-Hong Cai, "Speech Emotion Classification with the Combination of Statistic Features and Temporal Features," IEEE International Conference on Multimedia and Expo , vol. 3, pp. 1967-1970, 2004.

[7] Bjorn Schuller, Gerhard Rigoll, Manfred Lang, "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine – Belief Network Architecture," IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 577-580, 2004.

[8] Feng Yu, Eric Chang, Ying Qing Xu, et al., "Emotion Detection from Speech to Enrich Multimedia Content," IEEE Pacific Rim Conference on Multimedia, pp. 550-557, 2001.

[9] Björn Schuller, Gerhard Rigoll, Manfred Lang, "Hidden Markov Model-based Speech Emotion Recognition," IEEE International Conference on Acoustics, Speech, and

Signal Processing, vol. 2, pp. 1-4, 2003.

[10] Yasunari Yoshitomi, Sung-Ill Kim, Takako Kawano, et al., “Effect of Sensor Fusion for Recognition of Emotional States Using Voice, Face Image and Thermal Image of Face,” IEEE International Workshop on Robot and Human Interactive Communication, pp.173-183, 2000.

[11] Ahonen, T., Hadid, A., Pietikäinen, M., “Face Description with Local Binary Patterns: Application to Face Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, pp. 2037-2041, 2006.

[12] Meng, J., Yang, Y., “Symmetrical two-dimensional pca with image measures in face recognition,” International Journal of Advanced Robotic Systems, vol. 9, 2012.

[13] Wen-Chung Kao, Ming-Chai Hsu, Yueh-Yiing Yang, “Local contrast enhancement and adaptive feature extraction for illumination-invariant face recognition,” Pattern Recognition, vol. 43, pp. 1736-1747, 2010.

[14] Ramji M. Makwana, “Illumination invariant face recognition: A survey of passive methods,” Procedia Computer Science, vol. 2, pp. 101–110, 2010.

[15] Song, J., Chen, B., Chi, Z, et al., “Face recognition based on binary template matching,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 4681, pp. 1131-1139, 2007.

[16] Goh, Y.Z., Teoh, A.B.J., Goh, K.O.M., et al., “Wavelet-based illumination invariant preprocessing in face recognition,” Journal of Electronic Imaging, vol. 18, 2009.

[17] Jean-François Connolly, Eric Granger, Robert Sabourin, “Evolution of heterogeneous ensembles through dynamic particle swarm optimization for video-based face recognition,” Pattern Recognition, vol. 45, pp. 2460–2477, 2012.

[18] 人臉資料庫 http://pics.psych.stir.ac.uk/2D_face_sets.htm

[19] 人眼偵測方法
http://csie.ntut.edu.tw/labvsp/Chinese/docdownload/2010_11_24_human_eyes.pdf

[20] Yu-Shun Cho, “Intelligent Breast Tumor Detection System with Iterative Fuzzy Linear Discriminant Analysis,” Department of Applied Information Sciences Chung Shan Medical University, 2012.