摘要

在現今視訊監控的應用中，基於臉部特徵的表情偵測為較具有挑戰性且相對重要議題。本研究提出一套情緒辨識系統，主要利用影像辨識技術搭配臉部以及肢體的前處理，訓練出一個針對特定資料集的辨識架構，並解決部分臉部或部分肢體受到遮擋以及側臉等問題。透過各個部位的深度學習架構分別對臉部表情及肢體動作進行情緒分析，並利用 Fully Connected Layer 整合兩者的 Feature Map，最後判斷受測者當下的情緒為何。本研究以 GEMEP 資料集作為主要研究對象，並跟他篇論文做比較，在資料集官方的十二種情緒中，本研究共有八個情緒得到提升；整體而論，單一臉部的準確率提升 7%，單一肢體提升 18%，臉部結合肢體共提升 9%。

關鍵字：深度學習、臉部情緒辨識、肢體情緒辨識、GEMEP 資料集

Abstract

Facial expression recognition (FER) is an important and challenging problem for automatic inspection of surveillance videos. In recent years, with the progress of hardware and the evolution of deep learning technology, it is possible to change the way of tackling facial expression recognition. Our research proposed an emotion recognition system that we mainly used image recognition technique and preprocessing of facial and posture to train a recognition structure concerning the specific dataset. We also cleared up the occlusion problems of facial or posture and head pose problems by using the different parts of the body of deep learning structure to analyze emotion of independent facial, independent posture respectively. Then use a fully connected layer to learn and concate each feature map and judge the present emotion ultimately. Our study researched GEMEP dataset specifically and compared it with other papers. In the 12 official data emotions, our research enhanced 8 emotions. Overall, the accuracy of single face enhanced 7%, single posture enhanced 18%, hybrid part enhanced 9% respectively.

Keywords: Deep learning, Facial emotion recognition, Posture emotion recognition, GEMEP dataset

目錄

圖目錄

表目錄

第一章　緒論

1.1 研究背景

隨著科技的進展日新月異，圖形處理器(Graphics Processing Unit, GPU)的硬體計算有了大幅度的提升，且各個領域的資料集(Dataset)也日趨完善，導致以往需要大量硬體運算支援的演算法及網路逐漸成熟，訓練得出的數據也取得跳躍式的進展，並漸漸應用於日常生活中。近年來，人們開始將 GPU 應用在深度學習的領域，以完成各式各樣不同的任務，如物件偵測、臉部資訊分析、影像切割等；而在臉部資訊分析的領域中人臉情緒辨識的問題一直是一個具有挑戰性且值得探討與研究的議題；相對的，肢體情緒的發展較少人涉足，因此若是利用現今硬體設備的補足和深度學習盛行的優勢，臉部與肢體情緒辨識之間的問題與進展也能取得突破，未來在許多領域中也將發揮至關重要的角色。

1.2 研究動機

情緒會影響人的思維及行為[1]，會因為當下的情緒起伏而做出不同的反應，因此若能即時判斷人類的情緒，將可以實踐在社交機器人、醫療服務、駕駛員疲勞監測等許多人與電腦視覺的交互系統，許多神經系統相關之論文[2]對於此議題的研究，也有相當程度的發展性，由此可知此方向之研究對於未來的許多科技都有正向的影響。

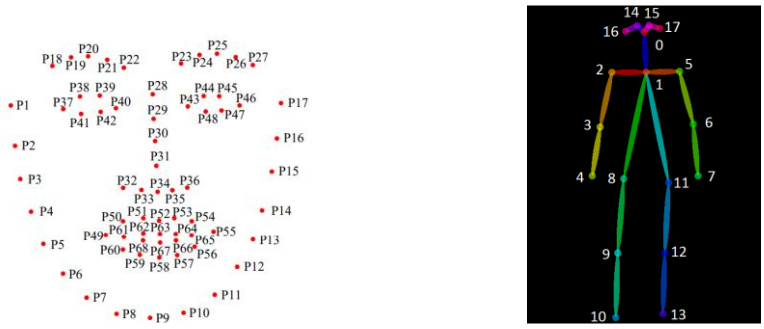
在二十世紀初期，由 Ekman 和 Friesen 的研究[3]，定義了六種最基本的情緒類別[4]，包含生氣(Anger)、厭惡(Disgust)、恐懼(Fear)、歡喜(Happiness)、難過(Sadness)與驚喜(Surprise)。然而，早期的資料集大多是在實驗室中自行蒐集並利用，到了近代因為網路技術的進步與使用者習慣的改變，才讓資料集的數量透過網路而增加。許多近期的論文[5-7]研究方向主要是在解決資料集的不足，因在網路學習的過程中會出現過擬合(Overfitting)的現象與表情無關的變量所造成學習上的困難，例如：光源變化(Illumination Variance)、非正臉的頭部姿勢(Non-Frontal Head Pose)以及身份誤差(Identity Bias)。有鑑於上述原因，本研究將嘗試針對靜態照片的面部表情辨識(Facial Expression Recognition)系統，加入人類肢體動作來輔助目前在人臉表情辨識所遇到的瓶頸。

根據文獻[2]中的說明，對於臉部表情偵測系統(Facial Expression Recognition System)的分類共可大致分為兩種：靜態照片 FER 與動態連續 FER。早期主要擷取特徵的方式為手動擷取或利用淺層學習(Shallow Learning)的方式。例如：Shan 等人[8]利用局部二值模式(Local Binary Pattern)處理靜態照片 FER 系統；Zhao 等人[9]則是利用在三維正交平面上的局部二值模式(LBP-TOP)處理動態連續 FER 系統。而到了 2013 年 EmotiW[10]從生活中的場景蒐集足夠多的訓練資料集，Kahou 等人[11]及 Fan 等人[12]才分別提出以 CNN 及 CNN-LSTM 為架構的網路作為訓練模型。此外，能夠達到以深度學習為主體的學習方式是由於現今硬體技術的進步與大量的資料數據(Data)才得以實現。在文獻[2]中更強調現今對於臉部表情辨識所遇到的瓶頸，除上述之外還包含臉上的障礙物(Occlusion)、有效資料集的不足(Insufficient Qualitative Data)與動作模糊(motion blur)等。同時，於 2018 年由 Barbosa 等人[13]提出的"基於數位圖片並透過身體姿勢分析情緒"的研究動機皆帶給本研究相當大的啟發，其因在於身體也是直接表達情緒的另一種方式，因此以身體為輔的偵測系統主要是利用特徵點(landmarks)擷取技術進行後續分析，透過身體的手勢、站姿、正在從事的動作等，可以讓網路模型更明確且更有效率的學習人臉表情與身體姿勢之間的關係，讓網路的學習能更趨近理想。而另

外，於 2019 年由 Filntisis 等人[14]提出的"將臉部與身體姿勢結合，混合識別兒童與機器人之間的互動"更是與本研究之方向大抵相同，其中提到：「有些情緒，例如驕傲等，其肢體的動作比臉部更容易偵測到」，文獻當中也提到了資料集的製作方式與相關的實驗步驟；實驗的結果也顯示，透過整體的模型(臉部與肢體)對受測者進行預測及分析，所得到的結果均相對於單一臉部偵測或肢體偵測較佳，讓本研究的實驗假設更明確。最後，有鑑於自動計算機視覺的人臉表情檢測方法系統對現階段檢測表情領域的幫助、需求與應用，故啟發了研究此題目的興趣。

1.3 文獻回顧

目前相關文獻對於情緒分析系統之研究，大致可分為 3 個部分，其包含前處理 (Pre-processing)、特徵擷取 (Feature extraction)與分類器(Classifier)之組合。前處理(Pre-processing)的部分主要在於如何快速定位臉部位置與肢體位置，再利用有效且快速的方法進行臉部與肢體位置的特徵點擷取，如圖 1 所示，並透過臉部與肢體位置之特徵點，快速擷取出相對應之感興趣的區域(Region of Interest, ROI)(圖 2)，再利用擷取出的區域萃取出特徵。接著，將特徵進行特徵篩選與組合，稱為特徵擷取。另外，特徵擷取不一定僅從臉部資訊擷取，也有可能從受測者的肢體動作、腦電圖(Electroencephalography, EEG)、呼吸、血液含氧量、汗液等等進行不同特徵之組合[15-17]。最後，將所得到的特徵擷取放入分類器(Classifier)進行受試者的情緒分類。



(a) (b)

圖 1、(a)臉部特徵點(68 點)示意圖 (b)人體姿勢偵測

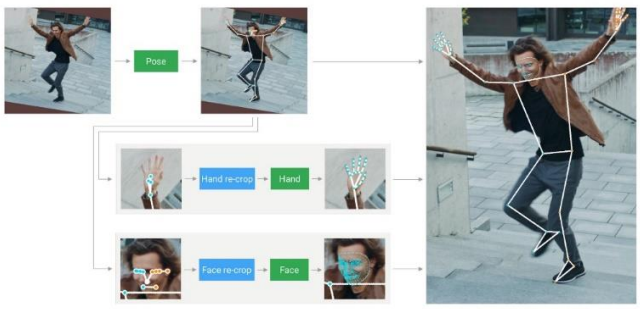


圖 2、臉部與人體姿勢之 ROI 區域示意圖

在前處理 (Pre-processing)中，又可以細分成人臉偵測(Face Detection)、肢體偵測(Pose Estimation)、特徵點擷取(Landmarks Extraction)及感興趣之區域擷取(Region of Interest, ROI)。臉部偵測(Face Detection)與人像擷取的方式有許多種，其中較經典的方式是由 Viola-Jones[18]提出的哈爾特徵(Haar-Like Features) 與串

接(Cascade)結構的自適應增強(Adaptive Boost)進行偵測，其利用一個與目標物體同尺寸的偵測視窗在偵測畫面上依據給予的步伐不斷進行計算，最終將會得出數個哈爾特徵，並將結果進行閥值的比較得出臉部位置，其最大的優點在於快速的計算，但實際測試後精準度還是不如目前主流的方法，例如：dlib、MTCNN(facenet-pytorch version)、MTCNN(Tensorflow version)等。基於網路架構和語言平台的不同，人臉辨識及特徵萃取的速度也會有所差異，透過圖 3，可以看出目前基於深度學習的網路，在不同照片解析度下的辨識速度，其中 facenet-pytorch 版本帶給本研究很大的幫助，原因在於該版本在人臉偵測速度快速且裁切精準度高(如圖 7)，當偵測大量圖片時，不論原圖亮暗、人物是否至中或正臉甚至臉部有無障礙物等，皆不需擔心照片誤判的問題導致花大量時間反覆檢查；而肢體偵測(Pose Estimation)則主要是透過(Skeleton-Texture Model) 與串接(Cascade)結構的自適應增強進行偵測。文章[19]提到將卷積網路(Convolutional Networks)應用在 Pose Machine 上，透過多階段的結構，不斷學習上一個階段的 Belief Map 以及原始圖片的特徵，並通過卷積層提取資訊產生新的 Belief Map，最終取得較為準確的結果；文章[20]提出利用 PAF(Part Affinity Fields)的向量分析，解決早期無法處理的多人重疊的問題，且在即時影像的分析中達到 30 幀以上。近年來，深度學習已被廣泛運用於電腦科學領域中，其因在於此方法透過大量閱覽資料，對數據進行統計以及分析且從中歸納出資料的規律性，並透過這些規律建構網路模型，用以對未知的資料進行預測以及推論。

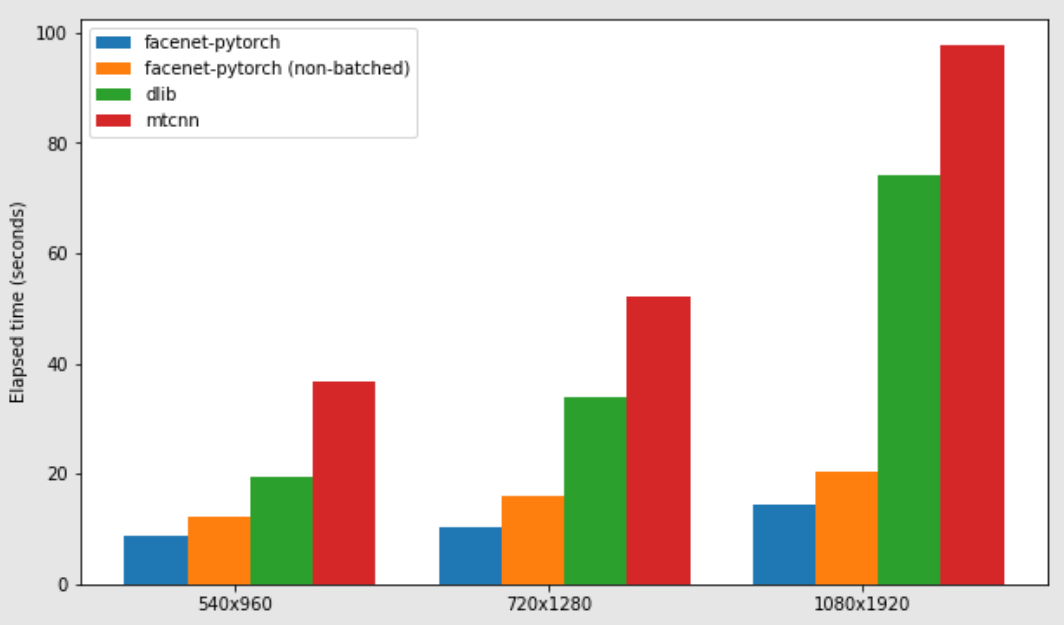


圖 3、各種臉部辨識方法的速度差異

在一般的情緒辨識網路中，通常以人臉情緒作為主要偵測的依據，相關的資料集也非常豐富，而其中偵測情緒的方法有許多種，例如由 K. Wang 等人所提出的 RAN(Region Attention Networks)網路[21]，利用圖片縮放及擷取擴增資料集(Data Augmentation)並突顯人臉的五官，同時使用 Self Attention 找出每張圖片的特徵，再藉由 Relative Attention 分享 Self-Attention 間的關聯性，解決部份臉部被遮蔽(Occlusion)及臉部角度傾斜(Pose-Variation)的問題；由 D .Meng 等人提出 FAN(Frame Attention Networks)模型[22]，先把每部影片切成前中後三段，並從中

各隨機挑選一張照片，透過 Self-Attention 獲得該照片各別資訊並標註重要的部分，透過 Relative-Attention 取得照片間的關聯性，用來判斷該影片的情緒，主要應用在分析影片的主要情緒；文獻[14]提出將人臉與肢體的情緒辨識做整合，臉部使用 Frame Level Face Alignment 框取人臉，並透過 Resnet50[23]學習人臉特徵，肢體則是使用 Frame Level Pose Detection 獲得肢體的節點(Keypoints)，使用 DNN(Deep Neural Network)學習肢體特徵，最終透過 FC(Fully Connected)把臉部及肢體的分數結合並判斷最終結果；文獻[24]提出 PSR(Pyramid With Super Resolution)網路，先透過 STN[25]將臉部校正，並利用 Down-Scale 和 SR 模擬遇到低解析度時還原高解析度的情境，因此輸入之影像大小與解析度不受限制，更提出 PDLS(Prior Distribution Label Smoothing)解決資料集不平衡的問題；文章[26]提出 AMR(Amend Representation Module)模型，照片會先進入 CNN(Convolution Neural Network)萃取特徵，再利用 FA(Feature Arrangement Block)重組新的 Feature Map，透過 DA(De-albino Block) 重新取得邊緣細節的特徵，並由 SA(Sharing Affinity Block)做整體的加權，最終透過 FC(Fully Connected)判斷結果，其優勢在於降低圖片模糊化、透過 FA 技術使得邊緣特徵不會遺漏。

最後，將上述擷取出的不同特徵引入分類器進行分類，可以採用 Softmax 、SVM(Support Vector Machine)、FCL(Fully Connected Layer)等。Softmax 跟 SVM 是一種機器學習的方法，可用來處理資料的二分類或多分類(Multi-Class)問題。而 FCL 是一種透過深度學習(Deep Learning)的方式，尋找神經元之間的關係並進行分類，並可以良好的解決非線性的問題，所以當面對真實影像(Wild Image)時，FCL 分類效果會比機器學習來的好。

1.4 專題貢獻

本研究將專注於臉部及肢體情緒變化，提出一項利用人臉情緒(Facial Emotion)和肢體情緒(Body Posture Emotion)的情緒辨識系統，以人臉為主肢體為輔的方式有效提升情緒辨識的精準度，當臉部或是肢體受到外力影響時，例如：臉部常見的障礙物、光線的變化、人臉位置等等，仍然可以精準判斷此張照片受測者的主要與次要情緒為何，並應用於未來的各種場合或產品，例如：與現代人息息相關的車用駕駛專心與否之議題，根據美國國家高速公路運輸安全管理局(NHTSA)的統計資料顯示[27]，多數的交通事故肇因於不專心駕駛，因此若能應用於此，相信會對於車禍的肇事率獲得一定程度的改善。

第二章　理論背景

2.1 研究方法

在本章節中將介紹本研究採取之臉部情緒結合肢體情緒的混合偵測結構，圖 4 為臉部情緒與肢體情緒偵測的完整流程圖，本研究將使用多種不同的資料庫進行預訓練(Pretrained)，臉部與肢體分別採用 RAF-DB 資料庫與 BRED(BabyRobot Emotion Database) 進行預訓練，最後將針對 GEMEP Corpus 資料庫做訓練及測試，並嘗試與他人數據進行比較。

首先進行資料前處理，並利用 MTCNN 與 OpenPose 分別進行人臉偵測與肢體特徵點擷取，以提高模型對於臉部及肢體表情的學習準確性；人臉部分還會加入 STN 技術使人臉可以對齊正臉(Alignment)，讓模型學習的更好。有了臉部與肢體的資料後，即可以採用這些資訊，對受測者的情緒資料進行訓練及判斷，人臉主要以 ARM 模型進行訓練，肢體則採用 DNN 模型；接著將兩者的特徵圖(Feature Map)進行一維的連結，送入 Fully Connected Layer 進行最後的分類，即可判斷受測者的主要情緒為何。

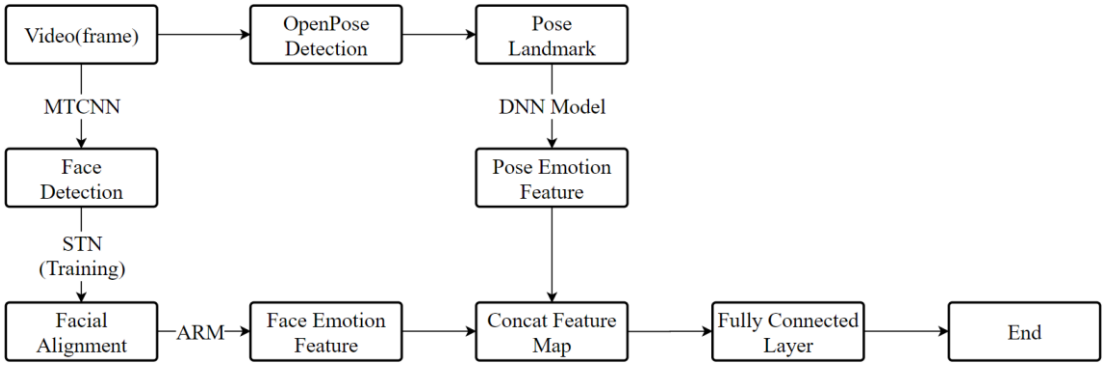


圖 4、整體架構流程圖

2.2 研究步驟

2.2.1 人臉偵測

基於深度學習的人臉偵測演算法 MTCNN(Multi-task Cascaded Convolution Networks)[28]，偵測流程如圖 5。透過 P-Net(Proposal Network)、R-Net(Refine Network)與 O-Net(Output Network)網路來偵測臉部和特徵點(如圖 6)。首先 P-Net(Proposal Network)淺層 CNN(Convolution Neural Network)快速尋找每個區域是否包含人臉，若包含人臉，則以 Bounding Box 框取，通過非極大值抑制(Non-Maximum Suppression, NMS)保留分數最高分的 Bounding Box 以及合併高度重複框取的 Bounding Box；接著，R-Net 透過 FCN(Fully Convolution Networks) 判斷 P-Net 所框出的 Bounding Box 是否有人臉，同時回歸相對應的 Bounding Box，並經過 NMS 過濾；最後 O-Net 透過 FCN 再次判斷 R-Net 所框取的 Bounding Box，並輸出 5 個 landmark(左右眼、嘴巴、左右嘴角)，透過五點框出臉部的位置。另外，部分臉部被物品遮蔽或是只露出側臉如圖 7，MTCNN 同樣可以精準偵測人的臉部位置。

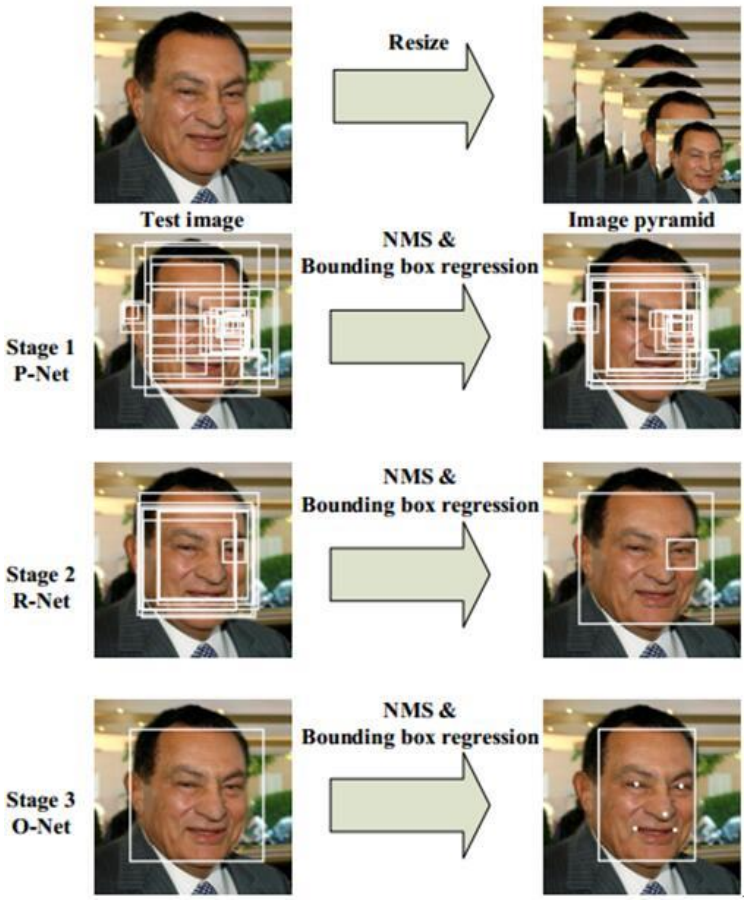


圖 5、MTCNN 框取人臉流程圖

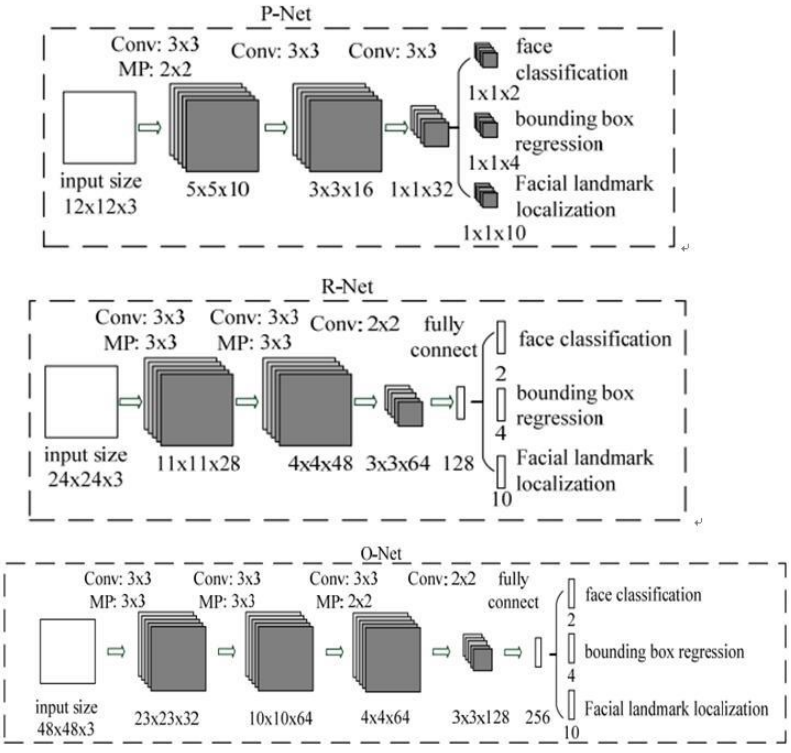
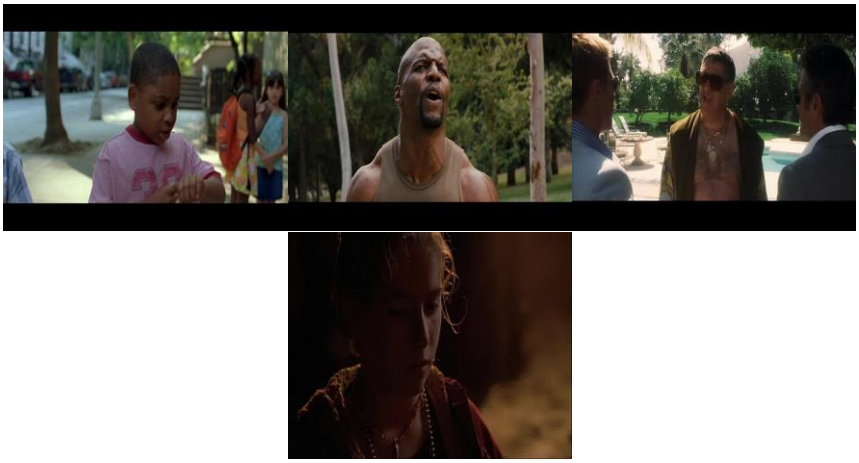


圖 6、MTCNN 網路架構圖



(a)

(b)

圖 7、(a)原圖 (b)facenet-pytorch 臉部辨識與裁切結果

2.2.2 人臉情緒訓練與判斷

文章[26]提出 AMR(Amend Representation Module)模型用以解決照片邊緣的特徵因為卷積特性而白化的缺點(Albino)，示意如圖 8。

1	2	3	3	3	2	1
2	4	6	6	6	4	2
3	6	9	9	9	6	3
3	6	9	9	9	6	3
3	6	9	9	9	6	3
2	4	6	6	6	4	2
1	2	3	3	3	2	1

圖 8、邊緣特徵忽略示意圖

該網路架構如圖 9，一個 Batch 的照片會先進入 ResNet18 萃取特徵，進入全連接層(Fully Connected Layer)之前做輸出並分成上下兩層作輸入，上層主要是對單張照片做萃取，下層主要是依據 Batch 大小對所有照片做萃取，接著上下兩塊各自會經過 3 個部分，分別是 FA(Feature Arrangement Block)、DA(De-albino Block)和 SA(Sharing Affinity Block)以取代原本的全連接層。

FA 主要是將 ResNet18 所萃取出的特徵從 512*7*7 的重組成 2*112*112 的

Feature Map，排列方式是將每張 Feature Map 的相對位置匯聚於同一區塊，示意如圖 10，其目的在於讓之後的 DA Block Kernel 萃取特徵時，可以更佳容易萃取邊緣被忽略掉的特徵。

DA 主要是利用新的 Kernel 做完整的特徵萃取，但是新的 Feature Map 不會做 Padding，目的是降地特徵模糊化，且重新擷取的 Kernel 會比原本在 Resnet 中的大上數倍，而步伐(Stride)大小則為至關重要的因素，小的步伐雖然有較少的資訊損失，但也限制了 DA Block 的效率，因此 Stride 大小為 Kernel 尺寸的一半是經過權衡的結果；而經過圖 11 的實驗結果證明，Kernel Size 為 5 時的準確率(Accuracy)最高，這些參數皆是為了更好的讓機器學習容易被忽略的邊緣特徵。

SA Block 的設置在於優化前期由網路學習到的特徵，當該圖片的情緒表達較冷門時，可以透過其他照片作參考，並有效提升準確率。首先我們從 FA 及 DA Block 學習到一個 Batch 所有的特徵集合$R = \{a_i\}_{i=1}^{N}$，N 及a_i各自代表 Batch 的數量和第i張圖片的原始特徵，而 Affinity 藉由 Average Global Representation 加權平均這些特徵當作整體 Batch 的整體特徵，公式如下：

$$a_{GAR} = \frac{\sum_{i=1}^{N} a_i}{N} \tag{1}$$

並把上式結果與每一張照片的特徵結合，公式如下：

$$a_{i,SA} = a_i + \lambda a_{GAR} \tag{2}$$

$a_{i,SA}$代表整體與個別的特徵經由 SA Block 相合併的結果；λ則代表 Affinity 並控制 Average Global Representation 的重要程度，也是可調的參數之一。最終透過 FC(Fully Connected)判斷結果。

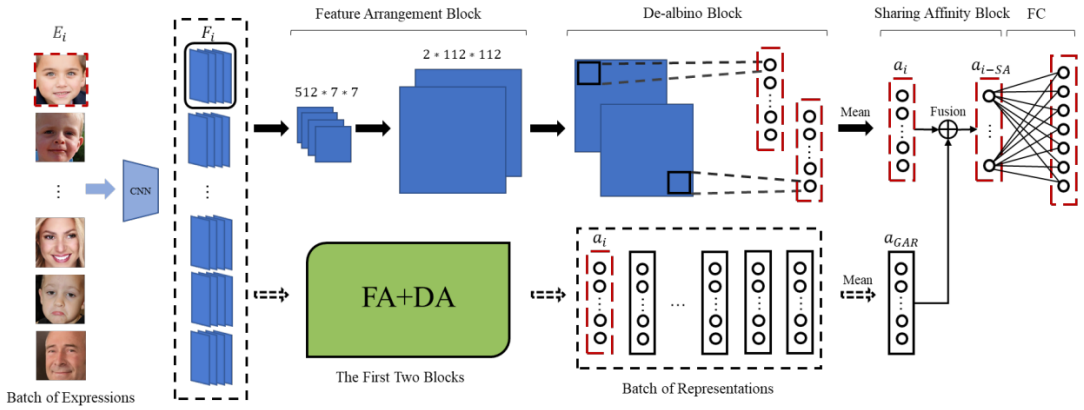


圖 9、ARM 架構圖

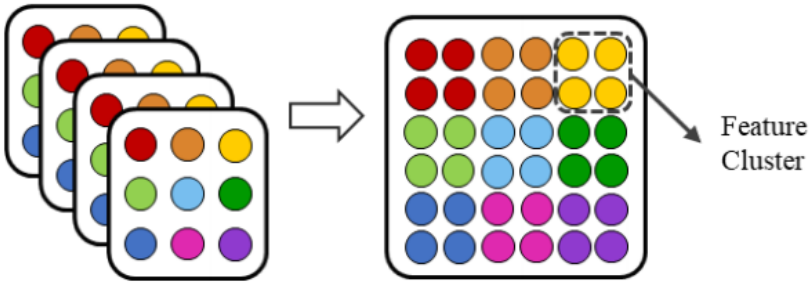


圖 10、FA 示意圖

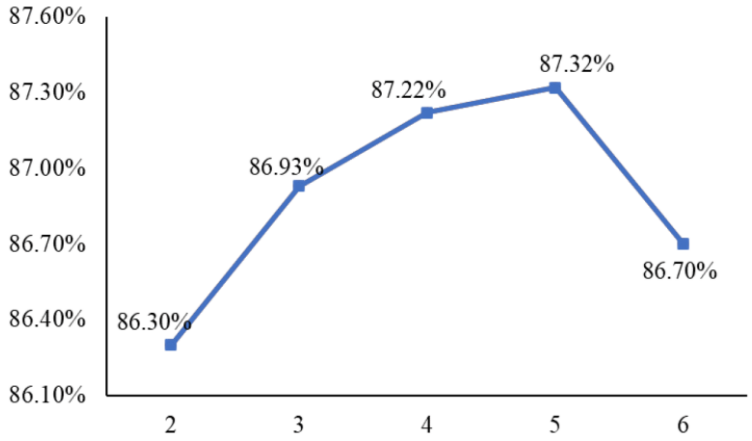


圖 11、Kernel Size 實驗數據圖

2.2.3 肢體特徵點偵測

透過多篇文章[14, 29, 30]發現，肢體辨識與特偵點的偵測皆是透過 OpenPose[20]來獲得身體的節點(Keypoints)，其主要的貢獻為應用 PAF 向量辨識在多人場景下的各個肢體特徵點。整體流程如圖 12，圖 12.a 為輸入圖片，接著由模型同時預測關節點的 Confidence Maps(圖 12.b)和 PAF 向量(Part Affinity Fields)(圖 12.c)，接著將第二步驟的預測做後處理(圖 12.d)，最後得到整體的姿態(圖 12.e)。

模型架構可分為兩個部分，分別是用於預測 PAF 的藍色區塊以及用於預測 Confidence Maps 的橘色區塊，如圖 13。首先圖 13 中的 F 為原始影像經過 VGG-19 前十層後萃取的特徵，並分別送入分支一(Branch 1)以及分支二(Branch 2)，前者主要是預測每一個關節的 Confidence Map，後者則是預測每一個軀幹的 PAF 二維向量，兩者的輸出都會各自對應與輸入相同數量的 Feature Map，接著透過多個 Stage 的迭代提升關節與軀幹的精準度。

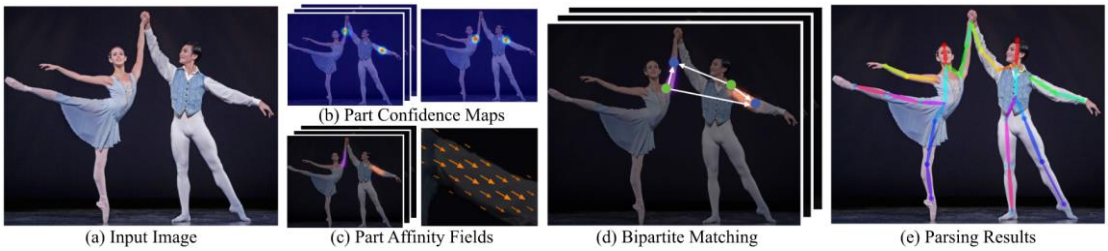


圖 12、OpenPose 流程圖

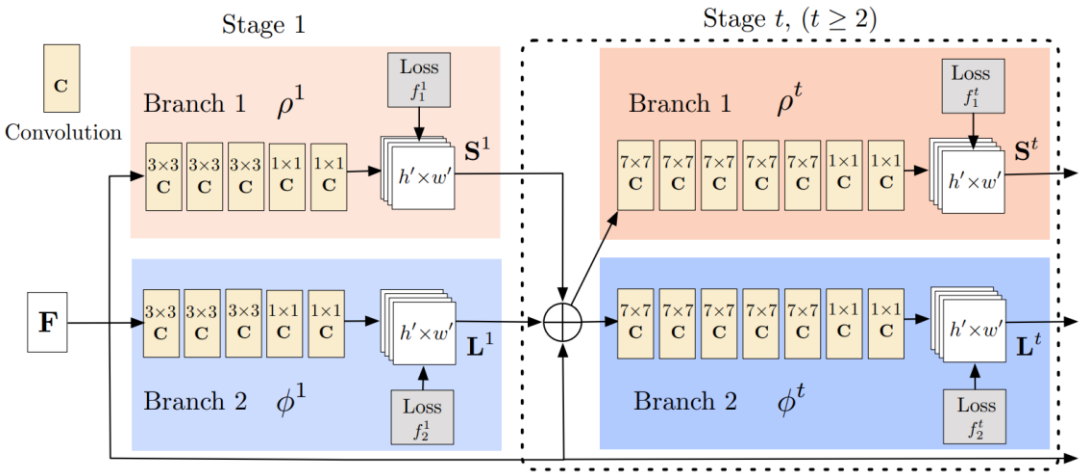


圖 13、OpenPose 網路架構圖

PAF 向量主要是偵測肢體的方向， 圖 14 為第 k 個人的前臂，$x_{j1,k}$為手肘關節點，$x_{j2,k}$為手腕關節點，兩點之間的區域即為此人的手臂 PAF，若是在這之間的像素則給定一個值$L_{c,k}^*(p) = v$，反之若在範圍外的像素則給定$L_{c,k}^*(p) = 0$，如公式(3)，其中 p 代表該圖片中的每一個點；$L_{c,k}^*$則為一個判斷 p 點是否在手臂上的涵式。而 v 則代表手臂的單位向量，其計算式如公式(4)，並判斷是否符合不等式(5)，若不等式成立，則 p 點位於手臂的 PAF 向量中，換言之，$\overrightarrow{x_{j1,k}p}$投影到$\overrightarrow{x_{j1,k}x_{j2,k}}$的向量及法向量都必須符合手臂像素的長度與寬度。

從上述技術可以知道，透過模型的可以找到人的關節點與 PAF 向量資訊，而此技術在針對多人情境下也能正確偵測並完成徵點之間的連結，如圖 15 所示。

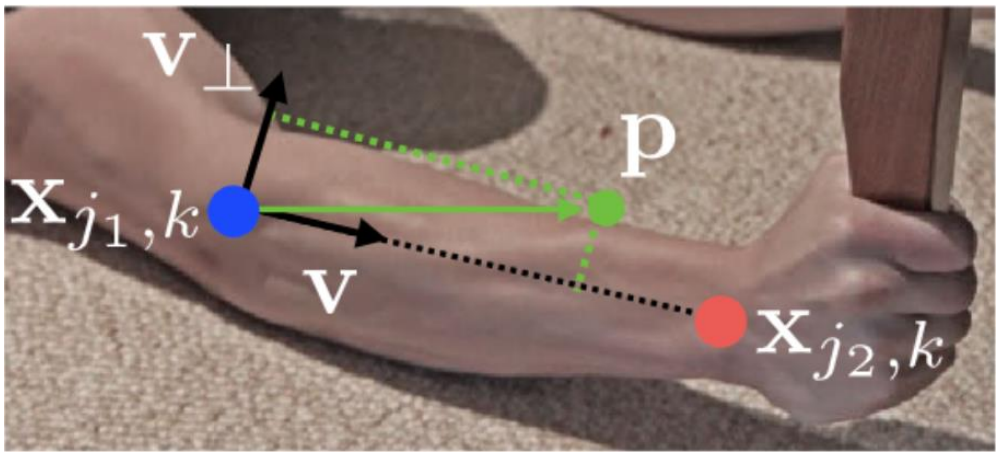


圖 14、PAF 定義示意圖

$$L_{c,k}^*(p) = \begin{cases} v \text{ if p on limb c,k} \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

$$v = \frac{(x_{j2,k} - x_{j1,k})}{\| x_{j2,k} - x_{j1,k} \|_2} \tag{4}$$

$$0 \leq v \cdot (p - x_{j1,k}) \leq l_{c,k} \text{ and } |v_\perp \cdot (p - x_{j1,k})| \leq \sigma_l \tag{5}$$

圖 15、多人情境的肢點偵測

2.2.4 肢體情緒訓練與判斷

本研究的肢體透過 OpenPose 取得特徵點並以 json 檔案的形式儲存，這些特徵點並非為圖片，沒有空間相對關係，因此本研究採取 DNN(Deep Neural Network)的方式尋找每個特徵點之間的關係並萃取其特徵。

本研究參考文章[14]中關於肢體特徵萃取的模型，其流程如圖 16 中的紅框處，在資料前處理的部分，是把在影片中所有 Frame 的特徵點先疊加後再送入模型，因此模型會根據所有 Frame 的特徵進行預測，輸出一個答案。在 DNN 模型的部分則是採用一層的 Linear 搭配 ReLU 激勵函數，並分別對資料進行 Average Pooling 及 Batch Normalization 的運算後做最後的分類。

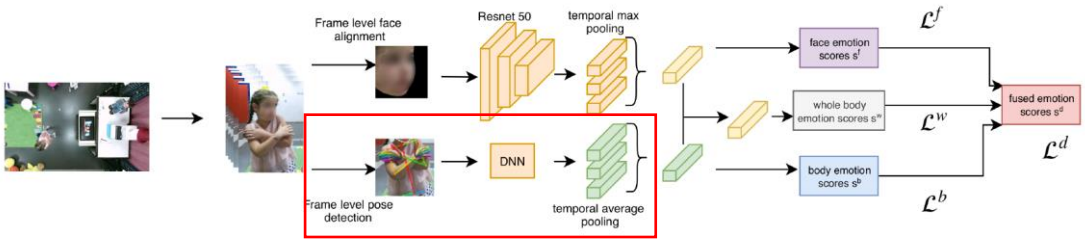


圖 16、Fusing Body 流程圖

本研究使用的 DNN 隱藏層(如圖 17)即為許多的線性關係和激勵函數組合而成的，公式如下：

$$y = a(W \cdot X + b) \tag{6}$$

其中 W 代表該神經元的權重，X 代表輸入的資料，b 代表偏移量，a 則代表激勵函數。為了找到輸入與輸出之間的關係式，需要藉由中間隱藏層的各個權重及偏移量改變線性的方程式，但若是只有單純的線性關係，可能很難找到完全符合兩者關係的方程式，因此特別加入激勵函數使方程式變成非線性，達到在多維空間中可彎曲的變化。

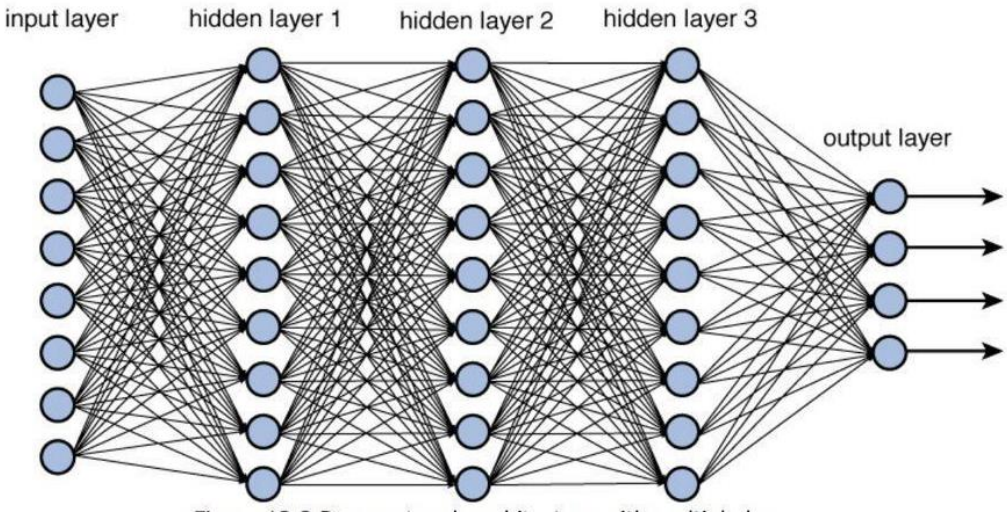


圖 17、DNN 隱藏層示意圖

同時為了使模型更加穩定及增加學習的效率，加入平均池化層(Average Pooling Layer)及 Batch Normalization。前者(如圖 18)主要做下採樣(Down-Sampling)並創立新的 Feature Map，除了可以減少後續 Layer 所需的參數量以加快學習的速度，對於鄰近的特徵偏差還具有抗干擾作用同時減少過擬合(Over-Fitting)的情況發生；而後者則是將每一層的輸入都做規一化(Normalize)，使每一個 Mini-Batch 達到平均值為零、標準差為一的狀態，如此可以將數據統一，幫助模型加速收斂，同時減緩梯度消失[31]。最後經由全連接層輸出各個情緒的預測值。

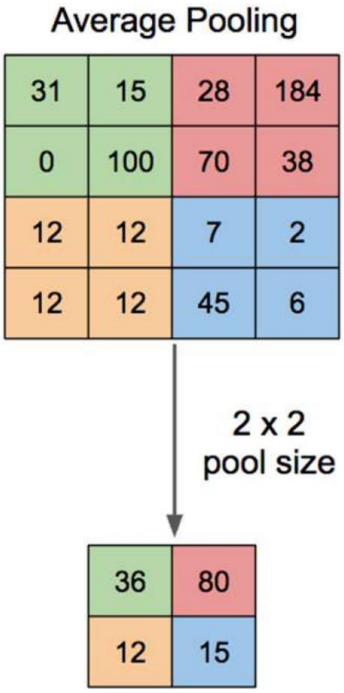


圖 18、Average Pooling 運作原理

2.2.5 Fully Connected Layer

透過上述兩個模型對於資料的預測，可以得到同一組資料的臉部以及肢體特徵圖，並利用 Concat 語法使兩者做一維的向量合併，最後送入一層的 Fully

Connected Layer 讓萃取出的 Feature 可以對映(Mapping)到樣本空間中(Label Space)完成分類的任務。此皆段的輸入為 1*24 的特徵，而輸出則為 1*12 的情緒特徵，整體的操作示意如圖 19。在訓練階段，為了避免全連接層出現過擬合(Overfitting)的問題，根據文章[32]的證明，我們適當的加入了 Dropout 機制，避免模型過於依賴訓練資料的權重。

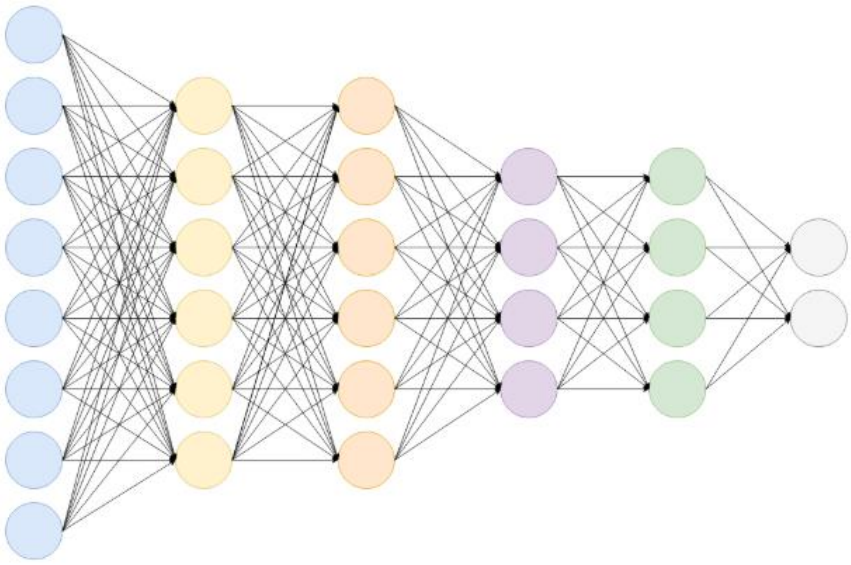


圖 19、全連接層示意圖

第三章　實驗數據

3.1 資料集

為了使模型能更好的學習訓練資料，我們分別利用 RAF-DB 與 BRED 資料庫進行模型的預訓練，並同時訓練由 GEMEP 影片裁切過後的每一幀，在臉部情緒辨識的演算法學習與驗證中，輸入採用各別的幀獨立學習；對於肢體情緒模型而言，其輸入為序列的幀所形成的連貫特徵點，兩者皆套用十次 Leave-One-Out 交叉驗證 (Cross Validations)的方法來進一步證實臉部與肢體情緒辨識系統的強健性，此節將針對三種所使用的資料庫進行詳細的分析與介紹

RAF-DB 資料庫[33]，是由 Shan Li 等人所提供的真實情感人臉資料庫，並透過 40 位專業人士單獨標記。資料集共分成兩個不同的子集，單標籤集(Single-Label Subset, Basic Emotions)和兩個標籤集(Two-Tab Subset, Compound Emotions)，總共 30,000 張臉部圖片。本實驗將使用具有 7 類情緒的單標籤集，其中的情緒包含：驚喜、恐懼、厭惡、快樂、悲傷、憤怒和中性。RAF-DB 單標籤集的訓練集和測試集數量分別為 12271 和 3068 張，各情緒之數量分佈如表 1。由於 RAF-DB 是多方來源的真實人臉情感資料集，增加許多臉部的多樣性與變化性(範例如圖 20)，包括臉部角度、有無障礙物、光影變化及照片大小的分佈不一(如圖 21)等等，大大增加訓練和辨識的難度。

圖 20、RAF-DB 資料範例

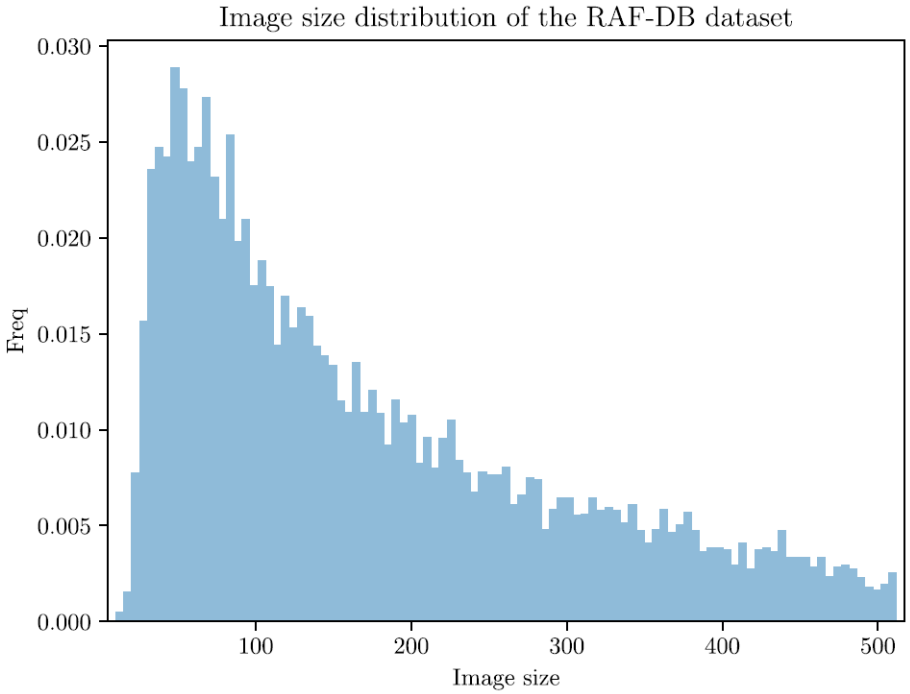


圖 21、RAF-DB 圖片大小分佈圖

BRED(BabyRobot Emotion Database)資料集[14]，透過小朋友與人問答的方式以及與機器人(Zeno and Furhat)的互動，收集過程中呈現的情緒資料(如圖 22)。共有 30 位年齡介於 6 歲到 12 歲的孩童參與數據的收集。在收集數據的過程中不會給孩子任何表達情緒的指示，完全由孩子自行發揮。資料集的情緒包含:生氣、開心、害怕、難過、噁心、驚喜，而這些情緒將會經由 3 個階段判斷臉部或是肢體是否有呈現該情緒，若 3 個階段中有 2 個階段判定"False"，將會把該情緒判斷為中性。本實驗將會使用 BRED 資料集中的肢體資料集，做為肢體情緒模型的預訓練資料，其資料包含 215 部的影片，且影片已經事先經過處理。

圖 22、BRED 資料收集範例

GEMEP(Geneva Multimodal Emotion Portrayals)資料集[34]是由 10 位法國演員演出 17 種情緒(範例如圖 23)，包含:欣賞(Admiration)、娛樂(Amusement)、憤怒(Anger)、焦慮(Anxiety)、蔑視(Contempt)、絕望(Despair)、厭惡(Disgust)、恐懼

(Fear)、興趣(Interest)、煩躁(Irritation)、喜悅(Joy)、快樂(Pleasure)、驕傲(Pride)、寬慰(Relief)、悲傷(Sadness)、驚喜(Surprise)和溫和(Tenderness)。資料集包含全身和上半身，本研究將會使用全身性的資料，從中針對 7 總情緒當作訓練和測試資料，分別是驚喜、恐懼、厭惡、喜悅、悲傷、憤怒以及、蔑視。

圖 23、GEMEP 資料集範例

表 1、各資料集情緒數量分佈

	RAF-DB		**BRED**		**GEMEP**
Emotion	Train	Test	Pre-Game	Game	Video
Anger	705	162	9	19	10
Disgust	717	160	18	26	5
Fear	281	74	11	22	10
Happiness	4772	1185	21	24	-
Neutral	2524	680	-	-	-
Sadness	1982	487	12	23	10
Surprise	1290	329	10	20	5
Contempt	-	-	-	-	5
Anxiety	-	-	-	-	10
Joy	-	-	-	-	10
Pleasure	-	-	-	-	10
Amusement	-	-	-	-	10
Tenderness	-	-	-	-	5
Despair	-	-	-	-	10
Irritation	-	-	-	-	10
Relief	-	-	-	-	10
Interest	-	-	-	-	10

Admiration	-	-	-	-	5
Pride	-	-	-	-	10
Total	**12271**	**3068**	**81**	**134**	**145**

3.2 實驗結果

3.2.1 實驗方法

本研究於測試階段採用的資料集為 GEMEP(Geneva Multimodal Emotion Portrayals)，由於此資料集沒有事先切割訓練資料集(Training Dataset)、驗證資料集(Validation Dataset)與測試資料集(Testing Dataset)，為了不失公允與爭議的問題發生，本研究將以 Leave-One-Out 的方式(如圖 24)，對 GEMEP 資料集做測試。

Leave-One-Out 的方式是將 GEMEP 資料集的每位演員獨立當成測試資料集，其他演員的所有資料則為訓練資料集，而 GEMEP 資料集是由 10 位演員演出，因此會分成 10 組，最後將會有 10 個結果，而本研究將會把 10 個結果對於做平均當作最終的準確值(accuracy)。

圖 24、Leave-One-Out Cross Validation 示意圖

3.2.2 GEMEP 七類情緒數據

從表 2 可以看出，臉部情緒為主肢體情緒為輔，使得整體的精準度有顯著的提升。對於臉部的混淆矩陣而言(如圖 25)，可以發現歡樂(Pleasure)以及難過(Sadness)的情緒學習的較好，其因為透過臉部的影像，這兩者的特徵較明顯且獨特；而對肢體來說(如圖 26)，害怕(Fear)以及生氣(Anger)的情緒較臉部佳，說明兩者情緒透過肢體的展現較能表達，也證實前面提到的某些情緒的特徵，肢體較臉部更明顯。

本研究最後透過 Fully Connected Layer(FCL)將臉部情緒及肢體情緒的 Feature Map 結合並進行分類。透過圖 27 可以看出， FCL 把臉部情緒及肢體情

緒中較弱的情緒，透過彼此的彌補提升該情緒的準確性，甚至在兩方都較差的厭惡(Disgust)情緒也有發揮作用。

表 2、GEMEP 各演員於七類情緒下之測試結果

Performer Number	1	2	3	4	5	6	7	8	9	10	Total
Face Part	66.7	60	50	80	80	66.7	66.7	50	80	60	66
Body Part	50	60	66.7	60	40	50	50	66.7	80	60	58.33
Fusion	66.7	60	66.7	80	60	66.7	66.7	66.7	80	60	67.34

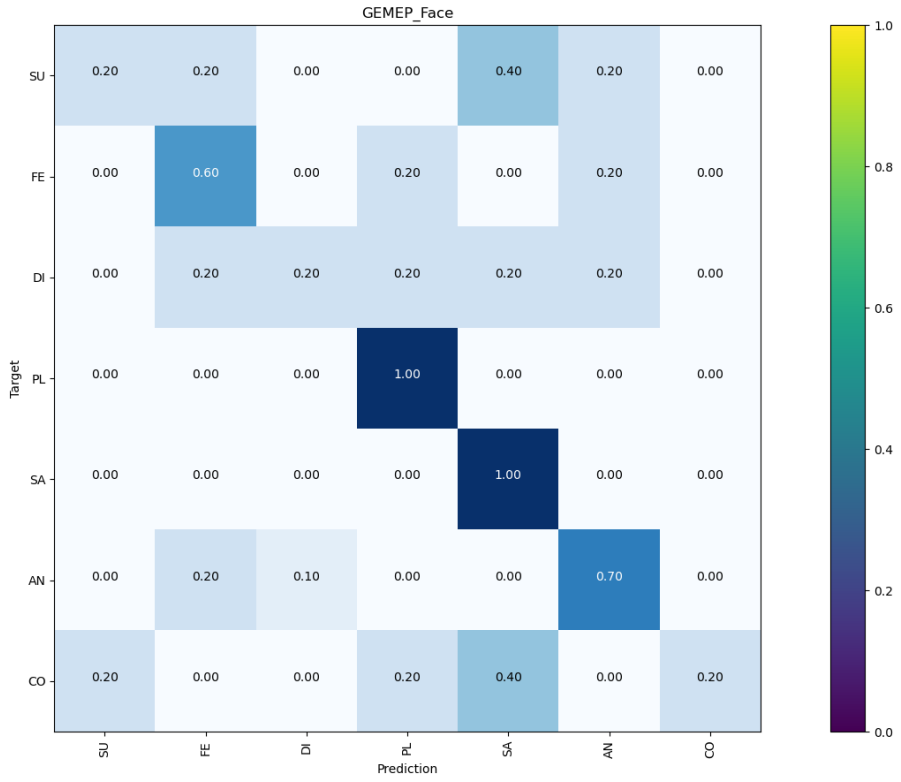


圖 25、GEMEP 七類臉部混淆矩陣

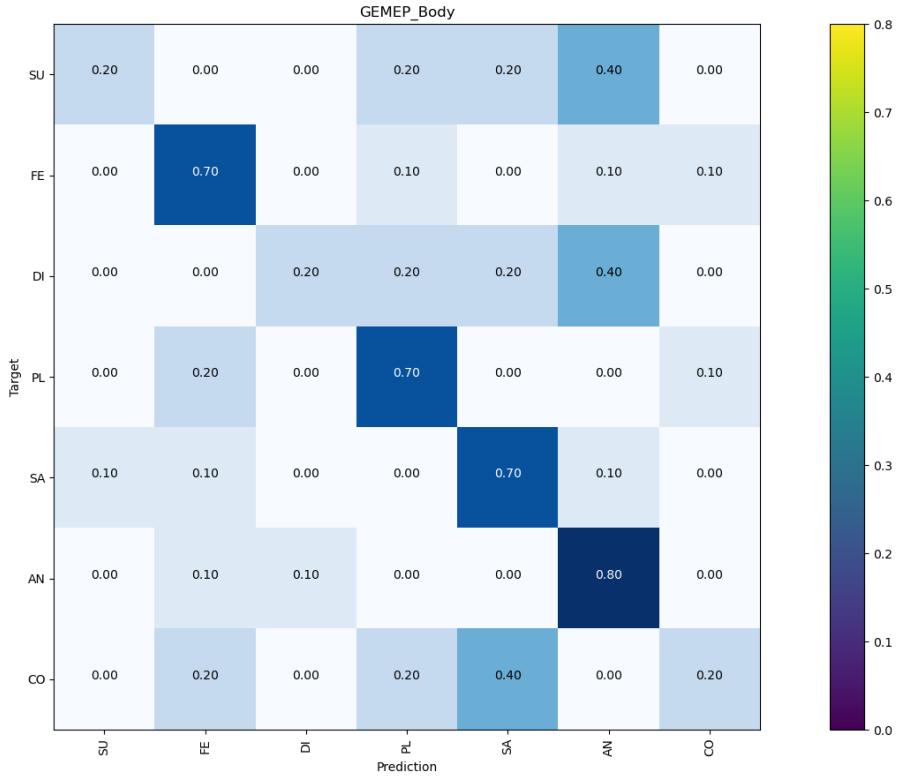


圖 26、GEMEP 七類肢體混淆矩陣

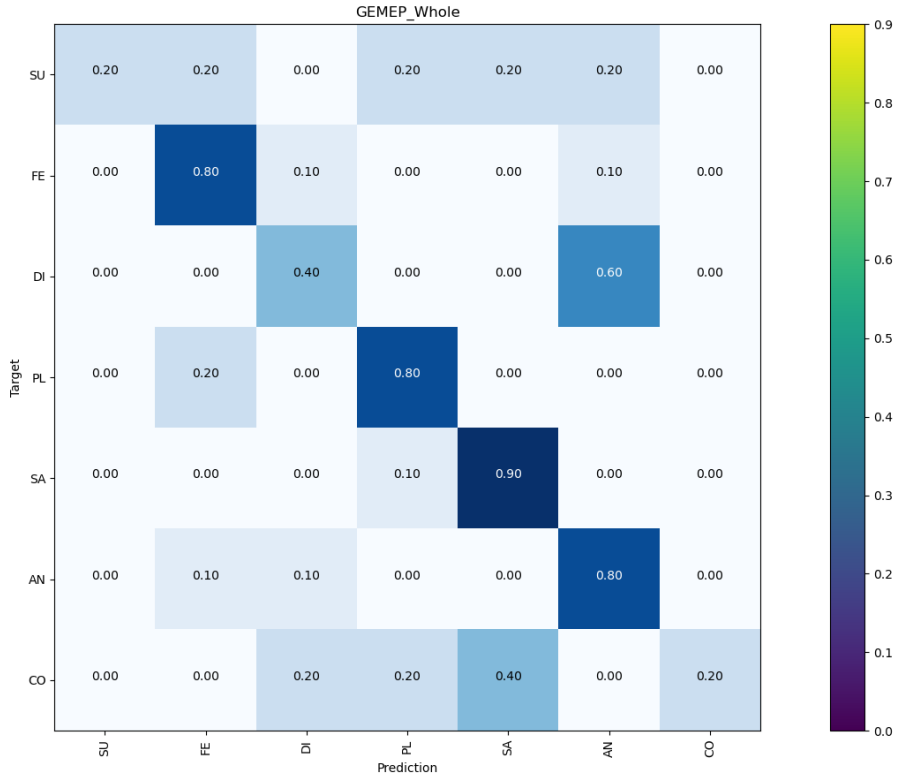
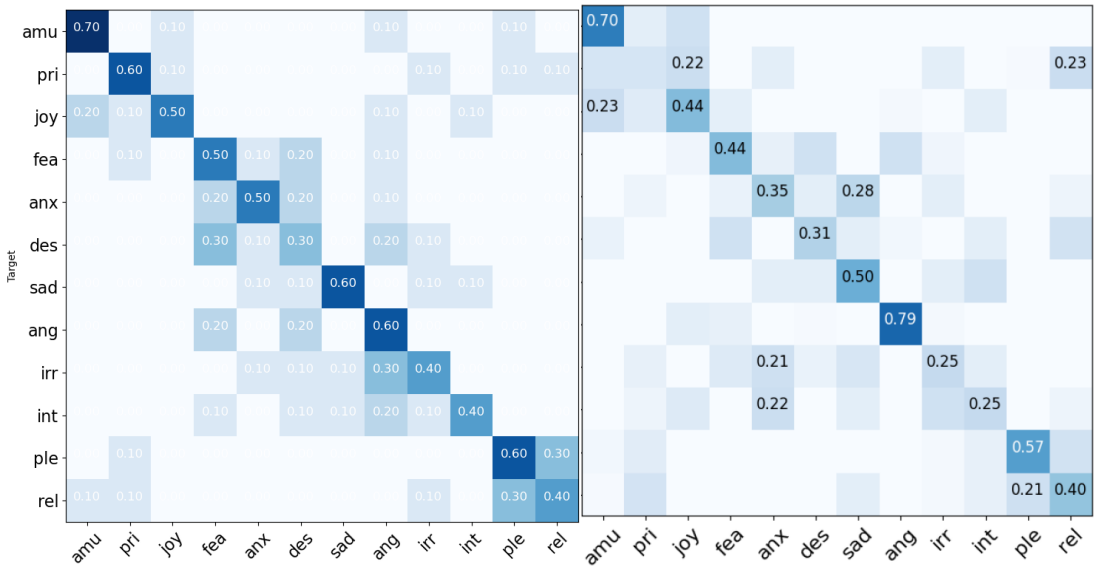


圖 27、GEMEP 七類整體混淆矩陣

3.2.3 他人技術比較

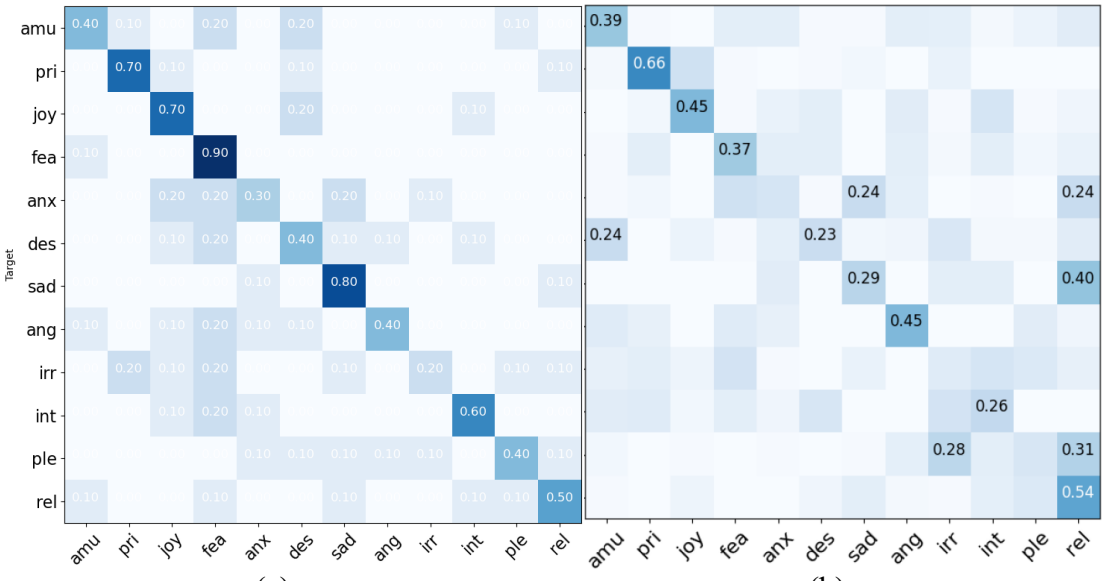
本研究與文章[14]所提出的 HMT(Hierarchical Multi-Label Training)架構做數

據比較。以臉部及肢體來說(如圖 28 和圖 29)，大多與比較對象相同甚至超越，尤其是臉部的驕傲(Pride)情緒以及肢體的害怕(Fear)、傷心(Sadness)、有趣(Interest)、歡樂(Pleasure)等情緒皆有大幅提升；透過結合臉部及肢體的特徵，可以從圖 30 發現共有八個情緒皆高於比較對象，且透過表 3 可以發現不論是單一部位或整體的準確率，皆高於比較對象。



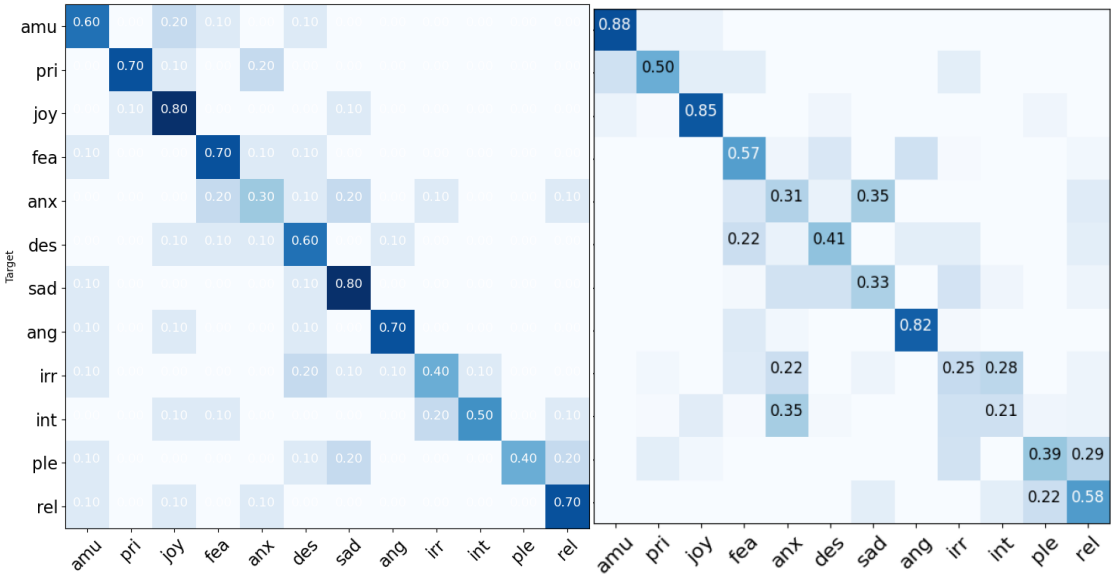
(a) (b)

圖 28、(a)本研究臉部混淆矩陣 (b)比較對象臉部混淆矩陣



(a) (b)

圖 29、(a)本研究肢體混淆矩陣 (b)比較對象肢體混淆矩陣



(a) (b)

圖 30、(a)本研究整體混淆矩陣 (b)比較對象整體混淆矩陣

表 3、個別部位與整體之數據比較

	Face Part	Body Part	Whole Part
Fusing Body[14]	43	34	51
本研究	50.83	52.5	60

3.2.4 實驗數據及設備

表 4 為本研究在實踐臉部情緒以及肢體情緒訓練的詳細資訊，表 5 為本研究使用的平台各硬體設備。

表 4、實驗參數

	Face Part	Body Part
Batch	64	64
Epochs	30	100
Learning rate(lr)	0.001	0.05
Optimizer	Adam	Adam
Step	10	50
Gamma	0.8	0.8

表 5、實驗設備

	Device 1	Device 2	Device 3
Environment	Window	Window	Ubuntu
GPU	RTX3070	RTX2060	RTX980*2
RAM	32G	24G	32G

第四章　建議與結論

本研究提出一個基於類神經網路架構的臉部整合肢體情緒辨識系統。此情緒辨識系統在 GEMEP 資料集的結果已超越[14]的研究成果。透過實驗結果可以證明，臉部情緒與肢體情緒的特徵結合，相對於只有單一臉部或肢體，更有顯著的

提升。從實驗可以看出，結合的情緒往往都是繼承臉部或肢體較好的情緒。本實驗在結合臉部與肢體情緒特徵的過程中試過多種分類器，例如：SVM、K-mean、Softmax 和 FCL。最終發現 FCL 在資料集的分類相對於其他的分類器的效果更好。本研究在整體的架構中不是端對端(End to End)的模式，因此之後能夠以此研究為基礎，設計出包含前處理、模型預測、分類等一氣呵成的端對端架構，讓其更貼近真實生活中的應用。

第五章　未來發展方向

本研究的測試資料是由專業演員以誇大的形式進行演出，所以模型較容易學習各個情緒間的差異，使得特徵萃取與分類相較容易，但在現今社會中，大部分的人都會隱瞞自我當下的情緒，尤其是較負面的情緒，例如，生氣、難過、害怕等。所以本系統在偵測現實狀況時，準確度會明顯的下降，原因在於現實狀況的人臉和肢體的變化，在大多時間裡不會有太多的改變，為此本研究認為透過學習微表情，偵測人類非自主意識的動作，可以有效地提升情緒辨識；另外未來也可以加入 Valance 和 Arousal 的資訊，使情緒辨識系統更加完善。

第六章　參考資料

[1] N. Schwarz, "Emotion, cognition, and decision making," *Cognition & Emotion,* vol. 14, no. 4, pp. 433-440, 2000.

[2] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions on Affective Computing,* pp. 1-1, 2020, doi: 10.1109/taffc.2020.2981446.

[3] P. Ekman, "Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique," 1994.

[4] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of personality and social psychology,* vol. 17, no. 2, p. 124, 1971.

[5] Y. Cheng, B. Jiang, and K. Jia, "A Deep Structure for Facial Expression Recognition under Partial Occlusion," presented at the 2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2014.

[6] M. Xu, W. Cheng, Q. Zhao, L. Ma, and F. Xu, "Facial expression recognition based on transfer learning from deep convolutional networks," in *2015 11th International Conference on Natural Computation (ICNC)*, 2015: IEEE, pp. 702-708.

[7] M. A. Ranzato, J. Susskind, V. Mnih, and G. Hinton, "On deep generative models with applications to recognition," in *CVPR 2011*, 2011: IEEE, pp. 2857-2864.

[8] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *Image and Vision Computing,* vol. 27, no. 6, pp. 803-816, 2009, doi: 10.1016/j.imavis.2008.08.005.

[9] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans Pattern Anal Mach Intell,* vol. 29, no. 6, pp. 915-28, Jun 2007, doi: 10.1109/TPAMI.2007.1110.

[10] A. Dhall, O. V. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and Image based Emotion Recognition Challenges in the Wild," presented at the

Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, 2015.
[11] S. E. Kanou *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," presented at the Proceedings of the 15th ACM on International conference on multimodal interaction - ICMI '13, 2013.
[12] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," presented at the Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016.
[13] B. Barbosa, A. J. Neves, S. C. Soares, and I. D. Dimas, "Analysis of Emotions from Body Postures Based on Digital Imaging," *SIGNAL 2018 Editors,* p. 81, 2018.
[14] P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, "Fusing Body Posture With Facial Expressions for Joint Recognition of Affect in Child–Robot Interaction," *IEEE Robotics and Automation Letters,* vol. 4, no. 4, pp. 4011-4018, 2019, doi: 10.1109/lra.2019.2930434.
[15] D. O. Bos, "EEG-based emotion recognition," *The influence of visual and auditory stimuli,* vol. 56, no. 3, pp. 1-17, 2006.
[16] M.-H. Grosbras, P. D. Ross, and P. Belin, "Categorical emotion recognition from voice improves during childhood and adolescence," *Scientific Reports,* vol. 8, no. 1, pp. 1-11, 2018.
[17] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," in *Tutorial and research workshop on affective dialogue systems*, 2004: Springer, pp. 36-48.
[18] M. Jones and P. Viola, "Fast multi-view face detection," *Mitsubishi Electric Research Lab TR-20003-96,* vol. 3, no. 14, p. 2, 2003.
[19] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724-4732.
[20] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291-7299.
[21] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing,* vol. 29, pp. 4057-4069, 2020.
[22] D. Meng, X. Peng, K. Wang, and Y. Qiao, "Frame attention networks for facial expression recognition in videos," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019: IEEE, pp. 3866-3870.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
[24] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with Super Resolution for In-the-Wild Facial Expression Recognition," *IEEE Access,* vol. 8, pp. 131988-132001, 2020.
[25] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," *Advances in neural information processing systems,* vol. 28, pp. 2017-2025, 2015.
[26] J. Shi and S. Zhu, "Learning to Amend Facial Expression Representation via De-albino and Affinity," *arXiv preprint arXiv:2103.10189,* 2021.
[27] T. A. Dingus *et al.*, "The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment," United States. Department of Transportation.

National Highway Traffic Safety …, 2006.
[28] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters,* vol. 23, no. 10, pp. 1499-1503, 2016.
[29] M. Nakai, Y. Tsunoda, H. Hayashi, and H. Murakoshi, "Prediction of basketball free throw shooting by openpose," in *JSAI International Symposium on Artificial Intelligence*, 2018: Springer, pp. 435-446.
[30] F. M. Noori, B. Wallace, M. Z. Uddin, and J. Torresen, "A robust human activity recognition approach using openpose, motion features, and deep recurrent neural network," in *Scandinavian conference on image analysis*, 2019: Springer, pp. 299-310.
[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015: PMLR, pp. 448-456.
[32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research,* vol. 15, no. 1, pp. 1929-1958, 2014.
[33] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing,* vol. 28, no. 1, pp. 356-370, 2018.
[34] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception," *Emotion,* vol. 12, no. 5, p. 1161, 2012.