Live Emotional Resonance Application Based on Facial Expression Recognition Technology

基於面部情緒辨識技術的實況直播情緒共鳴應用

指導教授:戴碧如

組員: 李昶勳 B10915065 余修辰 B10915036 許鎮承 B10915059 楊登傑 B10915060

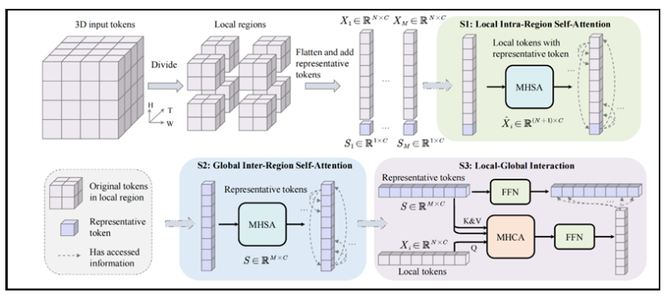
Introduction

本專題旨在讓觀眾之間能產生共鳴，藉此提高觀賞體驗，而實況主也能依據觀眾的反應去調整實況內容。為了能即時且有效地捕捉觀眾面部情緒，在方法上，我們基於動態面部情緒辨識的模型上進行改良，透過結合微表情識別模型，以找出性能表現最佳的模型。

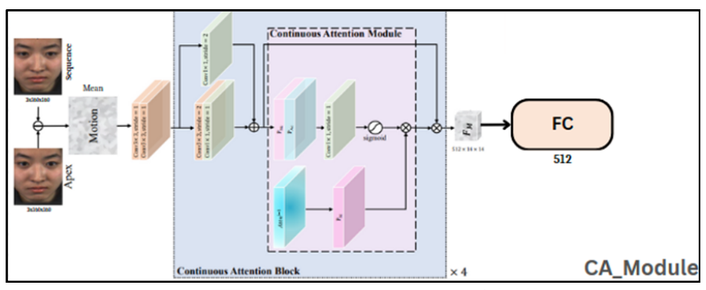
結果顯示模型能夠即時有效地偵測出觀眾的面部情緒，接著我們透過Google Sheets API將所有觀眾的情緒資訊進行彙整，並對這些資訊進行處理，來讓結果變的淺顯易懂，最終讓觀眾可以實時地看到其他人的情緒反應，增加觀賞體驗。

Model Architecture and Improvement

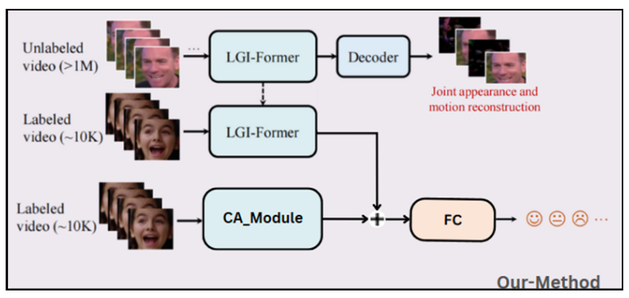
本次研究中我們在一開始採用Former-DFER[1]，但經過實驗後，發現在Disgust和Fear上表現特別差，其原因在於這兩個情緒在資料集中資料量十分稀少。考量到supervised(監督式學習)仰賴資料量的問題，最後我們將模型改成採用self-supervised(自監督式學習)的模型MAE-DFER[2]，該模型基於VideoMAE[3]進行改良，將ViT更改為LGI-Former，大幅減少運算成本。透過大量無標籤的資料進行預訓練，該模型在FERV39K[4]的準確率為52.07%，為目前表現最為優異的模型。在實驗途中，我們發現到模型會因為情緒強度太小，而誤判其結果，因此我們在之後也對其做出相對應的改良。



▲圖一 LGI-Former架構



▲圖二 CA_Module結構



▲圖三 改良後模型結構

為了改善上述所提到，模型會因為情緒強度太小而誤判其結果的問題。在我們觀察到微表情辨識模型能夠較有效地偵測出面部微小變化後，我們決定將微表情辨識模型和動態情緒辨識模型進行結合。在本次研究中我們所選用的模型為MMNET[5]，在模型結合上，為了使MMNET能與MAE-DFER進行結合，我們將MMNET的CA_Module獨立出來，希望能藉由CA_Module萃取肌肉變化特徵的能力，改善MAE-DFER辨認微小表情能力較差的問題，並在原本的Module後加上全連階層使其能與LGI-Former萃取的特徵進行結合。接著，在特徵結合的方式上，我們嘗試了兩種方式進行結合，分別是concatenate和add，最後再將結果丟入Fully connected layer後取得最終結果。

Experimental Results

在實驗上，我們選擇FERV39K[2]所提供的訓練和測試資料集進行驗證，該資料集具有38935部影片，共七個情緒分別是：Happiness、Anger、Neutral、Sadness、Surprise、Disgust、Fear。

從結果可以看到在MAE-DFER[2]加上MMNET[5]中的CA_Module後，雖然造成運算量些微提升，但在UAR與WAR上我們的模型分別提升了0.21%與0.58%，更為接近SOTA模型。

Method	#Params(M)	FLOPs(G)	UAR	WAR
Supervised methods				
C3D[11]	78	39	22.68	31.69
3D ResNet-18[12]	33	8	26.67	37.57
Former-DFER[1]	18	9	37.20	46.85
IAL[2]	19	10	35.82	48.54
POSTER-V2[7]	58	26	40.89	49.79
Self-supervised methods				
VideoMAE[9]	86	81	**43.33**	52.39
MAE-DFER[3]	85	50	42.09	51.82
MAE-DFER+MMNET(Ours)	114	52	42.30	**52.40**

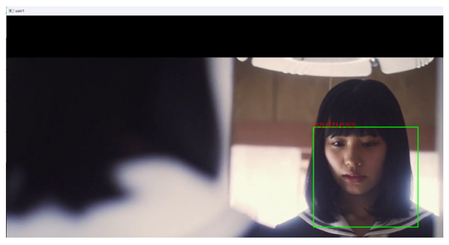
▲表一 實驗數據

表二為我們使用兩種方式將模型萃取出的特徵進行結合後的結果可以看到直接相加會比串接表現稍佳。

Method	UAR	WAR
add	**42.30%**	**52.40%**
concatenate	42.20%	52.30%

▲表二 結合方式比較

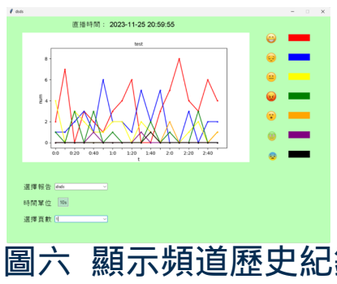
User Interface



▲圖四 情緒辨識結果

▲圖五 同步顯示所有觀眾情緒

在使用者介面的部分，我們讓使用者在看到結果的同時，可以看到其他使用者的情緒反應，而右方會呈現當下各個情緒出現的次數，讓觀眾可以快速了解當前所有觀眾的情緒，產生共鳴增加觀賞體驗。



▲圖六 顯示頻道歷史紀錄

在實況直播結束後，我們利用折線圖的方式去呈現各個時間觀眾情緒的變化，目的是希望可以讓實況主在結束直播後，可以透過這個圖去了解觀眾對於實況內容的情緒反應，藉此改善自己的實況內容。

Conclusion

在目前的結果我們能在有一定的準確度下即時地將觀眾的臉部情緒預測出來，將結果顯示在介面上。讓觀眾可以在觀賞直播的同時，也能知道其他觀眾的反應，產生共鳴進而提升觀賞體驗。另外，目前我們也完成了歷史紀錄的功能，透過折線圖呈現各個時間觀眾的情緒反應，讓實況主在結束直播後，能夠透過這個功能去了解自己實況內容帶給觀眾的反應，藉此去改善實況內容。

在未來我們會希望能夠持續優化模型，來讓模型在遇到情緒強度低的表情時能夠表現的更好。並優化使用者與直播者介面，讓觀眾和主播的使用體驗上升。

Reference:

[1] Zhao, Zengqun, and Qingshan Liu. "Former-dfer: Dynamic facial expression recognition transformer." Proceedings of the 29th ACM International Conference on Multimedia. 2021.

[2] Sun, L., Lian, Z., Liu, B., & Tao, J., "MAE-DFER: Efficient Masked Autoencoder for Self-Supervised Dynamic Facial Expression Recognition," in Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 6110–6121.

[3] Tong, Zhan, et al. "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training." Advances in neural information processing systems 35 (2022): 10078-10093.

[4] Wang, Yan, et al. "Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[5] Li, H., Sui, M., Zhu, Z., & Zhao, F., "MMNet: Muscle motion-guided network for micro-expression recognition," in Proceedings of the 31th International Joint Conference on Artificial Intelligence, 2022.