

Enhancing Emotion Recognition in Video Characters Through Multi-Modal Deep Learning Approaches

Mingyuan Yang *

School of Information Engineering, Huzhou University, Huzhou, Zhejiang, 313000, China

* Corresponding Author Email: 2022082105@stu.zjhu.edu.cn

Abstract. Human emotion recognition is a classic problem for Convolutional Neural Networks (CNNs) in classification tasks, and a plethora of research has emerged in recent years. This paper presents a video character emotion analysis model utilizing the The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset, exploring the integration of audio-visual data for emotion recognition. CNNs were employed to analyze image and audio data separately, then the models were fused using linear summation and greedy selection methods. The multi-modal approach significantly enhances performance over single-modal models, with the linear summation method showing the best results at a 0.4 weight ratio. The model's predictions are visualized to demonstrate the composite emotional analysis of video characters. The findings indicate that multi-modal models can improve accuracy, precision, recall, and F1-score by 30-40 percent compared to single-modal models. Future work will focus on optimizing model performance and leveraging confusion matrices to further refine the greedy selection algorithm.

Keywords: Convolutional neural network; emotion recognition; multi-modal model.

1. Introduction

Emotion recognition is a classic deep learning problem with significant applications in real-world scenarios. Convolutional neural networks (CNNs), which play a key role in the development of deep learning, are often used to solve emotion recognition tasks due to their ability to observe image details and improve model recognition [1,2]. Additionally, most current research is based on single-network approaches, such as real-time expression recognition on individual frames and emotion recognition on the Short Time Fourier Transform (STFT) of audio [3]. These schemes choose to analyze the emotions expressed by characters from either the auditory or visual dimension alone [4]. However, a person's emotions are clearly reflected in both their voice and facial expressions, akin to using ears and eyes. Therefore, it is meaningful and significant to explore methods of combining sight and hearing models to understand how they work together to identify a character's emotions and whether this approach performs better than single-model analysis. Moreover, in the field of deep learning, an increasing number of models crafted with multi-modal combination methods are gradually being integrated into classification tasks and playing more significant roles. Thus, the experiments were conducted on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) video dataset, which includes both image and audio files, with the hope of constructing a deep learning model that uses multi-modal combination methods for emotional analysis of video characters who are telling the same sentences in different tones and displaying different facial expressions [5]. Its performance was compared with single-modal models by observing several metrics. By conducting this experiment, a whole new perspective on emotion recognition in deep learning for the classification tasks of video can be discovered.

2. Method

2.1. Dataset

This research was tested on RAVDESS dataset [5]. During model training, the well-known Facial Expression Recognition 2013 (FER2013) dataset was used to train the expression recognition model

[6]. In the experiment, audio files should be used for the audio model training, and video files should be used for the final testing.

The RAVDESS dataset contains a total of 7,356 files, including recordings of 24 professional actors (12 male and 12 female) speaking two lexically matched statements with a neutral North American accent. The emotions include calm, happy, sad, fearful, angry, surprised, and disgusted. It contains three sub-folders: a video folder, an audio folder, and a silent video folder. In addition, to make it easier to distinguish the files, the creators of RAVDESS numbered each file in the dataset with seven digits. The first digit represents the type of document, the second digit represents the voice, the third digit represents the emotion, the fourth digit represents the emotional intensity, the fifth digit represents the content of the speech, the sixth digit represents the document number, and the seventh digit represents the actor's number. By encoding these seven digits in series, a file is uniquely identified within this large dataset.

The FER2013 dataset is a widely used facial expression recognition dataset containing 35,887 facial images of 48x48 pixels for facial expression recognition tasks. When training the Image Model, the FER2013 dataset was used, employing transfer learning for emotion recognition in video characters. Training the Audio Model required selecting the first 20 out of 24 actors from the RAVDESS dataset, excluding the 'neutral' and 'calm' expressions, totaling 960 videos for learning, with the training set accounting for 80 percent. The videos of the last four actors were used as the final test set.

2.2. Data Preprocessing

The FER2013 dataset consists of black and white images of 48x48 pixels; therefore, when testing the RAVDESS dataset, the images are also pre-processed to the same format as FER2013.

For video images, this work extracts two features: Mel Frequency Cepstral Coefficients (MFCC) and twelve-tone equal temperament and chord color (Chroma) [7]. The MFCC method first performs a short-time Fourier transform on the audio, then conducts Mel-scale cepstral operations. The formula for MFCC is defined as the following one. In the formula, π represents the ratio of the circumference of a circle to its diameter. n denotes the current frequency point, k is the index of the Mel-Frequency Cepstral Coefficient (MFCC), and N is the number of points in the Fast Fourier Transform (FFT), which corresponds to the number of sample points in the audio signal.

$$MFCC(K) = \sum_{n=0}^{N-1} \log(s(n)) \cdot \cos\left(\frac{\pi \cdot (2n+1) \cdot k}{2N}\right) \quad (1)$$

Chroma utilizes the twelve-tone equal temperament to distribute the frequencies in the audio across different pitches, and it is typically used to analyze the musicality of audio files. Here, it is used to analyze the pitch variations in speech, as demonstrated in Fig. 1.

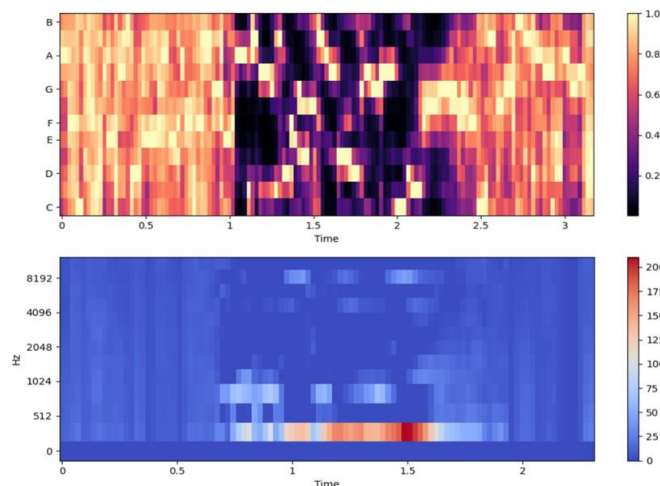


Fig. 1 Representative example of chroma and MFCC (Figure Credit: Original).

2.3. Face Detection

Due to the large size of the images in the RAVDESS dataset, the subjects only occupy about fifty percent of the image area. Therefore, a model for target recognition using transfer learning is constructed to identify the facial area of the subjects in the image frames and to eliminate irrelevant parts such as hair, ears, and jewelry. This allows the machine to focus on the key areas of the facial features and the distorted skin, as displayed in Fig. 2.

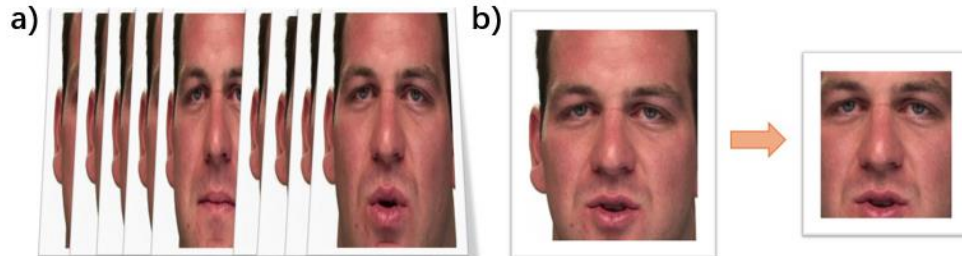


Fig. 2 Example of video frame extraction and face detection (Figure Credit: Original).

Assuming a frame rate of 20 frames per second for the RAVDESS dataset (where each video is 3 seconds long), this would result in 60 images. Obviously, not all of these images have high reference value, as some images may have blurred expressions, close eyes, or belong to exaggerated expressions. A pre-prediction elimination method was adopted, where the image is first classified to obtain the output vector, and then a threshold θ is set. If the variance is less than θ , it indicates that the reference value of the image is low, and it is therefore eliminated. The criterion for selecting this parameter is to ensure that no less than 50 percent of the images are discarded. The remaining images are then predicted one by one, and the arithmetic average of the results is taken. The steps are as follows: (1) Initialize an empty set Y for storing accumulated expression scores and the elements y_i within the set represent the cumulative probabilities of various expressions. (2) Initialize an empty set X for storing each image's expression scores and the elements x_{ij} within the set represent the possibility that the model considers the j -th image to be the i -th expression. (3) For each X , if $Var(X) < \theta$, discard X . Else for each element j in X , then $y_i \leftarrow y_i + x_{ij}$. (4) For each element j in Y : $y_j \leftarrow y_j / count(y_j)$.

2.4. Models

CNNs are primarily used for training on images and audio. CNNs acquire partial areas of images through the use of convolutional layers and reduce the feature matrix through the use of pooling layers. Due to their superior ability to recognize details and high accuracy, CNNs have become a common model for image and speech recognition. In this context, preprocessed images and audio were turned into image files, which are then classified through the model. Here, two typical CNN models were used: the Residual 18-layer Network (ResNet-18) and the Visual Geometry Group 19-layer model (VGG-19) [8].

The VGG-19 network has achieved very good results in the field of image recognition. It deepens the layers of VGG-16 [9]. In the process of training VGG networks, to prevent the gradient disappearance problem, the Scaled Exponential Linear Unit (SELU) function is used as the activation function in the convolutional layer. The SELU activation function is defined as:

$$SELU(x) = \lambda \begin{cases} x, & x > 0 \\ \alpha e^x - \alpha, & x \leq 0 \end{cases} \quad (2)$$

ResNet-18 has bottleneck layers that breaks the strict order in traditional convolutional networks so it helps to alleviate the over-fitting problem and also helps to solve gradient disappearance [10].

2.5. Model Combination Methods

After obtaining the outputs of the two models, it is necessary to perform dimensionality fusion first. The labeled FER2013 dataset outputs data for 7 types of expressions, while the labeled RAVDESS dataset outputs labeled data for 8 types of expressions. Different output dimensions will cause the input of the subsequent fusion model to not match; therefore, for the 'neutral' expression in RAVDESS, it is simply removed as its reference value is not high. Apart from this, the expressions in the two datasets are completely the same. Next, some model fusion schemes will be introduced.

The linear summation method is the most basic type of fusion method. It treats the output of each model as a vector and performs linear addition with the output vector of another model (with the same dimensions), and multiplies by weights to simulate the degree of preference for a certain model. This method simulates the method of mutual reference in the process of working together between the human ear and the human eye; if a person tends to believe their eyes (or ears) more, then the weight they receive will be greater. Here, alpha and beta represent weights, respectively, two vectors represent the analysis results of two models for an expression, which are probability values less than 1.0.

$$y = \alpha v_1 + \beta v_2 \quad s.t. \quad \alpha + \beta = 1 \quad (3)$$

Greedy Selection is a combination method that uses simple judgment statements to perform greedy algorithm evaluations. It can be observed that each model has different prediction abilities for different expressions, and most of the time, the opinions of the two models do not align. Therefore, when the models' opinions are not unified, select the result of the model that has higher recognition precision for the expression it is confident in, and completely discard the result of the other model's output.

2.6. Metrics

The experiments used a total of four evaluation metrics to assess the performance of the model's output: Accuracy, Precision, Recall, and F1-score. It is important to note that accuracy is for the entire sample set, while the latter three values are for individual samples (here specifically referring to a particular emotion); therefore, when assessing the model as a whole, the latter three indicators are averaged arithmetically.

3. Experiments and Results

3.1. Experimental Details

In object detection, transfer learning is utilized by loading pre-trained facial recognition models. Additionally, two model architectures are trained with a selection of epochs set to 10, a batch size of 64, and a learning rate of 0.01. The optimization methods used are Stochastic Gradient Descent (SGD) for VGG-19 and Adam for ResNet-18. After comparing the recognition accuracy, ResNet-18 is chosen as the model for facial expression recognition.

In the model trained to recognize the audio dataset, a simpler structure is used. By extracting features and training the MLP-Classification model, the recognition accuracy fluctuates greatly during the training process. Therefore, the hyperparameters are set as follows: epochs of 100, a learning rate of 0.01, a batch size of 32, and the optimizer is Adam.

3.2. Performance Comparison

First, an appropriate model was selected for the image model from two candidate models: VGG-19 and ResNet-18, which have different structures. It could be observed that as the training epochs progress, both the training and validation loss of the models are decreasing, as demonstrated in Fig. 3 and Fig. 4. However, a significant overfitting problem has arisen, where the training performance is much better than the validation performance, and this phenomenon occurs in both models. The

final validation accuracy has reached 0.6 and has stabilized, so the model can be considered quite good, as displayed in Table 1. Here, a comprehensive comparison of the performance of the two models was made based on several assessment values mentioned in the previous text and found that the latter edged out the former in terms of overall test set accuracy, precision of positive samples, and false negative rate. Therefore, ResNet was chosen as the image model.

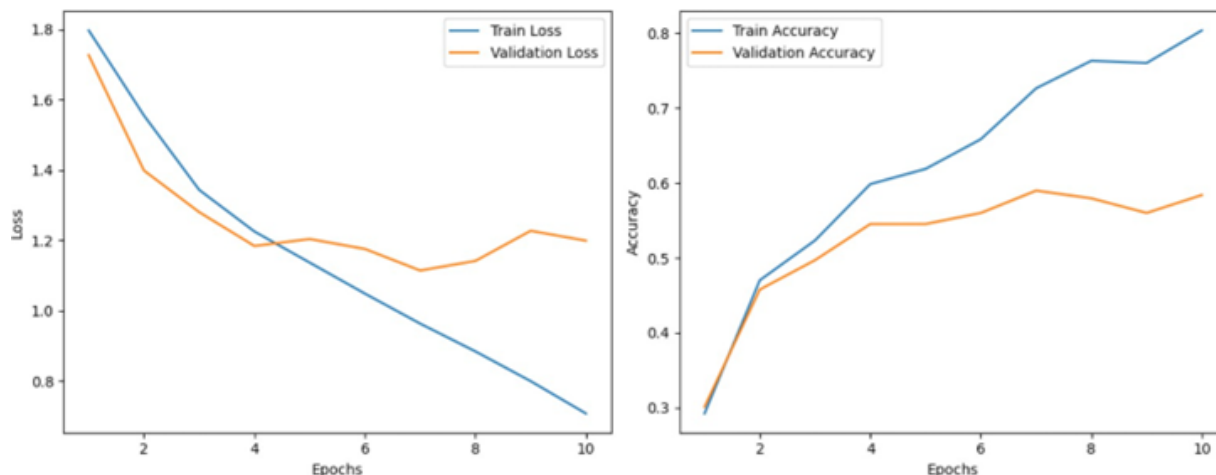


Fig. 3 Training curve of VGG-19 (Figure Credit: Original).

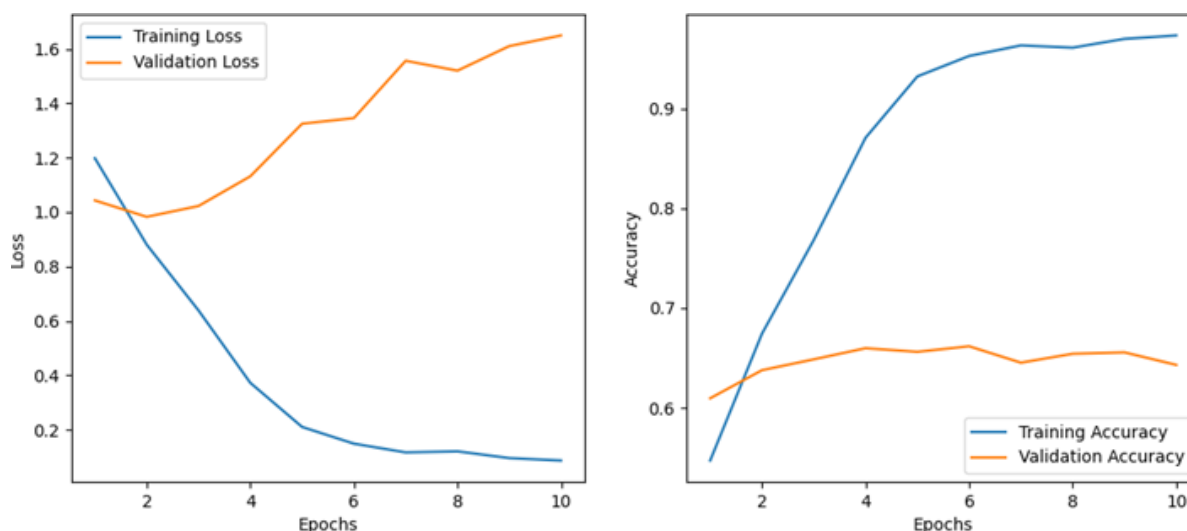


Fig. 4 Training curve of ResNet-18 (Figure Credit: Original).

Table 1. Model performance of image model on FER-2013

Name	Test Accuracy	Precision	Recall	F1 Score
VGG-19	0.6236	0.6135	0.6236	0.6185
ResNet-18	0.6461	0.6461	0.6461	0.6521

When training the audio model, a more lightweight model was utilized, resulting in larger fluctuations in the model's accuracy curve as it increased, as shown in Fig. 5.

Next, the selected model was going to be integrated into the framework, and it was planned to predict all videos within the RAVDESS dataset according to different combination methods, comparing the levels of assessment values, as displayed in Table 2.

Table 2. Model performance of audio model on RAVDESS.

Name	Test Accuracy	Precision	Recall	F1 Score
MLP-Classification	0.5463	0.5771	0.5436	0.5601

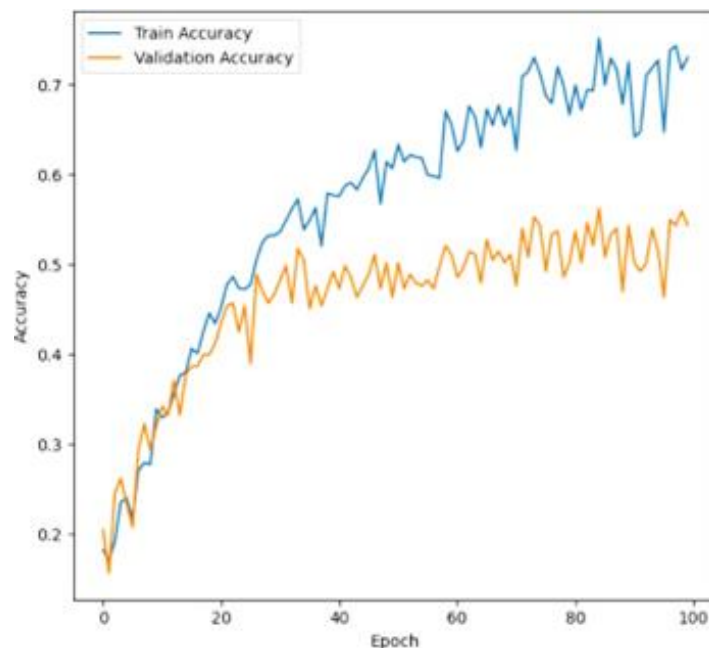


Fig. 5 Fluctuation in training and validation curve (Figure Credit: Original).

Using the linear summation method, this work iterated through different weight values, selecting the largest as the ideal value for this method, for comparison with other methods. The performance of each weight is shown in Table 3. The closer the ratio between the two models is to 1:1, the better the final performance of the model.

Table 3. Performance comparison using different weights.

Weight(audio / 1.00)	Test Accuracy	Precision	Recall	F1 Score
0.00	0.3085	0.3469	0.3085	0.2687
0.10	0.4219	0.4377	0.4219	0.3930
0.20	0.4531	0.4777	0.4531	0.4350
0.30	0.5104	0.5096	0.5104	0.4956
0.40	0.5104	0.5078	0.5104	0.4992
0.50	0.4740	0.4806	0.4740	0.4687
0.60	0.4323	0.4404	0.4323	0.4245
0.70	0.3802	0.3787	0.3802	0.3648
0.80	0.3490	0.3352	0.3490	0.3250
0.90	0.3490	0.3356	0.3490	0.3246
1.00	0.3490	0.3356	0.3490	0.3246

The Greedy Selection method requires the preliminary determination of the individual models' recognition capabilities for different expressions. The results can be seen in Fig. 6. It can be found that the two models each have different strengths in recognizing expressions, as displayed in Table 4 and Fig. 7. The Image Model performs better in recognizing Disgusted, Fearful, and Happy expressions, especially Happy, while the Audio Model is better at recognizing Angry, Sad, and Surprised expressions. However, the average values of Precision, Recall, and F1-score for both models are not significantly different, which can be seen from the above Table 5 (weights 0.00 and 1.00).



Fig. 6 Precision heatmap for image and audio models (Figure Credit: Original).

Table 4. Performance comparison of two methods

Method	Test Accuracy	Precision	Recall	F1 Score
Linear Summation(weight 0.4)	0.5463	0.5771	0.5436	0.5601
Greedy Selection	0.4792	0.4277	0.4792	0.4615

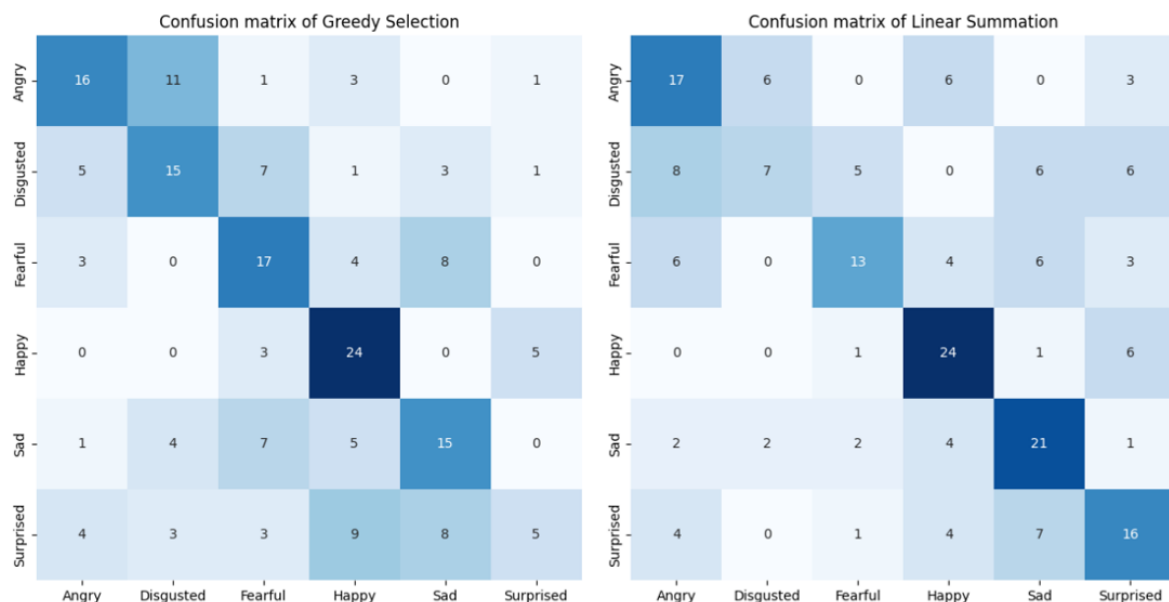


Fig. 7 Confusion matrix of image and audio models (Figure Credit: Original).

3.3. Performance Comparison with Single-Modal Recognition

Compared to single-modal recognition, the combination of audio-visual models has significantly improved the accuracy, precision, recall, and F1-score of recognition, as displayed in Table 5. It can be observed that the linear summation method outperforms the greedy selection algorithm in overall metrics; both models have their focuses: for instance, the greedy algorithm is good at recognizing 'Disgusted' but clearly fails at recognizing 'Surprised'; the latter is just the opposite.

Table 5. Performance comparison of different methods

Method	Test Accuracy	Precision	Recall	F1 Score
Linear Summation(weight 0.4)	0.5104	0.5078	0.5104	0.4992
Greedy Selection	0.4792	0.4277	0.4792	0.4615
Image Model	0.3085	0.3469	0.3085	0.2687
Audio Model	0.3490	0.3356	0.3490	0.3246

3.4. Visualizaiton Analysis

When visualizing the output of a model's predictions, it is necessary to use the linear summation method to display the possible values of each emotion that the model considers. Here, a weight of 0.4 is selected. It can be observed from Fig. 8 that, for example, in the first video below, the model believes that the emotion of the person is primarily happiness, while for the second video, the model is confident that she is sad. By displaying it in this way, it could be observed the composition of the model's final opinion.

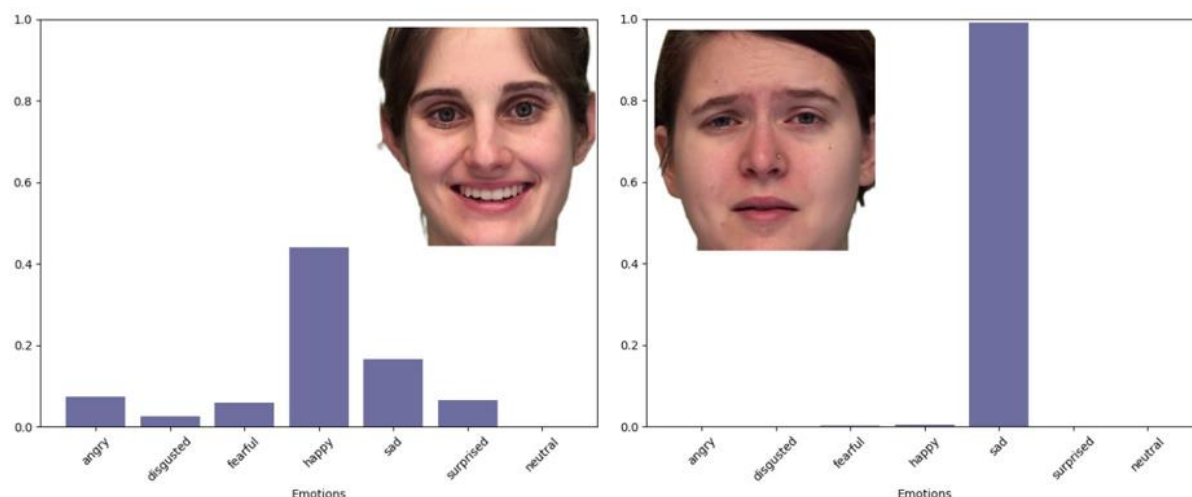


Fig. 8 Probability prediction of representative examples (Figure Credit: Original).

4. Discussion

Using multi-modal predictions indeed significantly enhances model performance compared to single-modal predictions. The two selected methods, linear summation and greedy selection, both greatly improved the model's final test set metrics of accuracy, precision, recall, and F1-score. The author can visualize the possibility sequences obtained through linear summation to observe the emotional composition of the video characters.

The Image Model and Audio Model themselves do not have high test accuracies on their respective datasets (slightly above 0.30); this makes the subsequent choice of modality fusion result in a significant improvement compared to the single model, but it also cannot reach 0.60. Therefore, in the future, the author can focus on selecting Image and Audio models with better performances to make the final result better. In addition, the greedy algorithm itself can also be further optimized. In fact, in the confusion matrix, the information between the combinations of expressions is ignored, and the model can be taught some special cases, telling it which two expressions are easily confused, rather than just comparing the size of the two precisions and making a judgment.

5. Conclusion

In this project, the performance comparison between multi-modal and single-modal approaches in predicting facial expressions of video characters was explored. First, single models for frame image and audio emotion recognition were trained, and then the two models were combined using linear summation and greedy algorithm methods. It was found that when using the linear summation method, the closer the ratio between the two models is to 1:1, the better the final performance of the model, and its graph shows a camelback shape. Finally, the final performance of each model was compared. It was found that the use of multi-modal models can bring improvements of 46.19%, 51.26%, and 53.78% in accuracy (recall), precision, and F1-score over the single-modal models, respectively, using both the linear summation method and the greedy selection algorithm. This indicates that using both visual and auditory models can indeed optimize the performance of emotion recognition for

video characters to a certain extent. In the future, the focus will be on improving the performance of the models, and in addition, greater utilization of the confusion matrix will be made to better optimize the greedy algorithm, resulting in better model performances.

References

- [1] Li Zewen, Liu Fan, Yang Wenjie, et al. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 2021, 33(12): 6999-7019.
- [2] Gu Jiuxiang, Wang Zhenhua, Kuen Jason, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 2018, 77: 354-377.
- [3] Durak Lutfiye, Arikan, Orhan. Short-time Fourier transform: two fundamental properties and an optimal implementation. *IEEE Transactions on Signal Processing*, 2003, 51(5): 1231-1242.
- [4] Dzedzickis, Andrius and Kaklauskas, Arturas and Bucinskas, Vytautas. Human emotion recognition: Review of sensors and methods. *Sensors*, 2020, 20(3): 592.
- [5] Livingstone Steven R, Russo Frank A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, 2018, 13(5): e0196391.
- [6] Goodfellow Ian J, Erhan Dumitru, Carrier Pierre Luc, et al. Challenges in representation learning: A report on three machine learning contests. *The International Conference on Neural Information Processing*, 2013: 117-124.
- [7] Garg Utkarsh, Agarwal Sachin, Gupta Shubham, et al. Prediction of emotions from the audio speech signals using MFCC, MEL and Chroma. *International conference on computational intelligence and communication networks*, 2020: 87-91.
- [8] Zhou Yitao, Ren Fuji, Nishide Shun, et al. Facial sentiment classification based on resnet-18 model. *2019 International Conference on electronic engineering and informatics*, 2019: 463-466.
- [9] Simonyan Karen, Zisserman Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, 2014:1409.1556.
- [10] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016: 770-778.