MULTIMODAL EMOTION RECOGNITION USING SPEECH AND FACIAL INFORMATION BASED ON LATE FUSION

JIN-JING JIANG[1], ERI SATO-SHIMOKAWARA[1]

[1]Graduate School of Systems Design, Computer Science, Tokyo Metropolitan University
6-6, Asahigaoka, Hino, Tokyo 191-0065, Japan
E-MAIL: jiang-jinjing@ed.tmu.ac.jp, eri@tmu.ac.jp

Abstract:

Emotion recognition plays a key role in human-computer interaction(HCI) and intelligent systems. This study proposes a multimodal approach that combines facial expressions and speech information to improve classification accuracy. VGG16-based CNN and 1D-CNN are used for facial and speech recognition, respectively. A weighted late fusion strategy integrates both outputs to make the final prediction. Experiments using KDEF and SAVEE datasets demonstrate that the proposed method outperforms unimodal models, particularly for neutral, fear, and disgust emotions, while reducing gender bias. The results highlight the effectiveness of multimodal fusion in enhancing emotion recognition.

Keywords:

Facial expression; Speech expression; Emotion recognition; Late fusion; Multimodal

1 Introduction

Emotion recognition is a key component of human-computer interaction (HCI), aiming to automatically classify human emotions using machine learning techniques.

While language-based communication has been extensively studied, understanding emotional signals such as facial expressions and speech information, remains a significant challenge. Recent advances in signal processing and machine learning have enabled multimodal emotion recognition systems that combine visual and auditory cues to improve classification accuracy [?].

Since the recognition accuracy of conventional single-modal approaches has reached a bottleneck in the increasingly advanced field of HCI, multimodal emotion recognition has garnered significant attention in recent years. By integrating both explicit and implicit information, the accuracy of emotion recognition can be effectively enhanced. Based on this idea, the present study employs multimodal recognition techniques that combine facial expression and speech information,representing explicit and implicit modalities, respectively to improve emotion classification performance and enable the system to handle more complex real-world scenarios.

2 Related Work

Most existing emotion recognition studies have focused on single-modal approaches, such as facial expressions or speech information. Facial expression recognition is essential in non-verbal communication and is widely applied in social robotics and neuromarketing. However, due to the subtle and complex nature of expressions, achieving high accuracy remains challenging. Convolutional Neural Networks (CNNs) are commonly used for this task, as they automatically learn discriminative features and classify expressions into basic categories [?].

Speech emotion recognition identifies emotions from acoustic features that reflect changes in vocal organ movements. CNNs are also effective in this context,as they capture emotional cues from spectrograms or MFCCs through local time-frequency pattern learning and dimensionality reduction [?].

Since facial expressions and speech information often occur simultaneously, multimodal approaches have emerged to improve recognition performance. By combining visual and acoustic signals, these methods offer richer emotional information and greater robustness than unimodal systems. This study adopts a late fusion strategy, where facial and speech modalities are processed independently and integrated at the decision level using weighted probability fusion [?].

3 Datasets and Preprocessing

In order to realize multi-modal emotion recognition, datasets are selected for different models to maximize the learning effect and generalization ability of the models.

3.1 Dataset Selection

To ensure effective multimodal emotion recognition, this study strategically selected different datasets tailored to each modality, maximizing the learning potential and generalization ability of the models. For facial expression recognition, the Karolinska Directed Emotional Faces (KDEF) dataset was employed. This dataset consists of facial images captured under strictly controlled conditions, featuring uniform lighting and seven distinct expressions recorded from five different angles. The diversity in angles and the natural appearance of expressions contribute to improving the robustness of the model, reducing overfitting, and enhancing generalization when applied to real-world scenarios [?].

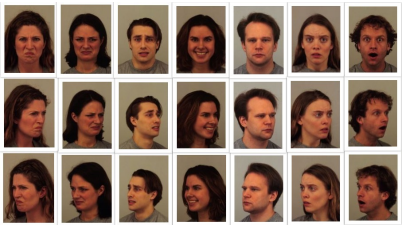

FIGURE 1. KDEF Datasets [?]

In the speech recognition task, the Surrey Audio-Visual Expressed Emotion (SAVEE) dataset was utilized. SAVEE comprises 480 utterances recorded by four male actors speaking in British English, each delivering sentences associated with seven emotional states. Recorded with high-quality equipment under controlled conditions, this dataset offers clear and noise-minimized audio samples. Moreover, the use of phoneme-balanced sentences derived from the TIMIT corpus ensures consistency across different emotions, making SAVEE a reliable and standardized resource for training models focused on speech-based emotion recognition [?].

To evaluate the performance of the proposed multimodal system in a more realistic and challenging context, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was adopted as the testing dataset. RAVDESS is a validated, gender-balanced multimodal dataset comprising recordings from 24 actors, each performing eight emotions at two levels of intensity. Its inclusion of both speech information and facial expression data allows for comprehensive testing of multimodal models. Furthermore, its emotional diversity and intensity variation enable rigorous assessment of the model's robustness across a wide range of expressive scenarios [?].

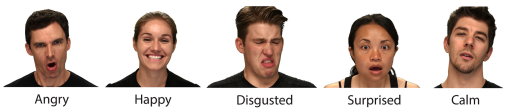

FIGURE 2. RAVDESS Datasets [?]

3.2 Preprocessing

In this research, image pre-processing began with converting all images to RGB format to ensure input consistency. Then, images were resized to 224×224 pixels using letterbox resize, which maintains the original aspect ratio and adds padding to avoid distortion. Pixel values were normalized from a 0–255 range to 0–1 to enhance training stability and accelerate convergence. Finally, a batch dimension was added to fit the input format required by deep learning models. These pre-processing techniques collectively enhance the consistency and quality of the input data, thereby optimizing the performance and accuracy of the CNN-based model in the facial emotion recognition task.

For the speech modality,Librosa was used to preprocess audio data in four steps: loading, feature extraction, normalization, and data splitting. After loading the audio files, zero-padding was applied to ensure consistent length. Key acoustic features were extracted, including MFCCs, pitch, Mel spectrogram, spectral centroid, spectral flatness, zero-crossing rate, and spectral contrast. These features were then standardized using a Standard scaler to ensure uniform scaling. Finally, the data was divided into training and testing sets for model evaluation. These steps improved input consistency and enhanced the model's accuracy in speech emotion recognition.

4 Methodology

4.1 Facial Expression Recognition Model

Model Selection: This study employs a convolutional neural network (CNN) architecture based on VGG16,

which was adopted for the facial expression recognition task. After comparing multiple models, VGG16 was chosen over MobileNet for the following reasons:

The VGG16 model achieved lower loss and higher accuracy compared to MobileNet. Although both models exhibited stable training according to their loss curves, VGG16 demonstrated a more favorable overall performance (3).

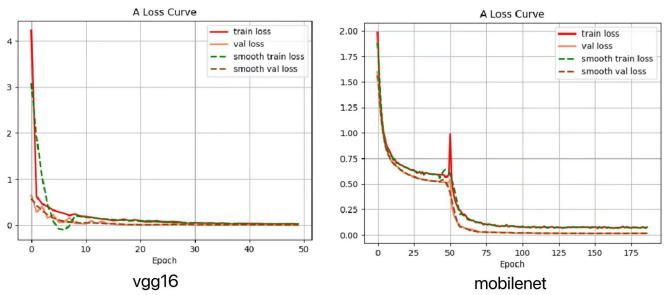

FIGURE 3. Loss-comparison; VGG16 (left) and MobileNet (right)

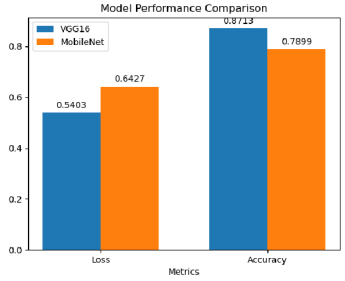

FIGURE 4. Comparison of Loss(left) and Accuracy(right)

Transfer Learning: To accelerate model convergence and improve classification accuracy, pre-trained weights from ImageNet were introduced into the model. During the initial training phase, the convolutional layers were frozen and only the fully connected layers were trained. Subsequently, the convolutional layers were gradually unfrozen for fine-tuning. Hyperparameter Settings: The input image size is 224×224 pixels with three channels (RGB). The Adam optimizer is used with a learning rate of 0.0001. The loss function is Categorical cross-entropy to calculate the error in the multi-class classification task.

4.2 Speech Recognition Model

In the phonetic emotion recognition task, this study conducted a comparative analysis between Long Short-Term Memory (LSTM) networks and one-dimensional Convolutional Neural Networks (1D-CNN). After evaluating their performance, the 1D-CNN architecture was ultimately selected due to several notable advantages. One key reason was its superior loss convergence. The 1D-CNN model demonstrated a smoother decline in loss during training and testing, resulting in consistently lower final loss values. In contrast, the LSTM model showed a rapid decrease in training loss; however, its test loss plateaued or even increased slightly, suggesting a tendency toward overfitting (Fig. 5).

Another important factor in the decision was generalization ability. The 1D-CNN exhibited only a small gap between training and test losses, indicating stable performance across different data samples. Conversely, the LSTM model displayed a larger discrepancy between training and testing results, implying reduced generalizability and reliability when dealing with unseen data.

Additionally, model complexity and training efficiency were also taken into account. The 1D-CNN model features a simpler architecture and offers higher computational efficiency, which is especially beneficial when processing input features such as Mel spectrograms and MFCCs. Owing to its strength in capturing localized temporal-frequency patterns, CNN proved to be more effective in this task than the sequential modeling capabilities of LSTM.

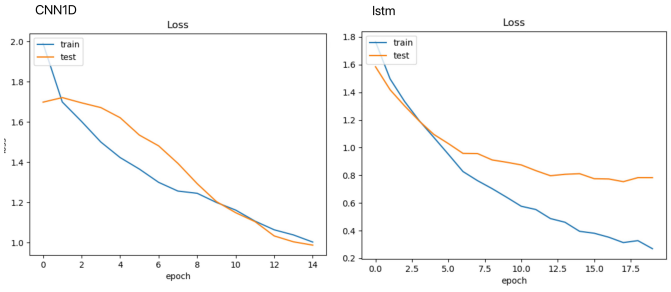

FIGURE 5. Comparison of 1D-CNN (left) and LSTM (right)

As for the model's hyperparameters, the 1D-CNN was configured with 32 filters, using kernel sizes of 5 and 7 to extract relevant features. A dropout rate of 0.5 was

applied to mitigate overfitting, and the Adam optimizer was employed with a learning rate of 0.001 to ensure stable and efficient convergence during training. These design choices contributed to the model's strong performance in recognizing emotions from speech information.

4.3 Late Fusion Method

In this research, the Late Fusion Method was employed to combine the output results of the independently trained VGG16 facial expression recognition model and the 1D-CNN speech emotion recognition model, thereby enhancing the overall accuracy of the emotion recognition. The specific methodology is as follows:

To begin with, both the VGG16-based facial model and the 1D-CNN speech model independently process input data and generate probability distributions across emotion categories using a Softmax activation function. The output of the facial expression model can be represented as

$$P_v = [p_{v1}, p_{v2}, ..., p_{v8}]$$

and the speech model output be

$$P_a = [p_{a1}, p_{a2}, ..., p_{a8}]$$

where 8 denotes the number of emotion classes.

Following this, a weighted probability fusion technique is applied to combine the outputs from the two models. The speech model is assigned a weight of 0.7 and the facial expression model a weight of 0.3. These weights were selected based on their respective performance in unimodal classification tasks. The final probability distribution is then calculated using a weighted average, expressed as

$$P_f = w_a P_a + w_v P_v$$

Where $w_a = 0.7$ (the weight for the speech model) and $w_v = 0.3$ (the weight for the facial expression model).

Based on the fused probabilities, the emotion category associated with the highest probability value in P_f is selected as the final predicted result. In other words, the classification decision is made by computing

$$\text{Predicted-Emotion} = \arg\max(P_f)$$

Finally, the effectiveness of this fusion approach was validated by comparing the results of the unimodal models with the fused model. The Late Fusion method led to notable improvements in recognition accuracy, especially for emotions that are typically difficult to classify, such as Neutral, Fear, and Disgust. These findings affirm that integrating multiple modalities through a weighted fusion strategy can significantly enhance the reliability and robustness of emotion recognition systems.

5 Experiments and Results

One of the authors must submit a signed copyright form to the Publications Chair before the final version of the paper can be accepted for publication. The copyright form is available on the conference website.

5.1 Experimental Setup

This study uses KDEF, SAVEE, and RAVDESS datasets for facial and speech emotion recognition. A VGG16-based model is used for facial expressions, and a 1D-CNN for speech. To evaluate multimodal fusion, both models were trained separately, and their predictions were combined using weighted probability fusion (0.7 for speech, 0.3 for facial), based on unimodal performance.

The training process utilized the Adam optimizer with a learning rate of 0.001 and a batch size of 32. All models were trained on the same dataset to ensure experimental fairness and consistency. The experiment was designed to comprehensively evaluate the effectiveness of the proposed system and consisted of three main components. First, the performance of the unimodal models—specifically, the speech recognition and facial expression recognition models—was assessed individually to understand their standalone capabilities. Second, the performance of the multimodal fusion model, which integrates both modalities through the Late Fusion strategy, was evaluated to examine its overall classification accuracy and robustness. Lastly, a gender-based analysis was conducted to investigate how recognition performance varied between male and female subjects, thereby revealing any potential biases and the impact of fusion on mitigating them.

5.2 Ablation Study

Contribution Analysis of Unimodal Models: Evaluated the performance of the speech information and facial expression models independently. The speech model performed well on emotions such as angry, disgust, and fear, while the facial model achieved the highest accuracy

(95.83%) on the happy category. By comparing accuracy before and after fusion, the effectiveness of the multimodal fusion strategy was validated. Model Performance on Different Emotion Categories: Analyzed the confusion matrix to explore model performance on various emotion categories. It was found that the facial model had a higher misclassification rate when handling similar expressions, such as angry and disgust. However this issue was significantly reduced with the aid of the speech model.

5.3 Results Analysis

Comparison of Unimodal and Multimodal Models: The speech model demonstrated stable performance in most emotions, particularly in angry (91.67%) and fear (86.11%) categories. However, its performance on sad emotions was relatively low at 54.17%. The facial model achieved the highest accuracy in the happy category (95.83%), but struggled significantly with angry (10.35%) and fear (19.44%) categories. Through multimodal fusion, the accuracy for the happy category improved to 100%, and the accuracy for the disgust category accuracy increased from 84.38% (using the speech model) to 90.63%, thereby enhancing model stability. Gender Difference Analysis: In the gender-based experiments, male subjects demonstrated higher recognition accuracy in angry, disgust, and happy categories, while female subjects performed better in fear and sad categories. The speech model demonstrated stable performance in male groups , while in female groups, the happy category achieved 100% recognition rate. The facial model generally underperformed in female groups. However, multimodal fusion significantly improved accuracy in sad (75.00%) and neutral (56.25%) categories.

TABLE 1. Emotion classification accuracy

Emotion	Speech	Facial expression	Fusion
Neutral	61.11%	22.22%	58.33%
Disgust	84.38%	57.81%	90.63%
Fear	86.11%	19.44%	86.11%
Happy	86.11%	95.83%	100.00%
Sad	54.17%	15.28%	58.63%
Angry	91.67%	10.35%	91.67%

Effectiveness of Fusion Method: The Late Fusion method not only improved the overall model performance but also effectively mitigated the overfitting problem common in unimodal models. By employing different weighting strategies (Weighted Fusion), the model performance was fine-tuned for specific emotion categories, resulting in superior recognition of emotions such as angry, fear, and disgust compared to unimodal models.

Conclusion: The multimodal fusion model outperformed unimodal models in all experimental scenarios, demonstrating the feasibility and effectiveness of combining facial and speech information in emotion recognition tasks. Particularly in complex and easily confused emotions such as neutral, fear, and disgust, the multimodal model showed higher robustness and accuracy.

6 Discussion

6.1 Contributions

This study proposes a multimodal emotion recognition method using Late Fusion, integrating speech recognition (1D-CNN) and facial expression recognition (VGG16) to improve emotion classification accuracy. The main contributions of this research are as follows:

Validation of the effectiveness of multimodal fusion: The experiments demonstrate the complementary nature of speech information and facial expressions, significantly improving recognition accuracy for emotions that are difficult to classify using a single modality, such as neutral, disgust, and fear.

Analysis of gender differences: The study finds that multimodal fusion reduces the impact of gender differences on emotion recognition, leading to more balanced performance across male and female groups. Notably, recognition accuracy for sad and neutral emotions improved significantly.

Comparison of unimodal and multimodal models: The experimental results indicate that multimodal fusion outperforms unimodal models (speech-only or facial expression-only) in terms of overall performance, generalization ability, and stability.

6.2 Limitations

Despite its achievements, this study has several limitations: Limited dataset size: The datasets used (KDEF, SAVEE, RAVDESS) are relatively small, particularly in categories such as disgust and fear, which may limit model generalization. Limited Modalities: This study only utilizes speech information and facial expressions, omitting potentially valuable modalities like text or physiological

TABLE 2. Gender Difference

Accuracy Rate	Male(Speech)	Male(Expression)	Male(After fusion)	Female(Speech)	Female(Expression)	Female(After Fusion)
neutral	65.00%	35.00%	60.00%	56.25%	6.25%	56.25%
disgust	84.38%	75.00%	96.88%	84.38%	40.63%	84.38%
fear	80.00%	32.50%	80.00%	93.75%	3.13%	93.75%
happy	77.50%	95.00%	100.00%	96.88%	96.88%	100.00%
sad	57.50%	12.50%	27.50%	50.00%	18.75%	75.00%
angry	87.50%	10.35%	87.50%	96.88%	10.35%	96.88%

signals, which may limit performance in complex scenarios. Subjective weight selection in the fusion process: Fusion weights ($w_a = 0.7$, $w_v = 0.3$) were manually set. Future work should explore adaptive weighting methods to improve flexibility and accuracy.

6.3 Future Work

To further improve the practicality and performance of this model, future research can focus on the following areas: Real-time detection: Transition from offline analysis to real-time emotion recognition for use in applications such as brilliant customer service and mental health monitoring. Adaptive Weighting: Replace manually set fusion weights with adaptive strategies to enhance flexibility and robustness across different contexts. Dataset Expansions: Enhance data diversity and scale by incorporating new data collection or augmentation to improve generalization. Multimodal Integration: Incorporate additional modalities, such as text and physiological signals to further boost recognition accuracy.

7 Conclusion

This study proposes a multimodal emotion recognition method that integrates speech information and facial expressions using the late Fusion strategy, in which weighted probability fusion is applied to determine the final emotion classification results. The experimental findings demonstrate the advantages of this approach in several key areas.

Multimodal fusion proved to be highly effective in enhancing emotion recognition accuracy, particularly for categories that are typically difficult to classify, such as neutral, fear, and disgust. The fusion of modalities leveraged the complementary strengths of both speech and facial data to improve performance.

In terms of individual modality contributions, the speech model showed strong recognition capability for angry and neutral emotions, while the facial expression model achieved the highest accuracy in the happy category, reaching 95.83%. When these two modalities were combined through fusion, the accuracy for the happy category was further improved to 100%, highlighting the synergy of combining different sources of emotional information.

Moreover, the multimodal fusion approach contributed to mitigating gender-based biases in emotion recognition. For example, recognition performance improved significantly in categories such as sad and neutral, with post-fusion accuracy reaching 75.00% and 56.25%, respectively. This indicates that fusion not only improves accuracy but also enhances fairness and balance across different demographic groups.

Although this study has certain limitations—such as the relatively small dataset size and the manual assignment of fusion weights—the results validate the practicality and effectiveness of multimodal fusion. Future research may explore more adaptive fusion strategies, expand the dataset size, and incorporate additional modalities such as text or physiological signals to further enhance the robustness of emotion recognition systems.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP23H00503.