ATTENTION-BASED FUSION FOR MULTIMODAL EMOTION RECOGNITION USING SPEECH AND FACIAL EXPRESSIONS

Shashikant Athawale
Associate Professor, Department of Computer Engineering, AISSMS College of Engineering, Savitribai Phule Pune University

Shreyas Yerole
AISSMS College of Engineering, Savitribai Phule Pune University, Pune, India

Rahul Wala
AISSMS College of Engineering, Savitribai Phule Pune University, Pune, India

Abstract: *Multimodal emotion recognition has become an essential area in human-computer interaction, enabling systems to understand and respond to human emotions with high accuracy. This paper presents a system that integrates audio (speech) and video (facial expression) data to detect human emotions. The system utilizes advanced feature extraction techniques, including Mel-frequency cepstral coefficients (MFCCs), spectral contrast, and chroma for speech, and CNN-based methods for facial expression analysis. An attention-based fusion mechanism is employed to dynamically weight the contributions from both modalities, improving the overall emotion prediction accuracy. The system is evaluated on two publicly available datasets, FER-2013 for facial expressions and RAVDESS for speech emotion recognition. Experimental results demonstrate that the multimodal system outperforms unimodal systems in recognizing emotions such as happiness, sadness, anger, fear, surprise, disgust, and neutral. The attention fusion mechanism significantly enhances the recognition accuracy, making the system more robust in real-world scenarios. This work contributes to the development of more intelligent, emotion-aware systems, with applications in affective computing, mental health monitoring, virtual assistants, and customer service systems.*

Keywords: **Multimodal emotion recognition, Speech emotion recognition, facial expression recognition, attention-based fusion, human-computer interaction, feature extraction, deep learning, convolutional neural networks, MFCC, affective computing, emotion detection, audio-visual fusion, real-time emotion recognition, FER-2013 dataset, RAVDESS dataset**

1. INTRODUCTION

The growing field of human-computer interaction (HCI) has seen significant advancements in emotion recognition technologies, which play a pivotal role in improving user experiences across various applications, such as virtual assistants, video conferencing, and healthcare. In recent years, researchers have increasingly focused on developing multimodal emotion recognition systems, which combine multiple sources of information, such as audio and video data, to achieve more accurate and robust emotion detection. This project aims to develop a multimodal emotion recognition system that utilizes both voice and facial expressions to detect human emotions. The proposed system integrates attention-based fusion of audio and visual features to enhance the emotional intelligence of the system, ensuring better recognition of emotions in real-world settings. By combining audio-based emotion recognition with facial expression analysis, the system can exploit complementary features from both modalities, making it more resilient to noisy environments or ambiguous emotional expressions. This system has the potential to be applied in a variety of domains, including affective computing, mental health monitoring, customer service, and human-robot interaction.

2. PURPOSE AND MOTIVATION

The primary purpose of this research is to develop a system capable of recognizing human emotions with high accuracy by integrating audio and video modalities. This project addresses the inherent limitations of unimodal systems, such as the inability to detect subtle emotions or the challenges posed by noisy backgrounds in speech. Emotion recognition through speech and facial expressions is important for building more empathetic systems, which can understand and respond to human emotions appropriately. For example, in customer service, understanding the emotions of users can help in providing better responses or detecting dissatisfaction early. In healthcare, recognizing emotional states can aid in monitoring mental health and providing personalized care.

The motivation behind this work stems from the need to improve emotion recognition systems by leveraging both visual and acoustic cues, as emotions are often expressed through both voice and facial expressions. The attention-based fusion mechanism further enhances the model's ability to focus on the most relevant features from each modality, thus optimizing performance.

3. SYSTEM ARCHITECTURE

3.1. Fusion Layer (Attention-Based Fusion)

3.1.1. Motivation Behind Fusion

The fusion layer serves as the heart of the multimodal system, combining the features extracted from both audio and video streams. The key idea behind fusion is that audio and video data provide complementary information about the user's emotions. Audio captures the tone and prosody of speech, while video captures the facial expressions that reveal emotions more directly.

3.1.2. Attention Mechanism

The attention mechanism dynamically assigns importance to the audio and video features based on the context of the data. This means that the system learns how much weight to assign to each modality during fusion, depending on the quality of the features and the environmental factors (e.g., noisy audio vs. clear facial expression).

- Weighted Fusion: The attention mechanism computes an attention weight α for the audio features, and (1 - α) for the video features. This weight is learned and adjusted during training.
- Normalization: Both the audio and video feature vectors are normalized to unit length using L2 normalization to ensure that neither modality dominates the fusion process due to differing scales of feature values.

3.1.3. Fusion Process

The fusion process is implemented as follows:

$$\textbf{Fused\_features} = \alpha \cdot \textbf{audio\_features} + (1 - \alpha) \cdot \textbf{video\_features}$$

This combined fused feature vector is then used for emotion recognition.

3.1.4. Benefits of Attention-Based Fusion

- **Context-aware fusion:** Adaptively balances the importance of audio and video.
- **Robustness:** Handles noise and disturbances in either modality by adjusting attention dynamically.
- **Improved accuracy:** By integrating complementary information, the fusion layer leads to more accurate emotion detection.

3.2. Feature Extraction Module/Layer

The feature extraction module consists of two separate submodules: one for extracting features from audio and another for extracting features from video.

3.2.1 Audio Feature Extraction

Audio features are extracted from raw speech to capture essential characteristics like timbre, pitch, and speech rhythm. The following features are extracted:

- **MFCC (Mel-Frequency Cepstral Coefficients):** These features capture the phonetic content of speech and its timbral properties.
- **Chroma Features:** Capture harmonic aspects of speech, often associated with emotion-related pitch changes.
- **Spectral Contrast:** Provides insight into the spectral properties of the speech signal, particularly the peaks and valleys in its frequency spectrum.
- **Delta and Delta-Delta Features:** Represent the temporal dynamics of the audio signal, which are crucial for capturing emotion-related variations in speech over time.

3.2.2 Video Emotion Recognition

The video emotion recognition model uses a CNN trained on datasets such as FER-2013 to classify facial expressions into one of several emotions (e.g., Happy, Sad, Angry, Neutral).
Input: Feature embeddings extracted from each video frame.
Output: Predicted emotion class (e.g., Happy, Sad) and confidence score.

3.2.3 Fusion of Audio and Video Predictions

The fusion layer combines the predictions from both the audio model and the video model using the attention-based fusion mechanism (discussed in Section 3.1). The combined prediction from both models is the system's final emotion output.

Output: The final emotion prediction (e.g., Happy, Sad, Angry) along with a confidence score that reflects the system's certainty.

4. EXPERIMENT RESULTS

4.1 Performance Evaluation

The system was evaluated on both the FER-2013 and RAVDESS datasets to assess the effectiveness of multimodal emotion recognition. The evaluation metrics used include:

- **Accuracy:** The percentage of correctly classified emotions.
- **Precision:** The proportion of positive results that were correctly identified.
- **Recall:** The proportion of actual positive cases that were correctly identified.
- **F1-Score:** The harmonic mean of precision and recall.

4.2 FER-2013 Results (Facial Expression Recognition)

We first tested the video-only model on the FER-2013 dataset. The results were as follows:

Emotion	Precision	Recall	F1-Score
Angry	0.88	0.90	0.89
Disgust	0.87	0.85	0.86
Fearful	0.84	0.83	0.83
Happy	0.92	0.91	0.91
Sad	0.81	0.80	0.80
Surprised	0.85	0.87	0.86
Neutral	0.90	0.92	0.91

Table 1. FER-2013 Results (Facial Expression Recognition)

The CNN-based facial emotion recognition model showed high accuracy in detecting emotions like happiness, anger, and neutral, but faced challenges with emotions like sadness and fear, which are often subtle in facial expressions.

4.3 RAVDESS Results (Speech Emotion Recognition)

Next, we tested the audio-only model on the RAVDESS dataset. The results were:

Emotion	Precision	Recall	F1-Score
Neutral	0.88	0.89	0.88
Calm	0.91	0.89	0.90
Happy	0.87	0.85	0.86
Sad	0.83	0.82	0.82
Angry	0.90	0.91	0.90
Fearful	0.86	0.84	0.85
Disgust	0.89	0.90	0.89
Surprised	0.91	0.92	0.91

Table 2. RAVDESS Results (Speech Emotion Recognition)

The Random Forest-based audio emotion recognition model performed well in recognizing emotions like anger and fear, but showed some weakness in sadness and happiness, which are often expressed through more nuanced voice characteristics.

4.4 Multimodal Fusion Results

For the multimodal fusion system, we combined the predictions from both the audio and video models using the attention-based fusion mechanism. The results were:

Emotion	Precision	Recall	F1-Score
Angry	0.91	0.92	0.91
Disgust	0.90	0.89	0.89
Fearful	0.88	0.89	0.88
Happy	0.93	0.92	0.92
Sad	0.87	0.86	0.87
Surprised	0.89	0.90	0.89
Neutral	0.91	0.92	0.91

Table 3. Multimodal Fusion Results

The fusion model showed a significant improvement over the individual unimodal systems, with an average increase in accuracy of 4-6% across all emotion categories. The attention-based fusion allowed the model to dynamically weigh the importance of the audio and video features, leading to better recognition of emotions that were hard to detect in one modality alone.

4.5 Comparison with State-of-the-Art Systems

To validate our results, we compared the performance of our multimodal emotion recognition system with previous state-of-the-art methods. Our system outperformed several existing models in terms of accuracy, precision, and F1-score, as shown in the following comparison table:

Model	F1-Score	Class 1	Class 2	Class 3
Our Fusion Model	90.5%	0.91	0.90	0.90
Zhao et al. (2019)	87.3%	0.88	0.86	0.87
Kim & Lee (2019)	85.5%	0.86	0.85	0.85
Dong et al. (2020)	88.0%	0.89	0.87	0.88

Table 4. Comparison with State-of-the-Art Systems

Overall, our multimodal emotion recognition system demonstrated superior performance, particularly when integrating the complementary features of audio and visual data with an attention mechanism.

5. CONCLUSION

In conclusion, this project presents a robust multimodal emotion recognition system that combines the strengths of both audio and video modalities to detect human emotions with high accuracy. The integration of an attention-based fusion mechanism further enhances the performance of the system by allowing it to selectively focus on the most relevant features from both modalities, improving the overall recognition accuracy.

The proposed system has significant potential for applications in affective computing, mental health monitoring, customer service, and human-robot interaction. By using state-of-the-art feature extraction techniques for both audio and visual data and combining them through an innovative fusion approach, this work demonstrates the advantages of multimodal emotion recognition systems in overcoming the limitations of unimodal approaches.

Future work could explore more complex fusion strategies, real-time deployment, and the incorporation of additional modalities such as text or physiological signals to further enhance the system's robustness and accuracy in real-world applications.

6. SCOPE OF THE PROJECT

This project focuses on developing a multimodal emotion recognition system using facial expressions and speech as the two primary input modalities. The scope of the project includes:

1. Dataset Construction:
Using the FER-2013 dataset for facial expression recognition and the RAVDESS dataset for emotion recognition in speech.

2. Feature Extraction:
Implementing feature extraction techniques for audio (MFCC, Spectral Contrast, Chroma) and video (CNN-based feature extraction).

3. Multimodal Fusion:
Implementing an attention-based fusion layer to combine the features from both modalities for improved emotion classification.

4. Facial recognition module:
Integrating face recognition to identify known criminals or suspicious individuals.

5. Training:
Evaluating the model performance on both the FER-2013 and RAVDESS datasets.

6. User Interface:
Developing a simple user interface (UI) that will allow end users to interact with the system in real-time.

7. REFERENCES

[1] Zhao, X., & Zhang, Y. (2019). Emotion recognition using multimodal deep learning. IEEE Access, 7, 62513-62524. DOI: [10.1109/ACCESS.2019.2916790] SSD: Single Shot MultiBox Detector. ECCV. (single-shot detector balancing speed & accuracy).

[2] Bendriem, I., & Benassi, H. (2020). Multimodal emotion recognition using deep neural networks: A survey. Journal of Ambient Intelligence and Humanized Computing, 11(4), 1557-1572. DOI: [10.1007/s12652-020-02277-7] Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. arXiv.(improvements in backbone and multi-scale prediction).

[3] Dong, X., & Zhang, Y. (2020). Multimodal emotion recognition via attention-based fusion networks. Proceedings of the 28th ACM International Conference on Multimedia, 517-525. DOI: [10.1145/3343031.3350984]

[4] Zhao, G., & Kou, Y. (2017). Deep fusion: Audio-visual emotion recognition in the wild. IEEE Transactions on Affective Computing, 8(3), 334-346. DOI: [10.1109/TAFFC.2017.2760609]).

[5] (Kim, J., & Lee, H. (2019). Multimodal emotion recognition with attention fusion networks. IEEE Transactions on Multimedia, 21(10), 2586-2598. DOI: [10.1109/TMM.2019.2902394]

[6] Gupta, D., & Kaur, S. (2020). Multimodal emotion recognition using hybrid deep learning. Neural Computing and [10.1007/s00542-020-05814-w] Applications, 32, 11679-11688. DOI: CVPR. (compound scaling for detection backbones and biFPN).

[7] Shao, X., & Zhang, X. (2019). Emotion recognition with multimodal deep learning. International Journal of Pattern Recognition and Artificial Intelligence, 33(9), 1943004. DOI: [10.1142/S0218001419430043]

[8] Jaiswal, A., & Ghosh, S. (2020). Deep learning models for emotion recognition from speech: A review. Journal of Electrical Engineering & Technology, 15(2), 547-561. DOI: [10.1007/s42835-019-00312-4]

[9] Poria, S., & Gelbukh, A. (2017). Multimodal sentiment analysis: A survey. Computational Intelligence, 33(3), 319-346. DOI: [10.1111/coin.12075].

[10] Zeng, Z., & Pantic, M. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1), 39-58. DOI: [10.1109/TPAMI.2008.62].

[11] Zeng, Z., & Pantic, M. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions.