

Enhanced Multimodal Emotion Recognition using Deep Representation Learning and Cross-Modal Feature Fusion

Ramya. G^{1*}

¹Department of Information Technology, Thirumalai Engineering College, Kanchipuram, Tamil nadu, India.

Corresponding Author E-mail: btechitgramya@gmail.com

ABSTRACT: Emotion recognition is a crucial aspect of humancomputer interaction, enabling machines to understand and respond to human emotions more effectively. In this paper, we explore state-of-the-art models for multimodal emotion recognition, leveraging textual, auditory, and visual inputs. By integrating these diverse data sources, we develop an ensemble model designed to enhance the accuracy and robustness of emotion detection. Our approach aims to capture intricate emotional cues from speech patterns, facial expressions, and textual content, ensuring a more comprehensive understanding of human emotions. The proposed system aggregates insights from each modality and presents the results in a clear and interpretable manner, facilitating better decision-making in real-world applications. This research contributes to advancing emotion recognition technologies, with potential applications in healthcare, customer service, and humancomputer interaction.

Keywords: Emotion Recognition- Text - Audio - Video Affective Computing

1. Introduction

Emotion recognition plays a vital role in enhancing human-computer interaction by enabling machines to perceive, interpret, and respond to human emotions in a meaningful way. With the rapid advancements in artificial intelligence (AI) and deep learning, there has been a growing interest in developing systems that can accurately recognize emotions across different modalities, such as text, speech, and video. These modalities contain rich emotional cues that, when analyzed collectively, can lead to a more precise and holistic understanding of human emotions.

Traditional emotion recognition systems primarily focus on single-modality approaches, such as text-based sentiment analysis or speech-based emotion detection. However, emotions are complex and multifaceted, often expressed through a combination of verbal content, vocal tone, and facial expressions. Relying on a single modality

can lead to incomplete or inaccurate interpretations, as textual data alone may lack tonal variations, and speech may not always capture the full context of emotions. To address these limitations, multimodal emotion recognition has emerged as a promising research direction, integrating information from multiple sources to enhance the accuracy and robustness of emotion detection systems.

In this study, we explore state-of-the-art models for multimodal emotion recognition by leveraging textual, auditory, and visual inputs. We propose an ensemble model that effectively combines insights from each modality to improve emotion classification and ensure a more comprehensive understanding of human affective states. Our approach captures intricate emotional cues from speech patterns, facial expressions, and textual content, allowing for a more detailed

interpretation of emotions. The integration of multiple modalities enables the system to compensate for the limitations of individual modalities, leading to improved performance and reliability.

Furthermore, the proposed system is designed to present its findings in a clear and interpretable manner, making it suitable for real-world applications that require human-like emotional intelligence. Emotion recognition has vast applications across various domains, including healthcare, customer service, virtual assistants, and affective computing. In healthcare, it can assist in diagnosing and monitoring mental health conditions, while in customer service, it can improve user experiences by enabling emotionally aware interactions. Similarly, affective computing systems can enhance virtual reality (VR) environments, e-learning platforms, and humanrobot interactions by making them more responsive to users' emotional states.

Through this research, we aim to contribute to the advancement of multimodal emotion recognition technologies by developing a system that integrates and processes diverse emotional cues with high accuracy. By leveraging the synergy of textual, speech, and visual data, our study seeks to enhance the reliability and applicability of emotion recognition in realworld scenarios, ultimately bridging the gap between artificial intelligence and human emotions.

2. Related Works

Multimodal emotion recognition has been an evolving field in affective computing, aiming to bridge the gap between human emotional intelligence and machine perception. Traditionally, researchers have focused on single-modality emotion recognition techniques, where emotions were identified using either text, speech, or facial expressions individually. Text-based emotion recognition primarily leverages Natural Language Processing (NLP) techniques, employing methods such as lexicon-based

approaches and deep learning models like Long Short-Term Memory (LSTM) networks, Bidirectional Encoder Representations from Transformers (BERT), and other transformer-based architectures. These models analyze written or spoken language to detect emotional sentiment, but they often struggle with contextual ambiguity and lack of tonal cues. Similarly, speech-based emotion recognition utilizes acoustic features such as pitch, tone, prosody, and Mel-Frequency Cepstral Coefficients (MFCCs) to detect emotional variations in voice. Early methods relied on classical machine learning techniques like Support Vector Machines (SVM) and Hidden Markov Models (HMM), whereas modern approaches incorporate deep learning models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformerbased architectures like Wav2Vec for improved performance.

Facial expression analysis, a key component of videobased emotion recognition, has also seen significant progress over the years. Initially, researchers used handcrafted feature extraction methods such as Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP) to detect facial features. However, with the advent of deep learning, models such as VGGFace, ResNet, and FaceNet have been widely adopted for facial expression recognition, significantly improving accuracy. More recent studies integrate spatiotemporal models, such as 3D CNNs and LSTMs, to capture the dynamic nature of facial expressions over time, while transformer-based vision models further enhance recognition capabilities.

Despite the advancements in unimodal approaches, relying on a single modality presents inherent limitations. Textbased models may lack prosodic and visual context, speechbased models may fail in noisy environments, and video-based models may struggle with occlusions, poor lighting, or camera angle variations. To overcome these challenges, researchers have shifted toward

multimodal emotion recognition, which integrates multiple data sources—text, speech, and video—to create a more holistic understanding of emotions. Various fusion techniques have been explored, including early fusion (combining raw data before processing), late fusion (integrating predictions from separate models), and hybrid fusion (a combination of both). Deep learning architectures such as Multimodal Transformer Networks (MTN) and Multimodal Deep Belief Networks (MDBN) have demonstrated promising results in combining multiple modalities effectively.

However, multimodal emotion recognition still faces challenges, such as synchronization of different modalities, effective feature alignment, and handling missing or noisy data in real-world scenarios. Furthermore, cultural and linguistic variations pose additional difficulties in generalizing emotion recognition models across diverse populations. Addressing these issues requires more sophisticated approaches that can adapt to varying conditions while maintaining high accuracy and interpretability.

Building upon these existing studies, our research aims to develop an ensemble model that integrates advanced deep learning techniques for multimodal emotion recognition.

By combining insights from text, speech, and video data, our approach seeks to enhance emotion classification accuracy while ensuring interpretability and robustness. Our model leverages state-of-the-art fusion techniques to effectively align multimodal features, mitigating common challenges such as data inconsistencies and synchronization errors. Through this study, we contribute to the advancement of emotion recognition technologies, enabling more reliable applications in fields such as healthcare, customer service, virtual assistants, and human-computer interaction.

3. proposed System

The proposed system for multimodal emotion recognition integrates text, audio, and video inputs to enhance emotion classification accuracy. Utilizing state-of-the-art deep learning models, the system processes textual data through Natural Language Processing (NLP) techniques, including tokenization, word embeddings (Word2Vec, BERT), and classification using Long Short-Term Memory (LSTM) networks. Speech-based emotion recognition leverages audio processing techniques such as Mel-Frequency Cepstral Coefficients (MFCCs), spectral analysis, and feature extraction, followed by classification using Support Vector Machines (SVMs) or deep neural networks like CNN-LSTM hybrids. Video-based recognition employs facial expression analysis using Convolutional Neural Networks (CNNs) trained on datasets such as FER2013, extracting facial landmarks and micro-expressions. The system adopts a hybrid fusion approach, combining feature-level fusion, where extracted features from all modalities are concatenated, and decision-level fusion, where independent classifiers generate predictions that are aggregated through a voting mechanism or weighted averaging. The final classification is performed using an ensemble model that integrates these modalities for a more comprehensive understanding of human emotions. Designed for real-time applications, the system is deployed with TensorFlow.js, allowing users to interact via a web-based interface, making it applicable in domains such as healthcare, customer service, and human-computer interaction. By addressing challenges like data synchronization, missing modalities, and computational complexity, this approach enhances the robustness and interpretability of multimodal emotion recognition.

4. Methodology

The proposed multimodal emotion recognition system follows a structured methodology that includes data collection, preprocessing, feature

extraction, fusion techniques, and classification. This section outlines the step-by-step approach used to develop the system.

4.1 Data Collection

The system utilizes datasets from different modalities—text, audio, and video—to create a comprehensive emotion recognition model.

4.1.1 Text Data:

Collected from psychological studies, sentiment analysis corpora, and datasets such as the Big Five Inventory (BFI) and social media posts.

4.1.2 Audio Data:

Utilizes speech recordings from datasets like RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), which includes various emotional speech recordings (happy, sad, angry, fearful, etc.).

4.1.3 Video Data:

Facial expression datasets, including FER2013, are used for training deep learning models to recognize emotions through facial movements.

4.2 Data Preprocessing

Each modality undergoes independent preprocessing before feature extraction.

4.2.1 Text Preprocessing Tokenization & Lemmatization:

Splitting text into words and reducing them to their base forms.

4.2.2 Stopword Removal:

Filtering out commonly used words that do not contribute to emotion detection.

4.2.3 Embedding Techniques:

Word2Vec, TF-IDF, and BERT embeddings convert text into numerical representations.

4.2.4 Audio Preprocessing Pre-Emphasis Filtering:

Enhances high-frequency components in speech.

4.2.5 Framing and Windowing:

Segments speech into short-term analysis frames (e.g., 25ms frame size with a 15ms overlap).

4.2.6 Feature Extraction:

Mel-Frequency Cepstral Coefficients (MFCCs), Spectral Centroid, Zero-Crossing Rate, Pitch, and Energy are extracted to capture vocal emotions.

4.2.7 Video Preprocessing Face Detection:

Uses OpenCV, Dlib, or CNN-based models to detect faces.

4.2.8 Feature Extraction:

Deep learning models like ResNet, VGGFace, and Xception extract facial expression features.

4.2.9 Data Augmentation:

Techniques such as brightness correction, rotation, and noise addition improve model robustness.

4.3 Feature Extraction

Each modality undergoes independent feature extraction:

Modality	Feature Type	Example Features
Text	Word Embeddings	BERT, Word2Vec
Audio	MFCCs, Spectral Features	Pitch, Tone, Energy
Video	Facial Expression Features	Eye Gaze, Smile Intensity

4.3.1 Text Features:

Extracted using word embeddings and NLP techniques.

4.3.2 Audio Features:

Derived from time-domain and frequency domain analysis.

4.3.3 Video Features:

Extracted using deep CNN architectures.

4.4 Fusion Techniques

The system integrates multiple modalities using a hybrid fusion approach to improve emotion recognition accuracy.

4.4.1 Feature-Level Fusion

Extracted features from text, speech, and video are concatenated into a single feature vector. The combined feature vector is processed using a Multimodal Transformer Network (MTN) to learn interdependencies.

4.4.2 Decision-Level Fusion

Each modality has an independent classifier (LSTM for text, CNN for video, CNN-LSTM for speech). A weighted voting mechanism aggregates predictions from each model to determine the final emotion classification.

4.5 Classification Models

The final classification is performed using deep learning models:

4.5.1 Text:

LSTMs and Transformers capture contextual dependencies.

4.5.2 Audio:

CNN-LSTM models analyze sequential speech data.

4.5.3 Video:

CNN-based architectures detect facial expressions.

4.5.4 Ensemble Learning:

Combines multiple classifiers using SVMs or Decision Trees for final prediction.

4.6 Evaluation and Performance Metrics

To assess the system’s effectiveness, various performance metrics are used:

- Accuracy, Precision, Recall, and F1-Score for evaluating classification performance.
- Confusion Matrix to analyze prediction errors.
- Cross-validation techniques to ensure model generalizability.

4.7 Implementation & Deployment

The final system is designed to be deployed in real-time applications using TensorFlow.js,

enabling users to interact with the system via a web-based interface. The model takes live webcam input, processes text, speech, and video, and displays real-time emotion classification results.

5. Result and Discussion

The multimodal emotion recognition system was evaluated using textual, auditory, and visual data. The results indicate that integrating multiple modalities enhances emotion classification accuracy compared to unimodal approaches. This section presents the experimental findings, performance comparisons, and discussions on the strengths and limitations of the proposed model.

5.1 Performance Evaluation

The performance of the system was assessed using standard classification metrics, including accuracy, precision, recall, and F1-score. The evaluation was conducted for each individual modality and for the final fused model.

5.2 Unimodal Performance

The accuracy of emotion recognition using only one modality (text, speech, or video) is summarized in the table below:

Modality	Accuracy (%)	Precision (%)	Recall (%)
Text(LSTM/BERT-based Model)	75.6	74.9	75.2
Audio(CNN-LSTM+MFCCs)	78.4	77.5	78.1
Video(CNN based Facial Expression Recognition)	81.2	80.6	80.9

From these results, it is evident that the video-based model achieves the highest accuracy among unimodal classifiers, aligning with previous research indicating that facial expressions are strong emotional indicators. However, speech and text-based models also demonstrate reasonable performance, highlighting their importance in capturing emotional cues from different communication channels.

5.3 Multimodal Fusion Performance

To improve accuracy, feature-level and decisionlevel fusion techniques were applied to combine information from all three modalities. The final performance of the multimodal fusion model is summarized below:

The decision-level fusion model achieved the highest accuracy of 87.2%, demonstrating the effectiveness of combining predictions from different modalities. Compared to unimodal approaches, multimodal fusion significantly improves recognition accuracy, as it compensates for missing or ambiguous cues in any single modality.

5.4 Discussion of Results

The results support the hypothesis that integrating multiple modalities improves emotion recognition accuracy. While unimodal models capture emotions reasonably well, they face limitations:

Text-based models struggle with sarcasm, irony, and ambiguous phrasing, affecting their reliability.

Audio-based models are influenced by background noise and variations in speech patterns.

Video-based models may fail when faces are partially occluded, expressions are subtle, or lighting conditions are poor.

Model Type	Accurac y (%)	Precisio n (%)	Recall (%)
Feature-Level Fusion (Concatenation+ Multimodal Transformer)	85.4	84.7	85.1
Decision-Level Fusion(Voting-Based Ensemble)	87.2	86.8	87.1

By integrating text, speech, and video, the proposed system overcomes these limitations by providing a more comprehensive emotion analysis. The decision-level fusion technique, which combines predictions from individual classifiers, proved to be more effective than

feature-level fusion, as it mitigates the risk of feature misalignment across different modalities.

5.5 Limitations and Challenges

Despite the promising results, several challenges remain:

5.5.1 Data Synchronization:

Aligning text, audio, and video streams remains a challenge due to differing time resolutions.

5.5.2 Real-World Generalization:

The model’s performance might decrease when applied to spontaneous, real-world conversations, as most datasets contain acted emotions.

5.5.3 Computational Complexity:

The multimodal approach requires higher computational resources, making real-time deployment challenging.

5.6 Future Improvements

To further enhance the system, future work can focus on:

Incorporating more diverse datasets to improve generalizability across different demographics.

Exploring transformer-based fusion techniques to better capture cross-modal dependencies.

Optimizing the model for real-time applications by reducing computational overhead.

The experimental results confirm that multimodal emotion recognition significantly outperforms unimodal approaches, achieving an accuracy of 87.2% using decisionlevel fusion. The integration of text, audio, and video enables a more robust and reliable emotion classification system, making it suitable for applications in healthcare, customer service, virtual assistants, and human-computer interaction. While challenges remain, future improvements can further enhance the adaptability and efficiency of multimodal emotion

6. Conclusion

This study explored multimodal emotion recognition by integrating text, audio, and video

data to improve accuracy and robustness in emotion classification. Traditional unimodal approaches, while effective to some extent, struggle with capturing the full complexity of human emotions due to limitations such as textual ambiguity, speech noise, and facial occlusions. By implementing a hybrid fusion approach, the proposed system successfully combines linguistic, acoustic, and visual cues, leading to a more comprehensive emotion recognition model. The experimental results demonstrate that multimodal fusion significantly improves performance, with the decision-level fusion model achieving an accuracy of 87.2%, outperforming unimodal methods. This confirms that integrating multiple data sources enhances emotion classification by compensating for missing or ambiguous signals in any single modality. Despite these improvements, challenges remain, including data synchronization, real-world generalization, and computational efficiency.

To address these limitations, future research can focus on incorporating more diverse datasets, optimizing deep learning architectures for real-time applications, and exploring transformer-based fusion techniques. With advancements in affective computing and artificial intelligence, multimodal emotion recognition has the potential to revolutionize applications in healthcare, customer service, human-computer interaction, and AI-driven virtual assistants. The proposed system lays a strong foundation for emotion-aware AI systems, making interactions more intelligent, empathetic, and humanlike.

References

1. Bo Pang; Lillian Lee, Year: 2008, "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval*, Vol: 2, No: 1–2, pp. 1–135.
2. Jacob Devlin; Ming-Wei Chang; Kenton Lee; Kristina Toutanova, Year: 2018, "BERT: Pre-training of deep bidirectional transformers for language understanding", pp. 4171–4186.
3. Sepp Hochreiter; Jürgen Schmidhuber, Year: 1997, "Long short-term memory", *Neural Computation*, Vol: 9, No: 8, pp. 1735–1780.
4. Florian Eyben; Martin Wöllmer; Björn Schuller, Year: 2010, "openSMILE – The Munich versatile and fast open-source audio feature extractor", *Proceedings of the ACM Multimedia (MM)*, pp. 1459–1462.
5. S. Poria; E. Cambria; D. Hazarika; N. Majumder, Year: 2017, "A review of multimodal sentiment analysis", *Information Fusion*, Vol: 37, pp. 98–126.
6. Yao-Hung Hubert Tsai; Shaojie Bai; Paul Pu Liang; J. Zico Kolter; Louis-Philippe Morency; Ruslan Salakhutdinov, Year: 2019, "Multimodal transformer for unaligned multimodal language sequences", pp. 6558.
7. L. Sun; C. Jia; L. Zhao; M. Wu; J. Yu, Year: 2020, "Deep multimodal fusion for emotion recognition in video conversations", *IEEE Transactions on Multimedia*, Vol: 22, No: 11, pp. 2753–2766.
8. Alex Krizhevsky; Ilya Sutskever; Geoffrey E. Hinton, Year: 2012, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NeurIPS)*.
9. Dzmitry Bahdanau; Kyunghyun Cho; Yoshua Bengio, Year: 2015, "Neural machine translation by jointly learning to align and translate", *International Conference on Learning Representations (ICLR)*.
10. Paul Ekman; Wallace V. Friesen, Year: 1971, "Constants across cultures in the face and emotion", *Journal of Personality and Social Psychology*, Vol: 17, No: 2, pp. 124–129.
11. Tadas Baltrušaitis; Chaitanya Ahuja; Louis-Philippe Morency, Year: 2019, "Multimodal machine learning: A survey and taxonomy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol: 41, No: 2, pp. 423–443.
12. Kaiming He; Xiangyu Zhang; Shaoqing Ren; Jian Sun, Year: 2016, "Deep residual learning for image recognition", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.
13. Devamanyu Hazarika; Roger Zimmermann; Soujanya Poria, Year: 2020, "MISA: Modality-invariant and -specific

representations for multimodal sentiment analysis”, Proceedings of the 28th ACM International Conference on Multimedia (MM), pp. 1122-1131.

14. Florian Schroff; Dmitry Kalenichenko; James Philbin, Year: 2015, “FaceNet: A unified embedding for face recognition and clustering”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815-823.

15. Ken Chatfield; Karen Simonyan; Andrea Vedaldi; Andrew Zisserman, Year: 2014, “Return of the devil in the details: Delving deep into convolutional nets”, British Machine Vision Conference.