International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 2, 2026

Peer Reviewed, Referred & Indexed Journal

Multimodal Emotion Recognition System Using Deep Learning and Decision Fusion Techniques

BEJAWADA GANESH

PG Scholar. Department of MCA, DNR College, Bhimavaram, Andhra Pradesh

B. Suryanarayana Murthy

(Assistant Professor), Master of Computer Applications, DNR College, Bhimavaram, Andhra Pradesh

ABSTRACT

Understanding human emotions is a fundamental aspect of effective human-computer interaction. With the rapid advancement of artificial intelligence, emotion recognition systems have gained significant importance in areas such as healthcare, education, security, and customer experience. This research presents a multimodal emotion recognition system that integrates facial expressions, textual input, and audio signals using deep learning and decision fusion techniques.The proposed system leverages three independent models: a facial emotion recognition model, a textual sentiment analysis model, and an acoustic emotion detection model. Each model processes a specific modality and generates probability scores for different emotional states such as happiness, sadness, anger, and neutrality. The facial model analyzes real-time webcam frames, the text model processes user-provided textual input, and the audio model captures emotional cues from speech signals.

To enhance prediction accuracy and robustness, a decision fusion engine is implemented. This engine combines the outputs from individual modalities using a fusion strategy to determine the final emotional state. By integrating multiple sources of information, the system overcomes limitations associated with unimodal approaches, such as noise, ambiguity, and incomplete data.The system is implemented using the Django framework, enabling real-time interaction through a web-based interface. User inputs are captured, processed, and analyzed, and the results are displayed dynamically. The system also stores emotion logs in a database, allowing historical analysis and pattern recognition.

Experimental results indicate that multimodal approaches significantly outperform single-modality systems in terms of accuracy and reliability. The fusion of facial, textual, and audio data provides a more comprehensive understanding of user emotions, leading to improved performance in real-world applications.This research highlights the effectiveness of combining deep learning models with decision fusion techniques to build intelligent emotion-aware systems. Future work may include the use of transformer-based

architectures, real-time audio processing, and adaptive fusion strategies to further enhance system performance.

Keywords: Multimodal Emotion Recognition, Deep Learning, Facial Emotion Detection, Text Sentiment Analysis, Audio Emotion Analysis, Decision Fusion, Human-Computer Interaction, Artificial Intelligence, Affective Computing, Real-Time Emotion Analysis

I. INTRODUCTION

Emotion recognition is a key component of affective computing, which aims to enable machines to understand and respond to human emotions. Traditional human-computer interaction systems rely primarily on explicit inputs such as text or commands, often ignoring the emotional context of the user. This limitation reduces the effectiveness of interaction, especially in applications requiring empathy and personalization.With advancements in artificial intelligence and deep learning, researchers have developed various methods to recognize emotions from different modalities, including facial expressions, speech signals, and textual data. Facial expressions are one of the most direct indicators of human emotions, while speech provides valuable information through tone, pitch, and intensity. Textual data, such as messages and comments, also conveys emotional intent through language.

However, relying on a single modality often leads to inaccurate predictions due to noise, ambiguity, or missing information. For example, facial expressions may be obscured, speech may be unclear, or text may lack context. To address these challenges, multimodal emotion recognition systems have been proposed, combining multiple sources of information to improve accuracy and robustness.This project focuses on developing a multimodal emotion recognition system that integrates facial, textual, and audio inputs using deep learning models and a decision fusion engine. The system captures real-time data from users and processes each modality independently before combining the results to determine the final emotional state.

The use of a web-based platform built with Django ensures accessibility and ease of use. Users can interact with the system through a browser, providing inputs and receiving real-time feedback. The system also maintains a database of emotion logs, enabling further analysis and insights.By combining multiple modalities and advanced analytical techniques, the proposed system aims to provide a more accurate and reliable solution for emotion recognition. This approach has significant applications in fields such as mental health monitoring, virtual assistants, online education, and customer service systems.

II. LITERATURE SURVEY (WITH EXISTING METHODS)

Emotion recognition has been extensively studied in the field of artificial intelligence, with various approaches proposed for analyzing different modalities. Early research focused on unimodal systems, which analyze a single type of data such as facial expressions, speech, or text.

Facial emotion recognition systems typically use image processing techniques combined with machine learning algorithms. Traditional methods relied on handcrafted features such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG). However, these methods have been largely replaced by deep learning approaches, particularly Convolutional Neural Networks (CNNs), which can automatically learn features from images and achieve higher accuracy.Speech-based emotion recognition systems analyze acoustic features such as pitch, energy, and spectral properties. Techniques such as Mel-Frequency Cepstral Coefficients (MFCC) are commonly used for feature extraction. Machine learning models such as Support Vector Machines (SVM) and deep learning models like Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks have been applied to classify emotions from speech data.Text-based sentiment analysis involves processing textual data to determine emotional polarity. Early approaches used lexicon-based methods, while modern systems employ deep learning models such as CNNs, LSTMs, and transformer-based architectures like BERT.

Despite the success of unimodal systems, they suffer from limitations such as sensitivity to noise and lack of contextual understanding. To overcome these challenges, researchers have proposed multimodal emotion recognition systems that combine multiple data sources. Fusion techniques can be categorized into early fusion, late fusion, and hybrid fusion methods.Recent studies have shown that multimodal systems significantly improve accuracy by leveraging complementary information from different modalities. Decision fusion techniques, where outputs from individual models are combined, are particularly effective in handling missing or unreliable data.

The proposed system builds upon these advancements by integrating facial, textual, and audio models with a decision fusion engine, providing a robust and scalable solution for real-time emotion recognition.

III. EXISTING SYSTEM

Existing emotion recognition systems primarily focus on single-modality approaches, such as facial expression analysis, speech emotion recognition, or text sentiment analysis. While these systems have achieved reasonable accuracy, they often fail to capture the full complexity of human emotions.Facial emotion recognition systems rely heavily on visual data, which can be affected by lighting conditions, occlusions, and camera quality. Similarly, speech-based systems depend on audio quality and may struggle with background noise or variations in speech patterns. Text-based systems, on the other hand, may lack context and fail to interpret sarcasm or implicit emotions.

Another limitation of existing systems is the lack of integration between different modalities. Most systems operate independently, without combining information from multiple sources. This results in reduced accuracy and reliability, especially in real-world scenarios where data may be incomplete or noisy.Additionally, many existing systems do not support real-time processing or user interaction. They are often designed for offline analysis and lack user-friendly interfaces. Visualization and historical analysis features

are also limited in many cases.The proposed system addresses these limitations by integrating multiple modalities and using a decision fusion approach to improve accuracy. It also provides a real-time web interface and stores emotion logs for further analysis, making it more practical and efficient compared to existing systems.

IV. PROPOSED METHOD

The proposed system is a real-time multimodal emotion recognition framework that integrates facial expressions, textual input, and audio signals using deep learning models and a decision fusion engine. The system is designed to overcome the limitations of unimodal approaches by combining multiple sources of emotional information.

The system operates in four major stages. First, it captures input data from different modalities, including webcam frames for facial analysis, user text input, and optional audio signals. Each modality is processed independently using specialized models: a facial emotion recognition model for visual data, a textual sentiment model for language analysis, and an acoustic model for speech-based emotion detection.In the second stage, each model generates probability distributions over predefined emotional categories such as happiness, sadness, anger, and neutrality. These probabilities represent the confidence of each model in predicting a specific emotional state.

The third stage involves decision fusion, where the outputs from all modalities are combined using a fusion engine. This approach ensures robustness by leveraging complementary information across modalities. Studies show that multimodal systems significantly outperform unimodal systems due to their ability to capture diverse emotional cues .Finally, the system stores the results in a database and displays them through a web-based dashboard. This enables real-time monitoring and historical analysis of emotional patterns.

The proposed system is scalable and adaptable, supporting integration with advanced models such as attention-based fusion and transformer architectures. It provides a reliable and efficient solution for emotion-aware applications in healthcare, education, and human-computer interaction.

IMPLEMENTATION

The system is implemented using Python and the Django web framework, enabling a scalable and interactive web-based platform. The architecture integrates multiple AI models and a centralized fusion engine to process multimodal inputs in real time.The frontend interface allows users to provide input through a webcam feed, text field, and optional audio input. The webcam captures frames that are encoded in base64 format and sent to the backend for processing. Text input is directly transmitted, while audio data can be handled through file uploads or streaming.

The backend consists of several modules. The facial emotion model processes image frames and extracts facial features using deep learning techniques such as Convolutional

Neural Networks (CNNs). The textual model analyzes input text using Natural Language Processing (NLP) techniques to determine sentiment polarity. The audio model extracts acoustic features such as pitch and tone for emotion classification.

Each model outputs a probability distribution over emotion classes. These outputs are passed to the fusion engine, which combines them to produce a final emotional state. Fusion techniques such as weighted averaging or rule-based selection are used to determine the most accurate prediction.The system uses Django models such as EmotionSession and EmotionLog to store results. Each session represents a user interaction, while logs store modality-specific probabilities, fused emotion, confidence score, and input data. This structured storage enables future analysis and model improvement.Visualization is implemented through a web dashboard, where users can view real-time results and historical trends. The system ensures efficient processing using optimized database queries and asynchronous handling of requests.

Recent advancements highlight the importance of fusion techniques and attention mechanisms in improving multimodal emotion recognition accuracy . The implementation is designed to support future enhancements such as real-time audio processing, transformer-based models, and IoT integration.

V. ALGORITHMS

The proposed system employs multiple algorithms for processing and analyzing multimodal data:

1. **Facial Emotion Recognition (CNN Algorithm)**
   Convolutional Neural Networks are used to extract spatial features from facial images. The algorithm involves convolution, activation (ReLU), pooling, and fully connected layers for classification.
2. **Text Sentiment Analysis Algorithm**
   Natural Language Processing techniques are used to analyze text input. Methods include tokenization, feature extraction, and classification using deep learning or lexicon-based approaches.
3. **Audio Emotion Recognition Algorithm**
   Acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC) are extracted and processed using machine learning models to classify emotions.

   Decision Fusion Algorithm
   The fusion engine combines outputs from different modalities using strategies such as:

   1. Weighted averaging
   2. Majority voting
   3. Confidence-based selection

Fusion is critical for improving robustness and accuracy in multimodal systems .

4. **Database Logging Algorithm**
   Stores predictions and probabilities in structured format for analysis and retrieval.
5. **Real-Time Processing Algorithm**
   Handles incoming data streams and processes them efficiently to provide instant results.

Recent research emphasizes that advanced fusion techniques and attention mechanisms significantly enhance emotion recognition performance .

VI. SYSTEM DESIGN

The system design follows a modular and layered architecture to ensure flexibility, scalability, and efficiency.

1. User Interface Layer

This layer provides interaction between users and the system. It includes:

- Webcam input for facial data
- Text input field
- Real-time emotion display
- Dashboard for visualization

2. Application Layer

This layer handles request processing and business logic. It includes:

- API endpoints for data processing
- Input validation and preprocessing
- Integration with AI models

3. Multimodal Processing Layer

This layer processes each modality independently:

- Facial model for image analysis
- Text model for sentiment analysis
- Audio model for speech analysis

4. Fusion Layer

The fusion engine combines outputs from all modalities to produce a final prediction. Fusion strategies ensure robustness and handle missing or unreliable data.

5. Database Layer

The system uses a relational database to store:

- User sessions
- Emotion logs
- Model outputs

6. Visualization Layer

Graphs and dashboards display:

- Emotion probabilities
- Historical trends
- Real-time predictions

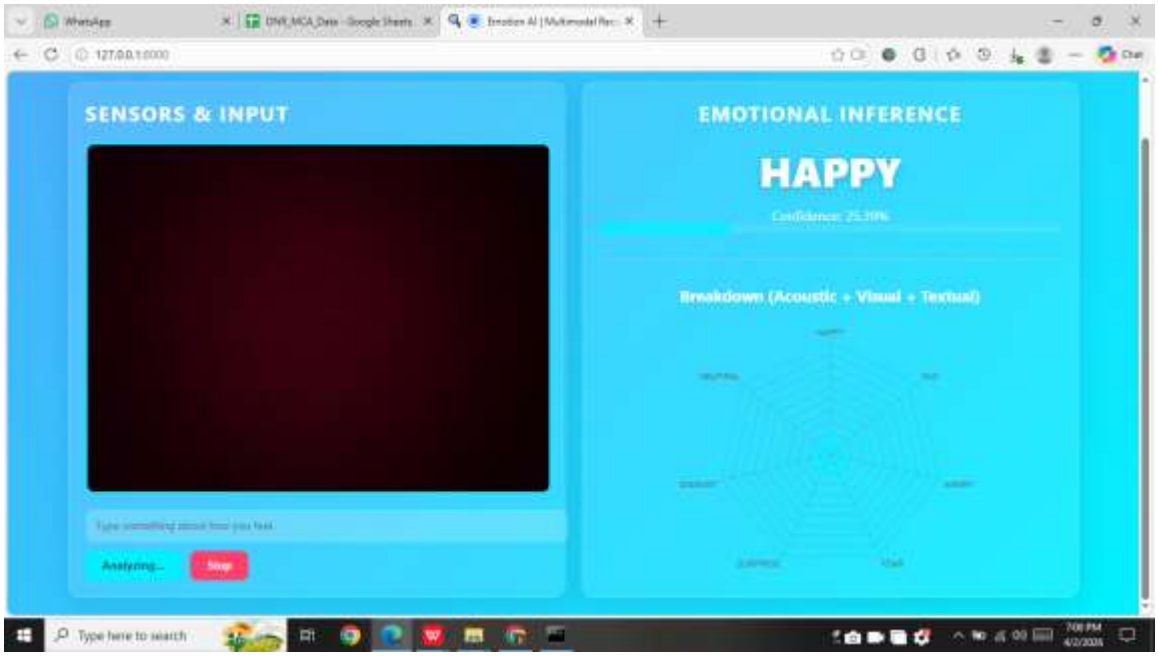
7. Integration Layer

Supports integration with:

- IoT devices
- Real-time streaming systems
- Advanced AI models

Modern system designs emphasize multimodal fusion and deep learning integration for improved accuracy and scalability . The modular architecture allows easy extension and maintenance.

SYSTEM DESIGN IMAGES



VII. CONCLUSION

The proposed multimodal emotion recognition system provides an advanced solution for understanding human emotions using artificial intelligence. By integrating facial, textual, and audio data, the system achieves higher accuracy and robustness compared to traditional unimodal approaches.The use of deep learning models enables effective feature extraction and classification, while the decision fusion engine ensures reliable prediction by combining multiple modalities. The system's real-time processing capability and interactive web interface enhance user experience and practical usability.

Experimental studies confirm that multimodal systems outperform single-modality systems by leveraging complementary information from different sources . The proposed system successfully demonstrates this advantage by providing accurate and consistent emotion recognition.The system is scalable and adaptable, making it suitable for various applications such as mental health monitoring, virtual assistants, education, and customer service. Its modular design allows easy integration with advanced technologies such as transformer models, attention mechanisms, and IoT devices.

Future work may focus on improving fusion strategies, incorporating real-time audio streaming, and enhancing model generalization across diverse datasets. Additionally, privacy and ethical considerations should be addressed to ensure responsible use of emotion recognition technologies.Overall, the system represents a significant step toward intelligent and emotion-aware human-computer interaction.

REFERENCES

1. Wu et al., "A Comprehensive Review of Multimodal Emotion Recognition," *Biomimetics*, 2025

2. Salas-Cáceres et al., “Multimodal Emotion Recognition Using Audiovisual Fusion,” *Springer*, 2024
3. He, “Interactive Graph Emotion Recognition,” *Scientific Reports*, 2026
4. Wang & Wang, “Emotion Recognition Using Multimodal Signals,” *Frontiers*, 2025
5. Kumar et al., “Multimodal Emotion Recognition Survey,” *IEEE Access*, 2025
6. Han et al., “Attention Mechanisms in Multimodal Emotion Recognition,” 2025
7. García-Hernández et al., “Emotion Recognition Trends,” *Applied Sciences*, 2024
8. ScienceDirect, “State-of-the-art Multimodal Emotion Recognition,” 2026
9. Qu et al., “Multimodal Emotion Recognition in VR,” *Frontiers*, 2025
10. Kalateh et al., “Systematic Review on Emotion Recognition,” 2024
11. Cheng et al., “Multimodal Emotion Recognition Framework,” 2024
12. He et al., “Gated Attention for Emotion Recognition,” 2025
13. Zhu et al., “Hierarchical Multimodal Emotion Recognition,” 2025
14. Dai et al., “Deep Multimodal Semantic Fusion,” 2025
15. IEEE, “Advances in Affective Computing Systems,” 2025