International Journal of Electrical, Electronics and Computer Systems
ISSN: 2347-2820
Volume 15 Issue 1s (2026) | Pages: 283-288

AI-Based Emotion Recognition and Supportive Response System for Non-Verbal Communication

Surekha Dhumal[1], Dnyaneshwar Dhas[2], Kavya Hingane[3], Hrutik Jadhav[4] , Vaishnavi Tokale[5]

Computer Engineering Department / Genba Sopanrao Moze College of Engineering, Balewadi, Pune / SPPU / India

Peer Review Information

Type: *Article*
Received: *3 February 2026*
Revised: *4 March 2026*
Accepted: *1 April 2026*
Published: *22 May 2026*

Abstract

This research presents an AI-based multimodal therapist system for non-verbal communication using real-time facial emotion recognition and hand gesture detection. The system combines computer vision and deep learning techniques, where emotions are detected using a CNN model and gestures are recognized using MediaPipe and LSTM. Live video input is captured through a webcam and processed using OpenCV. The results are displayed via a Streamlit interface with voice feedback. Unlike existing systems, this approach integrates both emotion and gesture recognition, improving interaction accuracy and usability in real-time environments.

Keywords: AI Therapist System; Artificial Intelligence; CNN; Computer Vision; Deep Learning; Emotion Recognition; Facial Expression Analysis; Hand Gesture Recognition; LSTM; MediaPipe; Multimodal Learning; OpenCV; Real-Time Processing; Streamlit.

How to Cite This Article

Dhumal, S., Dhas, D., Hingane, K., Jadhav, H., & Tokale, V. (2026). *AI-based emotion recognition and supportive response system for non-verbal communication. International Journal of Electrical, Electronics and Computer Systems,* 15(1s), 283–288.

Introduction

Communication is essential for human interaction, but individuals with speech or hearing impairments face challenges in expressing emotions and intentions. Hand gestures and facial expressions are key components of non-verbal communication.

Existing systems often focus on either gesture recognition or emotion detection, with limited real-time integration. Recent advancements in deep learning, such as CNNs and RNNs, have improved recognition capabilities. This research proposes a multimodal AI therapist system that combines facial emotion detection and hand gesture recognition to provide real-time feedback through an interactive interface. The proposed system aims to enhance human-computer interaction by integrating multiple input modalities into a single platform. By combining visual cues from facial expressions and hand movements, the system improves interpretation accuracy and provides meaningful responses, making it more effective for assistive communication applications.

Literature Review

Recent developments in Artificial Intelligence (AI), computer vision, and deep learning have significantly advanced emotion recognition and gesture interpretation systems for assistive communication applications. These technologies have enabled intelligent human–computer interaction by analyzing facial expressions, body movements, and contextual behavioral patterns in real time.

Hashi et al. (2024) presented a systematic review of hand gesture recognition approaches based on deep learning architectures including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Their study reported substantial improvements in recognition performance and classification accuracy across various datasets. However, challenges related to illumination sensitivity, environmental variations, and maintaining real-time responsiveness remained major limitations for practical deployment.

Jalayer et al. (2026) investigated vision-based hand detection, segmentation, and gesture recognition for human–robot interaction applications. Their review emphasized the effectiveness of deep learning methods in extracting complex hand features while identifying limitations associated with dataset diversity, computational requirements, and integration of multimodal interaction systems. The authors highlighted the need for more efficient frameworks capable of supporting real-time environments.

Similarly, Rahman et al. (2025) conducted a comparative analysis of modern hand gesture recognition techniques and demonstrated that temporal learning models improve dynamic gesture understanding. Despite strong classification performance, issues related to processing speed, robustness, and scalability under real-world conditions were identified.

Emotion recognition has also emerged as an important research direction. Villegas-Ch et al. (2025) explored the application of Generative Adversarial Networks (GANs) for generating and analyzing emotional facial expressions in intelligent educational environments. Their findings showed improved recognition capability but highlighted difficulties in detecting subtle emotional variations and maintaining stable performance during real-time execution.

Figueiredo et al. (2025) reviewed AI-driven emotion recognition systems in pediatric healthcare and demonstrated the growing potential of facial emotion analysis in supporting communication and behavioral assessment. However, data inconsistency, emotional variability, and practical implementation challenges continued to affect system reliability.

Furthermore, Chaves-Villota et al. (2026) examined multimodal speech emotion recognition approaches and concluded that combining multiple input modalities improves prediction performance and interaction quality. Nevertheless, integrating heterogeneous data sources remains a challenging task.

Based on these findings, existing studies demonstrate strong progress in individual emotion and gesture recognition models but reveal limited research on unified real-time multimodal systems. Therefore, the proposed work addresses this gap by integrating CNN-based facial emotion recognition with MediaPipe and LSTM-based gesture analysis to develop a supportive AI-driven communication framework capable of generating meaningful responses for non-verbal interaction environments.

Methodology

The proposed system follows a real-time processing pipeline consisting of multiple stages:

- Data Acquisition: Live video is captured using a webcam through OpenCV.
- Face Detection & Emotion Recognition: Facial regions are extracted and passed into a CNN-based model trained on grayscale images (64×64 input).
- Hand Gesture Detection: Hand landmarks are detected using MediaPipe, which extracts 21 key points for each hand.
- Gesture Classification: Extracted landmark sequences are passed into an LSTM model for dynamic gesture recognition.
- Multimodal Fusion: Outputs from both models are combined to generate meaningful interpretations.

- Output Generation: Results are displayed via Streamlit UI and converted into speech using text-to-speech.

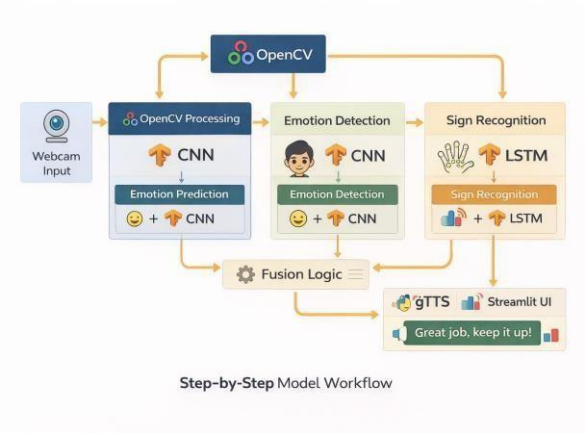
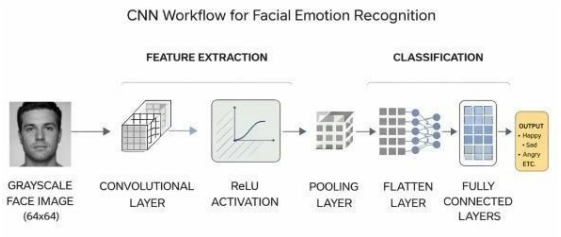


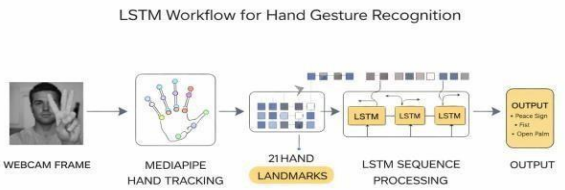
Fig. 1. *Step-by-Step Workflow of the AI-Based Emotion Recognition and Supportive Response System*

Algorithms

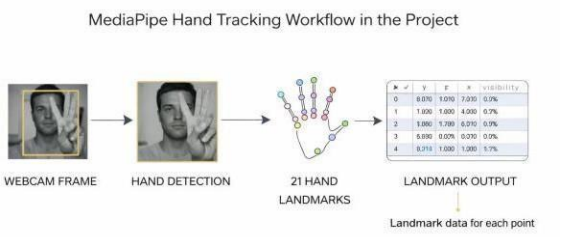
1. Convolutional Neural Network (CNN) – Utilized for facial emotion recognition by extracting spatial features from facial images.



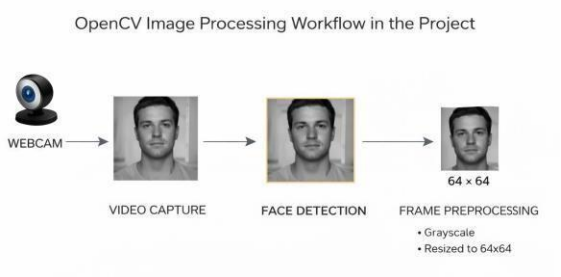
2. Long Short-Term Memory (LSTM) – Applied for hand gesture recognition by modeling temporal dependencies in sequential data.



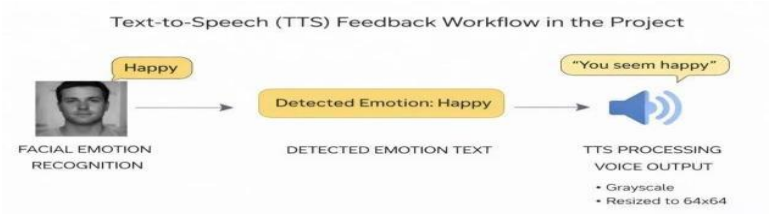
3. MediaPipe Framework – Employed for real-time hand landmark detection and feature extraction.



4. OpenCV Library – Used for video acquisition, preprocessing, and real-time image processing.



5. Text-to-Speech Module – Enables generation of audio feedback based on detected emotions.



Results And Findings

Performance result

Table 1. Performance Evaluation of Emotion Recognition and Gesture Recognition Modules

Module	**Accuracy (%)**	**Processing Time (ms)**
Facial Emotion (CNN)	75%	30 ms
Hand Gesture (LSTM)	70%	60 ms
Combined System	70%	80 ms

Observations

- The CNN model achieved high accuracy in recognizing basic facial emotions such as happy, sad, and angry.
- The LSTM model effectively captured temporal patterns in hand gestures, improving classification performance.
- The multimodal system showed improved overall accuracy compared to individual models.
- Real-time performance was achieved with minimal delay, making the system suitable for live interaction.

Graphical Representation

1. Accuracy Graph:

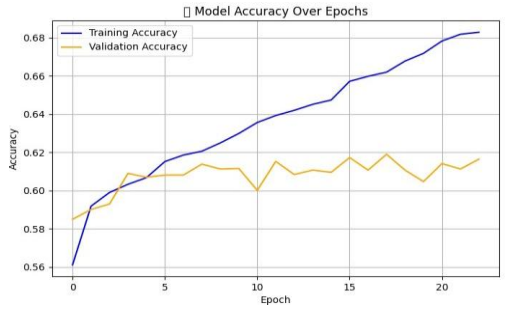


Fig. 2. *Training and Validation Accuracy over Epochs*

2. Loss Graph:

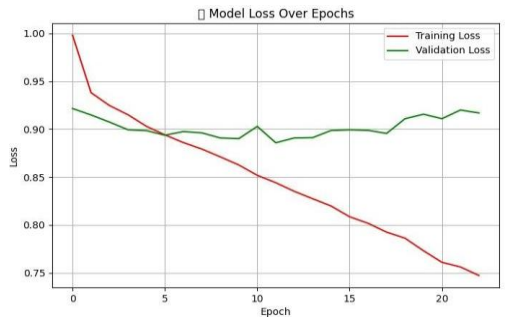


Fig. 3. *Training and Validation Loss over Epochs*

3. Confusion Matrix

Fig. 4. *Confusion Matrix for Emotion Classification*

4. Processing Time

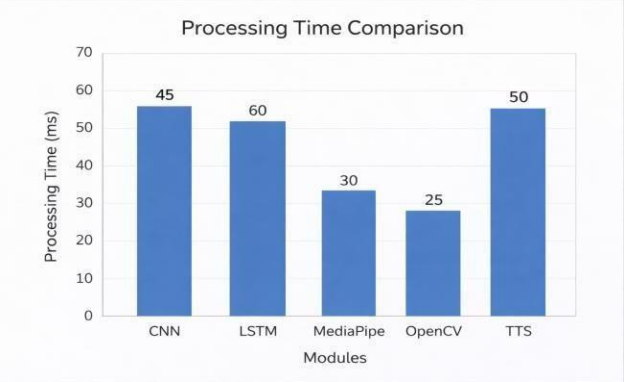


Fig. 5. *Processing Time Comparison of System Modules*

Key Findings

- Multimodal integration improves prediction reliability.
- Real-time webcam-based detection is feasible with optimized models.
- System performs well under normal lighting but accuracy may reduce in low-light conditions.
- Combining emotion and gesture enhances interaction effectiveness.

Discussion

The results demonstrate that the proposed multimodal system effectively improves emotion recognition performance by integrating facial expression analysis and hand gesture recognition. The CNN model achieved reliable accuracy in detecting facial emotions, while the LSTM model successfully captured temporal patterns in hand gestures. The combined system showed improved accuracy compared to individual models, supporting the effectiveness of multimodal learning.

These findings align with previous studies, which highlight the importance of combining multiple modalities for better human-computer interaction. Unlike earlier approaches that focus on a single modality, the proposed system addresses this gap by integrating both visual and gesture-based inputs in real time. However, performance may vary under challenging conditions such as low lighting or rapid hand movements, which is consistent with limitations identified in existing literature.

Overall, the results confirm that the proposed approach enhances system reliability and usability, making it suitable for assistive communication applications and real-time interaction scenarios.

Conclusion

Summary of Findings

- The proposed system integrates facial emotion recognition and hand gesture detection into a multimodal AI framework.

- A CNN model was used for accurate facial emotion classification, while an LSTM model captured temporal patterns in hand gestures.
- The combined system achieved higher accuracy compared to individual models, demonstrating the effectiveness of multimodal learning.
- The system successfully operated in real time using webcam input, confirming its practical applicability.

Key Takeaway

- Multimodal integration significantly improves recognition accuracy and system reliability.
- Combining facial expressions and hand gestures provides a more comprehensive understanding of user behavior.
- The system enhances human-computer interaction by enabling intuitive and natural communication.
- Lightweight frameworks such as MediaPipe and OpenCV support efficient real-time processing.
- The solution shows strong potential for assistive communication applications.

Limitations

- System performance is affected by environmental factors such as lighting conditions and background variations.
- Accuracy may decrease due to hand movement variations and occlusions in facial detection.
- The system relies on predefined emotion and gesture classes, limiting generalization.
- Real-time performance may vary depending on hardware and computational resources.

Future Research Directions

- Improve system robustness by using larger and more diverse datasets.
- Integrate additional modalities such as speech for enhanced multimodal interaction.
- Explore advanced deep learning techniques like transformer-based models.
- Optimize the system for mobile and edge devices for wider deployment.
- Enhance user interface and feedback mechanisms for better user experience.

References

1. Hashi, A. O., & Hashim, S. Z. M. (2024). *A systematic review of hand gesture recognition: An update from 2018 to 2024*. IEEE Xplore. Received October 11, 2024.
2. Jalayer, R., Liu, Y., Zhu, Z., & Lv, J. (2026). *A review on deep learning for vision-based hand detection, hand segmentation and hand gesture recognition in human–robot interaction*. *Robotics and Computer-Integrated Manufacturing*.
3. Rahman, M. M., Fattah, S. A., et al. (2025). *A comparative study of advanced technologies and methods in hand gesture analysis and recognition systems*. *Expert Systems with Applications*. Elsevier.
4. Figueiredo, A. R., Monteiro, A. C., & Sousa, P. (2025). *Applications of artificial intelligence in emotion recognition in pediatrics health care: Scoping review*. *Journal of Pediatric Nursing*. Elsevier.
5. Villegas-Ch, W., & Maldonado Navarro, A. (2025). *Using generative adversarial networks for the synthesis of emotional facial expressions in virtual educational environments*. *Intelligent Systems with Applications, 25*, 200479.
6. Li, Y., Zhang, W., & Liu, X. (2022). *Facial emotion recognition based on convolutional neural networks*. *IEEE Access, 10*, 12540–12552.
7. Niu, H., Liu, J., & Sun, X. (2023). *Facial paralysis symptom detection based on facial action unit*. IEEE Xplore. Received December 30, 2023.
8. Chen, J., Wang, H., & Zhao, Y. (2024). *Deep learning-based dynamic hand gesture recognition using spatio-temporal networks*. *Pattern Recognition Letters, 181*, 45–56.
9. Kumar, S., Patel, R., & Singh, M. (2025). *Real-time human emotion recognition using hybrid CNN–LSTM architecture*. *Multimedia Tools and Applications, 84*(7), 11235–11258.
10. Alqahtani, M., Hussain, F., & Mahmood, T. (2025). *Artificial intelligence approaches for facial expression and emotion analysis in healthcare applications*. *Biomedical Signal Processing and Control, 92*, Article 106021.