International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05965
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

AI Hand Gesture and Voice Controlled Interface

K. Esha[1], **N. Bhavani**[2], **S. Hemanth**[3], **D. Bhaskar**[4]

Supervisor: **B. Rajasekharam M-Tech (Ph.D)**, Assistant Professor, Dept. of CSE, VIET

[1]*Department of CSE (AIML), Visakha Institute of Engineering and Technology, Andhra Pradesh, India*
[2]*Department of CSE (AIML), Visakha Institute of Engineering and Technology, Andhra Pradesh, India*
[3]*Department of CSE (AIML), Visakha Institute of Engineering and Technology, Andhra Pradesh, India*
[4]*Department of CSE (AIML), Visakha Institute of Engineering and Technology, Andhra Pradesh, India*

Abstract - This paper presents an AI Hand Gesture and Voice Controlled Interface that enables users to interact with computers through real-time hand gestures and spoken commands, eliminating dependency on traditional input devices. The system uses computer vision via OpenCV and MediaPipe for hand landmark detection, and the SpeechRecognition library for processing voice commands. Recognized inputs are mapped to system actions using PyAutoGUI, enabling operations such as application control, media playback, scrolling, and volume adjustment. The system achieves an average gesture recognition accuracy of 90–95% with a response time of 50–100 ms, and voice recognition accuracy of 88–93% with a 1–2 second response window. Tested under varied conditions, the system demonstrates reliable, real-time performance on standard hardware without cloud dependency, making it suitable for assistive technologies, smart environments, and contactless computing.

Key Words: *Artificial Intelligence, Hand Gesture Recognition, Voice Control, Human-Computer Interaction, Computer Vision, Speech Recognition.*.

1. INTRODUCTION

Conventional human-computer interaction relies heavily on keyboards and mice, which impose physical constraints and limit accessibility, particularly in hands-free or contactless scenarios. The rapid growth of artificial intelligence and computer vision has enabled new paradigms of natural interaction, where users can communicate with systems using intuitive movements and speech.

This paper presents the design and implementation of an AI Hand Gesture and Voice Controlled Interface that integrates real-time gesture recognition and voice command processing into a unified multimodal system. Built using Python, OpenCV, MediaPipe, SpeechRecognition, and PyAutoGUI, the system runs on standard hardware without requiring cloud infrastructure, delivering a practical, accessible, and touchless interaction experience.

1.1 Problem Statement

Existing gesture or voice control systems typically address only one modality, lack real-time responsiveness, or suffer from poor accuracy under variable environmental conditions such as changing lighting or background noise. There is a clear need for a unified system that combines both gesture and voice inputs with high accuracy and fast response time. Moreover, most current solutions require specialized hardware or cloud connectivity, limiting their accessibility. The proposed system addresses these gaps by delivering a lightweight, offline-capable, multimodal interface that operates efficiently on standard consumer hardware.

1.2 Scope of the Project

The project encompasses three core modules: a Hand Gesture Recognition Module using a webcam to detect and interpret hand movements in real time; a Voice Command Processing Module that captures and processes spoken instructions via microphone; and a System Control & Execution Module that maps recognized inputs to system-level actions including application launching, media control, volume adjustment, and browser navigation.

The scope also includes performance evaluation under varying environmental conditions and a modular design that supports future enhancements such as expanded gesture sets, multilingual voice support, and deep learning-based recognition models.

1.3 Significance of Research

This research contributes to the advancement of natural, accessible, and intuitive human-computer interaction. The system is particularly relevant for assistive technology applications, smart environments, and contactless computing — areas where reducing physical input dependency improves both usability and hygiene. The integration of AI-based perception with rule-based execution demonstrates a practical implementation of machine learning in everyday computing.

2. LITERATURE REVIEW

Significant research has been conducted on gesture-based and voice-based human-computer interaction. Zhang et al. (2019) demonstrated the effectiveness of computer vision for real-time hand gesture recognition, while Smith et al. (2021) explored speech recognition for hands-free device control, both achieving improved accessibility. Rautaray and Agrawal (2015) provided a comprehensive survey of vision-based gesture recognition, emphasizing the importance of hand landmark detection. Mitra and Acharya (2007) further studied gesture and speech fusion, identifying integration challenges in multimodal systems.

However, most existing systems address gesture or voice control in isolation, lacking a unified multimodal approach. Challenges related to lighting, background noise, and environmental variability remain

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05965
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

largely unresolved. The work of Turk (2014) on voice and gesture-based HCI, and Wilson and Shafer on touchless interaction systems, underscore the necessity of combining both modalities into a single, robust, real-time interface — the core contribution of the proposed system.

3. SYSTEM DESIGN AND METHODOLOGY

3.1 System Architecture

The system follows a three-layer architecture: a Frontend UI Module, a Backend Logic Unit, and an AI Processing Core. The UI module displays real-time feedback including detected gestures, recognized commands, and system responses. The backend manages data flow and coordinates input processing and command mapping. The system operates in a continuous loop — capturing video and audio simultaneously, processing through AI models, and executing the corresponding action.

The AI Processing Unit forms the system's core intelligence. For gesture recognition, MediaPipe's hand landmark model detects 21 key points per frame, and Euclidean distance calculations between finger joints identify specific gestures. For voice processing, SpeechRecognition converts microphone audio to text using Google's pre-trained model. A rule-based decision engine then maps recognized inputs to system actions executed via PyAutoGUI, psutil, and win32gui.

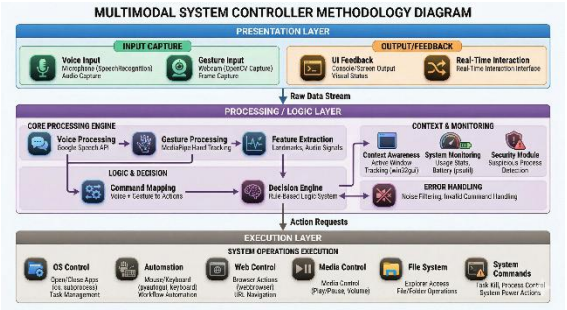

Fig. No .1 Methodology Diagram

3.2 Modules

The Gesture Recognition Module continuously tracks hand landmarks using OpenCV and MediaPipe, mapping gestures such as open palm, closed fist, pinch, and directional swipes to system commands. The Voice Command Processing Module captures audio via PyAudio and converts speech to text, filtering background noise to improve recognition reliability in varied environments.

The Command Execution Module is responsible for performing system-level actions including launching applications, controlling media playback, adjusting volume, scrolling, and opening web pages. It ensures smooth coordination between input modules and execution, with validation mechanisms such as multi-frame gesture confirmation and confidence thresholds for voice commands, reducing false triggers.

3.3 Technologies Used

Python serves as the primary language. OpenCV handles real-time video capture and frame preprocessing. MediaPipe provides lightweight pre-trained hand landmark detection. SpeechRecognition and PyAudio manage audio capture and speech-to-text conversion. PyAutoGUI simulates keyboard and mouse operations. Supporting modules include psutil for system information, win32gui for active window detection, the math module for geometric calculations, and the webbrowser module for web navigation.

The system requires no cloud integration or external database, running entirely on-device. Minimum hardware requirements are an Intel i3 processor, 8 GB RAM, a webcam, and a microphone, running Windows 10/11. This ensures that the system is lightweight, portable, and accessible to a broad range of users without specialized equipment.

4. RESULTS AND DISCUSSIONS

4.1 Gesture Recognition Accuracy

The gesture recognition module was evaluated across multiple test scenarios involving predefined gestures: open palm, closed fist, pinch, and directional swipes. Under standard conditions — proper lighting and uncluttered background — the system achieved an average accuracy of 90–95%, with a response time of 50–100 ms per frame. Multi-frame confirmation significantly reduced false detections.

Accuracy varied under adverse conditions. Low-light environments reduced hand landmark visibility, causing a measurable accuracy drop. Rapid movements and partial finger occlusion introduced occasional misclassifications. Despite these challenges, the system maintained acceptable performance owing to MediaPipe's robust landmark detection model, confirming suitability for real-time touchless interaction in standard environments.

4.2 Voice Recognition Accuracy

The voice command module was tested with multiple users exhibiting varied tones, accents, and speaking speeds. In quiet environments, the system achieved an average accuracy of 88–93%, with a response time of 1–2 seconds from speech input to command execution. Confidence-based matching ensured that low-confidence inputs were discarded rather than incorrectly executed.

In moderately noisy environments, a slight reduction in accuracy was observed due to audio signal degradation. Commands with similar phonetic structures also introduced minor misrecognitions. Future enhancements including advanced noise cancellation, adaptive noise thresholds, and multilingual model support are expected to address these limitations and improve performance under challenging acoustic conditions.

4.3 Command Execution Performance

The command execution module demonstrated high precision, achieving an execution accuracy of 95–98% under normal operating conditions. Gesture-based execution was near-instantaneous, while voice-based execution maintained a 1–2 second response window acceptable for real-time use. The system handled sequential and simultaneous command inputs without conflicts, owing to efficient module coordination.

User experience evaluation revealed that the system was intuitive and easy to operate with minimal prior training. Participants particularly appreciated hands-free control in contactless scenarios.

Minor usability challenges were noted in noisy or low-light environments; however, overall user satisfaction remained high. The system's lightweight design ensured stable performance on standard hardware throughout all test scenarios.

4.4 AI Model Performance Summary

The overall AI model demonstrated stable, efficient performance across all test conditions. Gesture recognition utilized MediaPipe's pre-trained landmark model enhanced with rule-based gesture classification, while voice recognition leveraged Google's speech-to-text API with confidence thresholding. Both models are lightweight and optimized for real-time execution, requiring no GPU or high-end hardware.

The modular design supports future integration of deep learning models for improved robustness, personalized gesture patterns, and extended command vocabularies. The current implementation confirms that AI-based multimodal interaction can be achieved with high accuracy and responsiveness on standard consumer devices, validating the practical applicability of the proposed system.

5. CONCLUSIONS

This paper presented the design, development, and evaluation of an AI Hand Gesture and Voice Controlled Interface enabling real-time, touchless human-computer interaction through integrated computer vision and speech recognition technologies. The system achieved gesture recognition accuracy of 90–95% and voice recognition accuracy of 88–93% under standard conditions, with response times suitable for practical real-time applications.

The modular architecture, lightweight implementation, and independence from cloud infrastructure make the system readily deployable on standard hardware. Future work will focus on expanding the gesture and command vocabulary, integrating deep learning models, supporting multilingual voice inputs, and enhancing robustness under challenging environmental conditions. The system provides a strong foundation for next-generation touchless interface research and development in assistive technology and smart computing.

ACKNOWLEDGEMENT

The authors express gratitude to B. Rajasekharam, Associate Professor, Dept. of CSE-AIML, VIET, for expert guidance throughout this project. Sincere thanks to Dr. P. Lalitha Kumari, Professor & HOD, for providing necessary facilities and institutional support, and to Dr. G. V. Pradeep Varma, Principal of VIET, for a conducive research environment. The authors also thank all faculty members and lab staff of the department for their cooperation and timely assistance during the course of this project.

REFERENCES

[1] Google Research, "MediaPipe Hands: On-device Real-time Hand Tracking," 2020.

[2] G. Bradski, "OpenCV: Open Source Computer Vision Library," 2000.

[3] S. S. Rautaray and A. Agrawal, "Real-Time Hand Gesture Recognition Using Machine Learning," Artificial Intelligence Review, 2015.

[4] S. Mitra and T. Acharya, "Vision-Based Hand Gesture Recognition: A Survey," IEEE Transactions, 2007.

[5] G. Hinton et al., "Speech Recognition Using Deep Neural Networks," IEEE Signal Processing Magazine, 2012.

[6] L. Rabiner and B. Juang, "An Overview of Speech Recognition Technology," IEEE Signal Processing Magazine, 1993.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning for Computer Vision," Nature, 2015.

[8] M. Turk, "HCI Using Voice and Gesture," Communications of the ACM, 2014.

[9] A. Wilson and S. Shafer, "Advances in Touchless Interaction Systems," Microsoft Research.

[10] J. Preece, Y. Rogers, and H. Sharp, "Human-Computer Interaction: Beyond the Desktop," Wiley, 2015.

[11] S. Russell and P. Norvig, "Artificial Intelligence: A Modern Approach," Pearson, 2016.

[12] A. Zhang, "SpeechRecognition Library Documentation," 2022.

[13] D. Wigdor and D. Wixon, "Designing Natural User Interfaces Using Gestures and Voice," Morgan Kaufmann, 2011.

[14] R. Szeliski, "Real-Time Interaction Systems Using Computer Vision," Springer, 2010.

[15] Pothuraju V V Satyanarayana et al., "AI-Powered Recommender Systems for E-Commerce," IEEE Proceedings of the International Conference on Recent Innovations in Science, Engineering and Technology (ICRISET-2025), ISSN: 0094-243X, E-ISSN: 1551-7616, 2025.

[16] Pothuraju V V Satyanarayana et al., "Next-Generation National Voting Framework Using Aadhaar-Integrated Decentralized Blockchain and Analytics," Global Journal of Research in Engineering & Computer Sciences, ISSN: 2583-2727.

[17] Pothuraju V V Satyanarayana et al., "Smart Contract-Based Decentralized Voting System for Transparent Elections on Ethereum Blockchain," International Scientific Journal of Engineering and Management, ISSN: 2583-6129.