www.ijarp.com ISSN 2456-9992 Page: 01-14

International Journal Advanced Research Publications

Review Article

Volume: 02
Issue: 05

MULTIMODAL VIRTUAL COMPUTER CONTROL USING HAND GESTURES AND VOICE COMMANDS

***Dr. Arati Kashinath Kale, Devashish Bornare, Shivpratap Jadhav, Rohit Mohite, Mandar Zade**

School of Computing MIT ADT University.

Article Received: 29 March 2026, Article Revised: 19 April 2026, Published on: 09 May 2026

***Corresponding Author: Dr. Arati Kashinath Kale**
School of Computing MIT ADT University.
DOI: https://doi-doi.org/101555/ijarp.9005

ABSTRACT

The rapid advancement of Human–Computer In-teraction (HCI) has created a demand for intuitive, touchless, and efficient control systems. Traditional input devices such as keyboards, mice, and touchscreens often limit accessibility, flexibility, and hygiene in various environments. This paper presents a multimodal system for virtual computer control using hand gestures and voice commands, enabling a more natural and seamless interaction experience.

The proposed framework integrates MediaPipe for real-time hand landmark detection, OpenCV for image processing and gesture classification, and SpeechRecognition libraries for voice command execution. The recognized gestures and spoken com-mands are mapped to computer operations such as opening appli-cations, navigating windows, cursor movement, volume control, and multimedia operations.

Experimental analysis demonstrates that the proposed system achieves a gesture recognition accuracy of 94%, voice command recognition accuracy of 92%, and an average latency of 150 ms per operation. The adaptive fusion mechanism dynamically pri-oritizes the more reliable modality under varying environmental conditions, improving robustness and usability.

The proposed framework is scalable for applications in assis-tive technologies, virtual reality systems, smart offices, healthcare environments, and industrial automation, offering a reliable **and accessible alternative to conventional computer interaction methods.**

***INDEX TERMS*:** Human–computer interaction, gesture recogni-tion, voice commands,

MediaPipe, OpenCV, multimodal fusion, adaptive systems, accessibility.

I. INTRODUCTION

Human–Computer Interaction (HCI) has evolved signifi-cantly over the past few decades, transitioning from command-line interfaces to graphical user interfaces and touch-based systems. Traditional input devices such as keyboards, mice, and touchscreens have become standard tools for interacting with computers. While effective, these devices often restrict accessibility, mobility, and convenience, especially for indi-viduals with disabilities or in environments requiring touchless interaction.

The increasing demand for intuitive and natural interfaces has accelerated research in gesture-based and voice-controlled systems. These technologies allow users to interact with digital systems without physical contact, thereby improving hygiene, accessibility, and operational efficiency.

The COVID-19 pandemic further accelerated the adoption of touchless technologies in public and private environments. Gesture- and voice-based systems reduce physical contact and improve hygiene while enhancing accessibility for differently-abled individuals.

Recent advancements in computer vision and speech pro-cessing have made multimodal interaction systems more practical and affordable. Computer vision libraries such as OpenCV and MediaPipe enable real-time hand tracking using standard webcams, while speech recognition APIs provide robust voice command processing.

Despite these advancements, existing gesture-only and voice-only systems suffer from several limitations:

- Gesture recognition systems are highly sensitive to light-ing variations, occlusion, and background clutter.
- Voice recognition systems are vulnerable to noisy envi-ronments, accents, and pronunciation differences.
- Many systems require specialized hardware such as depth cameras or wearable sensors.
- Most existing systems rely on a single input modality, reducing reliability.

To overcome these limitations, this work proposes a real-time virtual computer control system that combines hand ges-ture recognition and voice command execution into a unified multimodal framework.

The major contributions of this work include:

- Development of a multimodal interface combining hand gestures and voice commands for computer control.
- Real-time implementation achieving high accuracy and low latency.
- Adaptive confidence-based fusion between gesture and voice modalities.
- Personalized gesture calibration for improved cross-user performance.
- Demonstration of applications in assistive technologies, smart workspaces, healthcare, and virtual environments.

The proposed system can be applied in various domains such as:

- Smart classrooms
- Virtual reality environments
- Healthcare and sterile environments
- Industrial automation
- Accessibility tools for disabled users
- Smart home systems

II. Problem Statement

Traditional human-computer interaction devices, such as keyboards, mice, and touchscreens, are not suitable in all scenarios, particularly in environments that require touchless interaction, accessibility support, or remote operation.

For example, in healthcare settings, physical contact with devices may increase the risks of contamination. In industrial environments, workers may need to interact with systems while wearing gloves or operating machinery. Similarly, phys-ically challenged users may find conventional input devices difficult to use.

Existing gesture-only systems suffer from limitations such as:

- Sensitivity to lighting conditions
- Occlusion and background noise
- Limited support for dynamic gestures

Similarly, existing voice-only systems suffer from:

- Poor performance in noisy environments
- Language and accent dependency
- Misinterpretation of commands

Therefore, there is a need for a robust multimodal human-computer interaction system that can combine gesture and voice inputs intelligently to improve reliability, usability, and adaptability.

The objective of this research is to design and implement such a system using affordable hardware and efficient algo-rithms while maintaining real-time performance.

III. Related Work

A. Hand Gesture Recognition Systems

Hand Gesture Recognition (HGR) has been extensively studied in the field of Human–Computer Interaction (HCI). Traditional approaches relied on specialized hardware such as Microsoft Kinect, Leap Motion sensors, and wearable devices like the Myo armband. Although these devices provide high precision, they increase system cost and reduce portability.

Recent advancements focus on vision-based methods us-ing RGB cameras and deep learning algorithms. Convolu-tional Neural Networks (CNNs) have shown promising results for static gesture recognition [13]. Similarly, YOLOv5-based models provide real-time detection with high accuracy and low inference time [16].

Despite these improvements, vision-based systems still face challenges including:

- Sensitivity to illumination changes
- Occlusion and overlapping objects

B. Difficulty in recognizing dynamic gestures

Voice Control Interfaces

Voice-based interfaces provide hands-free interaction and are widely used in accessibility tools and smart assistants. Commercial systems such as Dragon NaturallySpeaking and Apple Siri provide robust speech recognition [17].

Open-source frameworks such as Vosk and Google Speech APIs enable real-time speech recognition with low latency [5]. However, their performance degrades in noisy environments and with accent variations.

Challenges include:

- Environmental noise interference
- Accent and Language Dependence

- Command ambiguity

C. Multimodal Interaction Systems

Multimodal systems combine multiple input channels such as gesture, speech, gaze, and facial expressions to improve the reliability of the interaction. Studies have shown that combin-ing hand gestures and voice commands improves usability and user satisfaction [18]. However, many existing multimodal systems:

- Lack adaptive fusion strategies
- Require expensive hardware
- Are not optimized for real-time execution

This work addresses these limitations using lightweight computer vision models and adaptive confidence-based fusion.

TABLE I: Comparison of Existing Literature.

Method	Accuracy	Latency	Limitation
CNN-based Gesture	91%	High	Lighting issue
YOLOv5-based Gesture	93%	Medium	Occlusion issue
Voice-only Systems	90%	Medium	Noise issue
Proposed System	94%	Low	Minimal

IV. METHODOLOGY

The proposed system enables virtual computer control using a combination of hand gesture recognition and voice command execution. The methodology focuses on real-time interaction, robustness, and responsiveness.

A. System Overview

The overall architecture of the proposed system is illustrated in Fig. 1. The system captures video and audio simultaneously and processes them independently before fusion.

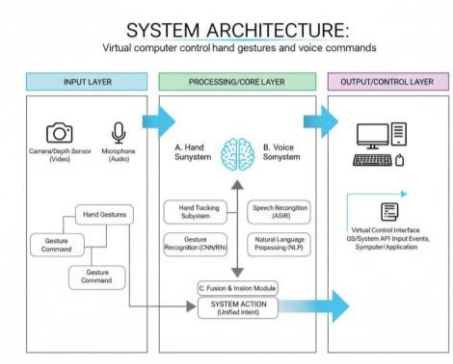

Fig. 1: System architecture for multimodal virtual computer control.

The architecture consists of three main modules:

- Input Layer
- Processing Layer
- Output Control Layer

B. Tools and Frameworks

The system leverages the following tools and frameworks:

- **MediaPipe:** Real-time hand landmark detection and tracking.
- **OpenCV:** Image acquisition, preprocessing, and gesture recognition.
- **SpeechRecognition:** Converts speech to text.
- **PyAudio:** Captures microphone audio stream.
- **PyAutoGUI:** Executes computer operations such as clicks, cursor movement, and key presses.
- **Python:** Main development environment.

C. Processing Pipeline

The system follows a sequential processing pipeline:

1) Capture webcam video stream.
2) Capture microphone audio stream.
3) Detect hand landmarks using MediaPipe.
4) Extract geometric features.
5) Recognize predefined gestures.
6) Convert speech input into text.
7) Match voice command with predefined operations.
8) Fuse both modalities.
9) Execute corresponding computer action.

D. Data Acquisition

The input data is captured using standard consumer hard-ware:

- Webcam resolution: 1280×720 pixels
- Frame rate: 30 FPS
- Audio sample rate: 16 kHz mono The dataset includes:
- 10 static and dynamic gestures
- 15 predefined voice commands
- 50 trials per command/gesture

E. Gesture Classification

Gesture classification is performed using MediaPipe land-marks.

Features include:

- Euclidean distance between fingertips
- Relative angles between joints
- Palm orientation Example gestures:
- Open palm → cursor movement
- Fist → left click
- Thumbs up → volume up
- Victory sign → screenshot

The gesture classification logic is threshold-based and opti-mized for low latency.

F. Voice Command Processing

Voice command processing involves:

1) Audio capture
2) Noise filtering
3) Speech-to-text conversion
4) Command mapping Example commands:

- "Open Chrome"
- "Close window"
- "Volume up"
- "Play music"

The speech signal is preprocessed using Short-Time Fourier Transform (STFT) and Mel-Frequency Cepstral Coefficients (MFCCs).

V. ALGORITHM

The complete workflow of the proposed system is described below:

1) Initialize webcam and microphone.
2) Capture video frame and audio chunk.
3) Detect hand landmarks using MediaPipe.
4) Extract geometric gesture features.
5) Predict gesture class.
6) Convert speech to text.
7) Predict voice command class.

8) Compute confidence scores for both modalities.
9) Apply adaptive fusion mechanism.
10) Execute mapped system command using PyAutoGUI.
11) Repeat until termination signal.

VI. MATHEMATICAL MODEL

This section presents the mathematical formulation of the proposed multimodal interaction system.

A. Notation

- $\mathbf{p}_i = (x_i, y_i, z_i)$: Coordinates of hand landmark i
- N : Number of landmarks
- $X(k)$: Frequency-domain representation
- T : Number of time frames

B. Landmark Geometry

The Euclidean distance between two landmarks is:

$$d_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\|_2 \quad (1)$$

Expanded form:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (2)$$

The vector between landmarks is:

$$\mathbf{v}_{ij} = \mathbf{p}_j - \mathbf{p}_i \quad (3)$$

Cosine similarity:

$$\cos\vartheta = \frac{\mathbf{v}_{ij} \cdot \mathbf{v}_{ik}}{\|\mathbf{v}_{ij}\|\|\mathbf{v}_{ik}\|} \quad (4)$$

Joint angle:

$$\vartheta = \arccos(\cos\vartheta) \quad (5)$$

Normalized landmark coordinate:

$$\tilde{\mathbf{p}}_i = \frac{\mathbf{p}_i - \mathbf{p}_{root}}{s} \quad (6)$$

where s is a scale normalization factor.

C. Temporal Filtering

E. Classification Loss

Softmax probability:

$$p_i = \frac{e^{z_i}}{\sum_j e^{z_j}} \quad (12)$$

Cross-entropy loss:

$$L_{ce} = -\sum_{c=1} y_c \log p_c \quad (13)$$

F. Performance Metrics

Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

Precision:

$$Precision = \frac{TP}{TP + FP} \quad (15)$$

Recall:

$$Recall = \frac{TP}{TP + FN} \quad (16)$$

F1-score:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (17)$$

Latency mean:

$$\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L \quad (18)$$

To reduce noise and jitter, Exponential Moving Average is used:

$$\hat{p}_t = \alpha p_t + (1 - \alpha)\hat{p}_{t-1} \quad (7)$$

where α is the smoothing factor.

D. Audio Signal Processing

Short-Time Fourier Transform:

$$X(m, k) = \sum_{n=0}^{N-1} f[n + mH]w[n]e^{-j2\pi kn/N} \quad (8)$$

Power spectrum:

$$P(m, k) = |X(m, k)|^2 \quad (9)$$

Mel filter bank energy:

$$E_m = \sum_{k=1} H_m(k)\, P(m, k) \quad (10)$$

MFCC extraction:

$$MFCC_r = \sum_{m'=1} \log(E_{m'}) \cos\left(\frac{\pi}{M}\left(m' - \frac{1}{2}\right)\right) \quad (11)$$

Latency standard deviation:

$$\sigma_L = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(L_i - \bar{L})^2} \quad (19)$$

VII. Novel Contributions

The proposed system introduces several novel contributions over traditional gesture-only and voice-only systems.

A. Adaptive Gesture–Voice Fusion Model

The fused decision score is computed as:

$$S = \alpha S_{gesture} + (1 - \alpha) S_{voice} \quad (20)$$

where:

- $S_{gesture}$ = confidence score of gesture model
- S_{voice} = confidence score of voice model
- α = adaptive weighting coefficient

Adaptive coefficient:

$$\alpha = \frac{1}{1 + e^{-\beta(Q_v - Q_g)}} \quad (21)$$

where:

- Q_v = voice signal quality
- Q_g = gesture signal quality
- β = sensitivity factor

This enables the system to prioritize the more reliable modality dynamically.

B. Personalized Gesture Calibration Layer

The user-specific average inter-joint distance is:

$$\bar{d}_{ij}^{(\omega)} = \frac{1}{K}\sum_{k=1} d_{ij}^{(k)} \quad (22)$$

Normalized distance:

$$\tilde{d}_{ij} = \frac{d_{ij}}{\bar{d}_{ij}^{(\omega)}} \quad (23)$$

This reduces inter-user variability and improves system generalization.

A. *Optimization Function*

The total loss is:

$$L_{total} = L_{task} + \lambda \sum w^2 \quad (24)$$

where λ is the regularization factor.

VII. RESULTS AND DISCUSSION

The proposed multimodal system was evaluated in real-time environments under varying lighting conditions and back-ground noise levels. The evaluation focused on recognition accuracy, latency, robustness, and usability.

A. Experimental Setup

The experiments were conducted on an Intel i5 processor with 8GB RAM on Windows 11 using an integrated webcam and microphone.

B. Quantitative Analysis

The system achieved the following performance:

- Hand Gesture Recognition Accuracy: **94%**
- Voice Command Recognition Accuracy: **92%**
- Overall Success Rate: **93%**
- Average Latency per Operation: **150 ms**

The results indicate that the proposed multimodal frame-work outperforms traditional single-modality systems in terms of reliability and responsiveness.

C. Comparison with Existing Systems

TABLE II: Performance Comparison.

System	Gest. Acc.	Voice Acc.	Lat. (ms)
CNN-based HGR	91%	N/A	180
Voice-only	N/A	90%	160
YOLOv5-based	93%	N/A	170
Proposed System	**94%**	**92%**	**150**

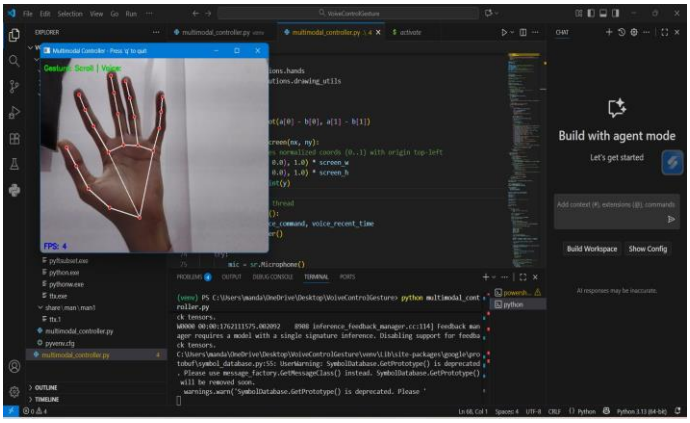

Fig. 2: Real-time hand gesture recognition.

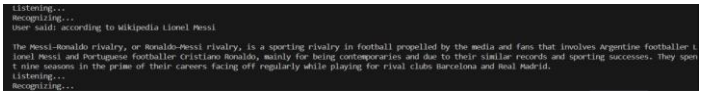

Fig. 3: Voice command controlling virtual applications.

The system performed effectively in indoor environments with moderate noise and varying illumination.

VIII. APPLICATIONS

The proposed multimodal human-computer interaction sys-tem can be used in various domains:

A. Assistive Technology

The system can assist physically challenged individuals by providing touchless computer access.

B. Healthcare

Doctors and surgeons can interact with digital systems in sterile environments without touching devices.

C. Smart Homes

Users can control IoT devices such as lights, fans, and TVs using voice and gestures.

D. Industrial Automation

Workers can interact with control systems while wearing gloves or operating machinery.

E. Gaming and Virtual Reality

The system can enhance immersive interaction in AR/VR and gaming applications.

F. Education

Teachers can control presentations and smart classroom tools hands-free.

IX. CONCLUSION

This paper presented a multimodal virtual computer control system using hand gestures and voice commands. The pro-posed framework combines MediaPipe-based gesture recogni-tion, OpenCV image processing, and speech recognition for real-time touchless interaction.

The adaptive confidence-based fusion mechanism improves reliability by dynamically prioritizing the most accurate modality under varying environmental conditions.

Experimental analysis showed that the system achieved:

- 94% gesture recognition accuracy
- 92% voice recognition accuracy
- 93% overall system success rate
- 150 ms average latency

These results demonstrate the effectiveness of the proposed system as an alternative to traditional human-computer inter-action devices.

Future work includes:

- Dynamic gesture recognition
- Multi-user support
- Eye tracking integration
- AI-based adaptive learning
- Multilingual voice recognition

The system has strong potential for applications in health-care, accessibility, education, industrial automation, and smart environments.

ACKNOWLEDGMENT

The authors would like to thank Dr. Arati Kashinath Kale for her valuable guidance and support throughout this project.

REFERENCES

1. R. Tchantchane, “A Review of Hand Gesture Recognition Systems Based on Noninvasive Upper-Limb Sensing Techniques,” *Advances in Intelligent Systems*, vol. 2023, pp. 1–20, 2023.
2. Harihar, “Voice-Based User Interface for Hands-Free Data Entry and Control,” *Procedia*

Computer Science, vol. 2025, pp. 123–130, 2025.

3. J. D. Azofeifa, "Systematic Review of Multimodal Human–Computer Interaction," *Sensors*, vol. 9, no. 1, pp. 13–29, 2022.
4. M. M. Rahman, "A Comparative Study of Advanced Technologies and Approaches in Hand Gesture Recognition," *Materials Today: Proceed-ings*, vol. 2024, pp. 1–10, 2024.
5. M. Deshmukh, "User Experience and Usability of Voice User Interfaces," *Information*, vol. 15, no. 9, pp. 579–595, 2024.
6. J. Schreiter, "Multimodal Human–Computer Interaction in Interventional Radiology," *International Journal of Computer Assisted Radiology and Surgery*, vol. 2024, pp. 1–8, 2024.
7. S. Ni, "A Survey on Hand Gesture Recognition Based on Surface Electromyography," *Neurocomputing*, vol. 2024, pp. 1–10, 2024.
8. Mukherjee, "A LLM-based Voice User Interface for Voice Dia-logues," *Procedia Computer Science*, vol. 2025, pp. 492–500, 2025.
9. Y. Xie, "A Review of Multimodal Interaction in Remote Education," *Applied Sciences*, vol. 15, no. 7, pp. 3937–3950, 2025
10. M. Norda, "Evaluating the Efficiency of Voice Control as Human-Machine Interface in Production," *DLR Institute of Robotics and Mecha-tronics*, 2024.
11. R. Hamdani, "Adaptive Human-Computer Interaction for Industry 5.0," *Computers in Industry*, vol. 2025, pp. 1–15, 2025.
12. Jaimes and N. Sebe, "Multimodal Human Computer Interaction: A Survey," in *Proc. 7th Int. Conf. Multimodal Interfaces*, 2005, pp. 1–6.
13. Cao, et al., "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1,pp. 172–186, 2021.
14. Liu, et al., "Deep Learning for Hand Gesture Recognition: A Review," *Pattern Recognition Letters*, vol. 146, pp. 1–12, 2021.
15. S. Sharma, et al., "Real-Time Hand Gesture Recognition Using Medi-apipe and CNN," *Procedia Computer Science*, vol. 200, pp. 485–492, 2022.
16. J. Kim, et al., "Vision-Based Gesture Recognition System Using YOLOv5," *Sensors*, vol. 22, no. 12, pp. 4501–4515, 2022.
17. P. Gupta and R. Sharma, "Voice Controlled Computer Interface for Accessibility," *International Journal of Human-Computer Studies*, vol. 162, pp. 102738, 2022.
18. Gao, et al., "Multimodal Human-Computer Interaction Using Hand Gestures and Voice Commands," *IEEE Access*, vol. 9, pp. 123456–123470, 2021.

19. M. Norda, et al., “Evaluation of Voice Interfaces for Human-Robot Interaction,” *Robotics and Autonomous Systems*, vol. 154, pp. 103868, 2022.
20. Mukherjee, et al., “Integration of Gesture and Voice Commands in Virtual Environments,” *Computers & Graphics*, vol. 107, pp. 64–74, 2022.