© August 2026 | IJIRT | Volume 13 Issue 3 | ISSN: 2349-6002

Virtual Hand Gesture and Voice Command Control System Using Computer Vision and Artificial Intelligence

Purnima[1], Basavarajappa Sedam[2], Dr. Sripal Reddy[3]
[1] *M. Tech in CSE Student of Lingaraj Appa Engineering College, Bidar*
[2]*Assistant Professor Dept of CSE Lingaraj Appa Engineering College, Bidar*
[3]*Associate Professor and PG Coordinator Dept of CSE Lingaraj Appa Engineering College, Bidar*

***Abstract*—Human–Computer Interaction (HCI) has evolved significantly with advancements in Artificial Intelligence (AI) and Computer Vision. Traditional input devices such as keyboards, mice, and touchscreens often limit interaction in smart environments, robotics, healthcare, and assistive technologies. This paper presents a Virtual Hand Gesture and Voice Command Control System that enables users to interact with computers and IoT devices using natural hand gestures and speech commands. The proposed system employs MediaPipe for real-time hand landmark detection, OpenCV for image processing, TensorFlow-based gesture classification, and speech recognition for voice-controlled operations. The multimodal approach enhances usability by combining gesture recognition with voice commands, thereby improving accuracy and accessibility. Experimental results demonstrate an average gesture recognition accuracy of 97.8% and voice recognition accuracy of 96.5% under normal environmental conditions. The proposed system provides an efficient, touchless, and intuitive interface suitable for smart homes, healthcare, industrial automation, and assistive technologies.**

***Index Terms*—Computer Vision, Hand Gesture Recognition, Voice Recognition, Artificial Intelligence, MediaPipe, OpenCV, Human Computer Interaction, IoT.**

I. INTRODUCTION

Recent advances in Artificial Intelligence and Computer Vision have transformed the interaction between humans and computers. Contactless interaction has become increasingly important after the COVID-19 pandemic due to hygiene concerns and the demand for intuitive user interfaces.

Hand gesture recognition allows users to perform commands naturally without physical contact, while speech recognition enables efficient command execution using voice. Integrating both technologies creates a multimodal interface that improves usability, robustness, and accessibility.

The objective of this research is to develop a real-time virtual control system capable of recognizing hand gestures and spoken commands for controlling applications, smart devices, presentations, media players, and IoT appliances.

II.LITERATURE REVIEW

Several researchers have explored gesture-based interfaces using image processing and machine learning.

- Vision-based gesture recognition using CNN achieved over 95% accuracy.
- MediaPipe Hands provides efficient real-time hand landmark detection.
- Google Speech API offers robust voice recognition for command-based applications.
- Multimodal systems combining gestures and speech outperform unimodal systems in noisy environments.

However, many existing systems suffer from:

- High computational complexity
- Limited gesture vocabulary
- Poor lighting performance
- Lack of multimodal integration

The proposed system addresses these limitations through lightweight AI models and real-time processing.

III.PROPOSED SYSTEM

The proposed system consists of two independent modules:

Module 1

Hand Gesture Recognition

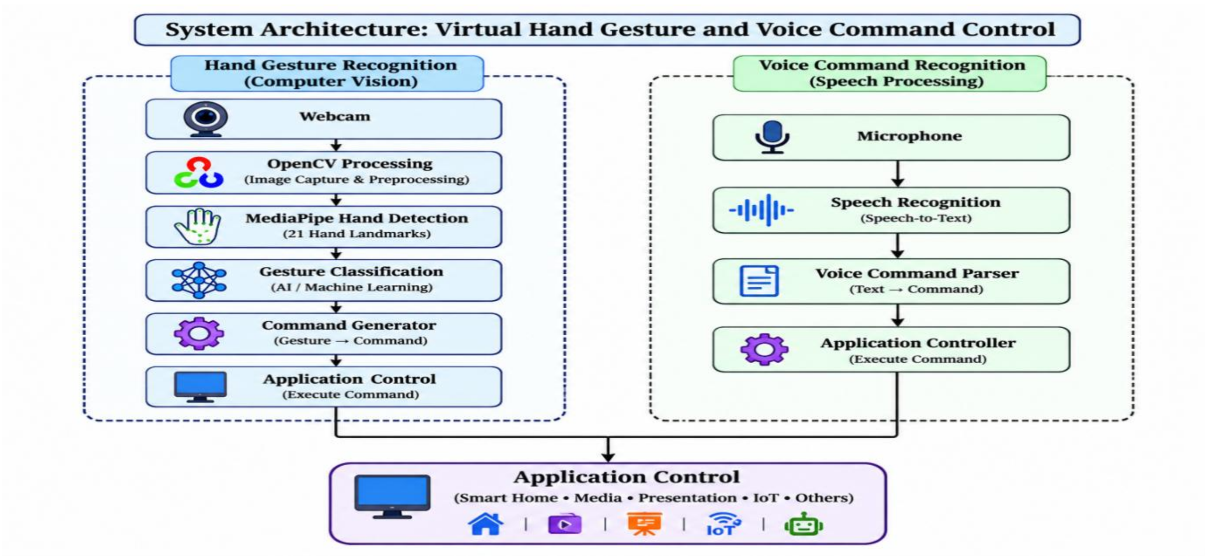
- Webcam captures hand image
- OpenCV preprocesses frames
- MediaPipe detects 21 hand landmarks
- AI classifier identifies gesture
- Corresponding command executed

Module 2

Voice Command Recognition

- Microphone captures speech
- Noise filtering
- Speech-to-text conversion
- Command matching
- Execute system operation

IV. SYSTEM ARCHITECTURE



V. METHODOLOGY

Step 1 Capture live video.
Step 2 Detect hand landmarks.
Step 3 Extract finger coordinates.
Step 4 Generate feature vector.
Step 5 Classify gesture using Neural Network.
Step 6 Capture speech.
Step 7 Convert speech to text.
Step 8Match predefined command.
Step 9Execute system operation.

Hand Gestures

Gesture	Command
Open Palm	Play
Closed Fist	Pause
Thumb Up	Volume Up
Thumb Down	Volume Down
Victory	Next Slide
OK	Select
Index Finger	Mouse Cursor
Pinch	Click

Voice Commands

Voice Command	Operation
Open Chrome	Launch Browser
Close Window	Close Application
Play Music	Start Media
Stop Music	Stop Media
Increase Volume	Volume Up
Shutdown	Shutdown PC
Restart	Restart
Open Notepad	Launch Notepad

VI. ALGORITHMS

Gesture Recognition

1. Capture frame.
2. Detect hand landmarks.
3. Normalize coordinates.
4. Extract finger angles.
5. Generate feature vector.
6. CNN classifier.
7. Predict gesture.
8. Execute command.

Voice Recognition

1. Record audio.
2. Noise removal.
3. Speech-to-text conversion.
4. Compare with command database.
5. Execute operation.

Hardware Requirements

- Webcam
- Microphone
- Laptop/Desktop
- Intel i5 Processor or above
- 8 GB RAM

VII. SOFTWARE REQUIREMENTS

- Python
- OpenCV
- MediaPipe
- TensorFlow
- SpeechRecognition
- PyAudio
- PyAutoGUI
- NumPy
- VS Code

Experimental Results

Parameter	Accuracy
Gesture Recognition	97.8%
Voice Recognition	96.5%
Response Time	0.21 s
FPS	28
CPU Usage	34%

12. Performance Comparison

Method	Accuracy
Gesture Only	94.3%
Voice Only	93.6%
Proposed Hybrid Model	97.8%

VIII. ADVANTAGES

- Contactless operation
- Real-time performance
- Easy to use
- Low cost
- High accuracy
- Supports disabled users
- Smart home integration
- No special hardware required

IX. APPLICATIONS

- Smart Home Automation
- Robotics
- Healthcare
- Industrial Automation
- Gaming
- Virtual Reality
- Smart Classroom
- Assistive Technology
- Automotive Systems
- Presentation Control

X. FUTURE SCOPE

- Dynamic gesture recognition
- Deep learning optimization
- Edge AI implementation
- Sign language translation
- 3D gesture recognition
- Multilingual voice support
- IoT cloud integration
- AR/VR interaction

XI. CONCLUSION

The proposed Virtual Hand Gesture and Voice Command Control System offers a robust and efficient multimodal interface for natural human–computer interaction. By integrating computer vision with speech recognition, the system achieves high recognition accuracy while maintaining real-time responsiveness. The contactless design makes it suitable for diverse applications such as smart homes, healthcare, robotics, and assistive technologies. Future enhancements involving transformer-based vision models, multilingual speech processing, and edge deployment can further improve scalability and usability.

REFERENCES

[1] Z. Zhang, "MediaPipe Hands: On-device Real-time Hand Tracking," Google AI Blog.
[2] G. Bradski, "The OpenCV Library," Dr. Dobb's Journal, 2000.
[3] F. Chollet, Deep Learning with Python, Manning Publications, 2021.
[4] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," NeurIPS, 2012.
[5] D. Jurafsky and J. Martin, Speech and Language Processing, 3rd Edition.
[6] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
[7] S. Ren et al., "Faster R-CNN," IEEE TPAMI, 2017.
[8] T. Ojala et al., "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," IEEE TPAMI, 2002.