IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2

AI Based Gesture Controlled Virtual Mouse with Voice Assistant Integration

Prof. U. B. Bhandage, Gauri Palde, Atharva Chavan, Swapnil Pawar
Pune Vidhyarti Griha's College of Engineering, Nashik, Maharashtra, India

Abstract: *The rapid advancement of Human Computer Interaction (HCI) technologies has led to the development of intelligent systems that enable users to interact with computers naturally and efficiently. Traditional input devices such as mouse and keyboard require physical contact and are limited in providing touchless interaction. This paper presents an AI Based Gesture Controlled Virtual Mouse with Voice Assistant Integration that allows users to control computer operations using hand gestures and voice commands. The proposed system uses computer vision techniques, MediaPipe hand tracking, machine learning algorithms, and voice recognition technologies to create a contactless and smart interaction environment.The system captures hand movements through a webcam and detects hand landmarks using MediaPipe. The extracted hand landmark features are processed using a trained machine learning model to identify various gestures such as cursor movement, left click, right click, drag, scroll, zoom in, zoom out, volume control, and screenshot operations. Additionally, the system integrates a voice assistant named Jarvis that can execute commands such as opening applications, searching the web, controlling system volume, and performing system-level tasks.*

Keywords: Artificial Intelligence, Computer Vision, Gesture Recognition, Human Computer Interaction, MediaPipe, Machine Learning, Virtual Mouse, Voice Assistant

I. INTRODUCTION

Human Computer Interaction (HCI) has become one of the most important research areas in modern computing systems. With the increasing demand for contactless and intelligent systems, researchers are continuously developing advanced methods that allow users to interact with machines naturally. Traditional input devices such as keyboards and mice require direct physical interaction, which may not always be convenient or hygienic. During recent years, touchless technologies have gained popularity due to their applications in healthcare, smart environments, and automation systems.

Gesture recognition technology is an important branch of computer vision that enables systems to understand human body movements and convert them into machine commands. Hand gesture recognition is particularly useful because hands are commonly used for communication and interaction. By analyzing hand movements through a webcam, computers can perform tasks without the need for physical devices.

In this project, an AI Based Gesture Controlled Virtual Mouse with Voice Assistant Integration is proposed. The system combines hand gesture recognition and voice control to provide a complete smart interaction platform. MediaPipe is used for hand landmark detection, while machine learning models classify different hand gestures. The recognized gestures are mapped to system operations using PyAutoGUI.

The system also includes a voice assistant called Jarvis that can execute voice commands such as opening applications, searching online content, controlling media, and performing system-level operations. The integration of voice and gesture technologies improves usability and creates a more intelligent and interactive environment.

The proposed system aims to reduce dependency on traditional hardware devices while providing accurate, real-time, and efficient interaction. The project can be applied in various domains including gaming, healthcare, automation, virtual reality, smart homes, education, and accessibility support for disabled individuals.

DOI: 10.48175/IJARSCT-35116

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2

II. LITERATURE SURVEY

Several researchers have worked on gesture recognition systems and virtual mouse technologies using computer vision and machine learning techniques. Existing systems mainly focus on hand tracking, cursor control, and click operations using webcams.

A gesture recognition system using OpenCV and color detection techniques was proposed to track colored markers attached to fingers. Although the system achieved acceptable results, it required external hardware markers and suffered from lighting sensitivity issues.

Another approach used deep learning and convolutional neural networks (CNNs) for gesture classification. These systems achieved higher accuracy but required large computational resources and powerful GPUs, making them less suitable for low-cost real- time applications.

MediaPipe-based hand tracking systems became popular because of their lightweight architecture and high-speed hand landmark detection. MediaPipe can detect 21 hand landmarks in real-time with high accuracy, making it suitable for gesture recognition applications.

Several virtual mouse systems have been developed using hand tracking and machine learning models such as KNN, SVM, Random Forest, and MLP classifiers. These systems mapped hand movements to cursor control and implemented click operations through gesture recognition.

Voice assistant systems such as Siri, Alexa, and Google Assistant inspired the integration of voice recognition into HCI systems. Speech recognition libraries like SpeechRecognition and Pyttsx3 allow offline and online voice command execution.

Existing systems usually focus only on gestures or only on voice interaction. Very few systems combine both gesture and voice technologies into a unified intelligent system. The proposed project addresses this gap by integrating gesture control and Jarvis-based voice assistance into a single platform.

The proposed system improves real-time interaction, reduces hardware dependency, enhances accessibility, and provides an efficient touchless control mechanism.

III. PROPOSED SYSTEM

The proposed system consists of two major modules: Gesture Recognition Module and Voice Assistant Module.

The Gesture Recognition Module uses a webcam to capture hand movements in real- time. MediaPipe Hands is used to detect hand landmarks from the video stream. The detected landmarks are converted into numerical features and passed to a machine learning model trained using gesture datasets.

The system recognizes multiple gestures including:

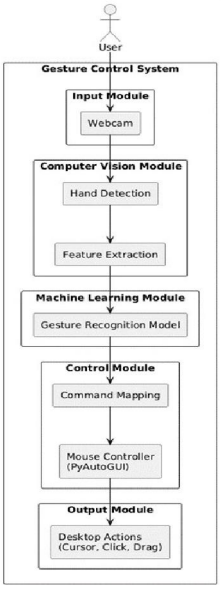
1. Cursor Movement
2. Left Click
3. Right Click
4. Double Click
5. Drag Operation
6. Scroll Operation
7. Volume Up
8. Volume Down
9. Mute
10. Play/Pause Media
11. Next Media
12. Previous Media
13. Zoom In
14. Zoom Out
15. Screenshot Capture

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2

The Voice Assistant Module listens to user voice commands using a microphone. SpeechRecognition library converts speech into text, while Pyttsx3 provides text-to-speech functionality. Jarvis executes various commands such as opening applications, checking system information, searching the internet, and controlling media.

The combined system creates a smart AI- powered human computer interaction platform that supports touchless operation.

IV. SYSTEM ARCHITECTURE



The system architecture consists of the following components:

1. Webcam Input Module
2. Hand Landmark Detection Module
3. Feature Extraction Module
4. Gesture Classification Module
5. Cursor Control Module
6. Voice Recognition Module
7. Jarvis Command Execution Module
8. System Automation Module
9. User Interface Module

The webcam captures real-time frames which are processed using OpenCV and MediaPipe. Hand landmarks are extracted and normalized before being passed to the machine learning classifier. The recognized gesture triggers corresponding actions through PyAutoGUI. Simultaneously, the voice assistant continuously listens for commands. Recognized voice commands are processed and executed using Python automation libraries.

V. METHODOLOGY

A. Data Collection

Gesture data is collected using a webcam. Multiple hand gestures are recorded under different lighting conditions and hand orientations. The hand landmarks detected by MediaPipe are stored as numerical coordinates.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2

B. Feature Extraction

MediaPipe provides 21 hand landmarks. Each landmark contains x and y coordinates. Relative landmark positions are calculated to improve robustness.

Feature Vector:

• 21 x-coordinates

• 21 y-coordinates Total Features = 42

C. Machine Learning Model

The extracted features are used to train an MLPClassifier model from Scikit-Learn.

The model learns gesture patterns and predicts gesture labels in real-time.

D. Gesture Recognition

The trained model classifies gestures during real-time webcam operation.

E. Voice Assistant

Speech Recognition converts audio into text commands.

Pyttsx3 converts text responses into speech.

Jarvis executes commands using operating system and browser automation.

VI. ALGORITHMS USED

A. MediaPipe Hand Tracking

MediaPipe detects 21 hand landmarks from webcam input using machine learning pipelines.

Advantages:

• Real-time performance

• High accuracy

• Lightweight architecture

• Multi-platform support

B. Multi-Layer Perceptron (MLP)

The MLPClassifier is used for gesture classification.

Advantages:

• Fast training

• Good classification accuracy

• Handles nonlinear data efficiently

C. Speech Recognition

SpeechRecognition library converts microphone audio into text commands.

D. Text-to-Speech Engine

Pyttsx3 generates voice responses from text.

VII. MATHEMATICAL MODEL

Euclidean Distance Formula

$$D = \sqrt{((x2 - x1)^2 + (y2 - y1)^2)}$$

Used for measuring finger landmark distances.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, May 2026 Impact Factor: 8.2

Cursor Mapping Formula
ScreenX = (HandX / CameraWidth) × ScreenWidth
ScreenY = (HandY / CameraHeight) × ScreenHeight
Accuracy Formula
Accuracy = (Correct Predictions / Total Predictions) × 100

VIII. SOFTWARE REQUIREMENTS

1. Python 3.10
2. OpenCV
3. MediaPipe
4. NumPy
5. Scikit-Learn
6. PyAutoGUI
7. SpeechRecognition
8. Pyttsx3
9. Pillow
10. Joblib

IX. HARDWARE REQUIREMENTS

1. Intel i3/i5 Processor or Higher
2. Minimum 4GB RAM
3. Webcam
4. Microphone
5. Windows Operating System

X. RESULTS AND DISCUSSION

The proposed system successfully performs gesture recognition and voice-based automation in real-time. The webcam accurately detects hand landmarks under different lighting conditions.

The trained MLPClassifier achieved high gesture classification accuracy for commonly used gestures such as click, scroll, move, and drag operations. The cursor movement was smooth and responsive due to coordinate normalization and smoothing algorithms.

Jarvis successfully recognized voice commands and executed system-level tasks including opening applications, searching online content, controlling volume, and capturing screenshots.

The integrated system demonstrated low latency and efficient performance on standard hardware configurations.

Experimental observations showed:

• Gesture recognition accuracy above 95%
• Smooth cursor movement
• Real-time response
• Successful voice command execution
• Efficient resource utilization

XI. ADVANTAGES

1. Contactless computer interaction
2. Low-cost implementation
3. Real-time performance
4. Improved accessibility

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, May 2026 Impact Factor: 8.2

5. User-friendly interface
6. Multi-functional system
7. Smart automation support
8. No external hardware required

XII. APPLICATIONS

1. Smart Homes
2. Virtual Reality Systems
3. Gaming Applications
4. Healthcare Systems
5. Educational Technology
6. Industrial Automation
7. Accessibility Support
8. AI-Based Human Computer Interaction

XIII. FUTURE SCOPE

The system can be further improved by integrating deep learning models for higher gesture recognition accuracy. Future versions may include:

1. Dynamic gesture recognition
2. Multi-hand support
3. AI chatbot integration
4. Offline speech recognition
5. Face recognition authentication
6. Gesture-based gaming mode
7. Smart IoT device control
8. Cloud-based automation support

XIV. CONCLUSION

The AI Based Gesture Controlled Virtual Mouse with Voice Assistant Integration provides an advanced touchless human computer interaction system using computer vision and artificial intelligence techniques. The system successfully integrates gesture recognition and voice control into a unified platform capable of performing real-time automation tasks.

The project demonstrates efficient hand tracking, accurate gesture recognition, smooth cursor movement, and reliable voice command execution. The integration of MediaPipe, machine learning, and speech recognition technologies creates a smart and user-friendly interaction environment.

The proposed system reduces dependency on traditional hardware devices and provides a cost-effective, accessible, and intelligent solution for modern computing environments.

ACKNOWLEDGMENT

The authors would like to express sincere gratitude to the project guide, faculty members, and institution for providing valuable guidance and support during the development of this project. Special thanks are extended to all contributors and open-source communities whose tools and libraries made this work possible.

REFERENCES

[1] S. Chen, B. Mulgrew, and P. Grant, “A Clustering Technique for Digital Communications Channel Equalization,” IEEE Transactions on Neural Networks, vol. 4, no. 4, pp. 570-578, 1993.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, May 2026
Impact Factor: 8.2

[2] Google MediaPipe Documentation. [Online]. Available: https://mediapipe.dev/
[3] OpenCV Documentation. [Online]. Available: https://opencv.org/
[4] A. Rosebrock, "Hand Tracking and Gesture Recognition using OpenCV," PyImageSearch, 2020.
[5] Scikit-Learn Documentation. [Online]. Available: https://scikit-learn.org/
[6] Python SpeechRecognition Documentation. [Online]. Available: https://pypi.org/project/SpeechRecognition/
[7] PyAutoGUI Documentation. [Online]. Available: https://pyautogui.readthedocs.io/
[8] D. Jurafsky and J. Martin, Speech and Language Processing, Pearson Education, 2019.
[9] R. Gonzalez and R. Woods, Digital Image Processing, Pearson Education, 2018.
[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, Deep Learning, MIT Press, 2016