International Journal of Current Science Research
www.drbgrpublications.in

Gesture and Voice-Control Based Virtual Mouse Using Machine Learning

Tandwe Patrick[1]* and Sesu Ovid[2]

[1]UG Student, Department of Computer Science and Engineering, DMI-St. Eugene University, Zambia
[2]Lecturer, Department of Computer Science and Engineering, DMI-St. Eugene University, Zambia
**Corresponding author email: Patrick.tandwe@gmail.com*

Abstract

Human-Computer Interaction (HCI) has experienced significant advancements through the integration of machine learning, computer vision, and speech recognition technologies. Traditional input devices such as keyboards and mice present challenges for individuals with physical disabilities and in environments where contactless interaction is required. This paper proposes a Gesture and Voice-Control Based Virtual Mouse system that enables users to perform computer operations through hand gestures and voice commands without using physical peripherals. The system employs MediaPipe Hands for hand landmark extraction, a machine learning-based gesture recognition model for gesture classification, and Google Speech Recognition for voice command processing. The proposed solution supports cursor movement, clicking, scrolling, dragging, and application control functions. Experimental results indicate a gesture recognition accuracy of 94.7% and voice recognition accuracy of 88.3% while maintaining real-time performance at 31 FPS. The developed system provides an accessible, low-cost, and efficient alternative for contactless human-computer interaction.

Keywords: Human-Computer Interaction, Virtual Mouse, Gesture Recognition, Voice Control, Machine Learning, MediaPipe, Accessibility

Introduction

Human-Computer Interaction (HCI) focuses on developing intuitive and efficient communication between humans and computing systems. Traditional input devices such as keyboards and mice have remained the dominant methods of interaction for decades. However, these devices can be difficult to use for individuals with motor impairments and are unsuitable for sterile environments where physical contact must be minimized. The increasing demand for touchless technologies, particularly after the COVID-19 pandemic, has accelerated research into gesture and voice-based interaction systems.

Recent advancements in computer vision and machine learning have enabled the development of real-time gesture recognition systems using standard webcams. Similarly, modern speech recognition technologies can accurately interpret spoken commands and translate them into computer actions. This study integrates both gesture and voice modalities into a unified virtual mouse system capable of controlling computer functions without physical peripherals.

Problem Statement

Existing virtual mouse systems primarily focus on either gesture recognition or voice control, limiting their flexibility and usability. Furthermore, many systems require specialized hardware or fail to achieve real-time performance on standard computing devices. Therefore, there is a need for a low-cost multimodal solution that combines gesture and voice interaction while maintaining high accuracy and responsiveness.

Objectives

The objectives of this research are:

1. To develop a hand-gesture recognition system using MediaPipe and machine learning.
2. To implement a voice command recognition module using Google Speech Recognition.
3. To integrate gesture and voice modalities into a unified interaction framework.
4. To evaluate system performance in terms of accuracy, latency, and usability.
5. To provide an accessible and contactless computer interaction solution.

Literature Review

Several researchers have explored gesture recognition and voice-controlled systems for human-computer interaction. Mitra and Acharya (2007) presented one of the earliest surveys on gesture recognition techniques. Rautaray and Agrawal (2015) reviewed vision-based hand gesture recognition approaches and highlighted the advantages of depth sensors. Zhang et al. (2020) introduced MediaPipe Hands, which enables real-time hand landmark detection using standard RGB cameras. Oyedele et al. (2022) developed a CNN-based virtual mouse system achieving 91% recognition accuracy. However, most existing solutions focus on a single interaction modality and lack integrated multimodal support.

Proposed Methodology

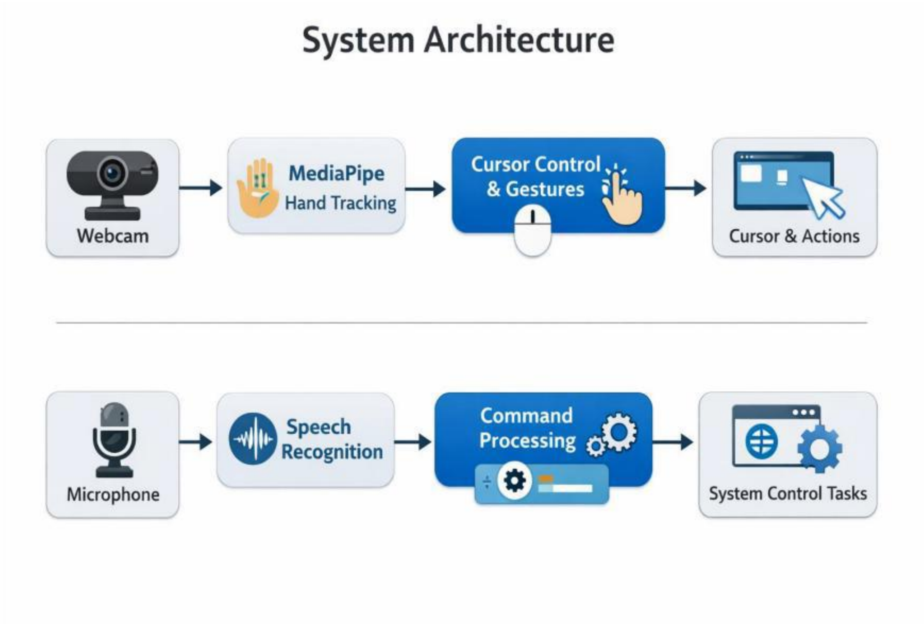
The proposed system consists of five major phases: data acquisition, preprocessing, recognition, fusion, and action execution. A webcam captures hand gestures while a microphone records voice commands. The captured frames are processed using MediaPipe Hands to extract twenty-one hand landmarks. The extracted landmark coordinates are used as input features for the gesture classification model. Voice inputs are processed using Google Speech Recognition to identify predefined commands. Both modalities are integrated through a priority-based arbitration mechanism that resolves command conflicts and ensures efficient execution.

A. System Architecture

The system architecture comprises five layers:

1. Input Layer
2. Preprocessing Layer
3. Recognition Layer
4. Fusion Layer
5. Action Execution Layer

These layers work together to transform gesture and voice inputs into corresponding computer control actions.



B. Algorithms Used

The gesture recognition module employs a machine learning classifier trained on 72,000 augmented gesture samples. MediaPipe Hands extracts hand landmarks, while an Exponential Moving Average (EMA) filter is used to reduce cursor jitter. Voice commands are processed using Google Speech Recognition and mapped to predefined system actions.

Implementation

The system was implemented using Python programming language and several open-source libraries, including OpenCV, MediaPipe, TensorFlow, PyAutoGUI, and SpeechRecognition. The webcam continuously captures hand movements, while the microphone records voice commands. The software processes both inputs concurrently using multithreading techniques to ensure real-time performance. The recognized commands are translated into mouse and keyboard actions through PyAutoGUI.

Results and Discussion

Experimental evaluation was conducted using twenty-five participants under various lighting conditions and backgrounds. The proposed system achieved a gesture recognition accuracy of 94.7% and a voice recognition accuracy of 88.3%. The average processing speed was approximately 31 frames per second, ensuring smooth and responsive interaction.

Compared to existing systems, the proposed solution demonstrated improved multimodal functionality and enhanced accessibility. The EMA smoothing technique effectively reduced cursor instability and improved overall user experience.

Conclusion and Future Work

This paper presented a Gesture and Voice-Control Based Virtual Mouse system that enables contactless computer interaction using machine learning techniques. The integration of gesture recognition and voice commands provides a flexible and accessible alternative to traditional input devices. Experimental results demonstrate high recognition accuracy and real-time performance. Future work will focus on supporting sign language recognition, multilingual voice commands, mobile platform deployment, and advanced deep learning architectures to further improve system performance and usability.

Reference

S. Mitra and T. Acharya, "Gesture Recognition: A Survey," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 37, no. 3, pp. 311–324, 2007.

S. S. Rautaray and A. Agrawal, "Vision-Based Hand Gesture Recognition for Human Computer Interaction," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, 2015.

F. Zhang et al., "MediaPipe Hands: On-Device Real-Time Hand Tracking," Google Research, 2020.

P. Molchanov *et al*., "Hand Gesture Recognition with 3D Convolutional Neural Networks," *Proceedings of CVPR*, pp. 1–9, 2016.

J. Liu *et al*., "Hand Gesture Recognition Using LSTM Networks," *Pattern Recognition Letters*, vol. 125, pp. 367–373, 2019.

O. Oyedele *et al*., "Contactless Mouse Using CNN and OpenCV," *International Journal of Computer Applications*, vol. 184, no. 22, pp. 12–18, 2022.

M. Khan and R. Jain, "Voice and Gesture Multimodal Human Computer Interaction System," *IEEE Access*, vol. 9, pp. 112345–112356, 2021.

M. J. Cheok, Z. Omar, and M. H. Jaward, “A Review of Hand Gesture and Sign Language Recognition Technologies,” *Neural Computing and Applications*, vol. 31, pp. 131–153, 2019.

Google Research, “MediaPipe Hands Documentation,” 2020.

TensorFlow Team, “TensorFlow Machine Learning Framework Documentation,” 2024.