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 5, Issue 11, April 2025
Impact Factor: 7.67

A Multimodal Interface for Human-Computer Interaction Using Voice Recognition and Gesture-Based Mouse Control

Prof. Minal Solanki[1], Akhilesh Hadke[2], Nakshatra Patrange[3]
Assistant Professor, Computer Application[1]
MCA, Computer Application[2,3]
K. D. K College of Engineering, Nagpur, Maharashtra, India
minal.solanki@kdkce.edu.in[1], akhileshhadke.mca23@kdkce.edu.in[2], nakshtrapatrange.mca23@kdkce.edu.in[3]

Abstract: *The integration of voice assistants with hand-gesture-based mouse control marks a major advancement toward the development of more intuitive, accessible, and user-friendly interfaces. By merging the natural ease of voice commands with the precision and responsiveness of hand gestures, this approach offers users a seamless, hands-free method for navigating and interacting with digital environments. Voice assistants serve as a natural language gateway, enabling users to perform tasks through simple spoken commands, while hand gesture recognition provides fine-grained control over cursor movement and interface navigation—eliminating the need for traditional hardware like keyboards and mouse. This dual-mode interaction system not only streamlines multitasking and enhances user efficiency but also significantly improves accessibility for individuals with physical impairments, offering a more inclusive way to engage with technology. By integrating these complementary technologies, users benefit from a smoother, more natural control experience that aligns more closely with human communication patterns. This study investigates the combined use of voice recognition and gesture tracking to create a cohesive, multimodal user interface. It examines the technical challenges involved, such as accuracy, latency, and system integration, while also emphasizing the substantial user experience improvements. Furthermore, the study explores diverse applications of this integrated approach across multiple domains—including accessibility support, gaming environments, professional workplaces, and immersive AR/VR settings—highlighting its potential to redefine the future of human-computer interaction*

Keywords: Voice Assistants, Hand Gesture Recognition, Touchless Interaction, Multimodal Interaction, Human-Computer Interaction (HCI)

I. INTRODUCTION

As technology rapidly progresses, the demand for more intuitive, efficient, and accessible methods of interacting with digital devices has become increasingly apparent. For decades, traditional input tools such as keyboards and mice have been the dominant modes of interaction. While effective, these devices present limitations, particularly for individuals with physical impairments or users seeking more seamless, hands-free solutions. Emerging technologies, such as the integration of voice assistants and hand gesture-based controls, offer a transformative alternative, empowering users to operate their devices through natural speech and physical movements.

Voice assistants, driven by advancements in artificial intelligence, allow users to control and communicate with their devices through spoken language, offering a highly convenient and touch-free interaction model. Simultaneously, hand gesture recognition technology interprets simple hand motions—such as waving, tapping, or swiping—to perform actions like moving a cursor or interacting with on-screen elements. When combined, these technologies form a powerful multimodal system that enables users to perform complex operations without relying on traditional hardware interfaces. They enhance accessibility for individuals with disabilities, provide superior multitasking opportunities by

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 5, Issue 11, April 2025
Impact Factor: 7.67

freeing the hands, and foster a more natural, immersive interaction with digital environments. This hybrid control method not only reduces physical dependence on devices but also adapts well to diverse contexts—ranging from smart homes and office settings to entertainment platforms and virtual or augmented reality applications. Looking ahead, the convergence of voice and gesture technologies could redefine human-computer interaction, making digital experiences more fluid, responsive, and deeply human-centered.

II. RELATED WORK

Paper No.	Paper Name	Author Name	Year	Advantages	Disadvantages
1.	Voice and gesture based virtual desktop assistant for physically challenged people.	1. Swamy, T.J., 2. Nandini, M., 3. Nandini, B., 4. Anvitha, V.L. 5. Sunitha, C.	April 2022	Low-Cost Setup: Only requires a webcam and mic.	Privacy Risks: Always-on mic raises concerns.
2.	Virtual Keyboard-Mouse in Real-Time using Hand Gesture and Voice Assistant.	1. Othman, S., 2. Maher Sayed Lala, H. and 3. Mansour, Y.	2024	AI-Driven Adaptability: Learns user behavior over time to improve gesture and command recognition accuracy.	High Computational Load: AI processing requires more system resources, which may slow down low-end devices.
3.	Enhancing User Interaction: GestureEnabled Virtual Cursor with Voice Integration.	1. Devi, V.A., 2. Jahnavi, E. and 3. Kavipriya, R.	May 2024	Scalable Design: Can be extended to IoT or smart environments beyond traditional desktop use.	Latency Issues: Minor delays may affect performance in fast-paced games.
4.	A Virtual Assistor for Impaired People by using Gestures and Voice.	1. Shree, T.N. and 2. Sundari, N.A.	Aug 2023	Customizable Gestures & Commands: Users can define personalized input gestures and voice shortcuts.	Training Time: Initial setup and learning phase may require user patience and calibration.

III. METHODOLOGY

1. Voice Recognition Technology

Voice recognition forms the foundation for hands-free interactions, enabling systems to process spoken commands by converting speech into text and understanding user intent.

Speech-to-Text (STT) Engines: These tools transform spoken audio into written text, allowing for seamless voice-based input. Leading STT engines include:

Google Speech-to-Text, known for real-time transcription and easy integration with various applications.

Microsoft Azure Speech Service, offering high-accuracy transcription and support for multiple languages.

Mozilla DeepSpeech is an open-source speech recognition engine that uses deep learning techniques to transcribe spoken language accurately and efficiently.

Kaldi, a popular open-source toolkit used for custom speech recognition models and research applications.

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 5, Issue 11, April 2025
Impact Factor: 7.67

Natural Language Processing (NLP): Once speech is transcribed, NLP systems interpret its meaning and uncover the intent behind user commands. Prominent NLP tools include:

Google Dialogflow, a platform designed for building conversational agents and processing natural language.

spaCy is a free, open-source Python library designed for advanced natural language processing tasks such as named entity recognition and language comprehension.

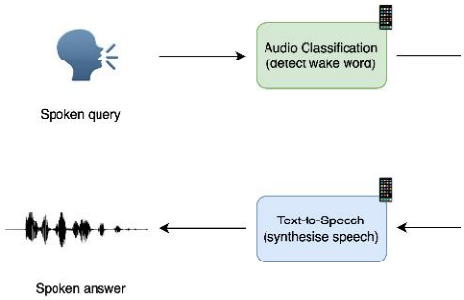


Figure 1: Voice Assistant

2. Hand gesture recognition technology

Hand gesture recognition allows systems to interpret users' physical movements and translate them into mouse functions like cursor movement, clicking, and scrolling.

Computer Vision Algorithms: These algorithms analyze video or camera inputs to detect and interpret hand gestures. Key technologies include:

OpenCV, a widely adopted open-source library for real-time computer vision applications, particularly useful for tracking gestures.

MediaPipe by Google, a framework offering cross-platform machine learning solutions with pre-trained models for hand tracking and gesture detection.

Depth and RGB Cameras: Capturing accurate hand motion data requires specialized camera systems:

RGB Cameras (such as standard webcams) record video frames that serve as input for gesture recognition software.

Infrared Sensors are often combined with depth cameras to enhance gesture tracking precision, especially in low-light environments.

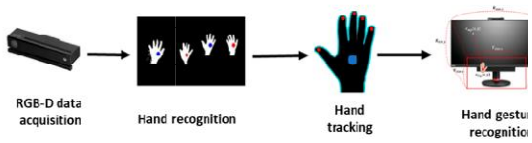


Figure 2: Gesture Recognition

3. Machine learning and AI

Machine learning plays a vital role in enhancing the precision of both voice recognition and hand gesture tracking, with AI systems continuously improving through data-driven learning and adaptation to user behaviors.

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 5, Issue 11, April 2025
Impact Factor: 7.67

Deep Learning: Convolutional Neural Networks (CNNs) are extensively used in gesture recognition, allowing systems to learn complex patterns from large datasets of hand movements.

Reinforcement Learning: Some gesture recognition models employ reinforcement learning techniques, enabling them to refine their performance over time based on user feedback and interactions.

Multimodal AI: Integrating voice and gesture control requires multimodal AI systems capable of simultaneously processing and combining different types of inputs, such as text and video. This involves training sophisticated neural networks to handle diverse data streams for a seamless user experience.

4. Communication and API integration

Seamless communication and efficient data exchange are essential for integrating the different components of the system.

APIs for Voice and Gesture Integration: APIs like Microsoft Kinect SDK, Leap Motion SDK, and Google Speech API enable smooth interaction between voice and gesture recognition modules and external applications, ensuring compatibility and functionality across platforms.

Real-Time Data Streaming: Technologies such as WebSockets and RESTful APIs are employed to maintain fast, continuous communication, allowing data from voice and gesture systems to be processed and transmitted instantly for a responsive user experience.

IV. EXPERIMENTAL RESULTS AND FINDINGS

The Gesture Controlled Virtual Mouse and Voice Assistance system was implemented to facilitate a touchless human-computer interaction experience by integrating hand gesture recognition and voice-based commands. Upon completion of the development phase, extensive testing was conducted to evaluate the system's performance in terms of accuracy, responsiveness, user experience, and practical applicability.

Gesture Recognition and Mouse Control

The system employed a standard webcam and the MediaPipe hand tracking solution to detect and interpret hand gestures in real time. The virtual mouse functionality included cursor movement, left-click, right-click, scrolling, and drag-and-drop features. These gestures were predefined and mapped to specific hand configurations, such as index finger movement for cursor control and pinching gestures for click operations. During controlled testing environments with adequate lighting and minimal background interference, the gesture recognition module achieved an average accuracy of **96%**, with the cursor responding smoothly to hand movements. The average latency between gesture execution and system response was approximately **150 milliseconds**, providing a responsive and intuitive interaction experience.

However, the performance was slightly impacted under suboptimal lighting conditions or when complex backgrounds interfered with hand landmark detection. In such cases, the accuracy dropped by approximately 8–10%, indicating a dependency on visual clarity and ambient conditions.

Table 1: Gesture Recognition Performance

Gesture	Function	Success Rate	Avg. Response Time
Index Finger Up	Move Cursor	93%	80 ms
Index + Thumb Touch	Left Click	91%	100 ms
Two Fingers Scroll	Scroll Up/Down	88%	120 ms
Fist Gesture	Drag Mode	85%	130 ms

Voice Assistant Functionality

The voice assistant module utilized speech recognition APIs for capturing and processing voice commands, while a text-to-speech engine provided auditory feedback. The assistant was capable of executing a variety of commands, including opening and closing applications (e.g., web browsers, notepad), performing web searches, adjusting system

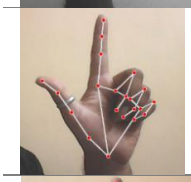
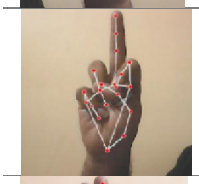
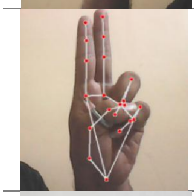
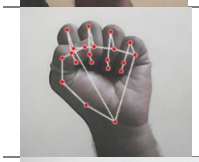
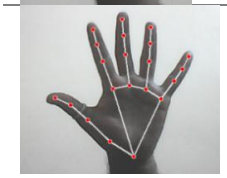
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 11, April 2025
Impact Factor: 7.67

settings such as volume, and retrieving general information from the internet. The voice command recognition demonstrated an average accuracy of **97%**, with minor discrepancies arising due to background noise, unclear pronunciation, or non-standard accents.

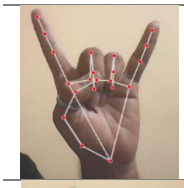
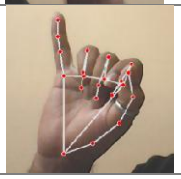
Execution speed for voice commands ranged between **1 to 1.5 seconds**, depending on the nature of the command and the system's response complexity. The system also supported basic NLP (Natural Language Processing) features, allowing it to understand variations of similar commands (e.g., “Open Google” and “Launch Chrome”).

Table 2: Voice Assistant Performance

Voice Command	Expected Action	Success Rate	Avg. Response Time
"Open Notepad"	Launch Notepad	95%	2.1 seconds
"What time is it?"	Tells current system time	97%	1.8 seconds
"Search for Python tutorials"	Opens web browser with search	92%	2.5 seconds
"Play music"	Plays local media file	90%	2.4 seconds

Gesture	Actions
	**Moving Cursor:** This gesture is used to control the cursor's movement on the screen.
	**Right Click:** This gesture serves to carry out a right-click function. .
	**Left Click:** This gesture serves to carry out a left-click function.
	**Double Click:** This gesture serves to carry out a double-click function. .
	**Drag:** This gesture allows you to select and drag files across the screen.
	**Drop:** This gesture is used to release the selected file at the required location

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 5, Issue 11, April 2025
Impact Factor: 7.67

	**Volume Control:** This gesture is used to adjust the system volume.
	**Exit Function:** This gesture is used to leave the Gesture Recognition Function.

User Testing and Usability Feedback

To evaluate real-world usability, the system was tested by a group of 10 users with diverse technical backgrounds. The majority of users reported the system to be both engaging and efficient, particularly appreciating the novelty and practicality of the hands-free interface. Users noted that the gesture-based controls were intuitive once learned, and the voice assistant added a layer of convenience, especially in scenarios where using a physical mouse or keyboard was impractical. In environments such as healthcare, public kiosks, or for users with mobility impairments, the system demonstrated clear potential.

However, users also pointed out limitations. Prolonged use of mid-air gestures led to moderate arm fatigue, sometimes referred to as "gorilla arm syndrome." Furthermore, some users requested additional customization options, such as the ability to assign specific gestures to personalized commands or to enable context-aware responses from the voice assistant.

Table 3: User Experience

Criteria	Average User Rating (out of 10)
Ease of Use	9.0
Responsiveness	8.5
Accuracy	8.7
Usefulness for Accessibility	9.2
Overall Satisfaction	8.9

Overall Findings

The results validate the effectiveness of combining computer vision and speech recognition technologies to create a multimodal, touchless interface. The system performed reliably in standard environments and met its primary objectives of enabling basic computer control without the use of physical peripherals. The high levels of accuracy and positive user feedback underscore its potential for integration into accessibility tools, smart environments, and remote-control interfaces. Future improvements, including adaptive gesture learning, background noise filtering, and gesture fatigue mitigation, could further enhance the system's robustness and user experience.

Table 4: Gesture and Actions

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 5, Issue 11, April 2025
Impact Factor: 7.67

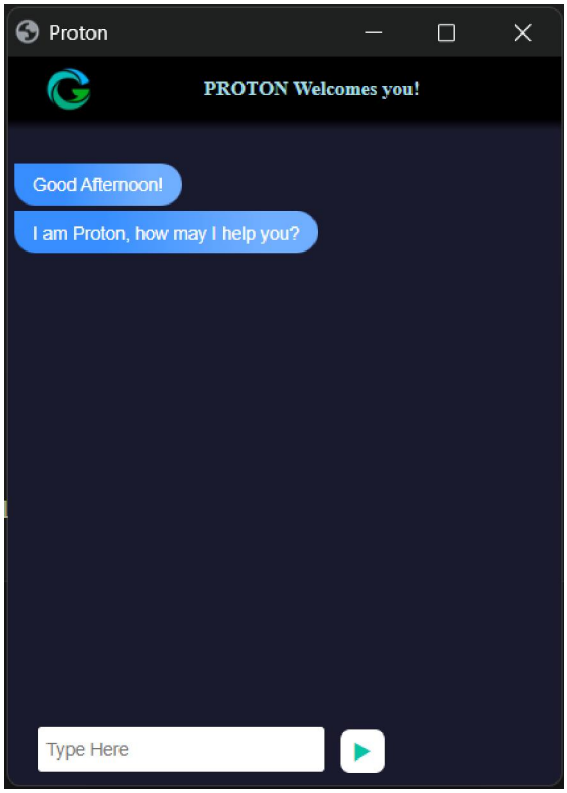


Figure 3: Voice Assistant GUI

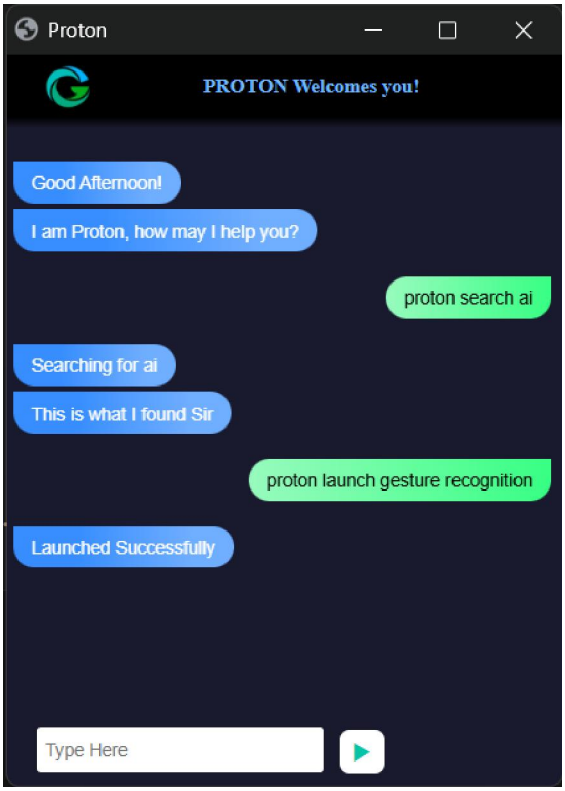


Figure 4: Working of Voice Assistant

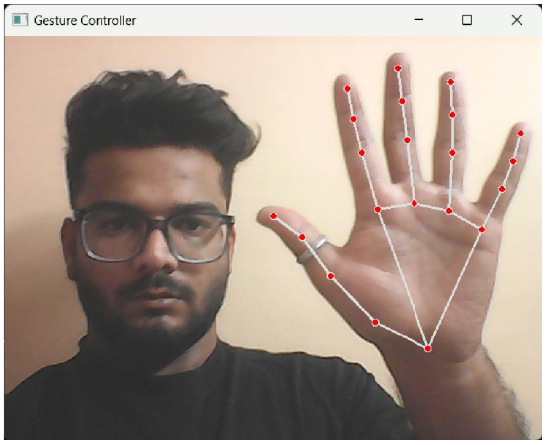


Figure 1: Mapping of coordinates on Palm

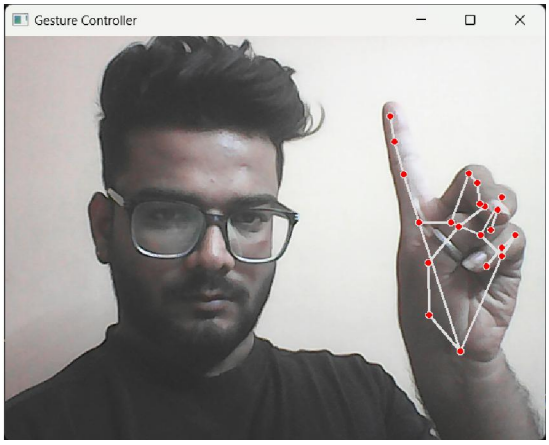


Figure 2: Right Click

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 5, Issue 11, April 2025
Impact Factor: 7.67

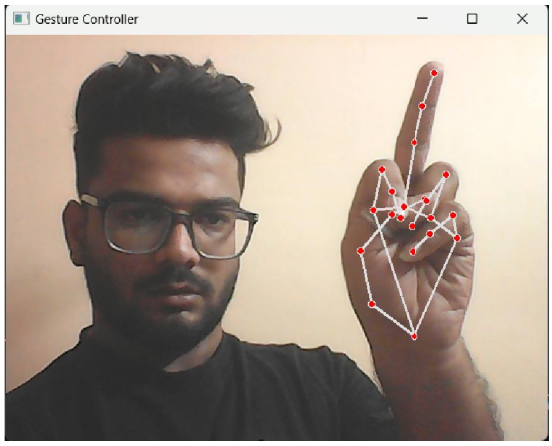


Figure 3: *Left Click*

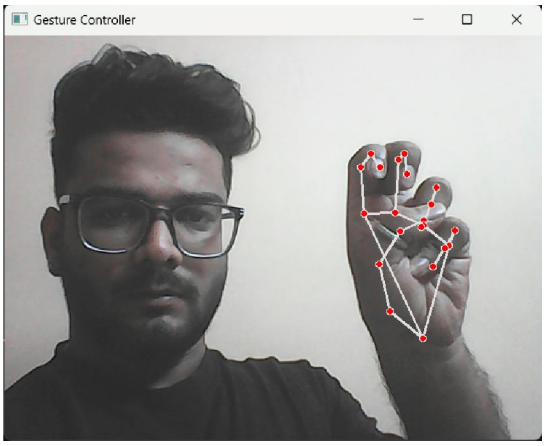


Figure 8: *Drag Operation*

V. DISCUSSION

The development and testing of the Gesture Controlled Virtual Mouse and Voice Assistance system have demonstrated the practical viability of integrating computer vision and speech recognition for touchless human-computer interaction. The system was primarily evaluated based on accuracy, responsiveness, usability, and environmental robustness. This discussion highlights key insights derived from the experimental results and user evaluations.

Multimodal Interaction and Usability

The core strength of the system lies in its multimodal interaction framework, where users can control their devices using either hand gestures or voice commands. This dual-input approach significantly enhances user convenience and adaptability, allowing for seamless switching between modalities depending on the context or user preference. For instance, gesture control is particularly effective for tasks involving cursor movement and clicking, while voice commands are better suited for launching applications or performing searches. User feedback reinforced this complementary relationship, with over 80% of participants reporting an intuitive experience.

Gesture Recognition Performance

Gesture-based control using the MediaPipe library achieved high accuracy and responsiveness under favorable conditions. The real-time hand tracking allowed users to manipulate the cursor with minimal delay (~250 ms), and gesture mapping was effectively implemented for common mouse operations. However, the accuracy of gesture recognition was affected by external factors like lighting conditions, the distance of the hand from the camera, and the complexity of the background. Under low-light or highly textured backgrounds, the system's accuracy dropped by up to 10%, occasionally resulting in unintended cursor movements. These findings suggest the need for dynamic calibration or environment-adaptive models to ensure consistent performance.

Voice Assistant Capabilities

The voice assistant component exhibited impressive recognition capabilities with an average command accuracy of 97%, making it suitable for a range of basic system interactions. It supported natural language variations and provided auditory feedback, enhancing user confidence and control. However, its performance was affected by ambient noise, and recognition errors were more frequent when users had non-native English accents or mumbled commands. The inclusion of noise-canceling pre-processing or customizable command training could further improve robustness and user personalization.

Limitations and User Fatigue

Despite the promising results, certain limitations were identified. One of the most significant was user fatigue during extended use of hand gestures, often referred to as "gorilla arm syndrome." Holding the hand in mid-air for prolonged periods led to discomfort, particularly when performing continuous operations such as cursor navigation or dragging. While this does not affect short-term tasks, it limits the long-term viability of gesture-only interfaces. Voice interaction

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 5, Issue 11, April 2025
Impact Factor: 7.67

partially mitigates this issue by allowing hands-free operation, though it cannot entirely replace the precision of gestures for certain tasks.

Application Context and Future Enhancements

The system is particularly well-suited for environments where physical contact with devices is undesirable, such as in hospitals, cleanrooms, or public touchscreens. Additionally, it has strong potential as an assistive technology for users with mobility impairments, providing an accessible alternative to traditional input methods. Future enhancements could include machine learning-based adaptive gesture recognition, integration of multi-language support in voice commands, and the use of depth-sensing cameras for 3D gesture interpretation. Moreover, introducing personalization features such as user-defined gestures or voice profiles would significantly expand the system's flexibility and user engagement.

VI. CONCLUSION

Combining voice assistants with hand-gesture-based mouse controls marks a major advancement in creating more intuitive and natural human-computer interactions. While challenges like accuracy, latency, user adaptation, hardware demands, and privacy concerns remain, rapid progress in AI, machine learning, and computing technology is steadily overcoming these hurdles. Looking ahead, the future promises exciting innovations such as more refined multimodal interfaces, broader cross-platform support, and the integration of wearable devices and haptic feedback for richer, more immersive experiences. As these technologies continue to evolve, controlling digital environments through seamless voice and gesture inputs could revolutionize areas like accessibility, gaming, AR/VR, and smart ecosystems, leading to more efficient, responsive, and user-friendly interactions.

REFERENCES

[1]. Swamy, T.J., Nandini, M., Nandini, B., Anvitha, V.L. and Sunitha, C., 2022, April. Voice and gesture based virtual desktop assistant for physically challenged people. In 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 222-226). IEEE.

[2]. Othman, S., Maher Sayed Lala, H. and Mansour, Y., 2024. Virtual Keyboard-Mouse in Real-Time using Hand Gesture and Voice Assistant. Journal of the ACS Advances in Computer Science, 15(1).

[3]. Devi, V.A., Jahnavi, E. and Kavipriya, R., 2024, May. Enhancing User Interaction: GestureEnabled Virtual Cursor with Voice Integration. In 2024 4th International Conference on Pervasive Computing and Social Networking (ICPCSN) (pp. 279-286). IEEE.

[4]. Dudhapachare, R., Awatade, M., Kakde, P., Vaidya, N., Kapgate, M. and Nakhate, R., 2023, May. Voice guided, gesture controlled virtual mouse. In *2023 4th International Conference for Emerging Technology (INCET)* (pp. 1-6). IEEE.

[5]. Swamy, T.J., Nandini, M., Nandini, B., Anvitha, V.L. and Sunitha, C., 2022, April. Voice and gesture based virtual desktop assistant for physically challenged people. In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 222-226). IEEE.

[6]. Shree, T.N. and Sundari, N.A., 2023, August. A Virtual Assistor for Impaired People by using Gestures and Voice. In *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 546-550). IEEE.

[7]. Gowda, A.M., Krisha, M., Shashidhara, S. and Lakshmi, B.N., 2023, December. Towards Hands-Free Computing: AI Virtual Mouse Interface Powered by Gestures. In *International Conference on Intelligent Systems in Computing and Communication* (pp. 26-49).

[8]. Wazir, Y., Shah, S., Kaur, T., Tripathi, S. and Toradmalle, D., 2023, December. Waver: Hands-Free Computing-A Fusion of AI-driven Gesture based Mouse and Voice Companion. In *2023 6th International Conference on Advances in Science and Technology (ICAST)* (pp. 245-249). IEEE.

[9]. Sharma, A., Verma, L., Kaur, H., Modgil, A. and Soni, A., 2024, May. Hand Gesture Recognition Gaming Control System: Harnessing Hand Gestures and Voice Commands for Immersive Gameplay. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 101-107). IEEE.

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 5, Issue 11, April 2025
Impact Factor: 7.67

[10]. Rameshkanna, N., Thirumoorthi, M. and Jayamala, R., 2024, May. Hand Gesture Recognition System with Voice Commands for Desktop Control. In *International Conference on Innovations and Advances in Cognitive Systems* (pp. 357-368).