Quanta Research
https://journals.tultech.eu/index.php/qr
ISSN: 2806-3279
Volume, Issue
DOI: https://doi.org/10.15157/qr.2026.4.1.162-173
Received: 03.07.2026; Revised: 15.08.2026; Accepted: 25.08.2026

Research Article

Comparative Evaluation of Gesture, Voice, and Mouse Modalities Using ISO 9241-9 in Controlled and Uncontrolled Environments

Sreya Nandi[1], Shreyani Saha[1], Satyaki Mondal[1], Rohit Kumar Choubey[1*], Sandipan Sahu[1], Abhishek Mukhopadhyay[1*]

[1]Computer Science and Engineering, School of Engineering and Technology, Amity University Kolkata, India

***abhishek.gmit@gmail.com**

Abstract

Human-robot interaction requires intuitive, efficient, and affordable interfaces for real-world deployment. This work presents a cost-effective comparative study of three interaction modalities such as gesture, voice, and mouse using the ISO 9241-9 pointing task. Twenty participants performed repeated trials in both controlled indoor and uncontrolled outdoor environments. Movement time was analyzed using Fitts' Law, while cognitive workload and usability were evaluated using NASA-TLX and the System Usability Score. Results show a significant effect of modality on performance, where mouse achieved the lowest movement time, gesture demonstrated comparable performance with no significant difference, and voice exhibited significantly higher delays. Subjective analysis indicates lower workload for mouse, while usability scores remain high across all modalities. Environmental conditions did not significantly affect performance. The findings highlight gesture as a low-cost and touchless alternative suitable for human-robot interaction, particularly in accessibility-focused and real-world scenarios where conventional input devices are not feasible.

Keywords: Human–Computer Interaction; Interaction Modalities; Voice Interaction; Gesture Interaction; Multi-modal interaction; ISO 9241-9 Pointing Task

INTRODUCTION

The increasing deployment of robots in healthcare, manufacturing, and assistive domains demand interaction techniques that are intuitive, efficient, and adaptable. Conventional input devices such as keyboard and mouse assume direct physical contact and sustained attention, which may not always be feasible in real-world human–robot interaction (HRI) scenarios. This motivates the exploration of alternative and complementary interaction modalities that can enhance accessibility, flexibility, and user experience across diverse environments and user groups.

Multi-modal interaction refers to the integration of multiple input and output channels to facilitate communication between humans and computer systems. Instead of relying on a single modality, systems combine speech, gesture, touch, gaze [1 - 5], haptic feedback, and visual interfaces to enable natural and efficient interaction inspired by human behavior. Such approaches are widely adopted in smart assistants and collaborative robotic systems. In industrial Human–Robot Collaboration, workers use voice commands, gestures, and touch panels to control Cobots, while in healthcare, systems such as the Da Vinci Surgical System [6] demonstrate the importance of advanced interaction mechanisms. Multi-modal interaction is therefore central to effective HRI [7, 8], allowing robots to interpret diverse inputs and adapt to user intent and context.

Prior research has investigated individual modalities and their combinations. Lee et al. [9] examined speech-based interaction in driving scenarios and reported reduced visual and manual distraction, though limitations included simulator-based evaluation and unaddressed noise factors. Doyle and Bertolotto [10] proposed a speech-and-pen multimodal framework for mobile GIS, highlighting flexibility across tasks but noting environmental sensitivity of speech input. Coronado et al. [11] explored Gesture-based robot control by developing an open-source wearable framework for wrist-motion mapping. Hettig et al. [12] and Bergh et al. [13] and compared gesture-based and conventional techniques in medical environments, observing that traditional inputs outperformed gestures for complex tasks. Saren et al. [7] compared between gaze, gesture and voice for interacting with robot using commercial off-the-shelf (COTS) devices.

Despite these efforts, systematic comparative evaluations of multiple modalities under different ambient conditions remain limited. Most studies tend to focus on a single modality or test them in controlled environments, without fully considering how factors like noise, lighting, and user context can affect performance. As a result, there is a lack of comprehensive understanding of how different interaction methods such as mouse, gesture, and voice perform relative to each other under controlled and uncontrolled conditions. This gap highlights the need for studies that evaluate multiple modalities together across diverse environmental conditions.

In this work, a comparative analysis of three interaction modalities is presented such as voice, gesture, and mouse with a focus on cost-effective implementation without reliance on expensive COTS hardware. The study evaluates their applicability in indoor and outdoor contexts using a standardized ISO 9241-9 pointing task under various lighting and sound conditions. Experimental results indicate that mouse interaction achieved the lowest movement time, gesture showed moderate and competitive performance with minor variability due to hand tracking, and voice performance was sensitive to ambient noise. These findings suggest that no single modality is universally optimal; instead, modality selection should be domain specific. Figure 1 illustrates the overall architecture and workflow of the system. The demonstration video showcasing the different interaction modalities is available at https://youtu.be/7uzuWb689aY.

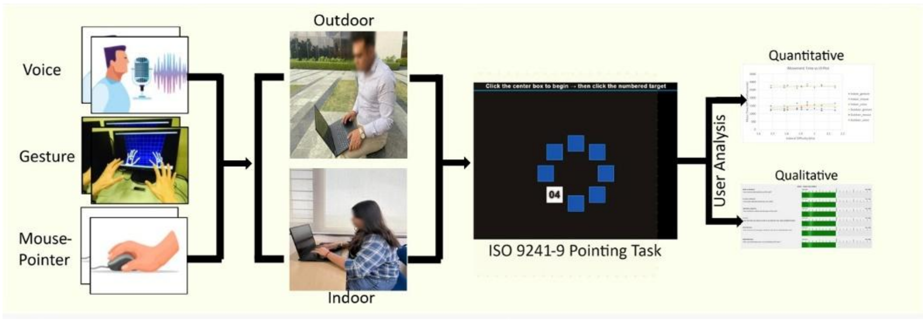

Figure 1. Schematic Diagram of the Proposed System

The primary objectives and contributions of this paper are as follows.

- To perform a systematic comparative evaluation of mouse, gesture, and voice interaction modalities in realistic environments.
- To conduct a user study analyzing objective performance and cognitive workload across modalities.
- To demonstrate a low-cost, hardware-independent framework suitable for human–robot interaction applications.

The rest of the paper is structured as follows. Next section presents study setup followed by the user study section. Penultimate section covers general discussion followed by concluding remarks and future scope in final section.

STUDY SETUP

This section describes different algorithms used to perform an ISO 9241-9 pointing task [14]. ISO 9241-9 is an international ergonomic standard that defines how to evaluate non-keyboard input devices used for pointing and selection. Part 9 specifically addresses non-keyboard input devices. The procedures are defined to evaluate user performance, comfort, and physical effort. It establishes uniform testing guidelines so that different pointing devices can be compared scientifically. The ISO 9241-9 is tested using three modalities such as voice, gesture, and mouse pointer. Each of these modalities is described briefly as follows. In this context, it is important to note that the primary objective of this study is to demonstrate a cost-effective interaction framework. While COTS gesture recognition devices such as Kinect can enhance system performance, they also significantly increase the overall cost of the system.

Gesture

For the gesture-based modality, hand movements are captured using a standard computer webcam integrated with the MediaPipe Hands framework [15]. The framework is used to track the hand in real time and extract 3D landmark coordinates of it. Cursor movement is controlled using the position of the index fingertip, while a pinch gesture is employed to simulate a click action. The 2D coordinates of the index fingertip are mapped to the corresponding screen coordinates to enable cursor control. To improve stability and reduce jitter, exponential smoothing is applied to the fingertip trajectory. Additionally, a debounce mechanism is implemented to prevent unintended multiple selections. The minimum duration of pinch hold is set to 85 ms, and the debounce interval is fixed at 600 ms to ensure reliable click detection.

Voice

For the voice-based modality, interaction is implemented using the browser-based Web Speech Recognition API [16]. This enables continuous speech recognition without the need for additional hardware. The API is configured in continuous listening mode to support uninterrupted interaction between multiple trials. The system continuously monitors spoken input and processed recognized commands in real time. The participants start each trial by issuing the start command, ***Go Ahead***, upon which the task is activated. Subsequently, multiple targets are presented in a circular arrangement around a central point, following a multi-directional tapping paradigm. Each target is labelled with a random number. Participants select the highlighted target by verbally stating the corresponding number.

Mouse Movement

The mouse-based modality serves as the baseline interaction technique in the user study, allowing participants to perform the ISO 9241-9 pointing task using a standard optical mouse.

USER STUDY

The user study is conducted in a semi-controlled environment to measure the average movement time of participants in the case of ISO 9241-9 pointing task. This study aims to compare and analyze pointing and selection between gestures, voice, and mouse-based interaction modalities. The flow of the study is shown in Figure 2. The experiments were conducted in controlled and uncontrolled environments. The controlled environment maintained ambient noise levels of approximately 40—50 dB and lighting conditions around 100—300 lux. In contrast, the uncontrolled environment included higher ambient noise levels (80—100 dB) and variable lighting conditions (2000—5000 lux) [as shown in Figure 1]. Voice input was captured using a built-in laptop microphone, while gesture input was performed using the built-in webcams of the laptops.

Participants

Twenty participants (ten male and ten female) with an average age of 23.2 (std: 6.48) years from our university volunteered to participate in this study. A free trial was conducted to familiarize participants with all interfaces.

Materials

The experimental software is implemented as a web-based application using HTML, CSS, and JavaScript. The participant interface includes structured input fields for demographic information (name, age, gender) and environment selection (indoor/outdoor), with client-side validation to ensure data integrity before proceeding. The system stores participant information locally using browser-based storage and redirects users to the pointing task module. The interface is designed with a dark theme, responsive layout, and interactive elements to maintain usability across different screen sizes.

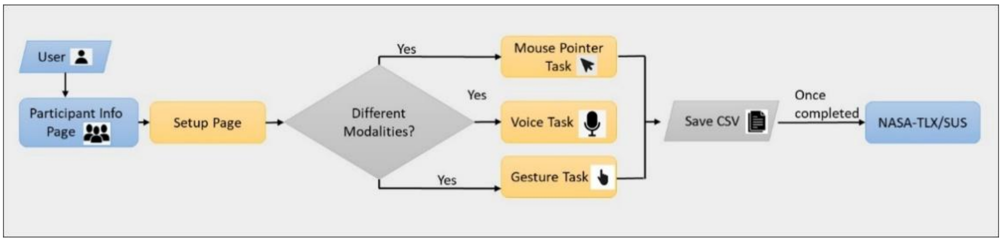

Figure 2. Block diagram of User Study

The software runs on 64-bit Windows operating systems and is tested on three hardware configurations: (1) an Intel Core i7-14650HX processor with 16 GB RAM, (2) an AMD Ryzen 5 5500U processor with 8 GB RAM, and (3) an Intel Core i5-13420H processor with 16 GB RAM. All systems operate on x64-based architecture without touch or pen input. The application does not require specialized hardware, ensuring accessibility and cost-effectiveness for controlled and uncontrolled experimental environments.

Design

Performance measurement begins when the participant initiates the trial when the user selects any of the modalities. The system records: (I) Start time at trial initiation, (II) End time upon correct target selection, (III) Movement trajectory from the centre to the target. We used the ISO 9241-9 multidirectional pointing task with the following target sizes and distances.

- Target sizes (W, in pixels): 70, 80, 90
- Distance of target from centre of screen (D, in pixels): 200, 220, 240

The pointing performance was analysed using Fitts' Law [17], which models the relationship between movement time (MT) and task difficulty. The Index of Difficulty (ID) is calculated as:

$$\mathrm{ID} = \log_2\left(\frac{D}{W} + 1\right) \quad (1)$$

According to Fitts' Law, movement time is linearly related to the index of difficulty.

$$\mathrm{MT} = \mathrm{a} + \mathrm{b} \times \mathrm{ID} \quad (2)$$

where a and b are empirically determined constants representing the intercept and slope of the linear regression model. This formulation allows us to quantitatively compare the efficiency of different interaction modalities.

Throughput (TP) was calculated as a combined measure of speed and accuracy, representing the overall efficiency of the interaction.

$$\mathrm{TP} = \frac{ID}{MT} \tag{3}$$

where ID is the Index of Difficulty and MT is the average Movement Time.

A repeated-measures study is designed with the following independent variables.

- Place of Study
  - Indoor
  - Outdoor (canteen environment with uncontrolled light and noise)
- Type of Modalities
  - Gesture
  - Mouse
  - Voice

The order of modalities was randomized for each participant to minimize learning effects and order bias. Each participant performed the pointing tasks under all combinations of target widths and distances across both environmental conditions. After completing the experimental trials for each modality, participants were asked to complete the NASA-TLX questionnaire to assess perceived workload and the System Usability Scale (SUS) to evaluate overall usability and user preference. In total, movement time is analysed for all combinations of target width and distance across modalities and environmental conditions, enabling a comprehensive comparison of motor performance, cognitive workload, and perceived usability.

Results

Figure 3 plots the movement times with respect to indices of difficulties for all four conditions. Correlation coefficient is found to be $r = 0.7$ between movement time and ID for the mouse indoor. However, for other modalities, the correlation coefficients were less than 0.3.

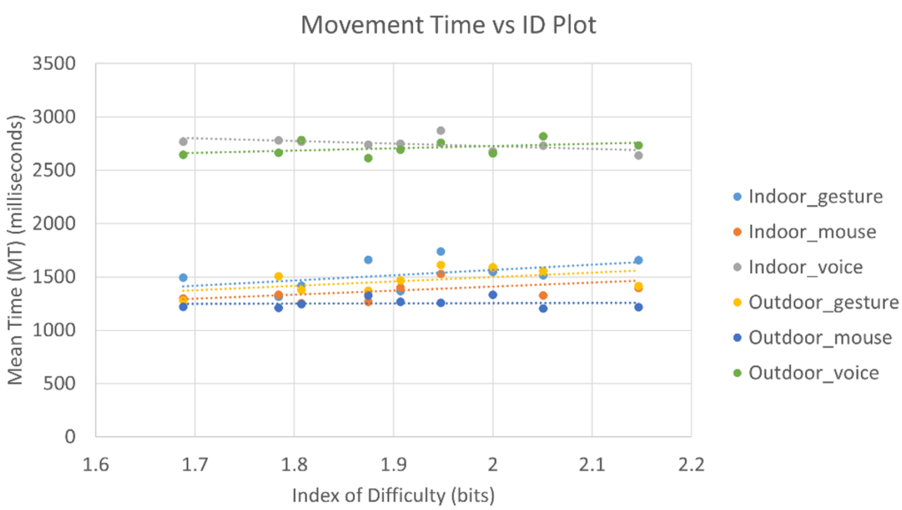

Figure 3. Movement time vs ID plotEffect for Different Modalities

The correlation analysis between Index of Difficulty (ID) and Movement Time (MT) shows varying trends across modalities and environments. For gesture interaction, a moderate positive correlation is observed in both indoor ($r = 0.49$) and outdoor ($r = 0.52$) conditions, indicating that MT increases with task difficulty as expected from Fitts' Law. Similarly, mouse interaction exhibits a moderate positive correlation in the indoor environment ($r = 0.7$), but this relationship weakens significantly in the outdoor condition ($r = 0.05$), suggesting reduced sensitivity to ID under uncontrolled settings. In contrast, voice interaction demonstrates an unexpected moderate negative correlation in the indoor environment ($r = -0.54$), indicating that MT does not follow the conventional Fitts' Law trend, likely due to variability in speech processing and command latency. However, in the outdoor environment, voice interaction shows a moderate positive correlation ($r = 0.43$), suggesting partial alignment with task difficulty.

An $Environment$ (2) $\times$ $Modalities$ (3) repeated-measures ANOVA was undertaken on the movement times. Mauchly's Test of Sphericity indicated that the assumption of sphericity was not violated for modality ($p = 0.132$), allowing the use of standard ANOVA results. A significant main effect of environment was observed ($F(1,19) = 5.606, p < 0.05, \eta^2 = 0.228$). A highly significant main effect of modality was found ($F(2,38) = 300.112, p < 0.05, \eta^2 = 0.940$). However no significant interaction effect was observed between environment and modalities.

Based on the significant main effect of environment and modalities reported earlier, post-hoc pairwise comparisons (LSD) were conducted to determine which modalities differed from each other. The estimated marginal means indicate that Mouse had the lowest movement time ($mean = 1313.99, sd = 70.01$), followed by Gesture ($mean = 1493.19, sd = 97.98$), while Voice showed substantially higher movement time ($M = 2727.17, sd = 50.94$) as shown in Figure 4. Pairwise comparisons revealed that voice produced significantly higher movement time than Gesture ($p < 0.05$) and mouse ($p < 0.05$). However, there was no significant difference between gesture and mouse.

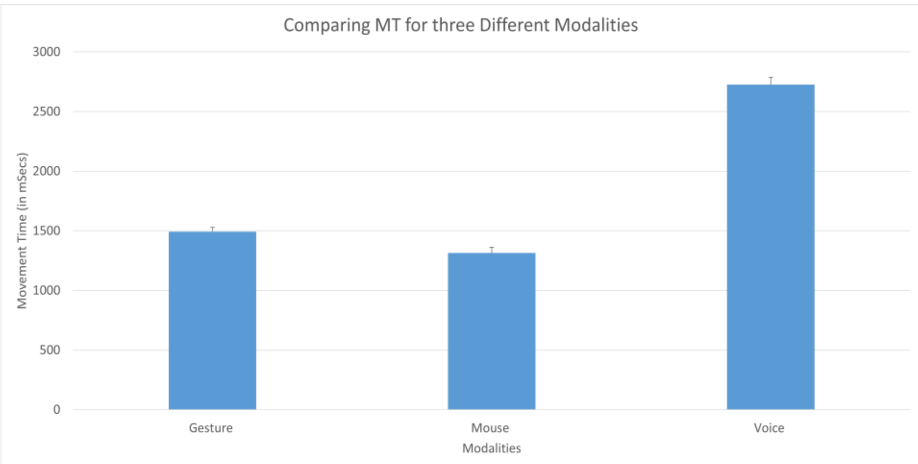

Figure 4. Effect of three modalities on Movement Time

To complement the performance analysis, subjective workload and usability were evaluated using NASA-TLX and the System Usability Scale (SUS), respectively. Paired-sample t-tests were conducted to compare modalities. Pairwise comparisons as shown in Figure 5 revealed that gesture imposed significantly higher workload than mouse, $t(19) = 3.09, p < 0.05$. Similarly, voice resulted in significantly higher workload than Mouse, $t(19) = -3.83, p < 0.05$. However, no significant difference was observed between Gesture and Voice, $t(19) = 0.67, p = 0.51$. The SUS scores were generally high across all modalities, indicating good overall usability. Mouse received the highest usability rating ($M = 91.69$), followed by gesture ($M = 87.31$) and voice ($M = 85.25$) as shown in Figure 6. The pairwise comparisons showed statistically significant differences between voice and mouse, $t(19) = 2.74, p < 0.05$. Gesture and Mouse did not differ significantly, nor did Gesture and Voice.

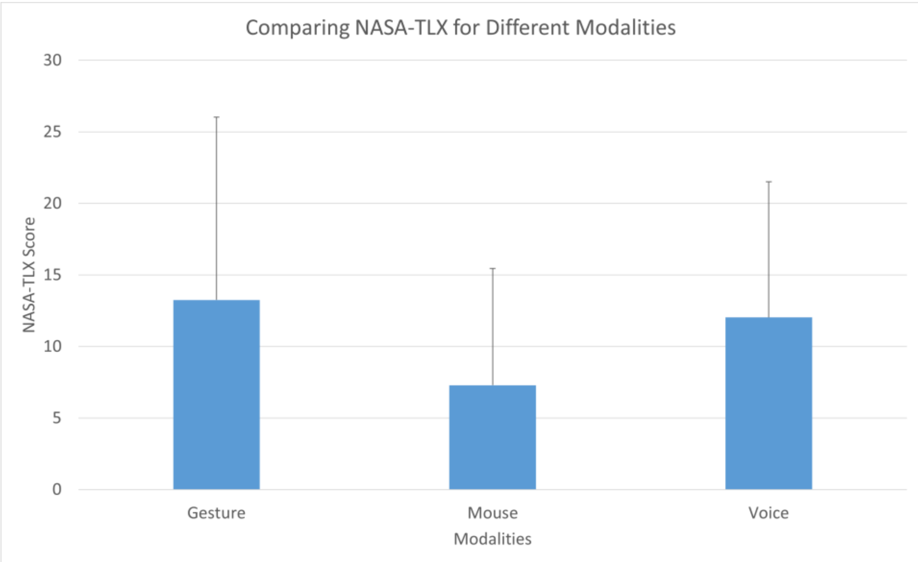

Figure 5. Comparing Cognitive Load for Different Modalities

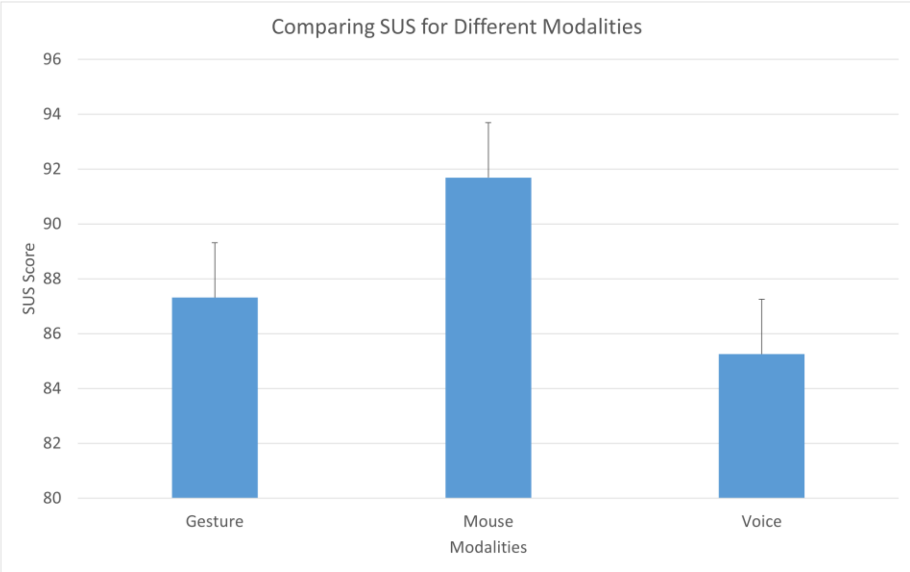

Figure 6. Comparing Acceptability of different modalities using SUS

Table 1 gives an overall comparison of the three interaction modalities based on movement time, throughput, workload, and usability. From the results, it can be observed that the mouse performs the best in terms of speed and efficiency, while both gesture and voice interactions tend to involve higher user workload.

Table 1. Summary of Interaction Metrics

Metric	Mouse	Gesture	Voice
Movement Time (ms) (↓)	1313.99 (70.01)	1493.19 (97.98)	2727.17 (50.94)
Throughput (bits/sec) (↑)	1.45	1.28	0.70
NASA-TLX (↓)	7.29	13.24	12.03
SUS (↑)	91.69	87.31	85.25

[*] ↑ Indicates a higher value is better. ↓ Indicates a lower value is better

DISCUSSION

The findings of this study provide important insights into the design of cost-effective human--robot interaction systems by comparing gesture, voice, and mouse-based modalities across controlled and uncontrolled environments. The primary objective of this work is to compare the movement time of the three interaction modalities in order to determine which modality performs most efficiently. Thus, factors such as user fatigue and physical obstructions are not considered within the scope of this study. The statistical analysis reveals a strong main effect of modality on movement time, with mouse interaction achieving the lowest movement time ($M = 1313.99$ ms), followed by gesture ($M = 1493.19$ ms), while voice shows significantly higher values ($M = 2727.17$ ms). Post-hoc comparisons indicate that voice is significantly slower than both mouse and gesture, whereas no significant difference is observed between mouse and gesture, suggesting comparable performance.

An important observation is that environmental conditions do not significantly affect interaction performance, as no interaction effect between environment and modality is found. This indicates that gesture and voice-based systems remain stable even under uncontrolled outdoor conditions, demonstrating robustness to variations in noise and lighting. Although mouse interaction shows the lowest cognitive workload based on NASA-TLX, gesture maintains acceptable workload levels and usability scores comparable to mouse. While mouse remains the most efficient conventional input device, gesture interaction emerges as a practical and low-cost alternative, particularly in scenarios where touch-based interaction is not feasible. Its contactless nature makes it suitable for users with limited mobility, sterile environments, and outdoor human--robot interaction settings. Voice interaction, despite its hands-free advantage, is affected by latency and environmental noise. These findings suggest that modality selection should be driven by application context, accessibility needs, and environmental constraints rather than relying solely on efficiency.

CONCLUSION AND FUTURE SCOPE

This work presents a systematic and cost-effective comparison of gesture, voice, and mouse-based interaction modalities using the ISO 9241-9 pointing task across controlled and uncontrolled environments. The results demonstrate that mouse interaction achieves the lowest movement time and cognitive workload, confirming its efficiency as a conventional input technique. Gesture interaction shows comparable performance with no significant difference in movement time relative to mouse, along with high usability scores, highlighting its potential as a practical alternative. Voice interaction, although beneficial for hands-free operation, exhibits higher movement time and sensitivity to environmental noise. Importantly, the absence of a significant interaction effect between environment and modality indicates that gesture and voice-based systems remain robust under varying real-world conditions. These findings support the need for context-aware modality selection in human--robot interaction systems.

Future work will extend this framework by incorporating additional modalities such as eye gaze, and haptic feedback to enable richer multimodal interaction. Adaptive systems that dynamically switch between modalities based on user context and task requirements will be explored. Integration with robotic platforms for real-time task execution and the inclusion of diverse user groups, particularly individuals with disabilities, will further enhance the applicability and impact of this work.

CONFLICT OF INTERESTS

All authors declare that they have no conflicts of interest.

REFERENCES

1. Biswas, P.; Langdon, P. Eye-Gaze Tracking Based Interaction in India. Procedia Computer Science 2014, 39, 59–66, doi:10.1016/j.procs.2014.11.010.
2. Hürst, W.; Van Wezel, C. Gesture-Based Interaction via Finger Tracking for Mobile Augmented Reality. Multimed Tools Appl 2013, 62, 233–258, doi:10.1007/s11042-011-0983-y.
3. Card, S.K.; Moran, T.P.; Newell, A. The Keystroke-Level Model for User Performance Time with Interactive Systems. Commun. ACM 1980, 23, 396–410, doi:10.1145/358886.358895.
4. Biswas, P.; Langdon, P. A New Interaction Technique Involving Eye Gaze Tracker and Scanning System. In Proceedings of the Proceedings of the 2013 Conference on Eye Tracking South Africa; ACM: Cape Town South Africa, August 29 2013; pp. 67–70.
5. Biswas, P.; Dv, J. Eye Gaze Controlled MFD for Military Aviation. In Proceedings of the 23rd International Conference on Intelligent User Interfaces; ACM: Tokyo Japan, March 5 2018; pp. 79–89.
6. Intuitive. Robotic Assisted Surgery: Da Vinci Surgical System. Available at: https://www.intuitive.com/en-in. Accessed on 11 February 2026.
7. Saren, S.; Mukhopadhyay, A.; Ghose, D.; Biswas, P. Comparing Alternative Modalities in the Context of Multimodal Human–Robot Interaction. *J Multimodal User Interfaces* 2024, *18*, 69–85, doi:10.1007/s12193-023-00421-w.
8. Mitra, M.; Patil, A.A.; Mothish, G.; Kumar, G.; Mukhopadhyay, A.; L R D, M.; Chakraborty, P.P.; Biswas, P. Multimodal Target Prediction for Rapid Human-Robot Interaction. In Proceedings of the Companion Proceedings of the 29th International Conference on Intelligent User Interfaces; ACM: Greenville SC USA, March 18 2024; pp. 18–23.
9. Lee, J.D.; Caven, B.; Haake, S.; Brown, T.L. Speech-Based Interaction with In-Vehicle Computers: The Effect of Speech-Based E-Mail on Drivers' Attention to the Roadway. *Hum Factors* 2001, *43*, 631–640, doi:10.1518/001872001775870340.
10. Doyle, J.; Bertolotto, M. Combining Speech and Pen Input for Effective Interaction in Mobile Geospatial Environments. In Proceedings of the Proceedings of the 2006 ACM symposium on Applied computing; ACM: Dijon France, April 23 2006; pp. 1182–1183.
11. Coronado, E.; Villalobos, J.; Bruno, B.; Mastrogiovanni, F. Gesture-Based Robot Control: Design Challenges and Evaluation with Humans. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA); IEEE: Singapore, Singapore, May 2017; pp. 2761–2767.
12. Hettig J., Saalfeld P., Luz M., Becker M., Skalej M. and Hansen C. Comparison of Gesture and Conventional Interaction Techniques for Interventional Neuroradiology. International Journal of Computer Assisted Radiology and Surgery, 2017; 12(9); pp. 1643--1653.
13. Van Den Bergh, M.; Carton, D.; De Nijs, R.; Mitsou, N.; Landsiedel, C.; Kuehnlenz, K.; Wollherr, D.; Van Gool, L.; Buss, M. Real-Time 3D Hand Gesture Interaction with a Robot for Understanding Directions from Humans. In Proceedings of the 2011 RO-MAN; IEEE: Atlanta, GA, USA, July 2011; pp. 357–362.
14. International Organization for Standardization (ISO). ISO 9241-9:2016 Ergonomics of Human-System Interaction -- Part 9: Requirements for Non-Keyboard Input Devices. Available at: https://www.iso.org/obp/ui/#iso:std:iso:9241:-9:ed-1:v1:en. Accessed on 11 February 2026.
15. Google AI. MediaPipe Hands Solution Documentation. Available at: https://mediapipe.readthedocs.io/en/latest/solutions/hands.html. Accessed on 11 February 2026.

16. Microsoft. Speech Recognition. Microsoft Learn, 2025. Available at: https://learn.microsoft.com/en-us/windows/apps/develop/input/speech-recognition. Accessed on 11 February 2026.
17. Fitts, P.M. The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement. *Journal of Experimental Psychology: General* 1992, *121*, 262–269, doi:10.1037/0096-3445.121.3.262.