International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

IRIS:GESTURE NAVIGATION CONTROL

Aswin R *1, Bhavya S Kumar *2,Boomika S *3, Thomas Jacob *4, Varsha Varghese *5, Linda Sebastian *6
College Of Engineering,Kidangoor
Kottayam, Kerala, India
aswinrjuly2004@gmail.com*1 , bhavyaskumar21@gmail.com*2 , boomika b22118cse b@ce-kgr.org*3 ,
thomasjacobtj2003@gmail.com*4 , varsha_b22104cse_b@ce-kgr.org*5 , lindasebastian@ce-kgr.org*6

ABSTRACT: *This project presents a real-time multi-modal human-computer interaction system that integrates gesture-based cursor control with voice-to-text conversion. The objective is to provide an intuitive, touch-free interface for navigating on-screen elements while enabling accurate speech transcription and translation for commands or dic tation. The system operates using only a device's built-in camera and microphone, requiring no external hardware. For gesture navigation, the camera continuously captures live video frames of the user's hand. These frames are pre-processed for noise reduction and color normalization, then analyzed by a machine-learning pipeline. Using advanced computer-vision frameworks such as MediaPipe Hands and deep neural networks, the system detects and tracks key hand landmarks in real time. Recognized gestures-such as pointer movement, click, drag, or scroll-are mapped to operating-system events through coordinate-mapping and motion-smoothing algorithms, delivering stable and respon sive cursor control. Audio signals are cleaned and normal ized before being processed by a lightweight speech-to-text en gine, which outputs editable digital text. This feature sup ports tasks like composing messages, executing commands, and controlling applications through natural voice input. By combining these two input modes, the system provides a fully contactless user experience suited for accessibility ap plications. Built with Python and open-source libraries such as TensorFlow, OpenCV, and PyAutoGUI, the architecture remains scalable and cross-platform, running efficiently on laptops or low-power edge devices.*

INDEX TERMS: *Human–Computer Interaction (HCI), Gesture Recognition, Voice Recognition, Multimodal Interaction, Hand Tracking, MediaPipe, Computer Vision, Speech-to-Text, Cursor Control, Touchless Interface, Real-Time Systems, OpenCV.*

1. INTRODUCTION

Traditional human–computer interaction (HCI) methods such as keyboards, mice, and touchscreens have long been used for system control. However, these interfaces limit hands-free usage and reduce accessibility, especially for individuals with physical disabilities. Users often face chal lenges when relying on these input methods, which creates a need for more natural, flexible, and contactless interaction techniques.

Recent advancements in computer vision and machine learning have enabled the development of gesture-based interaction systems, where hand movements are captured through a camera and mapped to system commands. Tech nologies such as MediaPipe and deep learning models al low accurate real-time hand tracking, enabling actions like cursor movement, clicking, dragging, and scrolling. At the same time, voice recognition systems convert speech into text or commands, making interaction faster and more user-friendly. However, gesture systems may be affected by light ing and background conditions, while voice systems may struggle in noisy environments.

To address these limitations, modern research focuses on multimodal systems that combine gesture and voice in puts. By integrating both modalities, systems can improve accuracy, reliability, and adaptability across different envi ronments.

Based on these advancements, this project proposes a real-time multimodal human–computer interaction system that integrates gesture-based cursor control with voice-to-text and translation capabilities. The system uses only a device's built-in camera and microphone, eliminating the need for additional hardware. Hand gestures are processed using MediaPipe and mapped to system actions, while voice input is converted into text and can be translated into mul tiple languages. This enables users to interact with the system naturally while also supporting multilingual com-munication.

Overall, the proposed system provides a touchless, intu itive, and accessible interface, making it suitable for appli cations such as assistive technologies, smart environments, and multilingual communication systems. It represents a step toward more inclusive, intelligent, and user-friendly human–machine interaction.

2. LITERATURE SURVEY

"Audio-Visual Speech and Gesture Recognition by Sen sors of Mobile Devices" [1] presents a deep learning–based multimodal system for recognizing speech and gestures us ing only a device's built-in camera and microphone. It includes two modules: Audio-Visual Speech Recognition (AVSR) and Gesture Recognition. The AVSR module uses a dual-stream architecture combining audio features (via PANN) and visual features (via CNN models like ResNet-18 and VGG), with BiLSTM for temporal modeling. The gesture module uses MediaPipe Holistic to extract hand and facial landmarks, along with attention mechanisms and BiLSTM for sequence learning. Dimensionality reduction techniques such as PCA, LDA, and t-SNE improve effi ciency. Evaluated on LRW and AUTSL datasets, the sys tem achieves high accuracy (98.76 percentage for AVSR and 98.56 percentage for gesture recognition), demonstrating its effectiveness for real-time multimodal interaction.

"Dynamic Visualization of VR Map Navigation Systems Supporting Gesture Interaction" [2] presents a VR-

based map navigation system that enhances user experi ence through natural gesture interaction and dynamic visual feedback. It aims to improve the intuitiveness and usability of navigation in virtual environments. The study follows a two-stage approach, beginning with a gesture elicitation experiment to identify intuitive gestures for tasks like zooming, panning, and rotation, which are then analyzed for consistency and user preference. In the second stage, a VR navigation prototype is developed using Unity, Map-box, and Leap Motion, integrating gesture-based controls with dynamic visualization for real-time interaction. The system is evaluated through user studies measuring task completion time, gesture consistency, and usability, along with Likert-scale feedback on comfort and ease of use. Re sults show that gesture-based interaction significantly im proves navigation efficiency and user engagement compared to traditional methods.

"Handtracking for clinical applications: Validation of the Google MediaPipe Hand(GMH)and the depth-enhanced GMH-D frameworks" [3] evaluates the performance of Google's MediaPipe Hand (GMH) and its depth-enhanced version, GMH-D, for clinical hand-tracking applications. Experiments were conducted on participants performing standard hand movements, with results compared against a motion capture system using metrics like RMSE, ICC, CCC, and Pearson correlation. Findings show that GMHD achieves higher accuracy and reliability due to the inte gration of depth information, making it suitable for clinical assessments. The study also illustrates different coordinate systems used for tracking, including image, world, and real-world coordinates.

"Leveraging Speech for Gesture Detection in Multi-modal Communication" [4] focuses on improving co-speech gesture detection by jointly modeling speech and visual data using a Transformer-based multimodal framework. Visual features are extracted using ST-GCN from skeletal move ments, while audio features are obtained using a VGGish CNN from Mel-spectrograms. These features are combined through early, late, and cross-modal fusion techniques to capture temporal relationships. Evaluated on the Rasen berg et al. dataset using metrics like F1-score and mAP, the model significantly outperforms unimodal approaches, highlighting the effectiveness of multimodal integration.

"Lightweight real-time hand segmentation leveraging MediaPipe landmark detection" [5] proposes a lightweight real-time hand segmentation method using MediaPipe land mark detection for AR/MR applications. It follows a six-stage pipeline combining hand localization, adaptive skin color modeling in CIELab space, and morphological refine ment to generate accurate hand masks. Evaluated on the Ego2Hands dataset, the system achieves an IoU of 0.869 and runs at around 90 FPS on a standard CPU, demon strating a strong balance between accuracy and efficiency for real-time applications.

"Multimodal Fusion of Speech and Gesture Recognition based on Deep Learning" [6] presents a deep learning–based multimodal system that combines speech and gesture recog nition to improve human–computer interaction. A CNN is used for speech recognition, while an LSTM processes ges ture sequences captured via a Leap Motion sensor. The out puts are fused using keyword matching and similarity mea sures to produce reliable commands. Experimental results show an accuracy of up to 96.67 percentage, demonstrating that multimodal fusion outperforms unimodal approaches in robustness and performance.

"Real-Time Hand Gesture Monitoring Model Based on MediaPipe's Registerable System" [7] proposes a real-time, registerable hand gesture recognition system using MediaPipe and deep learning. It introduces FingerNet, an enhanced ResNet-16 model with a FingerComb Block for im proved feature extraction, trained using a combination of Triple Loss and Cross-Entropy Loss. The system supports both predefined and user-defined gestures, improving flexi-bility. Evaluated on RGDS, AUTSL, and ChaLearn IsoGD datasets, it achieves accuracies of 87.8 percentage, 95.3 per-centage, and 57.2 percentage respectively, demonstrating strong performance and adaptability in gesture recognition. "Research Progress of Human–Computer Interaction Technology Based on Gesture Recognition" [8] provides a comprehensive review of gesture recognition technologies for HCI, comparing different sensing methods such as electro-magnetic, mechanical, EMG, and vision-based approaches. It analyzes various machine learning and deep learning mod els, including SVM, HMM, CNN, and YOLO, along with benchmark datasets like Widar 3.0 and SignFi. The sur vey reports high accuracy levels (up to 99.9 percentage) in controlled settings and highlights the strengths, limitations, and future potential of gesture recognition systems for real-world applications.

"Survey on Hand Gesture Recognition from Visual In put" [9] presents a comprehensive survey of hand gesture recognition using visual inputs such as RGB images, depth data, and videos. It reviews methods from traditional ap proaches like SVM and HMM to deep learning models such as CNN, LSTM, and GCN, along with key datasets like SHREC and EgoGesture. The study highlights major chal lenges including occlusion and generalization, while empha sizing the progress and future potential of robust, real-time gesture recognition systems for applications in HCI, health-care, and virtual environments.

"Interactive Design With Gesture and Voice Recogni tion in Virtual Teaching Environments" [10] explores the use of gesture and voice recognition to improve interaction in virtual teaching environments. Implemented using Unity on the HTC Vive Pro 2 platform, the system allows users to control virtual tools through hand gestures and voice com mands. Gesture recognition is handled by a strategy-based method, while voice commands are classified using a GRU network with attention. Evaluations using accuracy, F1-score, and user feedback show that combining gesture and voice significantly enhances engagement and interaction ef ficiency in virtual classrooms.

In conclusion, the reviewed papers show how much ges ture, voice, and multimodal interaction systems have im proved in recent years. While gesture and voice recognition work well on their own, each has its own limitations de pending on the environment. Combining both approaches helps overcome these issues, making the system more accu rate and reliable. Overall, the research clearly points toward more natural, user-friendly, and contactless ways of inter-acting with technology in the future.

3. METHODOLOGY

The proposed system IRIS, is an intelligent gesture controlled human-computer interaction system that enables users to interact with computing devices using hand ges tures, voice input, and translation features.

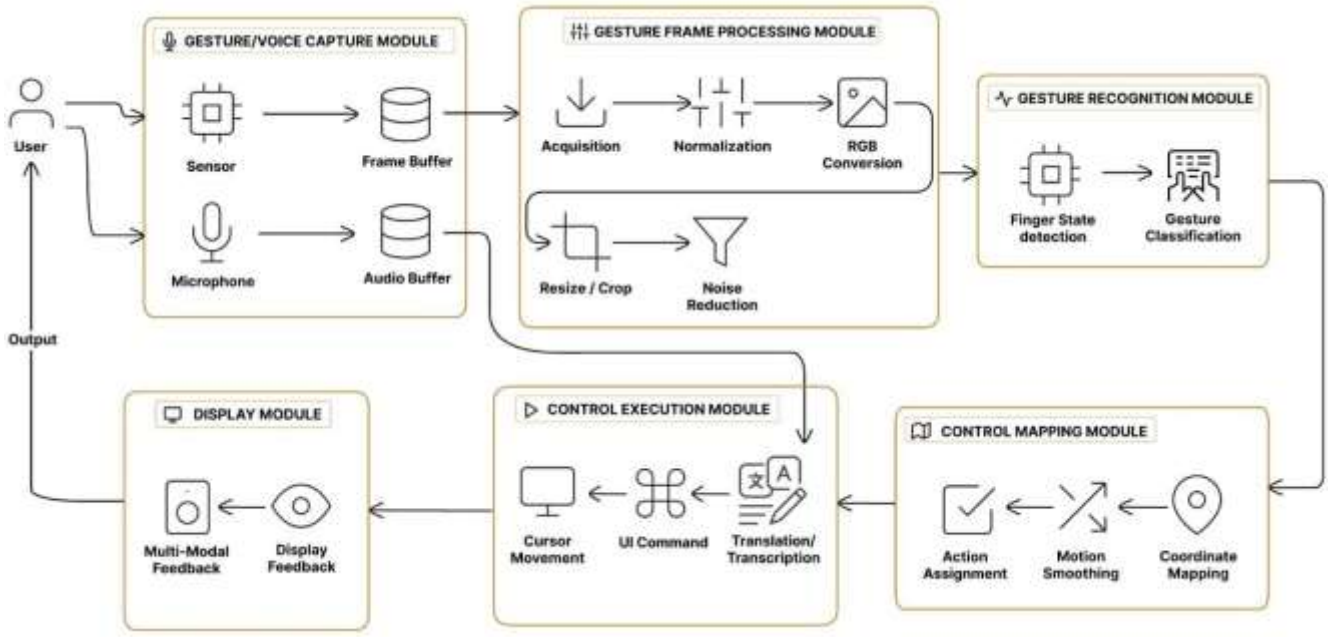

Figure 1: System Architecture

The system utilizes real-time hand tracking through com puter vision, geometric feature extraction, and rule-based gesture interpretation to perform various system-level operations. The methodology ensures seamless and efficient interaction between the user and the system. The overall workflow of the system is structured as follows:

1. System Initialization and User Interaction

The workflow of the IRIS system begins with initializing the camera and system modules. The user interacts with the system through a live video feed captured using a built-in camera. Unlike traditional systems requiring explicit au thentication, IRIS provides a direct interaction interface, where the user is immediately connected to the system upon launching the application.

The live camera feed acts as the primary input source, en abling continuous monitoring of hand movements. The sys tem processes each frame in real time, ensuring minimal latency and smooth interaction.

2. Mode Selection and Control Mechanism

Once the system is active, the user can control the system through different operational modes:

- Mouse Control Mode
- Keyboard Control Mode
- Voice Interaction Mode

Each mode is activated using predefined hand gestures de tected through the camera. The mode selection mechanism ensures that only relevant actions are performed, thereby reducing ambiguity and improving accuracy.

3. Hand Detection and Landmark Extraction

When the system receives input from the camera, it pro cesses each frame using a hand tracking module. The sys tem employs a real-time hand detection framework to iden tify hands and extract key landmark points.

Each detected hand consists of 21 landmark points, repre senting finger tips, joints, and the wrist. These landmarks are converted into pixel coordinates for further processing. The continuous extraction of these landmarks forms the foundation for gesture recognition.

4. Gesture Recognition Workflow

After detecting hand landmarks, the system performs ges ture recognition using geometric analysis. This process in volves:

- Finger State Detection: The system determines whether each finger is in an open or closed state by comparing the relative positions of landmarks.
- Distance Calculation: The system computes distances between specific landmark points to identify gestures such as clicks and drag operations.
- Multi-Hand Analysis: The system differentiates be tween left and right hands to assign specific function alities, improving control precision.

Based on these computations, gestures are classified using a rule-based approach.

5. Action Mapping and System Control

Once a gesture is recognized, it is mapped to a correspond ing system action. The action mapping module translates gestures into commands such as:

- Mouse movement and clicks
- Keyboard inputs (Enter, Backspace, Arrow keys)
- Scrolling and dragging
- Application launching

These actions are executed using system automation li braries, enabling real-time control of the operating system.

6. Voice Processing Workflow

In addition to gesture-based control, IRIS integrates a voice interaction module. When the user activates voice mode, the system captures audio input through the microphone. The audio is processed using a speech recognition model, which converts spoken input into text. The workflow in cludes:

- Audio capture and buffering
- Real-time transcription

- Output generation

7. Translation Module
The system also includes a translation module that enables language conversion. Once text is obtained from voice input, it can be translated between languages using an exter nal translation service.
8. Continuous Feedback and Real-Time Processing
The IRIS system operates in a continuous loop, where each frame from the camera is processed in real time. The system provides immediate feedback through cursor movement and executed actions, ensuring a responsive user experience.
9. System Integration
All modules in IRIS—including gesture recognition, voice processing, translation, and system control—are integrated into a unified framework.
Summary of Workflow:
Capture real-time video input, Detect hands and extract landmarks, Analyze gestures using geometric features, Classify gestures using rule-based logic, Map gestures to system actions, Execute actions in real time, Support additional interaction through voice and translation.
This structured methodology enables IRIS to function as a robust, real-time gesture-controlled system

4. RESULTS AND DISCUSSION

Figure 2: Real-Time Hand Landmark Detection (Single Hand)

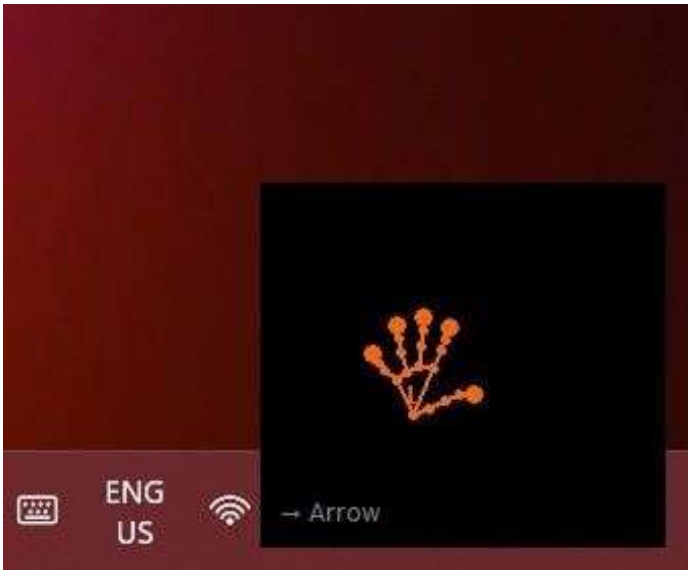

Figure 3: Real-Time Gesture Recognition and Action Dis-play

Figure 4: Dual-Hand Detection

Figure 5: Home Page

Figure 6: Home Page

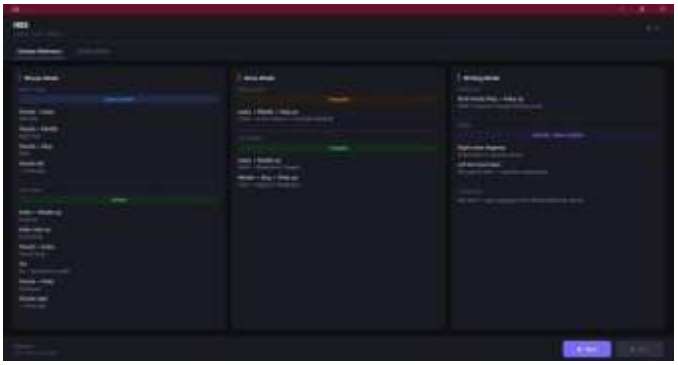

Figure 7: Desktop Application Interface – Gesture Reference Mode

Figure 8: Writing Mode with Custom Application Mapping Interface

IRIS demonstrates strong performance in enabling real-time, touchless human–computer interaction through the integration of gesture recognition and voice-based input. By utilizing MediaPipe Hands and efficient gesture-mapping al gorithms, the system is able to accurately detect and track hand landmarks in real time, ensuring smooth and respon sive cursor control. The results show that common actions such as cursor movement, clicking, dragging, and scrolling can be performed with high precision and minimal delay, making the interaction natural and intuitive.

Figure 2 illustrates real-time single-hand landmark detec tion, where the system successfully identifies key finger joints and tracks their positions continuously. This forms the core foundation for gesture recognition.

Figure 3 shows how these landmarks are interpreted into meaningful gestures, and how each gesture is mapped to a corresponding system action. The transition from detection to action is smooth, demonstrating the effectiveness of the gesture-mapping mechanism.

Figure 4 highlights the system's ability to detect and track both hands simultaneously. This dual-hand detection capa bility allows for more advanced interaction possibilities and improves the flexibility of the system.

Figures 5 and 6 present the home interface of the applica tion, which is designed with simplicity and clarity in mind. The clean layout ensures that users can easily understand and interact with the system without prior training.

Figure 7 displays the desktop application interface in ges ture reference mode, where users are provided with a list of available gestures and their corresponding functions. This feature enhances usability by helping users quickly learn and adapt to the system.

Figure 8 shows the writing mode with a custom mapping interface, which allows users to assign specific gestures or voice commands to different actions. This customization capability makes the system more adaptable to individual user preferences and application requirements.

In addition to gesture control, the system incorporates voice-to-text and translation features, enabling users to per form tasks such as typing, issuing commands, and commu nicating across different languages. The speech recognition module performs efficiently in low-noise environments, pro-ducing accurate and editable text output. The addition of translation further expands the usability of the system, making it suitable for multilingual applications.

Despite these positive results, certain limitations are ob served. Gesture recognition performance may be affected by variations in lighting, background complexity, and par-tial occlusion of the hand. Similarly, voice recognition ac curacy can decrease in noisy environments or with unclear speech input. However, the multimodal nature of the sys tem helps mitigate these issues, as users can switch between gesture and voice input depending on the situation.

Overall, the system achieves a balanced performance in terms of accuracy, speed, and usability while maintaining low computational requirements. It operates effectively on standard devices without the need for additional hardware, making it cost-effective and scalable. The results confirm that the proposed system is a practical solution for real-world applications, particularly in assistive technologies, smart environments, and modern touchless interaction sys tems.

5. .CONCLUSION

This project presents a practical and efficient real-time gesture-based navigation system that allows users to con trol the cursor and perform various UI actions using simple hand movements captured through a built-in camera. By utilizing MediaPipe Hands and optimized gesture-mapping algorithms, the system delivers smooth, accurate, and re sponsive interaction without requiring any additional hard ware, making it both cost-effective and easy to deploy.

The system is designed to be device-independent and user-friendly, ensuring accessibility for a wide range of users, including those who prefer or require touchless interaction. It maintains a good balance between accuracy, speed, and ease of use, which is essential for real-world applications. The ability to operate in real time with minimal computa tional requirements further enhances its practicality.

In addition, the framework is flexible and scalable, allowing future enhancements such as voice recognition, multilingual translation, and support for a wider range of gestures. These improvements can further expand its applications in areas like assistive technologies, smart environments, and advanced human–computer interaction systems. Overall, the project demonstrates the potential of combining sim plicity, efficiency, and innovation to create more natural and accessible ways of interacting with technology.

6. REFERENCES

[1] Ryumin, D., Ivanko, D., Ryumina, E. (2023). Audio-visual speech and gesture recognition by sensors of mobile devices. Sensors, 23(4),2284.

[2] Xiao, W., Lv, X., Xue, C. (2023). Dynamic visual-ization of vr map navigation systems supporting gesture in-teraction. ISPRS International Journal of Geo-Information,

12(3), 133.

[3] Amprimo, G., Masi, G., Pettiti, G., Olmo, G., Priano, L., Ferraris, C.(2024). Hand tracking for clinical applica-tions: Validation of the Google MediaPipe Hand (GMH) and the depth-enhanced GMH-D frameworks. Biomedical Signal Processing and Control, 96, 106508.

[4] Ghaleb, E., Burenko, I., Rasenberg, M., Pouw, W., Toni, I., Uhrig, P.... Fern´andez, R. (2024). Leveraging Speech for Gesture Detection in Multimodal Communication. arXiv preprint arXiv:2404.14952.

[5] S´anchez-Brizuela, G., Cisnal, A., de la Fuente-Lopez, E., Fraile, J.C., Perez-Turiel, J. (2023). Lightweight real-time hand segmentation leveraging MediaPipe landmark detection. Virtual Reality, 27(4), 3125-3132.

[6] Qiu, X., Feng, Z., Yang, X., Tian, J. (2020). Multimodal fusion of speech and gesture recognition based on deep learning. In Journal of Physics: Conference Series (Vol. 1453, No. 1, p. 012092). IOP Publishing.

[7] Meng, Y., Jiang, H., Duan, N., Wen, H. (2024). Real-Time Hand Gesture Monitoring Model Based on MediaPipe's Registerable System.Sensors, 24(19), 6262.

[8] Zhou, H., Wang, D., Yu, Y., Zhang, Z. (2023). Research progress of human–computer interaction technology based on gesture recognition. Electronics, 12(13), 2805.

[9] Linardakis, M., Varlamis, I., Papadopoulos, G. T. (2025). Survey on hand gesture recognition from visual input. arXiv preprint arXiv:2501.11992.

[10] Fang, K., Wang, J. (2024). Interactive design with gesture and voice recognition in virtual teaching environments. IEEE Access, 12, 4213-4224.