Review on Techniques of Gesture Navigation Control

Aswin R
Dept. of Computer Science and Engineering
College of Engineering Kidangoor
Kottayam, Kerala, India
aswinrjuly2004@gmail.com

Bhavya S Kumar
Dept. of Computer Science and Engineering
College of Engineering Kidangoor
Kottayam, Kerala, India
bhavyaskumar21@gmail.com

Boomika S
Dept. of Computer Science and Engineering
College of Engineering Kidangoor
Kottayam, Kerala,India *boomika_b22118cse_b@ce-kgr.org*

Thomas Jacob
Dept. of Computer Science and Engineering
College of Engineering Kidangoor
Kottayam, Kerala, India
thomasjacobtj2003@gmail.com

Varsha Varghese
Dept. of Computer Science and Engineering
College of Engineering Kidangoor
Kottayam,Kerala, India *varsha_b22104cse_b@ce-kgr.org*

Linda Sebastian
Dept. of Computer Science and Engineering
College of Engineering Kidangoor
Kottayam, Kerala, India
lindasebastian@ce-kgr.org

***Abstract*—Recent advancements in Human–Computer Interaction (HCI) have focused on developing multimodal systems that enable intuitive and contactless communication between users and machines. Traditional input devices such as keyboards and mice restrict accessibility and limit natural interaction, prompting research into gesture and voice-based interfaces. Numerous studies have explored vision-based gesture recognition using machine learning and computer vision frameworks like MediaPipe and OpenCV, enabling real-time detection of hand movements for cursor control, clicking, and scrolling actions. Similarly, speech recognition technologies leveraging deep learning have evolved to convert spoken language into digital text with high accuracy, facilitating command execution and dictation. Integrating these two modalities—gesture navigation and voice recognition—enhances system adaptability and usability, particularly for users with physical impairments or in touch-restricted environments. This survey indicates a growing shift toward hybrid interfaces that utilize built-in sensors such as cameras and microphones to achieve efficient, low-cost, and cross-platform operation. This convergence of gesture and speech modalities forms the foundation for developing real-time multimodal HCI systems capable of providing seamless, hands-free interaction without external hardware dependencies.**

***Keywords*—Human–Computer Interaction (HCI), Gesture Recognition, Voice Recognition, Touchless Interaction**

I. INTRODUCTION

Modern interaction systems are evolving beyond traditional input methods to improve accessibility and user experience. Conventional interfaces like keyboards, mice, and touchscreens, though widely used, restrict hands-free operation and can pose difficulties for individuals with physical disabilities or limited motor skills. These barriers have driven the development of alternative input systems such as gesture navigation, which utilizes computer vision and machine learning algorithms to detect and interpret human hand movements.

Through gesture navigation, each specific movement or pose—like waving, pinching, or pointing—is mapped to corresponding system commands, allowing users to control a cursor, perform clicks, or navigate applications without any physical contact. This technology enables natural, intuitive, and hygienic interaction, which is especially beneficial in scenarios requiring touchless operation, such as healthcare or public environments.

In addition to gestures, voice recognition enhances the interaction system by introducing a second mode of communication. Using microphone input, the system can recognize spoken commands to execute actions—such as opening files, adjusting settings, or performing searches—or convert speech into text for typing and transcription tasks. Combining both gesture and voice inputs creates a multimodal interface, offering flexibility, improved accessibility, and a more immersive user experience. This integration marks a significant step toward smarter, more inclusive human-computer interaction systems that adapt to users' needs and environments.

II. Literature Survey

A. Audio-Visual Speech and Gesture Recognition by Sensors of Mobile Devices

This work presents a deep learning–based multimodal system that recognizes speech and gestures using only the built-in camera and microphone of mobile devices [1]. The proposed system aims to enhance human–computer interaction (HCI) by enabling real-time, accurate, and noise-robust recognition of spoken words and body movements without requiring additional hardware.

The framework consists of two independent modules: Audio-Visual Speech Recognition (AVSR) and Gesture Recognition. The AVSR module employs a dual-stream neural network architecture, where audio and visual information are processed in parallel. The audio stream extracts acoustic features using Pretrained Audio Neural Networks (PANN), while the visual stream captures lip and facial movements and extracts spatial features using convolutional neural networks such as ResNet-18 and VGG. Feature-level fusion is performed at the model stage to integrate audio and visual representations, improving robustness in noisy environments. Temporal dependencies in speech are modeled using a Bidirectional Long Short-Term Memory (BiLSTM) network.

The gesture recognition module focuses on capturing dynamic body movements, particularly hand, face, and lip gestures. MediaPipe Holistic is used to extract precise spatio-temporal landmark features from video frames. An attention mechanism is incorporated to emphasize key motion frames while suppressing redundant information. Sequential modeling of gestures is achieved using BiLSTM, enabling effective learning of temporal patterns.

To enhance computational efficiency and reduce feature redundancy, dimensionality reduction techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and t-Distributed Stochastic Neighbor Embedding (t-SNE) are applied. These techniques help retain discriminative features while improving processing speed, making the system suitable for real-time mobile applications.

The proposed system is evaluated using benchmark datasets, namely LRW (Lip Reading in the Wild) for audio-visual speech recognition and AUTSL (Ankara University Turkish Sign Language) for gesture recognition. Experimental results demonstrate high performance, achieving 98.76 percentage accuracy for AVSR and 98.56 percentage accuracy for gesture recognition, thereby validating the effectiveness and robustness of the proposed multimodal recognition framework.

B. Dynamic Visualization of VR Map Navigation Systems Supporting Gesture Interaction

The paper "Dynamic Visualization of VR Map Navigation Systems Supporting Gesture Interaction" by Xiao et al. (2023) presents a VR map navigation system that enhances user experience through gesture-based interaction and dynamic visual feedback [2]. The primary objective of the study is to improve the immersiveness, intuitiveness, and usability of map navigation within virtual environments.

The authors adopt a two-stage methodology. In the first stage, a gesture elicitation and collection experiment is conducted to identify intuitive and naturally preferred gestures for common map navigation tasks, including zooming, panning, and rotation. The collected gestures are analyzed to determine consistency and user preference, forming the basis for interaction design.

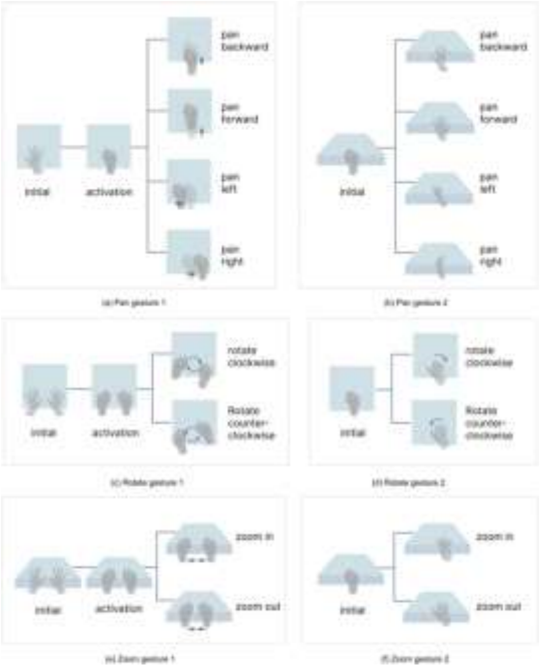

Fig. 1. Two groups of gestures. [2]

In the second stage, a functional VR navigation prototype is developed using Unity, Mapbox, and Leap Motion. The prototype integrates dynamic visualization techniques with gesture-based controls to enable real-time and responsive map interaction within a VR environment.

The system is evaluated through controlled user experiments focusing on task completion time, gesture consistency, and overall usability. Subjective user feedback is collected using Likert-scale questionnaires, assessing metrics such as comfort, ease of learning, fluency, and satisfaction. Experimental results indicate that gesture-supported dynamic visualization significantly improves navigation efficiency and user engagement compared to conventional interaction methods.

Figure 1 illustrates the two groups of gestures designed according to heuristic gesture library.

C. Handtracking for clinical applications: Validation of the Google MediaPipe Hand(GMH)and the depth-enhanced GMH-D frameworks

The paper titled "Hand Tracking for Clinical Applications: Validation of the Google MediaPipe Hand (GMH) and the Depth-Enhanced GMH-D Frameworks" evaluates the performance of Google's MediaPipe Hand framework and an enhanced variant, GMH-D, which incorporates depth information to improve tracking accuracy [3]. The primary objective of the study is to assess the suitability of these frameworks for clinical applications, particularly in the quantitative evaluation of motor symptoms associated with neurological disorders.

The authors conducted experiments using video recordings from ten healthy participants performing standardized clinical hand tasks, including finger tapping and hand opening–closing movements. Hand joint positions and motion features estimated by GMH and GMH-D were compared against measurements obtained from a gold-standard motion capture system to evaluate precision and reliability.

To quantitatively assess performance, the study employed multiple statistical evaluation metrics, including Root Mean Square Error (RMSE), Intraclass Correlation Coefficient (ICC), Concordance Correlation Coefficient (CCC), and Pearson correlation coefficient. These metrics were used to analyze both joint position estimation accuracy and the consistency of derived motion parameters.

Experimental results demonstrate that the depth-enhanced GMH-D framework achieves higher accuracy and stronger correlation with the reference motion capture system compared to the standard GMH model. The improved performance highlights the benefit of integrating depth data and indicates that GMH-D is a promising solution for precise and reliable hand-tracking in clinical assessment scenarios.

Figure 2 shows set of coordinates tracked by GMH and GMH-D: for GMH, in green Image Coordinates (pixels), centred in the upper left corner of the image; in blue, World Coordinates (metres) centred in the middle of the detected palm. In orange, the Real World Coordinates (metres) estimated by the GMH-D framework, centred in the RGB-D recording camera. For Image Coordinate of GMH, axis Zim expresses an adimensional depth parameter, relative to the wrist and scaled as the other two axes.

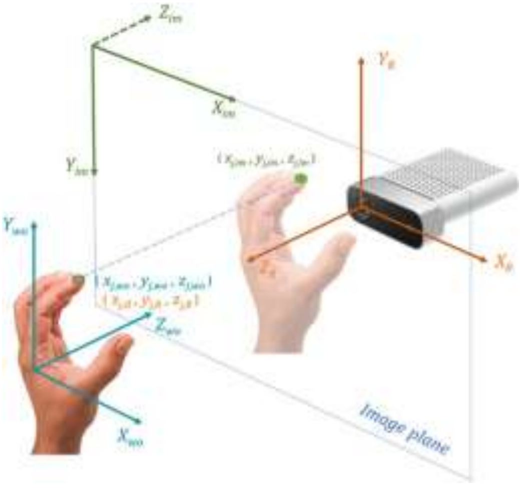

Fig. 2. Set of coordinates tracked by GMH and GMH-D. [3]

D. Leveraging Speech for Gesture Detection in Multimodal Communication

The work presented in [4] addresses the problem of co-speech gesture detection, which involves identifying natural hand and body movements that occur synchronously with spoken language. The primary objective of the study is to improve gesture detection accuracy by jointly modeling speech and visual modalities within a unified deep learning framework.

To achieve this, the authors propose a Transformer-based multimodal architecture that effectively captures temporal and cross-modal dependencies between speech and gesture data. Visual information is processed using Spatial–Temporal Graph Convolutional Networks (ST-GCN) to model structured body movement patterns based on skeletal representations. In parallel, the speech modality is analyzed using a VGGish convolutional neural network, which extracts high-level audio features from Mel-spectrogram representations.

The extracted audio and visual features are integrated using multiple fusion strategies, including early fusion, late fusion, and cross-modal fusion, within the Transformer encoder. This design enables the model to learn complex interactions between speech and gestures and to align them temporally for improved detection performance.

The proposed system is trained and evaluated on the Rasenberg et al. dataset, which consists of 19 dialogue sessions involving 38 participants and includes over 6,000 annotated gesture strokes. Performance evaluation is conducted using standard detection metrics such as F1-score, Mean Average Precision (mAP), and Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves. Experimental results demonstrate that the proposed multimodal Transformer framework significantly outperforms unimodal approaches, confirming the effectiveness of integrating speech and visual cues for accurate co-speech gesture detection.

E. Lightweight real-time hand segmentation leveraging MediaPipe landmark detection

The paper titled “Lightweight Real-Time Hand Segmentation Leveraging MediaPipe Landmark Detection” proposes an efficient and lightweight hand segmentation framework capable of operating in real time without the need for specialized hardware [5]. The primary objective of the study is to develop a computationally efficient segmentation method suitable for augmented reality (AR) and mixed reality (MR) applications, where low latency and high accuracy are critical.

The proposed approach follows a six-stage processing pipeline that integrates hand landmark detection, color-based segmentation, and post-processing refinement. Initially, a hand landmark detection model is used to localize the region of interest containing the hand. Based on this localization, the system dynamically estimates the skin color distribution in the CIELab color space, allowing it to adapt to variations in illumination conditions and skin tones. Subsequently, morphological and logical operations are applied to refine the segmentation output and generate accurate hand masks.

The method is evaluated using the Ego2Hands dataset, which contains challenging scenarios with diverse hand appearances and lighting environments. Experimental results demonstrate that the proposed system achieves a high Intersection over Union (IoU) score of 0.869 while maintaining a processing speed of approximately 90 frames per second (FPS) on a standard CPU. These results indicate that the approach provides an effective balance between segmentation accuracy and real-time performance, making it well suited for resource-constrained interactive applications.Figure 3 shows algorithm stages. a) Input image. b) MediaPipe hands output. c) Proposed segmentation solution results.

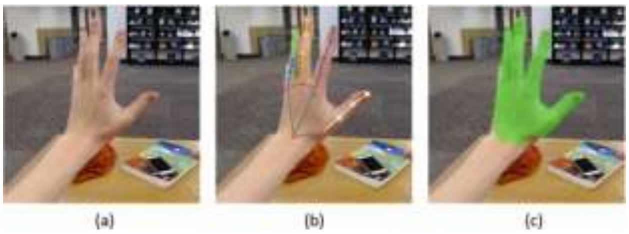

Fig. 3. Algorithm stages. [5]

F. Multimodal Fusion of Speech and Gesture Recognition based on Deep Learning

The paper titled “Multimodal Fusion of Speech and Gesture Recognition Based on Deep Learning” proposes a deep learning–based multimodal interaction system that integrates speech and gesture recognition to enhance human–computer interaction (HCI) [6]. The primary objective of the study is to address the limitations of unimodal recognition systems, such as speech degradation in noisy environments and ambiguity in gesture interpretation, by jointly exploiting complementary information from both modalities.

The proposed framework employs a Convolutional Neural Network (CNN) to process and recognize spoken commands, while hand gesture sequences are interpreted using a Long Short-Term Memory (LSTM) network. Gesture data are captured using a Leap Motion sensor, which provides precise motion information for dynamic gesture recognition. The outputs from the speech and gesture recognition modules are subsequently integrated using keyword matching and similarity comparison techniques, enabling the system to generate reliable and context-aware command decisions.

For training and evaluation, the speech recognition module utilizes data obtained through the Baidu API, while gesture recognition is trained on a custom-collected gesture dataset. Experimental results show that the proposed multimodal system achieves a recognition accuracy of up to 96.67 percentage, significantly outperforming systems that rely solely on speech or gesture input.

These findings demonstrate that multimodal fusion effectively improves robustness and reliability in interactive command recognition systems. Figure 4 shows an example of a gesture corresponding to four sets of actions.

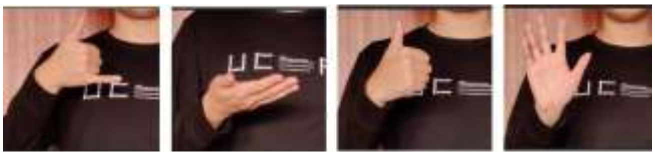

Fig. 4. Gesture examples. [6]

G. Real-Time Hand Gesture Monitoring Model Based on MediaPipe’s Registerable System

The work presented in [7] proposes a real-time and registerable hand gesture recognition system based on MediaPipe and deep learning techniques. The primary objective of the study is to develop a flexible recognition framework capable of accurately identifying predefined gestures while also allowing users to register and recognize new gestures, thereby improving adaptability in human–computer interaction (HCI) applications.

To achieve this, the authors introduce FingerNet, a modified deep neural network derived from ResNet-16, which incorporates a FingerComb Block to enhance feature extraction and accelerate model convergence. This architectural enhancement enables the network to better capture fine-grained hand and finger motion patterns obtained from MediaPipe-based landmark detection. The training strategy combines Triple Loss with Cross-Entropy Loss to improve feature discrimination and overall classification accuracy.

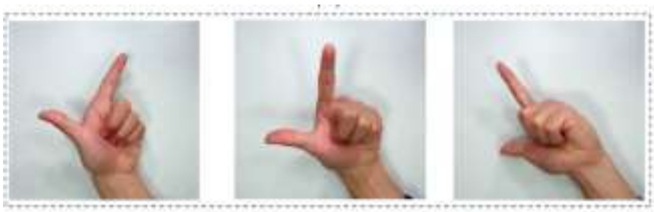

Fig. 5. Gesture data. It represent gesture photographs for different gestures. [7]

The proposed system is trained and evaluated using the Registerable Gesture Dataset (RGDS), which includes 32 gesture classes with 50 samples per class, and is further validated on benchmark datasets such as AUTSL and ChaLearn IsoGD to demonstrate generalization capability. Performance evaluation is conducted using recognition accuracy and loss convergence metrics.In figure 5 the RGDS dataset captures variations in gesture postures by including images with different finger joint bending degrees, enabling robust recognition of the same gesture across multiple poses.

Experimental results indicate that the system achieves an accuracy of 87.8 percentage on RGDS, 95.3 percentage on AUTSL, and 57.2 percentage on ChaLearn IsoGD, demonstrating the effectiveness, efficiency, and adaptability of the proposed approach in recognizing a wide range of hand gestures, including newly registered classes. In figure 6 the red dots are the 21 key points selected for the hand, which

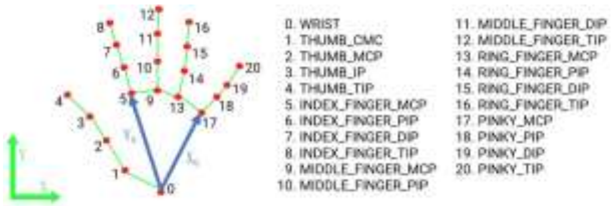

Fig. 6. MediaPipe finger landmark. [7]

are connected by a green line to form a complete line of identification of the hand.

H. Research Progress of Human–Computer Interaction Technology Based on Gesture Recognition

The study presented in [8] provides a comprehensive review and analysis of recent advancements in gesture recognition technologies for human–computer interaction (HCI). The primary objective of the research is to examine and compare existing gesture recognition approaches by analyzing their sensing mechanisms, recognition performance, and practical applicability in real-world HCI systems.

The paper categorizes gesture recognition systems based on the type of sensing technology employed, including electromagnetic sensing, mechanical (strain-based) sensing, electromyographic (EMG) sensing, and vision-based sensing. Each sensing modality is evaluated in terms of accuracy, robustness, user comfort, and deployment feasibility, highlighting the trade-offs associated with different hardware and sensing requirements.

A systematic review methodology is adopted to summarize gesture recognition principles, commonly used algorithms, and publicly available datasets. The study analyzes both machine learning and deep learning approaches, including Support Vector Machines (SVM), Hidden Markov Models (HMM), Convolutional Neural Networks (CNN), and object-detection-based methods such as YOLO, for gesture classification and recognition tasks.

Several benchmark datasets, including Widar 3.0, SignFi, Ninapro, and Leap Motion–based datasets, are examined to assess system performance using metrics such as recognition accuracy, F2-score, and robustness under varying conditions. Reported results indicate recognition accuracies of up to 99.9 percentage for certain controlled scenarios, demonstrating the effectiveness of advanced learning-based methods.

Overall, the paper offers a comparative and analytical perspective on gesture recognition technologies, identifying their strengths, limitations, and future research directions. The survey highlights the potential of integrating advanced sensing techniques with deep learning models to improve the scalability, reliability, and usability of next-generation HCI systems.

I. Survey on Hand Gesture Recognition from Visual Input

The paper titled "Survey on Hand Gesture Recognition from Visual Input" presents a comprehensive review of recent advancements in hand gesture recognition (HGR) based on visual input modalities, including RGB images, depth maps, and video sequences [9]. The primary objective of the survey is to analyze state-of-the-art techniques developed between 2018 and 2025 that enable accurate and real-time gesture understanding for applications such as sign language recognition, human–computer interaction (HCI), and virtual and augmented reality systems.

The authors adopt a systematic literature review methodology, augmented with topic modeling techniques, to categorize existing studies based on input modality, camera configuration, and recognition algorithms. The survey examines a wide range of computational approaches, encompassing traditional machine learning methods such as Support Vector Machines (SVM) and Hidden Markov Models (HMM), as well as deep learning architectures including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, Graph Convolutional Networks (GCN), and various hybrid models that combine multiple learning paradigms.

In addition, the paper reviews widely used benchmark datasets for training and evaluation, such as SHREC, EgoGesture, ChaLearn IsoGD, MS-ASL, and WLASL, highlighting their characteristics and relevance to different HGR tasks. The survey also discusses commonly used performance evaluation metrics, including accuracy, precision, recall, F1-score, and Mean Per Joint Position Error (MPJPE), which are employed to assess both recognition accuracy and pose estimation quality.

Overall, the survey highlights the significant progress achieved in visual-based hand gesture recognition, identifies existing challenges such as generalization, occlusion handling, and dataset bias, and outlines future research directions aimed at developing more robust, scalable, and generalizable HGR systems for real-world applications.Figure 7 shows indicative applications of hand gesture recognition. a) Medical Assistance b) Sign Language Interpretation c) Gaming and Virtual Environments.

Fig. 7. Indicative applications of hand gesture recognition. [9]

J. Interactive Design With Gesture and Voice Recognition in Virtual Teaching Environments

The paper titled "Interactive Design With Gesture and Voice Recognition in Virtual Teaching Environments" investigates the integration of gesture and voice recognition technologies to enhance interaction quality in virtual education systems [10]. The primary objective of the study is to provide a more natural, immersive, and intuitive interaction paradigm for teachers and students within virtual classroom environments.

The proposed system is implemented using the Unity engine and deployed on the HTC Vive Pro 2 virtual reality platform with support from the OpenXR SDK. The system enables users to interact with virtual teaching tools through hand gestures and spoken commands. For instance, a predefined fist gesture is used to activate the speech recognition module, allowing subsequent voice-based control of instructional elements.

Gesture recognition is achieved using a strategy-based recognition algorithm, while voice command classification is performed using a Gated Recurrent Unit (GRU) neural network enhanced with an attention mechanism to improve temporal feature modeling and recognition accuracy. The speech recognition component is trained on a dataset of approximately 4,000 Chinese voice samples, which are augmented using techniques such as pitch shifting and time stretching to improve model robustness. Figure 8 shows finger statement judgment example, how gesture is identified.

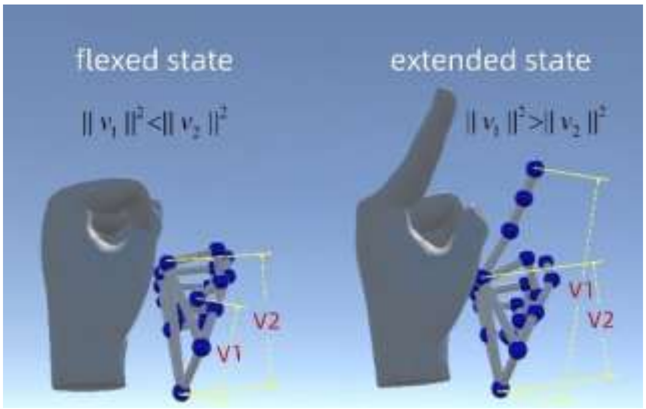

Fig. 8. Finger state judgment example. [10]

System performance is evaluated using standard metrics including accuracy, F1-score, and loss, along with user satisfaction analysis conducted through t-tests, ANOVA, and Likert-scale questionnaires. Experimental results demonstrate that the combined use of gesture and voice recognition significantly enhances user engagement and interaction efficiency, confirming the effectiveness of multimodal interaction in virtual teaching environments.

III. Comparative Study

Recent advancements in Human–Computer Interaction (HCI) have led to gesture, voice, and multimodal systems designed for different conditions and performance needs. Gesture-based systems use camera input with frameworks like MediaPipe, CNN, and LSTM to achieve high accuracy, though their performance can be affected by lighting changes, background variation, and occlusion.

Voice-based systems process audio using deep learning models such as DeepSpeech and Wav2Vec, enabling reliable command recognition in quiet environments, though their accuracy may decrease in noisy conditions.

Combining gesture and voice creates multimodal systems that integrate visual and audio data to improve accuracy and robustness across different environments. While they offer better adaptability and fewer errors, they require more computational resources due to the processing and fusion of multiple inputs.

TABLE I
Comparative Evaluation of Gesture, Voice, and Multimodal Interaction Systems

Metric	Gesture	Voice	Multimodal
Input Type	Visual frames	Audio signal	Visual and audio data
Main Models	MediaPipe, CNN, LSTM	DeepSpeech, Wav2Vec	CNN and Transformer
Accuracy (%)	92–97 (lighted)	94–98 (quiet)	96–99 (mixed)
Latency (ms)	40–60	70–100	90–120
Environment Impact	Light variation	Sound interference	Adaptive fusion
Hardware Need	Camera sensor	Microphone	Camera and mic
Key Uses	AR/VR, robotics	IoT, assistants	Smart and assistive tech

Gesture recognition provides a natural and intuitive interface but depends on proper lighting and camera placement. Voice recognition enables fast and convenient control, though its accuracy can decrease in noisy or multilingual environments. Multimodal systems combine both approaches to improve robustness and adaptability, but they require higher computational power, leading researchers to explore lightweight models and efficient fusion techniques for real-time performance on edge devices.

Conclusion

This survey highlights the rapid progress in multimodal Human–Computer Interaction (HCI) systems that combine gesture and voice recognition to enable natural, touch-free interaction. Research shows that computer vision and deep learning support accurate real-time gesture detection, while modern speech processing ensures reliable voice command recognition. Together, these technologies improve accessibility, efficiency, and user experience, especially in contactless and assistive environments.

Despite significant progress, challenges remain in improving accuracy under varying lighting and background conditions, reducing latency, and ensuring compatibility across different hardware platforms. Future research focuses on lightweight, edge-optimized models and personalized systems that can adapt to user-specific gestures and speech. Overall, combining gesture and voice interfaces is a promising step toward more intelligent, adaptive, and inclusive human–machine interaction.

Using MediaPipe and efficient gesture-mapping algorithms, the system achieves real-time performance without additional sensors, making it cost-effective and device-independent. It operates offline to ensure privacy and smooth performance in low-resource environments. Compared to traditional methods, it balances accuracy, speed, and usability, and its extendable design allows future enhancements like voice command integration and multi-gesture support for applications in smart homes, assistive technologies, and HCI systems.

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 visualization of vr map navigation systems supporting gesture interaction. 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 applications: 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., ... & Ferna´ndez, R. (2024). *Leveraging Speech for Gesture Detection in Multimodal Communication. arXiv preprint arXiv:2404.14952*.

[5] Sa´nchez-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). *Sur-vey 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*.