IRJCS: International Research Journal of Computer Science
Volume 12, Issue 08, October 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1208.05

Eyes, Voice and Hands: A New Era of Human-Computer Interaction

Shruthi K
The Oxford College of Engineering,
Bengaluru, Karnataka
Ayaan Ahmed, Chandu J
The Oxford College of Engineering,
Bengaluru, Karnataka
Faizan khan, Faraaz Khan
The Oxford College of Engineering,
Bengaluru, Karnataka

Publication History
Manuscript Reference: IRJCS/RS/Vol.12/Issue08/OCCS10084
Research Article | Open Access | Double-Blind Peer Reviewed Article ID: IRJCS/RS/Vol.12/Issue08/OCCS10084
Received:23,October 2025, Revised: 09, October 2025, Accepted: 31October 2025 Published Online: 21November 2025
https://www.irjcs.com/volumes/Vol12/iss-08/05.OCCS10084.pdf
Article Citation:Shruthi,Ayaan,Chandu,Faizan,Faraaz(2025), Eyes, Voice and Hands: A New Era of Human-Computer Interaction**,**IRJCS: International Research Journal of Computer Science, Volume 12, Issue 08 of 2025 pages 310-320 **doi:>** https://doi.org/10.26562/irjcs.2025.v1208.05
BibTeX Shruthi@2025Eyes
IRJCS papers should be cited as IRJCS (International Research Journal of Computer Science, AM Publications, India 2025, ISSN 2393-9842, https://doi.org/10.26562/ijrcs.2025.v1208.05 The journal's official abbreviation is IRJCS.
Orcid: https://orcid.org/0009-0004-9398-7488

Abstract: This project introduces a next-generation multimodal Human Computer Interaction (HCI) system that combines three natural input modalities eye-tracking, voice analysis, and hand gesture recognition to establish an intuitive, seamless, and highly efficient interaction framework between humans and digital systems. The core objective is to address the limitations of traditional input devices such as key boards, mice, and touch screens, particularly in environments where hands-free or contactless interaction is essential. The proposed system leverages a synergistic fusion of advanced technologies supported by machine learning models, including precise eye-tracking mechanisms to monitor gaze and attention, sophisticated voice recognition algorithms to interpret spoken commands, and real-time hand gesture interpretation for expressive and dynamic control. These models are trained to recognize complex patterns in user behaviours, enhancing the system's ability to interpret intent accurately and efficiently. Each modality operates in parallel and contributes uniquely to understanding user input, enabling flexible, adaptive, and reliable interaction with machines. By emulating natural human communication behaviours, the system substantially reduces interaction friction, lowers cognitive load, and enhances usability across diverse user groups. The integration of multiple input channels powered by machine learning also ensures greater accuracy, responsiveness, and robustness, especially under varying environmental and user conditions. This research advances the field of multimodal interaction by presenting a unified, extensible system architecture that supports more natural, inclusive, and intelligent human machine communication. The successful development and implementation of this machine learning based system represent a significant step toward the future of HCI, where machines can interact with users in a manner more closely aligned with human behaviors and intent.

1. INTRODUCTION

In recent years, the field of Human–Computer Interaction (HCI) has undergone a rapid transformation, driven by advancements in machine learning, sensor technologies, and user entered design. Traditional interaction methods such as keyboards, mice, and touch screens while effective in many scenarios, often fall short in dynamic, hands- free, or accessibility-focused environments. As digital systems become more integrated into everyday life, there is an increasing demand for interfaces that are natural, intuitive, and adaptable to diverse user needs. Multimodal HCI offers a promising solution by integrating multiple human input channels such as voice, gaze, and gestures to improve interaction efficiency, reduce cognitive effort, and enable seamless communication between users and machines. These modalities, when combined through intelligent systems, can interpret complex human behavior more accurately than any single in put method alone. This project aims to address these evolving requirements by developing a multimodal interaction system powered by machine learning, combining eye-tracking, voice recognition, and hand gesture detection into a unified platform. The goal is to create a user-friendly system that mirrors natural human communication and functions effectively in a wide range of contexts, enhancing both accessibility and user engagement.

2. RELATED WORK

1. Hand Gesture Recognition Using Machine Learning: A Comprehensive Review Authors: E.Wang,F.Zhang ,This comprehensive review explores the growing field of hand gesture recognition through machine learning.

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1208.05

2. The authors examine a wide range of traditional and modern classification algorithms including Support Vector Machines (SVM), decision trees, k-nearest neighbours (KNN), and neural networks for recognizing hand gestures from visual input.

3. The review also evaluates different datasets and feature extraction techniques such as contour modeling, skeletonization, and optical flow. A critical assessment of these algorithms' performance under varying environmental conditions, including different lighting, hand orientations, and occlusions, is provided. Relevance: This paper lays a strong foundation for understanding the challenges and best practices in hand gesture recognition. It is directly relevant to the multimodal HCI projects in cegesture input is one of the core modalities being integrated. Learning about algorithm performance and input preprocessing helps in selecting suitable models for real-time implementation. Limitations: Although comprehensive, the review mainly focuses on static datasets and lacks real-time implementation evaluations. It also treats gesture recognition as an isolated task, with no mention of integration with other modalities like voice or gaze.
4. Hand Gesture Recognition Using Machine Learning, Authors: Caminate Na Ranga, Paulo Jerónimo, Carlos Mora, Sandra Jardim. This study proposes a lightweight model for recognizing static Latin alphabet and gestures using RGB images. The authors use Media Pipe for hand landmark detection and Random Forest for classification. A small, custom dataset was created for training and testing the model. The focus was on ensuring low computational complexity for potential use in embedded or mobile systems. Relevance: This model's simplicity and efficiency make it highly relevant for systems with constrained processing power, such as embedded HCI applications. It demonstrates the feasibility of using hand gestures for textual input in multimodal systems. Limitations: The system is limited to static gestures and does not account for dynamic motion patterns. The gesture vocabulary is also quite restricted, making it less suitable for complex interactions. Integration with other modalities is not discussed.
5. Voice-Based Age, Gender, and Language Recognition Based on ResNet Deep Model and Transfer Learning in Spectro-temporal Domain, Author: Samira\Mavadatoti, This paper introduces a unified deep learning framework capable of recognizing age, gender, and language from speech input. The architecture employs ResNet34 and multi-attention mechanisms applied to spectro-temporal features extracted from voice signals. Transfer learning is used to enhance generalization across speakers. The model is trained on the Mozilla Common Voice dataset and validated with added Gaussian noise for robustness. Relevance: This system provides critical insights into how voice input can be used not only for commands but also for user profiling, which enhances personalization in HCI. Incorporating this capability into a multimodal system can improve system responsiveness and contextual awareness. Limitations: Despite its innovation, the system does not focus on real- time inference or integration with other input types like gestures or gaze. It remains a standalone module, limiting its application in broader multimodal environments.

6. Research on Click Enhancement Strategy of Hand–Eye Dual- Channel Human–Computer Interaction System Authors: Ya-feng Niu, Rui Chen, Yi-yan Wang, Xue-ying Yao, Yun Feng. This research tackles the accuracy issues in hand eye coordination systems used in HCI. Specifically, it addresses the Midas touch problem, where unintended clicks occur due to overlapping gaze and gesture regions. The paper evaluates various cursor techniques such as area cursors, delay clicks, and predictive models and how these affect task performance. Relevance: Highly applicable to any HCI system incorporating gaze and gesture modalities. It informs the design of user interfaces to improve reliability and reduce input errors, which is crucial for achieving seamless multimodal interaction. Limitations: The study remains limited to laboratory settings and fixed interface setups. It does not test adaptive or intelligent UI systems nor explore integration with voice-based input.

7. A Data-Driven Gaze Gesture-Based Interaction Framework for Multimodal Systems. Authors: J.Wang,L.Zhang,X. Li This work proposes a fusion framework combining gaze and gesture inputs for context-aware interaction in smart environments. Using a combination of machine learning and heuristic rules, the system maps user gaze patterns and gestures into specific commands or actions. The fusion model accounts for user intent, attention focus, and hand motion trajectory. Relevance: It contributes directly to the core objectives of a multimodal HCI system by exploring how gaze and gesture inputs can be fused. The context-aware design also informs decision-making logic within such systems. Limitations: The system's scalability is limited due to its dependence on extensive training data. The framework does not address audio input or natural language interaction.

8. Deep Learning for Multimodal Human–Computer Interaction: A Survey Authors: M.A.Jaiswal, S.Singh, R.Rawat This survey reviews the application of deep learning techniques in multimodal HCI systems. It categorizes systems based on input types visual, audio, and textual and evaluates fusion strategies (early, late, and hybrid).The paper discusses model architectures like CNNs, RNNs, and transformers and how they are applied to synchronize and interpret multiple input streams. Relevance: It offers a theoretical and technical background on multimodal fusion strategies that are vital for the current project. It helps guide the selection of deep learning architectures for integrating eye tracking, speech, and gestures. Limitations: The paper is primarily theoretical and lacks practical implementation in sights. Many systems discussed are either proof-of- concept or evaluated in simulated environments

IRJCS: International Research Journal of Computer Science
Volume 12, Issue 08, October 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1208.05

S. No	Title	Authors	Objective	Methodology Used	Research Gap
1	Mulmodal Emoon Recognion Using Deep Learning Architectures	Srinivasan,R.; Sundaram, S.; Roy,D.(2018)	Recognize user emo on through facial and voice input.	CNN and RNN models analyzing audio visual moon datasets.	Focused one moon analysis only; gesture/gaze not considered.
2	Real-Time Hand Gesture Recognion Based on Deep Learning for Contactless Control	Huang,J.; Wang,Q.; Wang,S.(2020)	Develop a touchless control system using hand gestures.	CNN-based hand gesture classifier with real-me processing.	Gesture-only focus; lacks integraon with gaze and voice input.
3	Whisper: Robust Speech Recognion via Large-Scale Weak Supervision	OpenAI(2022)	Create a mullingual and noise robust speech recognion system.	Transformer-based large-scale ASR model trained with weak supervision.	Speech-focused; does not combine other user modalies like gesture or gaze.
4	A Mulmodal Interface for Natural Human–Machine Interace on Based on Deep Learning	Papadopoulos, D.P.;Daras,P. (2021)	Buildamulmodal interface combining voiceand gestures.	Deep learning based voice and gesture fusion in smart environments.	Integrates only two modalies; Eye tracking notulized.
5	Eye Tracking for Everyone	Kraa,K.etal. (2016)	Make eye-tracking accessible using mobile devices.	CNN-based gazeesma on using front- facing cameras.	Focuses only on gaze; Lacks integraon with speech or gesture modalies.
6	Adaptive Multimodal Interfaces: Combining Voice, Gesture, and Gaze in Dynamic Environments	H. Chen, D. Kwon	Develop adaptive interfaces that adjust input weighting based on context	Real-time sensor fusion with user state modeling	Requires complex context modeling; lacks validation in long-term usage
7	Gaze-Augmented Speech Interfaces for Accessibility Applications	M. Farooq, N.Singh	Design accessible systems combining gaze and speech	Dwell-time optimization and gaze filtering	Focused narrowly on accessibility; not generalized
8	Fusion Strategies for Multimodal HCI Systems Using Reinforcement Learning	Y.Zhao, T.Mori	Optimize input fusion strategies dynamically	Reinforcement learning- based fusion weights	Requires extensive training and computational resources
9	Augmented Reality HCI Using Multimodal Input with Eye and Voice Tracking	D.Lin,K. Watanabe	Enable hands-free AR control using gaze and voice	Voice recognition +gaze direction mapping	Limited testing in non-AR environments
10	Intelligent Assistant Systems with Multimodal Dialogue Using Hand, Voice, and Eye Input	A.Becker, M.Rehm	Build a natural dialogue system using multimodal cues	Fusion of speech, gaze, and gesture to resolve ambiguity	Domain-specific; lacks general-purpose applicability
11	Multimodal Fusion Using Graph Neural Networks for HCI	R.Mehta, S.Agrawal	Improve input fusion accuracy in multimodal HCI	Graph Neural Networks to model modality relationships	Computationally intensive; not ideal for real-time systems
12	Adaptive Gaze-Voice Interfaces for Multi- User Environments	L.Han, B. Zhou, Y. Chen	Support gaze and voice command resolution in shared environments	Speaker recognition combined with gaze tracking	Accuracy reduces in dense user proximity

3. METHODOLGY

The development of the proposed multimodal Human–Computer Interaction (HCI) system follows a structured methodology consisting of eight stages: requirement analysis, data acquisition, preprocessing, feature extraction,

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1208.05

multimodal fusion, decision-making, personalization, and evaluation.
Each stage is designed to ensure modularity, scalability, and real-time performance.

Requirement Analysis

The initial phase involved defining the short comings of conventional input devices and identifying application domains requiring multimodal interaction (e.g., accessibility systems, industrial automation, immersive AR/VR).

- Hardware Requirements: A high-definition webcam or infrared (IR) sensor for eye tracking, a microphone array for speech acquisition, and RGB/depth cameras for gesture recognition. AGPU-enabled computing platform is required for real-time processing.
- Software Requirements: Python(≥3.7) with libraries such as OpenCV, MediaPipe, Tensor Flow/Py Torch for deep learning, and Whisper/Wav2Vec for speech recognition. This ensures the availability of computational resources necessary for machine learning–driven multimodal processing.

Data Acquisition

Three parallel input modalities are employed to capture natural user interaction:

- Eye-Tracking Module: A webcam/IR camera captures gaze direction, fixation points, and blink patterns.
- Voice Recognition Module: A microphone array captures spoken input in real time. Audio signals are digitized for feature extraction.
- Hand Gesture Module: RGB or depth cameras capture static and dynamic hand gestures through video frames. Simultaneous acquisition enables synchronized multimodal interaction in natural settings.

Preprocessing

Raw inputs contain environmental noise, motion artifacts, and redundancies. Preprocessing ensures robustness by filtering and normalizing data.

- Eye-Tracking: Noise filtering and extraction of ocular landmarks to estimate gaze vectors.
- Voice Recognition: Application of spectral analysis, noise suppression, and voice activity detection to prepare audio for feature extraction.
- Hand Gestures: Background subtraction, contour extraction, and hand landmark identification using MediaPipe and OpenCV.

This stage produces structured feature-ready data for each modality.

Feature Extraction and Model Training

Machine learning models are applied to capture modality-specific patterns:

- Eye-Tracking: Convolutional Neural Networks (CNNs) trained for gaze estimation.
- Voice Recognition: Transformer or Recurrent Neural Network (RNN) models (e.g., Whisper, Wav2Vec trained for Automatic Speech Recognition (ASR).
- Hand Gesture Recognition: CNN + LSTM and 3D-CNN architectures trained on gesture datasets (e.g., EgoHands, MediaPipe Hands). Each model outputs high-level features: gaze coordinates, transcribed speech commands, and gesture labels.

Multimodal Fusion

The extracted features are combined to form a unified understanding of user intent.

- Temporal Alignment: Synchronization of asynchronous signals from gaze, voice, and gestures.
- Confidence Scoring: Each modality is assigned a reliability ore depending on environmental conditions.
- Fusion Techniques:
  - Feature-level fusion for early integration.
  - Decision-level fusion using ensemble earning.
  - Attention-based fusion for context-sensitive prioritization.

This ensures robustness against noisy or missing inputs and reduces ambiguity.

Decision-Making and System Response

The Decision-Making Unit interprets fused multimodal data to infer final user intent.

- Conflict resolution is achieved by prioritizing inputs based on reliability and context.
- Commands are mapped to system actions such as selection, navigation, or activation.
- Outputs are conveyed through the Output Layer, providing graphical (GUI updates) or auditory feedback.

This ensures that user interactions are executed in real time with low latency.

Personalization and Adaptation

The system integrates adaptive learning mechanisms for long-term usability:

- Reinforcement learning techniques optimize command prediction accuracy.
- Personalized thresholds are set for voice accents, gaze sensitivity, and gesture motion ranges.
- Frequently used actions are stored as user-specific shortcuts. This adaptation reduces cognitive load and enhances user satisfaction.

Evaluation and Testing

The system is evaluated across multiple dimensions:

IRJCS: International Research Journal of Computer Science
Volume 12, Issue 08, October 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1208.05

- Performance Metrics: Recognition accuracy, system latency, error rate, and command execution success rate.
- Environmental Testing: Validation under different lighting conditions, background noise levels, and user motion.
- User Studies: Collection of subjective feedback regarding usability, comfort, and intuitiveness. Iterative refinement is conducted based on logged data and evaluation outcomes to improve reliability and robustness.

4. Architecture Diagrams.

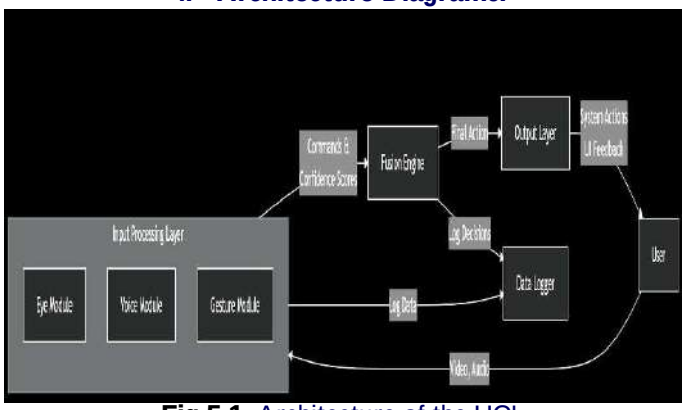

Fig 5.1: Architecture of the HCI

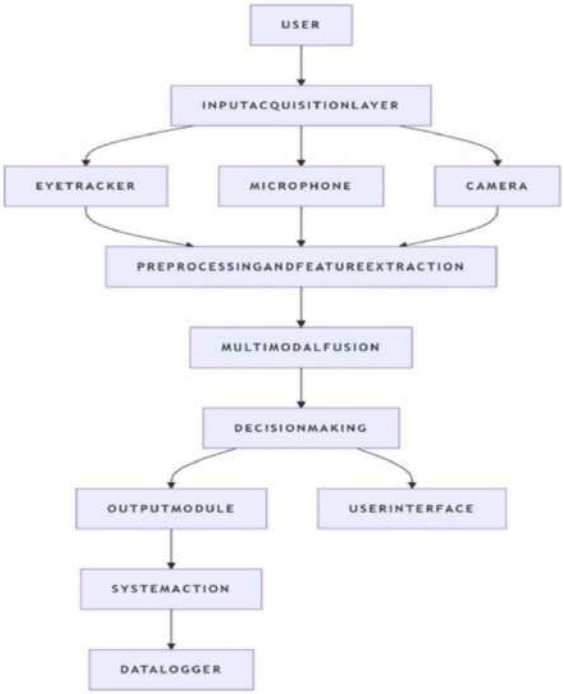

Fig 5.2: System Architecture

5. Algorithms and Implementation Techniques.

The proposed multimodal Human–Computer Interaction (HCI) system integrates multiple machine learning and computer vision algorithms across different modules. The algorithms are chosen for their robustness, efficiency, and suitability for real-time processing.

Eye-Tracking Module

- Algorithm: Convolutional Neural Networks (CNNs)
- Implementation: The gaze estimation pipeline uses

OpenCV for image acquisition and preprocessing, while Media Pipe is employed for eye landmark detection. CNN-based models are trained to map eye-region features to gaze vectors.

- Role: Enables cursor movement, fixation-based selection, and blink detection.

Voice Recognition Module

- Algorithm: Transformer-based Automatic Speech Recognition (ASR)
- Implementation: The system uses OpenAI Whisper and Wav2Vec 2.0 for speech-to-text conversion, supported by Speech Recognition library for integration. Noise reduce filters environmental noise to improve robustness.
- Role: Converts user commands into structured text and supports natural language queries.

IRJCS: International Research Journal of Computer Science
Volume 12, Issue 08, October 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1208.05

Hand Gesture Recognition Module

- Algorithm: CNN+LSTM/3D-CNNModels
- Implementation: The system employs MediaPipe Hands for real-time hand landmark detection and Tensor Flow/ PyTorch to classify static and dynamic gestures. Scikit-learn supports lightweight models for gesture classification in controlled environments.
- Role: Enables navigation, scaling, selection, and control through contactless gestures.

Multimodal Fusion Engine

- Algorithm: Hybrid Fusion (Feature-Level + Decision- Level)
- Implementation: Extracted features (gaze vectors, voice commands, gesture labels) are processed in Python 3.9+ using Tensor Flow/PyTorch. Attention-based weighting mechanisms prioritize reliable inputs under noisy conditions.
- Role: Resolves conflicts, reinforces intent, and infers the most probable user action.

System Integration and Automation

- GUI and Automation: PyAuto GUI, PyQt, Tkinter are used to build a responsive interface that reflects real-time system status.
- Development Tools: Visual Studio Code and Git provide the environment for prototyping, debugging, and version control.

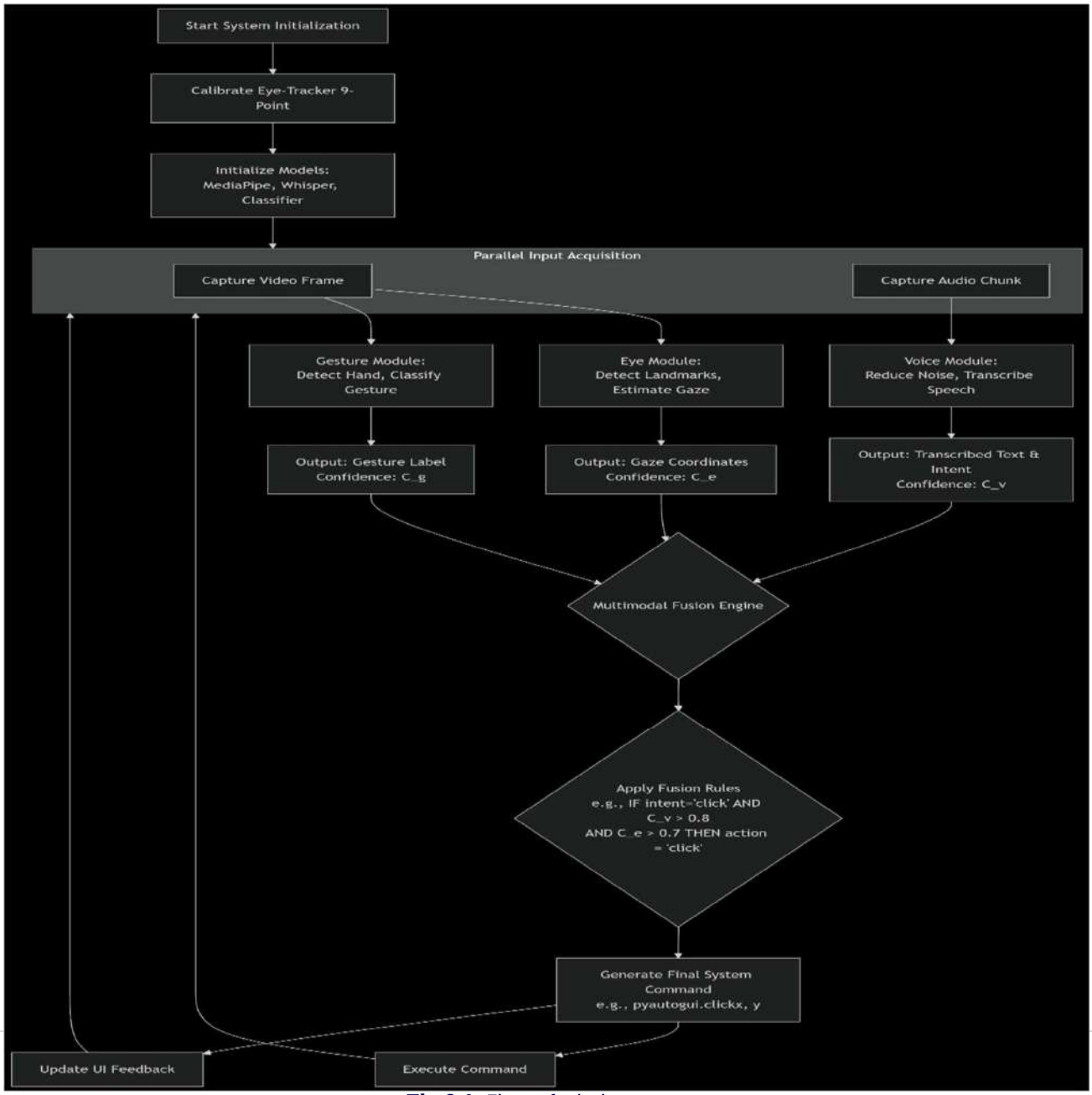

Fig 6.1: Flow of whole system

IRJCS: International Research Journal of Computer Science
Volume 12, Issue 08, October 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1208.05

6. RESULT

Fig7.1: Eye Mode Activation

Fig7.2: Voice Mode Activation

Fig 7.3: Hand Mode Activation

IRJCS: International Research Journal of Computer Science
Volume 12, Issue 08, October 2025
ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1208.05

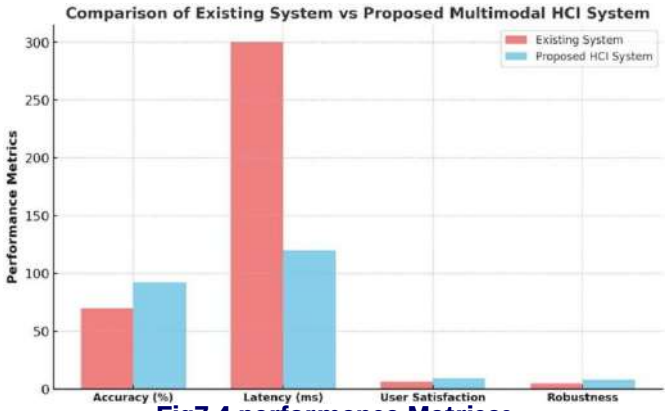

Fig7.4 performance Matrices

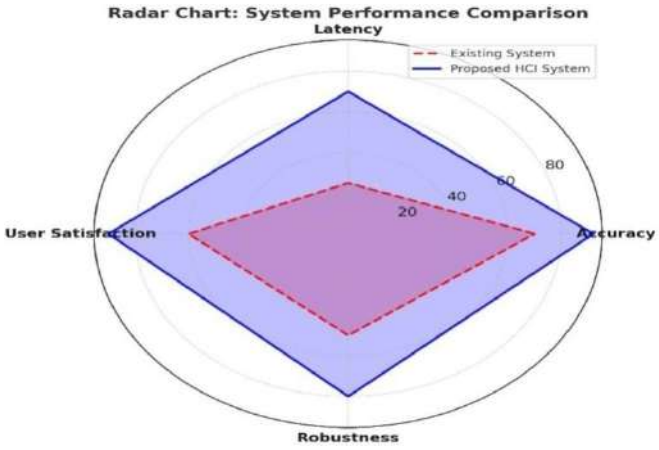

Fig7.5: System Performance Comparison Latency

7. FUTURE WORK

Although the proposed multimodal Human–Computer Interaction (HCI) system demonstrates promising results in integrating eye- tracking, voice recognition, and hand gesture recognition, several opportunities exist for enhancement and future research.

Integration of Emotion and Context Awareness

Future system scan in corporate affective computing by analyzing user facial expressions, voice tone, and physiological signals (e.g., heart rate from wearables). This would enable emotion-aware interfaces that adapt system responses based on user mood, stress, or cognitive load.

Expansion of Input Modalities

While the current system relies on three primary modalities, additional input channels can be integrated:

- Facial recognition for authentication and identity management.
- Body posture recognition for immersive AR/VR environments.
- Haptic feedback to support bi directional communication between user and machine.

Cloud and Edge Deployment

To ensure scalability and global accessibility, future work can explore:

- Cloud-based platforms for heavy model training and centralized updates.
- Edge computing for real-time inference on portable devices (e.g., AR glasses, smart phones, IoT).

Personalization through Adaptive Learning

Reinforcement learning and user profiling can be extended to create fully adaptive systems. This includes:

- Automatic adjustment of gaze sensitivity for different eye movement behaviors.
- Custom gesture recognition for users with physical limitations.
- Dynamic adaptation of speech models for diverse accents and languages.

Cross-Platform and Hardware Support

Expanding system compatibility across Windows, Linux, macOS, Android, and iOS would make the solution more versatile. Integration with wearables, smart home devices, and AR/VR headsets will widen real-world applications.

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1208.05

Security and Privacy Enhancements

As the system captures sensitive biometric data (gaze, gestures, speech), future versions should include:

- End-to-end encryption of multimodal data streams.
- Privacy-preserving machine learning techniques such as federated learning.
- User-controlled data storage and sharing policies.

Large-Scale Real-World Deployment

Extensive pilot testing in healthcare (assistive technologies), industrial automation (hands-free control), and education (interactive learning) will help validate the system's robustness. Real-world trials will generate datasets for improving recognition accuracy and usability.

8. CONCLUSION

This project successfully explores a next-generation, multimodal Human–Computer Interaction (HCI) system that seamlessly integrates eye-tracking, voice recognition, and hand gesture recognition into a unified interface powered by machine learning. The system addresses the growing limitations of traditional single- modal interfaces like keyboards, mice, and touch screens, particularly in environments where contactless, hands-free, or adaptive interaction is essential. The proposed system design demonstrates how multimodal input, when processed through intelligent data fusion techniques, can significantly enhance the intuitiveness, flexibility, and accessibility of HCI. By leveraging cutting-edge models such as convolutional neural networks (CNNs) for vision, Transformer-Based models like Whisper or Wav2Vec for speech, and 3D-CNNs for gesture recognition the project ensures robust interpretation of user intent even under complex or noisy conditions. A major strength of the system lies in its modular and scalable architecture, enabling independent upgrades and adaptations to suit different use cases and user needs. The real-time Multimodal Fusion Engine plays a pivotal role by aligning asynchronous inputs, resolving ambiguities, and reinforcing decisions using ensemble learning and confidence scoring. This capability not only ensures accurate command interpretation but also reflects the natural redundancy in human communication, where gaze, speech, and gestures often overlap to reinforce meaning. Furthermore, the project emphasizes usability and personalization, enabling the system to adapt to individual behavioral patterns and user preferences over time. This adaptability is key to enhancing long-term user experience and system responsiveness. The integration of a Data Logger and Backend Server further supports continuous system improvement through performance tracking and data-driven refinement. In conclusion, this project not only delivers a functional prototype of a multimodal HCI system but also lays down a forward-looking framework for the future of human–machine interaction. It demonstrates how intelligent, adaptive systems can bridge the gap between human intention and digital execution, paving the way for more inclusive, natural, and intelligent interfaces. As digital systems become increasingly embedded in daily life, such multimodal solutions hold the potential to revolutionize accessibility, productivity, and human empowerment across a wide range of applications—from assistive technologies and smart environments to industrial automation and immersive computing.

REFRENCES

1. S.Mavadatoti," Voice-based Age, Gender,and Language Recognition Based on ResNet Deep Model and Transfer Learning in Spectro-temporal Domain," Journal/Conference Name, Year.(Note: Please replace "Journal/Conference Name" and "Year" with the actual publication details if available.)
2. Y.-f. Niu, R. Chen,Y.-y. Wang, X.-y.Yao, andY. Feng, "Research on Click Enhancement Strategy of Hand–Eye Dual-Channel Human–Computer Interaction System, "Journal/Conference Name, Year. (Again, replace with actual publication details if known.)
3. X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in Proc. IEEE Conf.Computer Vision and Pattern Recognition(CVPR),2015,pp. 4511– 4520.

A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self supervised learning of speech representations," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 12449–12460.

4. P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW),2016, pp.1–7. [6]T. Baltrušaitis, C.Ahuja, and L.-P.Morency,"Multimodal machine learning: A survey and taxonomy," IEEETrans. Pattern Anal.Mach.Intell., vol.41,no.2, pp.423–443,Feb.2019.
5. S.Oviatt, "Ten myths of multimodal interaction,"Commun. ACM, vol. 42, no. 11, pp. 74–81, Nov. 1999. [8] R. Srinivasan, S. Sundaram, and D. Roy, "Multimodal emotion recognition using deep learning architectures," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 1–5.
6. J.Huang,Q.Wang,andS.Wang,"Real-time hand gesture recognition based on deep learning for contactless control," IEEE Access, vol. 8, pp. 189263–189273, 2020.
7. OpenAI, "Whisper: Robust speech recognition via large- scale weak supervision," OpenAI, 2022. [Online]. Available: https://openai.com/research/whisper
8. D.P.Papadopoulos and P.Daras,"A multimodal interface for natural human– machine interaction based on deep learning," Multimedia Tools and Applications, vol. 80, no. 10, pp. 15101– 15126,Apr.2021.
9. K.Krafkaetal.,"Eyetrackingfor everyone," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2176–2184.

10. Freeman, W. T., & Roth, M. (1994). Orientation histograms for hand gesture recognition.
11. Starner, T., Weaver, J., & Pentland, A. (1998). Real-time American Sign Language recognition using desk and wearable computer-based video.
12. Lee, H. K., & Kim, J. H. (1999).An HMM-based threshold model approach for gesture recognition. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints.
13. Simonyan,K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos.
14. Zhang,F., Bazarevsky, V., Vakunov, A., & Tkachenka, A. (2020). MediaPipe hands: On-device realtime hand tracking.
15. Zhang,F., Bazarevsky, V., Vakunov, A., & Tkachenka, A. (2021). MediaPipe: A Framework for Building Perception Pipelines.
16. Bradski,G.(2000).The Open CV Library.
17. Mittal, A., & Bhatia, K. (2012). Hand gesture recognition using OpenCV.
18. Sweigart,A.(2014).Automate the Boring Stuff with Python.
19. Sharma,N.,Singh,P.,&Gupta,R.(2013).Gesture-controlled presentation using OpenCV and PyAutoGUI.
20. Freeman, W. T., & Roth, M. (1994). Orientation histograms for hand gesture recognition.
21. Starner, T., Weaver, J., & Pentland, A. (1998). Real-time American Sign Language recognition using desk and wearable computer-based video.
22. Lee, H. K., & Kim, J. H. (1999).An HMM-based threshold model approach for gesture recognition. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints.
23. Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos.
24. Zhang, F., Bazarevsky, V., Vakunov, A., & Tkachenka, A. (2020). Media Pipe hands: On-device real time hand tracking.
25. Zhang, F., Bazarevsky, V., Vakunov, A., & Tkachenka, A. (2021). MediaPipe: A Framework for Building Perception Pipelines.
26. Bradski,G.(2000).The Open CVLibrary.
27. Mittal, A., & Bhatia, K. (2012). Hand gesture recognition using OpenCV.
28. Sweigart,A.(2014).Automate the Boring Stuff with Python.
29. Sharma,N.,Singh,P.,& Gupta,R.(2013).Gesture-controlled presentation using OpenCV and PyAuto GUI.
30. R.J.K. Jacob,The use of eye movements in human-computer interaction techniques: what you look at is what you get, ACM Trans. Information Syst. (TOIS)9 (2) (1991) 152–169.
31. Esteves, A., Velloso, E., Bulling, A., & Gellersen, H. (2015, November). Orbits: Gaze interaction for smart watches using smooth pursuit eye movements. In Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (pp.457-466).
32. Z.Ying, W.Kang, W.Fei-Yue, Research advances and prospects of eye tracking, Acta Automatica Sinica 48 (5) (2022)1173–1192,https://doi.org/10.16383/j.aas.c210514(In Chinese).
33. C.Z. Tang, J. Jia, Advances in occupational therapy for hand dysfunction after stroke, Chinese J. Rehab. Med. 29 (12) (2014) 1191–1195. In Chinese.
34. Encora. (2023, October 27). Beyond controllers: Apple's Vision Pro brings hand gestures and eye tracking to virtual worlds. Encora. https://www.encora.com/insights/apple-vision-probrings-handgestures-and-eye-tracking.
35. R.J.K. Jacob, What you look at is what you get: Eye movement-basedinteractiontechniques,in:Proceedingsofthe SIGCHI Conference on Human Factors in Computing Systems, ACM, 1990, pp. 11 18.
36. Blanch R, Ortega M. Rake cursor: improving pointing performance with concurrent input channels. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2009:1415-1418.
37. Wang J, Zhai S, Su H. Chinese input with keyboard and eye- tracking: an anatomical study. In: Proceedings of the SIGCHI conference on Human factors in computing systems. 2001: 349356.
38. Yamato,M.,Inoue,K.,Monden,A.,Torii, K., &Matsumoto, K.I.(2000,May).Button selection for general GUIs using eye and hand together. In: Proceedings ofthe working conference on advanced visual interfaces (pp. 270-273).
39. M.Parisay, C.Poullis,M.Kersten,Eyetap: a novel technique using voice inputs to address the midas touch problem for gaze-based interactions, arXiv preprint arXiv:2002.08455. (2020).
40. A.Ç¨oltekin, J. Hempel, A. Brychtova, I. Giannopoulos, S. Stellmach, R. Dachselt, Gaze and feet as additional input modalities for interacting with geospatial interfaces, ISPRS Ann. Photogrammetry, RemoteSens.SpatialInformationSci. 3 (2016) 113–120.
41. B.B.Velichkovsky, M.A.Rumyantsev, M.A.Morozov, New solution to the Midas touch problem: identification of visual commands via extraction of focal fixations, ProcediaComput. Sci. 39 (2014)75–82.
42. A.H. Lo, R.H. So, B.E. Shi (Eds.), Inferring Targets from Gaze, IEEE, 2014, pp. 1–6.[46]Istance, H., Bates, R., Hyrskykari, A., & Vickers, S. (2008). Snap clutch, a moded approach to solving the Mid a stouch problem.In: Proceedings oftheSymposium21Y.-F.Niuetal.onEyeTrackingResearch and Applications (ETRA'08) (pp. 221–228). NewYork, NY: ACM.
43. J.Liu,The research on dual-channel intera constrategy integrang eye control and hand control, Master's thesis, Southeast University, 2022.

ISSN: 2393-9842
https://www.irjcs.com/archives
https://doi.org/10.26562/irjcs.2025.v1208.05

44. Z.Lv,Y.Wang,C.Zhang,X.Gao,X.Wu,AnICA-based spatial filtering approach to saccadic EOG signal recognion, Biomed. Signal Process. Control 43 (2018) 9–17.
A. Pen nen,A.K.Ylitalo, Deducing self-interacon in eye movement data using sequenal spa al point processes, Spa al Stat. 17 (2016) 1–21.
45. F.Li,C.H.Lee,C.H.Chen,L.P.Khoo,Hybriddata-driven vigilance model in traffic control center using eye-tracking data and context data, Adv. Eng. Inf. 42 (2019) 100940.
46. Y.F.Niu,Y.Gao,Y.T.Zhang,C.Q.Xue,L.X.Yang, Improving eye-computer interacon interface design: ergonomic invesgaons of the opmum target size and gaze-triggering dwell me, J.EyeMov.Res.12(3)(2020),hps://doi.org/10.16910/je