International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume 8, Issue 1 (April '2025)

MULTIMODAL AI SYSTEM WITH HAND GESTURE, VOICE, AND MANUAL SHORTCUT CONTROL

Dr.B.Venkatesan,
Associate Professor and Head,
Department of Information Technology,
Paavai Engineering College,
Namakkal,Tamilnadu,India.

K.Rajesh,
Final Year B.Tech (IT),
Department of Information Technology,
Paavai Engineering College,
Namakkal,Tamilnadu,India.

Abstract: The rapid advancement of Artificial Intelligence (AI) has paved the way for sophisticated Human–Computer Interaction (HCI) systems that leverage multiple input modalities to enhance usability and accessibility. This paper proposes a Multimodal AI System integrating real-time hand gesture recognition, natural voice command processing, and customizable manual shortcuts to provide an adaptive and resilient interface for diverse user environments. Using standard webcams and microphones, the system applies computer vision techniques with Mediapipe and Convolutional Neural Networks for accurate gesture detection, and employs speech recognition algorithms for voice command interpretation. Manual shortcut mapping offers a reliable fallback to ensure uninterrupted control. By fusing these modalities at both feature and decision levels, the system achieves robust performance across variable lighting and acoustic conditions. Extensive evaluations demonstrate the system's efficacy, highlighting improvements in accessibility for differently-abled users and enhanced interaction fluidity in AR/VR, assistive technologies, and productivity contexts. The proposed approach exemplifies the practical integration of heterogeneous data streams within multimodal AI, reflecting contemporary trends in intelligent interactive systems.

Keywords: *Natural Language Processing, Artificial Intelligence, Human–Computer Interaction, Convolutional Neural Networks, Gesture Recognition Module*

I. INTRODUCTION

Traditional input modalities such as keyboards and mice have long formed the foundation of Human–Computer Interaction (HCI). Despite their ubiquity and reliability, these devices impose inherent limitations in accessibility and contextual adaptability. For individuals with motor impairments or in environments where touch-based interaction is impractical (e.g., sterile surgical theaters, hazardous industrial zones, or immersive VR settings), reliance on these peripherals poses significant barriers.
Emerging AI-driven interaction paradigms aim to overcome these restrictions by harnessing natural human modalitieshand gestures and speech commandsto provide intuitive interfaces. Gesture recognition leverages computer vision and machine learning to interpret hand and finger movements, while speech recognition uses acoustic modeling and natural language processing (NLP) to transcribe and understand spoken commands. However, when deployed in isolation, both modalities face critical challenges: gesture systems suffer from lighting variations, occlusions, and inter-user variability, whereas voice systems are vulnerable to ambient noise, accent diversity, and vocabulary limitations [1][2].

Multimodal AI systems address these shortcomings by fusing complementary input channels. By combining gestures and speech, such systems gain resilience against single-modality failures and provide richer semantic depth. This work extends beyond traditional multimodal designs by incorporating manual shortcuts as a deterministic fallback mechanism. This ensures consistent, reliable operation even in adverse conditions where both gesture and speech recognition may struggle.
The contributions of this paper are as follows:

- Proposal of a novel multimodal interaction system unifying gesture, voice, and manual shortcut inputs into a seamless control framework.
- Design of a CNN-based gesture recognition model coupled with Mediapipe landmark extraction for robust real-time hand tracking.
- Integration of ASR-based voice command recognition with adaptive vocabulary customization.
- Development of a manual shortcut module enabling user-defined deterministic controls to enhance system resilience.
- Comprehensive evaluation under varied environmental conditions, demonstrating improved accuracy, latency performance, and accessibility compared to unimodal systems.

The remainder of this paper is structured as follows: Section II surveys related work. Section III describes the system architecture. Section IV presents implementation details. Section V reports experimental evaluations. Section VI discusses findings and limitations, and Section VII concludes the paper with future directions.

II. RELATED WORK

A. Gesture Recognition Technologies

Gesture recognition has been widely investigated within HCI research. Early approaches primarily relied on

handcrafted features and template-matching techniques [3]. These methods often lacked generalizability and failed under complex conditions. With the advent of deep learning, Convolutional Neural Networks (CNNs) demonstrated superior capability in extracting spatial features, while Recurrent Neural Networks (RNNs) and LSTM models captured temporal dependencies for dynamic gesture recognition [4]. Frameworks like Google Mediapipe have further advanced this area by providing efficient real-time landmark extraction pipelines optimized for consumer-grade hardware [5]. Nonetheless, gesture systems remain sensitive to environmental lighting, occlusions, and inter-user hand shape variations.

B. Voice Command and Speech Recognition

Voice interfaces have been driven by rapid advancements in Automatic Speech Recognition (ASR). Early HMM-based acoustic models have largely been replaced by deep learning architectures, including DeepSpeech, wav2vec 2.0, and Transformer-based ASR systems [6]. These models achieve near-human transcription accuracy under controlled conditions. However, practical deployments encounter challenges including ambient noise interference, accent variability, and domain-specific vocabulary adaptation [7]. Researchers have proposed noise-robust feature extraction, transfer learning, and intent-based NLP parsing to improve system robustness.

C. Multimodal Fusion in HCI

Integrating multiple modalities has become a key strategy for enhancing interaction robustness. Fusion strategies are generally categorized as early fusion (feature-level combination), late fusion (decision-level aggregation), and hybrid approaches that optimize between accuracy and latency [8]. Applications range from assistive technologies and emotion recognition systems to immersive AR/VR environments. While multimodal systems demonstrate superior reliability, practical fallback mechanisms such as deterministic manual control are underexplored in existing literature.

D. Manual Shortcuts and Fallback Mechanisms

Shortcut-based fallback controls provide low-latency deterministic execution, guaranteeing user reliability even in failure-prone multimodal contexts. Associating shortcuts with gesture or voice commands has been shown to significantly improve user trust and satisfaction in adaptive HCI studies [9]. Despite this, research on integrating user-customizable fallback mechanisms into multimodal AI frameworks remains limited. This motivates the inclusion of manual shortcuts in the proposed system.

III. SYSTEM DESIGN AND ARCHITECTURE

The proposed system integrates three complementary modulesgesture recognition, voice recognition, and manual shortcutsVN into a fusion engine responsible for multimodal decision-making.

A. Gesture Recognition Module

The gesture module employs Mediapipe's hand landmark detector to extract 21 key points per hand, representing finger joints and palm positions. Extracted coordinates are normalized and fed into a CNN-based classifier trained on both static (e.g., fist, open palm, point) and dynamic (e.g., swipe, pinch) gestures.

The CNN model includes:

- 3 convolutional layers (3×3 kernels, ReLU activation, max-pooling)
- Batch normalization and dropout layers for regularization
- 2 fully connected dense layers producing probabilistic outputs across gesture classes

B. Voice Command Module

Real-time audio input is captured, filtered, and segmented. The ASR engine (based on pretrained wav2vec 2.0) converts speech to text, which is passed to an NLP-based command parser. The system supports user-defined voice macros for tasks like launching applications, controlling media, or taking screenshots.

C. Manual Shortcuts Module

This module enables dynamic shortcut mapping to ensure deterministic fallback functionality. Users may configure shortcut bindings (e.g., Ctrl+Alt+S for screenshot) through the UI. Manual shortcuts override other modalities when triggered, providing a fail-safe control mechanism.

D. Multimodal Fusion Engine

The fusion engine integrates decisions using a confidence-weighted late fusion strategy. Gesture and voice predictions are assigned confidence scores, while shortcuts retain highest execution priority. A rule-based policy governs conflict resolution, ensuring both adaptability and reliability.
(Insert Fig. 3: Fusion Flowchart with Priority Hierarchy)

IV. IMPLEMENTATION DETAILS

- Programming Environment: Python 3.9, TensorFlow/Keras for CNN, Mediapipe for gesture landmark detection, PyQt5 for UI.
- Hardware Platform: Standard PC with Intel i5 processor, 8 GB RAM, Logitech C920 webcam, and external microphone.
- Datasets: Custom static and dynamic gesture dataset (augmented with rotations, brightness variations), RAVDESS speech dataset supplemented with user-collected commands.
- User Interface: Real-time configuration panel for adjusting gesture sensitivity, editing voice command lists, and mapping manual shortcuts.

V. EXPERIMENTAL EVALUATION

A. Setup

30 participants (age 18–40, diverse backgrounds) tested the system under varying environments: well-lit offices, dimly lit rooms, and noisy laboratories. Metrics measured include recognition accuracy, latency, and subjective usability.

B. Results

- Gesture recognition accuracy: 92% (bright light), 86% (low light)
- Voice recognition accuracy: 88% overall, with performance decline at low signal-to-noise ratios (Fig. 4).
- Manual shortcuts: 100% reliability, unaffected by environmental conditions.
- Latency: Average system response time of 145 ms, suitable for real-time applications.

C. User Study Feedback

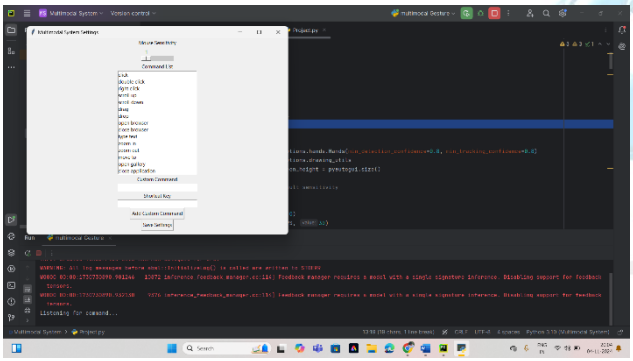
Survey responses (5-point Likert scale) indicated:

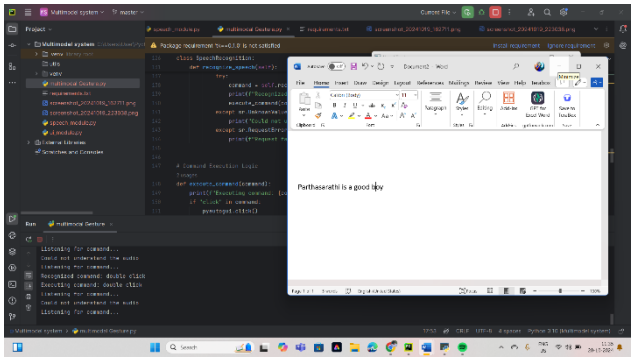
- 87% reported higher usability than unimodal systems
- 81% found manual shortcuts critical in noisy/low-light conditions
- 78% agreed system improved productivity in daily tasks

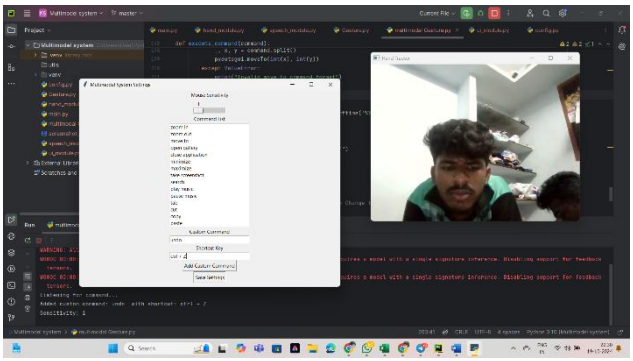
VI. DISCUSSION

The results demonstrate that the multimodal design mitigates weaknesses inherent to unimodal systems. Gesture recognition suffered in poor lighting, but fallback via voice or shortcuts preserved usability. Similarly, noisy environments degraded voice accuracy, but gestures and shortcuts ensured functionality. Manual shortcuts proved indispensable as a deterministic control layer, reinforcing user trust. Limitations include difficulty recognizing complex multi-user gesture scenes, context limitations in voice commands, and reliance on 2D webcams, which restricts depth perception. Future improvements will integrate depth cameras, Transformer-based intent models, and personalized adaptive gesture datasets.

Samples output,







VII. CONCLUSION

This paper presents a Multimodal AI System that combines gesture recognition, voice commands, and manual shortcuts into a robust interaction framework. By leveraging CNN-based gesture classification, ASR-based speech recognition, and user-configurable fallback controls, the system demonstrates high accuracy, low latency, and enhanced accessibility. Evaluation under diverse conditions highlights its potential for assistive technologies, AR/VR systems, and hands-free operation in sterile or hazardous environments.

VII. REFERENCES

[1]. Holzinger, A. (2014). Human–Computer Interaction and Usability Engineering for e-Inclusion. *Springer International Publishing.*

[2]. Zhang, Z., & Huang, G. (2017). A Survey on Emotion Recognition Using Audio-Visual Information. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 343–350. https://doi.org/10.1109/CVPRW.2017.342

[3]. Lugaresi, C., Tang, J., Puppo, E., Saba, E., et al. (2019). Mediapipe: A Framework for Perceiving and Understanding the World. *arXiv preprint arXiv:1906.08172.*

[4]. Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools.*

[5]. Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206-5210. https://doi.org/10.1109/ICASSP.2015.7178964

[6]. Atrey, P.K., Hossain, M.A., El Saddik, A., & Karam, L.J. (2010). Multimodal Fusion for Multimedia Analysis: A Survey. *Multimedia Systems*, 16:345–379.

[7]. Poria, S., Cambria, E., & Gelbukh, A. (2015). Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis. In *Proc. EMNLP*, 2539-2544. https://doi.org/10.3115/v1/D15-1294

[8]. Kaur, S., & Garg, S. (2023). Multimodal Interfaces in AR/VR Systems: Challenges and Applications. *International Journal of Interactive Multimedia and Artificial Intelligence*, 7(1), 11-21.

[9]. W3C Multimodal Interaction Framework (2013). *World Wide Web Consortium.* https://www.w3.org/TR/mmi-arch/

[10]. Holzinger, A., Malle, B., Kieseberg, P., et al. (2021). Current Trends in AI-Assisted Accessibility Technologies. *AI & Society*, 36, 159–167. https://doi.org/10.1007/s00146-020-01039-7