IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 12, April 2026
Impact Factor: 8.2

Design and Implementation of a Hand Gesture and Voice Command Interaction System for Understanding Human-Computer Interface

Prof. Rohan B. Kokate and Varsha R. Mohod
Department of Master of Computer Application
JD College of Engineering & Management, Khandala, Nagpur,
rbk7557@gmail.com and Varshamohod2002@gmail.com

Abstract: *HCI usually sticks to things like mice and keyboards, touchscreens too. They work okay. But they kind of hold back real natural ways to interact. Especially if someone has physical issues[1], [2]. Or if you need hands-free stuff in certain spots. So this paper talks about a multimodal HCI setup. It mixes hand gestures with voice commands[1], [3]. Users get a more natural feel when dealing with their computer. The gesture part uses MediaPipe in Python[14]. It tracks hand points pretty accurately[4]. Then it figures out different gestures. Those link up to regular computer tasks. Like moving the mouse. Or scrolling. Opening apps such as Notepad or Google[3], [4]. Even grabbing screenshots. At the same time the voice side runs on the Speech Recognition library[5], [6]. It handles what you say. So you can search online[1], [3]. Control music or videos. Mess with system tools. Putting these two together gives more options. Makes it accessible. Improves the whole experience in a way[1], [3].*
We go over the methods for gestures and voice. Plus some performance numbers. Results show multimodal HCI pushes things forward. It creates interfaces that feel natural. More efficient too. For all sorts of uses..

Keywords: Computer vision, gesture recognition, human-computer interaction, MediaPipe, multimodal interface, speech-to-text, voice commands

I. INTRODUCTION

Technology has changed a lot over time. It really shifted the way people deal with digital systems. Back in the day, we had punch cards. Then command-line stuff came along. Graphical user interfaces followed that. Now touchscreens are everywhere [8], [9]. Each step tried to make things easier to use. More intuitive too. Still, there this big gap. We haven't nailed interactions that feel natural [1], [2]. Like using gestures or just talking. Keyboards and mice work fine. But they get in the way sometimes. Especially if your hands are busy. Or in a clean room where you can't touch stuff. Multimodal interaction mixes different ways to input[1], [3]. At least two kinds together. That seems promising. Hand gestures don't need contact[4]. They let you express ideas clearly[4]. Voice commands keep your hands free[5], [6]. You just say what you want. Put them both in one system[1], [3]. You get something strong Flexible too. Its user-friendly. Handles all sorts of situations [1], [3].

This paper lays out a design. And how to build it. A multimodal HCI setup. It uses computer vision to spot hand gestures[4], [14]. Plus a speech-to-text tool for voice. We put together a system like that. It lets you control the computer smoothly. Through moves of your hand. And words you speak.

II. RELATED WORK

Dhamanskar and his team put together this system[13]. It mixes hand gestures with voice commands for controlling a computer[13]. Their way of doing things really shows off the good parts of combining different inputs like that. Voice

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 12, April 2026
Impact Factor: 8.2

commands turn out straightforward for stuff like playing media or handling system functions[5], [6]. All that sets up a solid base for blending those inputs in a way that works well[13], [1].

You see other research too on vision-based setups for all sorts of uses[4], [7], [10]. Zhang and Zhang came up with an interaction system for 3D TVs[10]. It relies on freehand gestures. They wanted to let people navigate virtual spaces naturally. No need for physical controllers or anything. The paper talks about a touch-based virtual interface. Still the main part on freehand gestures gives some useful ideas[10]. It covers mapping hand movements to actual controls. That ties right into what we're doing here[1],[3]. Kind of proves that just a few intuitive gestures can handle things pretty effectively[4], [7].

These earlier works pushed the whole area forward a lot[1], [2]. Even so building a full system that pulls together all kinds of gestures and voice commands for everyday computer tasks. That still poses real challenges. Our project takes those bases and builds a strong real-time setup on them. It brings in a wide range of gestures. Plus it adds in various voice commands[5], [6].

III. PROPOSED SYSTEM

The proposed multimodal HCI system provides a natural interface by combining two input modalities: hand gestures and voice commands[14].

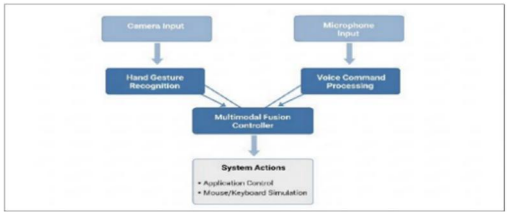

Figure : Performance Comparison of Gesture, Voice, and Integrated System

Hand gesture module. This thing takes care of pulling in video from the camera and sorting out what those hand moves actually mean[4], [14].

Video capture works simple enough. It just grabs a standard webcam to catch frames as they roll in[4]. Each frame gets scanned for a hand. MediaPipes model spots the key points on it, like the fingertips and all the joints[14].

The process looks at how those fingertips line up with the other joints. For instance, scrolling down triggers when the thumb points up and the fingers curl toward the palm[4].

Voice command module comes next. This part lets people tell the system what to do just by talking[5], [6]. Its all set up in Python with that speech recognition library. Audio capture happens through a microphone that picks up whatever the user says[5]. Speech to text then converts that sound into regular words on screen[5], [6].

Command mapping follows up. The system runs the text through a list of commands we have predefined.

Central controller ties it all together. It combines the gesture stuff and voice inputs to handle bigger tasks. Things like firing up apps, surfing the web, or running some basic math.

IV. IMPLEMENTATION AND RESULTS

Gesture-to-Action Mapping

Table 1. Gesture-to-Action Mapping

Gesture	Description	System Action
Fist (all fingers folded)	Saves screenshot	Take Screenshot
Thumb up	Scrolls down page	Scroll Down
Open hand	Scrolls up page	Scroll Up
Index + Middle + Ring up	Opens Notepad	Open Notepad

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 12, April 2026
Impact Factor: 8.2

Gesture	Description	System Action
Rock sign	Opens Google	Open Google
Claw gesture	Opens calculator	Open Calculator

B. Voice Command-to-Action Mapping

Table 2. Voice Command-to-Action Mapping

Voice Command	System Action
Start hand gestures	Activates gesture control
Open Camera	Opens Windows Camera
Open YouTube	Plays YouTube
Play Music	Plays local music
Weather in [City]	Speaks live weather update
Exit/Stop	Exits system

C. Performance Evaluation

Feature	Accuracy (%)	Latency (s)
Gesture Recognition	84%	1.5
Voice Commands	92%	1.2

V. DISCUSSION

People still talk about this multimodal HCI setup[1], [3]. It gives users a lot of flexibility since they can switch back and forth between gestures and voice commands pretty easily. Gestures work well for that precise kind of control you need sometimes[4], [7]. Voice commands handle the more complicated stuff better, like doing web searches or whatever[5], [6]. Thing is, mixing the two together really gets around the problems you run into with just one mode. It ends up being a stronger solution overall, and more accessible too[1], [3].

VI. CONCLUSION AND FUTURE WORK

This paper shows off a system for human-computer interaction that mixes hand gestures with voice commands[1], [3]. It works by linking those inputs right to what the computer does next. That setup makes things feel more natural. It also speeds up how people interact with machines[1], [3].

Down the line, there room to bring in machine learning for gestures that adjust on the fly[2], [5]. Support could stretch to phones and tablets too. And tying it into IoT gadgets seems like a solid next step[3], [5].

REFERENCES

[1] Raj, R. (2024) – Control Video With Gesture and Voice Recognition https://rjpn.org/ijcspub/papers/IJCSP24B1114.pdf

[2] Sujata, T. et al. (2025) – Gesture and Voice Controlled Virtual Assistant https://www.irjmets.com/upload_newfiles/irjmets70600219270/paper_file/irjmets70600219270.pdf

[3] Bodem, V., Kancharla, S., & Reddy, M. (2025) – Voice and Gesture-Driven IoT System https://foundryjournal.net/wp-content/uploads/2025/05/19.FJ25C628.pdf

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 12, April 2026
Impact Factor: 8.2

[4] Patel, D., & Chauhan, R. (2023) – Deep Learning-Based Human Interaction Using Hand Gestures and Voice https://thegrenze.com/pages/servej.php?association=GRENZE&fn=29.pdf&id=1572&issue=1&journal=GIJET&name=Deep+Learning+based+Human+Interactions+using+HandGestures+and+Voice+Commands&volume=9&year=2023
[5] Li, Z. et al. (2023) – Enabling Voice-Accompanying Hand-to-Face Gesture Recognition https://arxiv.org/abs/2303.10441
[6] Williams, A. S., & Ortega, F. R. (2020) – Understanding Gesture and Speech Multimodal Interactions https://arxiv.org/abs/2009.06591
[7] Xu, P. (2017) – A Real-Time Hand Gesture Recognition and Human-Computer Interaction System https://arxiv.org/abs/1704.07296
[8] Sen, A. et al. (2022) – Deep Learning-Based Hand Gesture Recognition System https://arxiv.org/abs/2206.03256
[9] Chaudhary, A. et al. (2013) – Intelligent Approaches to Interact with Machines Using Hand Gesture Recognition https://arxiv.org/abs/1303.2292
[10] Zhang, S., & Zhang, S. (2019) – A Novel Human-3DTV Interaction System Based on Free Hand Gestures and a Touch-Based Virtual Interface https://www.researchgate.net/publication/337298638_A_novel_Human-3DTV_Interaction_System_Based_on_Free_Hand_Gestures_and_a_Touch-based_Virtual_Interface
[11] Turk, M. (2014) – Multimodal Interaction: A Review https://doi.org/10.1016/j.patrec.2013.07.003
[12] Bolt, R. A. (1980) – "Put-That-There": Voice and Gesture at the Graphics Interface https://doi.org/10.1145/965105.807503
[13] Dhamanskar, P. et al. (2025) – Human–Computer Interaction Using Hand Gestures and Voice https://www.researchgate.net/publication/339980999_Human_Computer_Interaction_using_Hand_Gestures_and_Voice
[14] MediaPipe Documentation https://developers.google.com/mediapipe
[15] OpenCV Documentation https://docs.opencv.org/4.x/