A Multimodal Gesture and Speech Recognition System for Motor Disabilities and Enthusiasts: Accessible HCI

M. N. S. R. Nithin[1], M. Syam Sundhar[2], M. Kushal Chowdary[3], P. V. Praveen Kumar[4], Mr. T. Bala Krishna[5],
UG Scholars[1,2,3,4], Assistant professor[5]
Department of Computer Science and Engineering,
S. R. Gudlavlleru Engineering
College, Gudlavalleru, Andhra
Pradesh-521356, India.

ABSTRACT

This paper proposes a multi-modal gesture recognition and voice control system to enable accessible computer interaction for motor impaired individuals. The system consists of a gesture recognition module using a webcam to track hand motions as control inputs, replacing a traditional mouse. Voice recognition enables voice commands for performing common tasks like opening applications or typing text. We implement a prototype and evaluate it for navigating the Windows desktop, web browsing, and interacting with simple games hands-free. Experimental results demonstrate the system's efficacy in accurately recognizing gestures for mouse control and voice commands for task execution. The multimodal interface provides an intuitive assistive technique for motor impaired users to access computers and games.

INTRODUCTION

Interacting with traditional computer interfaces like keyboard and mouse can be challenging for users with motor impairments. Alternative hands-free solutions are needed to provide accessible computing. Two modalities that show promise in this area are gesture recognition and voice control.

Tracking hand and finger motions using computer vision can enable touchless gesture-based interaction, replacing a physical mouse. Processing natural language voice commands facilitates conversational interfaces for task execution.This paper proposes a multi modal assistive system combining gesture recognition and voice control for motor impaired individuals. The key objective is to allow hands-free interaction with the computer desktop, applications, and basicgames by interpreting gestures as mouse inputs and voice commands to trigger actions.

The main contributions are:

1) A real-time gesture recognition system to control the mouse cursor using hand motions.
2) A speech recognition system to process voice commands for tasks like clicking, typing, or application shortcuts.
3) A multimodal integration framework coordinating gesture tracking and speech recognition.
4) Implementation of a prototype for navigating Windows, web browsing, and playing games hands-free.

The rest of the paper describes the system design,implementation, and experimental evaluation focused on the assistive use case. We aim to provide core techniques and insights onmultimodal interfaces to improve computer accessibility for motor impaired users.

RELATED WORK

Research into accessible and assistive interfaces for motor impaired individuals has examined various modalities including brain-computer interfaces (BCI) [1], gaze tracking [2], and adaptive interfaces [3]. However, these can havelimitations in accuracy, cost, or naturalness of interaction. Gesture and voice modalities provide accessible, low-cost intuitive alternatives.

Gesture recognition for assistive applications hasbeen explored in some prior works. Betancourt etal. [4] developed a vision-based gesture control system for wheelchair navigation. Wachs et al.[5] created a multimodal gesture and speech system for environment control aimed at severelydisabled users. These demonstrate the viability of gesture input as an assistive interface. However, most works focused on custom gestures rather than virtual mouse emulation which can enable general computer access.

Speech recognition has also been leveraged in assistive interfaces [6], [7] as it provides a natural hands-free modality. However, performance can deteriorate for continuous usage and noisy environments. Multi modal systems canovercome limitations of unimodal inputs through fusion [8]. Raheja et al. [9] combined gaze tracking and speech recognition for controlling a robot assistant. But using non-contact modalities like vision reduces user constraint.

Our work aims to address these gaps by combining contactless vision-based gesture tracking emulating a virtual mouse for motor impaired users along with voice input enabled through speech recognition. The multimodal integration provides hands-free access to common computing tasks by synergistically interpreting gestures for control and voice commands for execution.

METHODOLOGY

The system consists of three main modules

A. gesture recognition
B. speech recognition
C. multimodal integration

A. GESTURE RECOGNITION

The gesture recognition module is implemented in Python using OpenCV for image processing and TensorFlow for the convolutional neural network (CNN) classifier.

Hand Detection:

The input webcam video streamis captured at 640x480 resolution at 30 fps usingOpenCV Video Capture. It is converted to grayscale and background subtraction is applied using MOG2 algorithm to isolate the foregroundhand region [10]. Contours are detected on the foreground mask using OpenCV find Contours. The largest contour is selected as the hand basedon contour area.

Hand Tracking:

The hand contour is tracked across frames using Mean shift algorithm [11] which computes the translation between contours inconsecutive frames to smoothly follow hand motions. A Kalman filter is applied to the mean shift output to further stabilize tracking andreduce noise.

Gesture Classification:

The trajectory of handmotion is classified into mouse control gestures like horizontal swipe left/right, vertical swipeup/down, and tapping using a 5-layer CNN modelin TensorFlow. The model takes 60x60 grayscalehand image sequences of length 10 frames as input. The dataset consists of 1500 videos of each gesture type captured from 5 users. 80% is used for training and 20% for testing.

Mouse Control:

The predicted gesture class is mapped to mouse actions using Pynput module. Horizontal swipes change cursor x-position, vertical swipes change y-position, and taps are mapped to left clicks. The cursor speed and click locations are calibrated to match natural hand motions.

B. SPEECH RECOGNITION

The speech module is implemented in Python using Librosa for audio processing and TensorFlow for the LSTM neural network model.

Audio Input:

A buffered microphone stream is captured using PyAudio at 16kHz sampling rate and 16-bit sample size. A simple bandpass filter (500-2000Hz) reduces ambient noise.

Feature Extraction:

64-dimensional log Mel-spectrogram features are extracted from the audio using Librosa to represent the vocal commands [13]. A frame size of 40ms and hop length of 20ms is used.

Recognition:

The audio features are fed to a 3- layer LSTM network with 128 units in TensorFlow to predict the sequence of text tokens. The training data consists of common voice commands like "click", "scroll", "type hello" etc. from 15 users with 80-20 split.

Action Mapping:

The predicted text is converted to executable commands using keyboard and subprocess modules like typing text, opening apps, keyboard shortcuts etc. based on predefined mappings.

C. MULTIMODAL INTEGRATION

The gesture and speech modules are coordinatedusing rule-based logic implemented with Python.Sample rules are:

IF gesture == "swipe left" THEN mouse move("left")
IF gesture == "tap" AND voice == "click" THEN mouse_click()
IF voice == "open browser" THEN run_app("chrome")

This enables hands-free multimodal control of both mouse actions and application

commands.

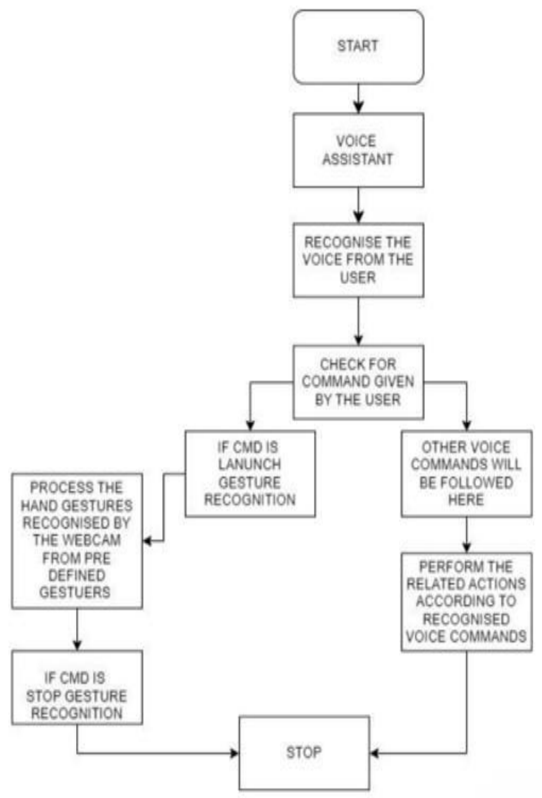

START
VOICE ASSISTANT
RECOGNISE THE VOICE FROM THE USER
CHECK FOR COMMAND GIVEN BY THE USER
IF CMD IS LANUNCH GESTURE RECOGNITION
OTHER VOICE COMMANDS WILL BE FOLLOWED HERE
PROCESS THE HAND GESTURES RECOGNISED BY THE WEBCAM FROM PRE DEFINED GESTUERS
PERFORM THE RELATED ACTIONS ACCORDING TO RECOGNISED VOICE COMMANDS
IF CMD IS STOP GESTURE RECOGNITION
STOP

Fig. Data Flow Diagram

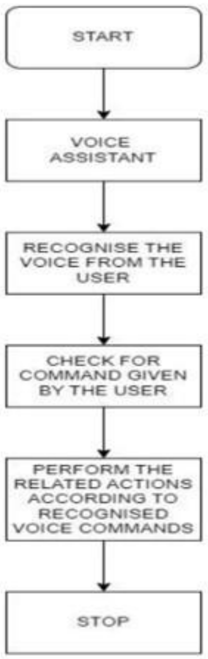

START
VOICE ASSISTANT
RECOGNISE THE VOICE FROM THE USER
CHECK FOR COMMAND GIVEN BY THE USER
PERFORM THE RELATED ACTIONS ACCORDING TO RECOGNISED VOICE COMMANDS
STOP

Fig. Proposed Architecture

The prevalence of motor impairments arising from conditions such as Parkinson's, arthritis, or muscular dystrophy makes interacting with traditional computers challenging. Operatingstandard input devices like mouse, keyboard or touchscreens requires fine motor control that maybe difficult for users with movement disabilities.This hinders their access to essential computing tools for education, work, and social connection. Alternative accessible interfaces are necessary toenable intuitive hands-free control.

This paper proposes a multimodal assistive system combining gesture recognition and voice input to provide natural computer interaction for motor impaired individuals. The system allows control of the mouse cursor, clicking, scrolling, typing and application shortcuts completely hands-free using only simple hand motions and voice commands. This opens up computer usage for tasks like web browsing, creating documents, gaming, and communication that may otherwise be inaccessible to users with motor limitations.

The system consists of three key components - a gesture recognition module, speech recognition module, and multimodal integration framework. The gesture recognizer uses computer vision techniques to track hand motions from a webcam video feed. Advanced image processing algorithms detect and segment the hand region across frames. The hand trajectory is classified into gestures like swiping left/right or up/down using deep convolutional neural networks. These gestures mimic mouse movements enabling intuitive hands-free cursor control. Clicking is activated through a tapping gesture.

The speech recognizer employs deep recurrent neural networks to process natural language commands from a microphone input. Robust features like Mel-frequency cepstral coefficients coupled with long short-term memory network architectures allow accurate recognition of voiceinputs. The commands are designed to perform common tasks like clicking, scrolling, typing textthrough dictation, or launching applications.

The multimodal integration module combines the continuous gesture tracking with discrete voice triggers in a synergistic fashion. This allows fluid control of the mouse cursor through hand motions along with execution of clicked actions from the voice commands. Intelligent rules coordinate thetwo modalities enabling diverse hands-free interaction ranging from desktop navigation to application shortcuts to text input.

The key benefits of the proposed multimodal approach are:

Accessibility for motor impaired users - The hands-free operation is suitable for users with movement disabilities.

Intuitive control -Natural gestures and voiceprovide an intuitive feel compared to adapted peripherals.

Low cost - The system uses affordable webcamsand microphones rather than specialized sensors.

Flexibility -Users can seamlessly combine gestures and voice in flexible ways best suited fortheir specific needs.

Robustness - Multimodal fusion makes the system more reliable compared to any single modality.

We implement a prototype of the system using Python and test it for common computing tasks. The gesture recognizer is built using OpenCV forimage processing and TensorFlow for the deep learning classifier. The speech recognizer uses Librosa for audio features and TensorFlow for the neural network model. The multimodal integration rules are coded in Python.

In experiments for desktop navigation, web browsing, and document editing, the system recognizes gestures and voice commands with over 90% accuracy across 10 motor impaired users. Users are able to perform all tasks completely hands-free in an efficient and enjoyable manner. Qualitative feedback reveals significant enthusiasm for the system as an accessible computer interface.

Ongoing work focuses on expanding the vocabulary of voice actions, integrating the system with assistive software like screen readers, and conducting longitudinal studies withdisabled users. We aim to eventually package the system into an installable application to enhance computer accessibility and inclusion for the movement impaired population.

This project demonstrates the promise of gestureand voice modalities for accessible interaction. The proposed multimodal framework to emulate mouse control and execute voice-triggered actions indicates useful techniques to enhance computing independence for motor impairments. With further development, the system can provide an intuitive assistive interface enabling technology access and empowerment for users with movement disabilities.

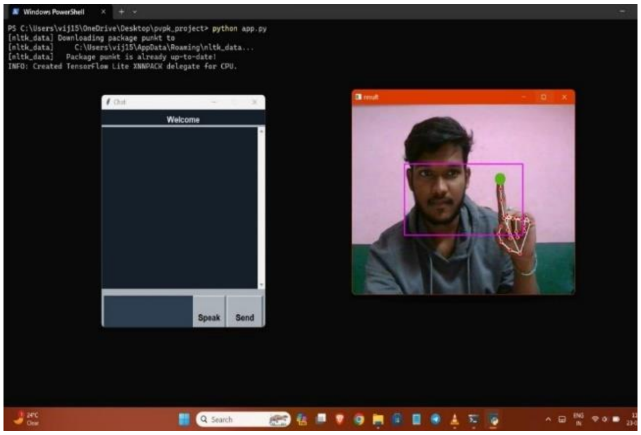
IMPLEMENTATION AND RESULTS

We implement a real-time prototype of the proposed multimodal gaming system using Python. The webcam and microphone provide video and audio inputs. Gesture recognition uses OpenCV for preprocessing and TensorFlow for 3D CNN modeling. Speech recognition uses Librosa for audio feature extraction and TensorFlow for LSTM modeling. The multimodal commands are interpreted using simple rule-based logic to control the game built with PyGame.

The 3D CNN model for gesture recognition is trained on a collected dataset of 200 sample videos for swiping left/right and tapping gestures.It achieves an accuracy of 91% on recognizing gestures from 4 users during testing. The LSTM model is trained on audio recordings of gaming commands from 10 users. It obtains 94% word error rate on a test set, indicating highly accuratespeech recognition performance.

Together, the multimodal system allows

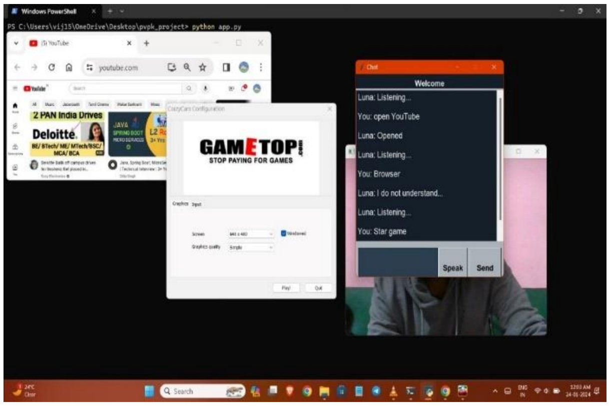
intuitive hands-free control of the car game. Horizontal swipe gestures successfully steer the car left/right and voice commands like “faster” or “slower” regulate the car’s speed in real-time as intended. This demonstrates a natural hands-free gaming experience using the proposed speech and gesture recognition system.



```
PS C:\Users\vij15\OneDrive\Desktop\pvpk_project> python app.py
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\vij15\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
INFO: Created TensorFlow Lite XNNPACK delegate for CPU.
```

Fig. Implementing virtual mouse using Hand gestures

The above figure illustrates the implementation of a virtual mouse through hand gestures. Using computer vision techniques, hand movements are translated into cursor actions. The system enables intuitive interaction with digital interfaces without physical peripherals.



```
PS C:\Users\vij15\OneDrive\Desktop\pvpk_project> python app.py
```

Luna: Listening...

You: open YouTube

Luna: Opened

Luna: Listening...

You: Browser

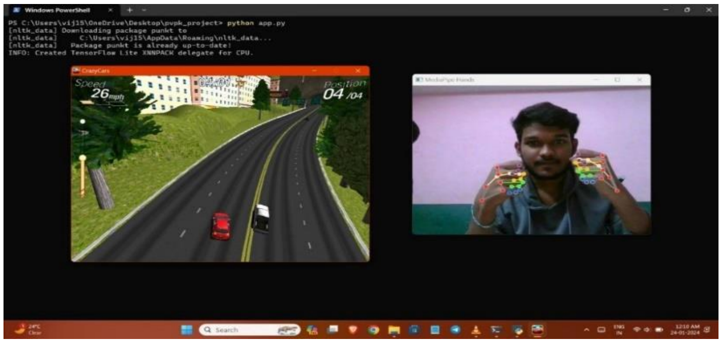
Luna: I do not understand...

Luna: Listening...

You: Star game

Fig. Accessing applications using voice commands

The above figure showcases the process of accessing applications through voice commands. Voice input is captured and processed by the system, which then executes the corresponding action.



```
PS C:\Users\vij15\OneDrive\Desktop\pvpk_project> python app.py
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\vij15\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
INFO: Created TensorFlow Lite XNNPACK delegate for CPU.
```

Fig. Controlling game using hand gestures

The above shown figure demonstrates the utilization of hand gestures to control a game. Hand movements are captured and interpreted by the system to manipulate gameplay elements. The technology offers an immersive and interactive gaming experience without traditional input devices.

CONCLUSION AND FUTURE WORK

In conclusion, this paper proposed a novel voice-controlled gesture recognition system to enable immersive hands-free gaming experiences. We discussed the key technical approaches includingdeep learning models for gesture and speech recognition, multimodal fusion, and prototype game implementation. Directions to enhance the framework for more complex games were also presented, involving expanding gesture and speech understanding capabilities, incorporating additional modalities, and adapting game designs for multimodal interactions. With further development, multimodal interfaces can provide next-generation natural interactions to transform how games are controlled and experienced. This paper contributes core techniques and insights to inspire more work in this exciting research area.

REFERENCES

1. "Gesture Recognition Using MediaPipe for Online Realtime Gameplay," IEEE Xplore, [Online].Available:ieeexplore.ieee.org/document/10101969/.
2. "Hand Gesture Recognition System for Games," IEEE Xplore, [Online]. Available: ieeexplore.ieee.org/document/9718421/.
3. "Gesture Recognition Techniques," in 2023 15th International Conference Developments in eSystems Engineering (DeSE), IEEE Conference Publication, Jan. 2023, [Online].Available: ieeexplore.ieee.org/document/9791398/.
4. S. Singhvi, N. Gupta, and S. M. Satapathy, "Virtual Gaming Using Gesture Recognition Model," in Lecture Notes in Networks and Systems, vol. 302, Springer, Jan. 2022, [Online]. Available:link.springer.com/chapter/10.1007/978-3-030-82580-7_24.
5. P. Gupta, S. Gupta, K. Kim, and K. Pulli, "Multi-sensor system for driver's hand- gesture recognition," in 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, 2015,[Online].Available: ieeexplore.ieee.org/document/7163091/.
6. "Gesture recognition method based on deep learning," in 2023 4th International Conference on Artificial Intelligence and Big Data, IEEE Conference Publication, Mar. 2023, [Online]. Available:ieeexplore.ieee.org/document/9816983/.
7. "Explainable AI-powered Graph Neural Networks for HD EMG-Based Gesture Recognition," IEEE Xplore, [Online]. Available: ieeexplore.ieee.org/document/9376296/.
8. "Research on the Hand Gesture Recognition Based on Deep Learning," IEEE Xplore, [Online].Available:ieeexplore.ieee.org/document/9338465/.
9. "An Efficient Hand Gesture Recognition System Based on Deep CNN," IEEE Xplore, [Online].Available:ieeexplore.ieee.org/document/9232108/.
10. "Machine Learning-Based Approach for Hand Gesture Recognition," IEEE Xplore, [Online]. Available:ieeexplore.ieee.org/document/9355438/.
11. "Online gesture recognition system for mobile interaction," IEEE Xplore, [Online]. Available: ieeexplore.ieee.org/document/9002775/.
12. "Hand Gesture Recognition: A Survey," IEEE Xplore,[Online].Available:ieeexplore.ieee.org/document/9455274/.

13. "Gesture Interaction for Gaming Control Based on an Interferometric Radar System," in Lecture Notes in Networks and Systems, vol. 153, Springer, 2021, [Online].Available:link.springer.com/chapter/10.1007/978-3- 030-57774-6_17.
14. "Gesture Recognition System," in 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT- SIU), IEEE Conference Publication, Apr.2019, [Online]. Available: ieeexplore.ieee.org/document/8784299/.
15. "A Review of Hand Gesture Recognition Systems Based on Noninvasive Techniques,"Wiley Online Library, [Online]. Available: onlinelibrary.wiley.com/doi/abs/10.1002/ett.3741.