Archives available at journals.mriindia.com

International Journal on Advanced Electrical and Computer Engineering

ISSN: 2349-9338
Volume 14 Issue 01, 2025

Gesture and voice-based personal computer control system

Mr. Deepak Bhonde[1], Mr. Khilesh Mongse[2], Mr. Limesh Naikwar[3], Mr. Nitesh Dwivedi[4] ,Mr. Om Mahulkar[5]

UG Student, Department of Computer Engineering, SCET, Nagpur, Maharashtra, India
[1]deepakbhonde8@gmail, 7558278392, [2]mongasekhilesh@gmail.com, 9766553941, [3]niteshdubeya@gmail.com, 9834264191, [4]limeshnaikwar07@gmail.com, 9322672300, [5]mahulkarom37@gmail.com, 7666071108

Peer Review Information

Submission: 07 Feb 2025
Revision: 16 Mar 2025
Acceptance: 18 April 2025

Keywords

Gesture Control
Voice Recognition
Human-Computer Interaction

Abstract

The rapid advancement of human-computer interaction technologies has led to the development of more intuitive and accessible control systems. This project introduces a gesture and voice-based control system that enables users to interact with their computers using hand gestures and voice commands, enhancing convenience in situations where traditional input methods are impractical. It utilizes Media Pipe Hands for real-time gesture recognition via a webcam and Speech Recognition in Python for processing voice commands, mapping them to actions such as opening applications, adjusting volume, taking screenshots, and closing programs. A PyQt5 graphical user interface (GUI) allows users to manage the system effortlessly. The results demonstrate accurate recognition, enabling seamless computer control and highlighting the potential of combining gesture and voice recognition for a more natural, hands-free computing experience. This technology has significant implications for accessibility, providing an alternative input method for individuals with physical disabilities. Future improvements may include expanding the command library, enhancing accuracy, and integrating the system with smart home devices or virtual assistants, showcasing the transformative potential of gesture and voice-based interactions in human-computer interaction.

Introduction

With the rapid advancement of technology, human-computer interaction has evolved significantly, moving beyond traditional input devices such as keyboards and mice. The rise of artificial intelligence (AI), machine learning, and computer vision has enabled the development of more natural and intuitive interaction methods, such as gesture and voice-based control systems. These systems allow users to communicate with their computers using hand movements and spoken commands, making computing more accessible, efficient, and seamless.

A Gesture and Voice-Based Personal Computer Control System is designed to enhance user experience by recognizing hand gestures and voice instructions to execute various tasks. This technology is particularly useful in scenarios where hands-free operation is essential, such as for people with disabilities, professionals in fields like healthcare and engineering, and users seeking a futuristic, touch-free computing experience. With the integration of gesture

recognition through cameras and voice processing through microphones, such a system can be used to open applications, control media, navigate documents, and perform system-level actions without requiring any physical contact.

The Role of Gesture and Voice Recognition

Gesture recognition is a technology that interprets human hand and body movements using computer vision techniques. It typically involves tracking hand positions, detecting finger movements, and identifying predefined gestures to perform corresponding commands. This can be achieved through hardware such as webcams, infrared sensors, and depth cameras or software-based solutions using AI-powered image processing. By analysing a user's hand movements, the system can trigger various operations like scrolling through a document, switching between applications, or adjusting system settings.

Similarly, voice recognition employs speech processing and natural language understanding (NLU) to recognize and interpret spoken commands. Modern voice recognition systems leverage deep learning models to convert speech into text and match it with predefined commands. This allows users to perform actions like opening software, searching for files, dictating text, and controlling smart home devices using only their voice. The integration of gesture and voice recognition creates a multi-modal interaction system, providing an alternative and efficient method of controlling a computer.

One of the key benefits of a gesture and voice-based control system is its ability to improve accessibility for individuals with physical disabilities who may struggle with traditional input devices. By enabling hands-free operation, it also enhances productivity for professionals working in fields where direct computer interaction is impractical, such as surgeons, factory workers, and mechanics.

Additionally, such systems offer a more natural and immersive user experience, making computing more intuitive. They can be integrated into smart homes, virtual reality (VR), and gaming to create a more engaging environment. Moreover, with the growing use of AI-driven assistants like Alexa, Siri, and Google Assistant, voice-based control is becoming a standard feature in personal computing, extending beyond mobile devices to desktop and laptop environments.

Despite its advantages, the implementation of gesture and voice-based systems faces challenges such as accuracy, environmental noise interference, and hardware limitations. Gesture recognition may struggle with poor lighting conditions or varying hand positions, while voice recognition can be affected by background noise and speech variations. However, advancements in AI, sensor technology, and machine learning algorithms continue to improve the efficiency and reliability of these systems.

In the future, gesture and voice-based computing could become the primary mode of interaction, eliminating the need for physical peripherals altogether. With further developments in wearable devices, AR/VR, and brain-computer interfaces, users will be able to seamlessly interact with computers using only their natural movements and speech, paving the way for a truly touch-free digital world.

LITERATURE SURVEY

The evolution of computer technology has heralded a paradigm shift in human-computer interaction, epitomized by the advent of Multimodal Interaction Systems. This innovative approach seamlessly amalgamates hand gesture recognition and voice recognition, forging a dynamic interface that redefines user engagement. Leveraging the power of low-resolution webcams and OpenCV, the system empowers users to navigate their digital realms effortlessly through intuitive gestures. From precise cursor manipulation to seamless clicking and dragging, users wield a newfound agency over their computing experience.

VoiceGesture Fusion (VGF) is like teaching computers to understand both our voices and hand movements so we can control them better. In our day-to-day activities, wireless gadgets are increasingly prevalent, with the computer mouse being a significant advancement in human-computer interaction. Even now, Bluetooth and wireless mice remain popular tools. However, it's important to note that these wireless devices still require hardware, such as batteries for power and a dongle for connecting to the computer. This study looks at how VGF works, the problems it faces, and what might be improved in the future. By combining how we talk and move, VGF helps us interact with devices like phones or computers in easier ways. It's useful for things like games, virtual reality

Hand gesture mouse control has garnered significant attention due to its versatile applications and seamless integration with machines through human-computer interaction. While traditional visual hand motion detection systems are limited by lighting conditions and complex backgrounds, advancements in computer vision and machine learning are driving the demand for enhanced human-machine interaction. The proposed methodology offers a simple yet effective solution for rapid manual tracking, overcoming the complexities of

the past. This system not only tracks hand movements and detects gestures but also addresses issues like motion blur

Elderly people face unique challenges when using conventional computer interfaces. Therefore, there is an essential need to model a system for such people for easy accessing of modern computer technologies. This paper presents an inventive solution, the "Gesture and Voice Controlled Virtual Mouse" designed to improve the digital interaction experience for older individuals facing challenges with traditional computer input peripherals. Employing advanced technology, this project establishes an intuitive interface using natural gestures and vocal commands. Gesture recognition employs cutting-edge computer vision and machine learning models, for accurate interpretation of automatic hand gestures extraction which is implemented using Media Pipe framework on top of pybind11 along with OpenCV.

METHODOLOGY AND WORKFLOW

A Gesture and Voice-Based Personal Computer Control System is an advanced human-computer interaction mechanism that enables users to operate their computers using hand gestures and voice commands. This approach eliminates the need for traditional input devices like keyboards and mice, making computing more intuitive, accessible, and efficient. It is particularly useful for people with physical disabilities, professionals who need hands-free operations, and users seeking a more immersive computing experience.

To develop such a system, a well-defined methodology and structured workflow are required. This includes data acquisition, preprocessing, feature extraction, classification, command mapping, and execution. By integrating computer vision for gesture recognition and speech recognition technologies, this system can interpret user inputs in real-time and translate them into actionable commands.

Methodology

1. Data Acquisition

The first step in developing a gesture and voice-based control system is acquiring real-time input data. This data comes from:

Gesture Input:

- Captured through a webcam, depth sensor (like Kinect), or infrared cameras (like Leap Motion Controller).
- Frames are extracted in real-time for further processing.

Voice Input:

- Captured via a microphone and processed using speech recognition libraries such as Google Speech API, CMU Sphinx, Deep Speech, or Microsoft Azure Speech.
- The system continuously listens for user commands while filtering out background noise.

2. Data Pre-processing

Raw input data is noisy and must be pre-processed before being analyzed.

Gesture Pre-processing:

- Background removal to isolate hand movements using techniques like background subtraction.
- Image filtering and smoothing (Gaussian blur) to reduce noise.
- Normalization and resizing to ensure consistency across different lighting conditions.
- Hand tracking and segmentation using OpenCV and MediaPipe Hand Tracking.

Voice Pre-processing:

- Noise reduction and filtering to enhance voice clarity.
- Conversion to Mel-Frequency Cepstral Coefficients (MFCCs), which represent the unique features of human speech.
- Segmentation of speech into phonemes for better accuracy in speech-to-text conversion.

Feature Extraction and Classification

Once the input data is pre-processed, features are extracted to classify gestures and recognize speech.

-Gesture Feature Extraction and Recognition:

- Key points of the hand (e.g., fingertips, palm position) are extracted using Media Pipe, Open Pose, or Tensor Flow.
- Feature vectors representing hand positions, movement direction, and shape are generated.
- Classification is performed using Machine Learning (ML) models like CNN (Convolutional Neural Networks), Support Vector Machines (SVM), or Decision Trees.

3. Voice Feature Extraction and Recognition:

- Spectrograms of audio signals are generated and analysed.
- Speech-to-text conversion is performed using deep learning models like RNN (Recurrent Neural Networks) or Transformer-based models like Whisper.

- The recognized text is compared with predefined voice commands.

4. Command Mapping and Action Execution

After recognizing the gesture or speech command, it must be mapped to a specific computer action.

- A predefined command dictionary stores all possible gesture-action and voice-action mappings.
- When a command is recognized, the system translates it into corresponding system actions such as:
- File Operations: Opening, closing, deleting files.
- Media Control: Play, pause, adjust volume, switch tracks.
- System Navigation: Switching applications, minimizing/maximizing windows, controlling the cursor.
- Custom User Commands: Executing user-defined automation tasks.

5. System Integration and Execution

The recognized commands are executed using system automation tools, including:

- PyAutoGUI: Automates mouse movements and keyboard inputs.
- Windows Speech Recognition API / Linux Voice Commands: For OS-level control.
- OpenCV + Tensor Flow Integration: For real-time hand gesture execution.

Workflow Of The System

The system follows a structured workflow from data collection to command execution.

Step 1: Capturing User Input

- The system starts by continuously capturing video frames for hand gestures and audio input for voice commands.
- Inputs are processed in real-time, ensuring low latency.

Step 2: Pre-processing the Data

- The captured gesture images are filtered, and the background is removed to focus only on the hand.
- Voice signals are denoised and converted into meaningful representations for analysis.

Step 3: Recognizing Gestures and Speech

- The system classifies the gesture using a deep learning model trained on various hand positions and movements.
- The speech input is converted into text and matched against predefined commands using NLP.

Step 4: Mapping to System Commands

- Recognized gestures and voice commands are compared to a command database.
- If a match is found, the system executes the corresponding action.

Step 5: Executing the Command

- The system uses automation libraries to trigger the appropriate command.
- The user receives visual feedback (GUI changes) or audio confirmation.

ADVANTAGES OF GESTURE AND VOICE-BASED PC CONTROL

1. Hands-Free Operation: Allows users to interact with their computer without touching a keyboard or mouse.

2. Improved Accessibility: Beneficial for people with physical disabilities or those who cannot use traditional input devices.

3. Enhanced Productivity: Professionals in healthcare, engineering, and industrial settings can operate computers more efficiently.

4. More Natural Interaction: Mimics real-world interactions, making computing more intuitive and engaging.

5. Multi-Modal Input: Combines gesture and voice commands for better flexibility.

CHALLENGES AND FUTURE SCOPE

Challenges

- Gesture Recognition Accuracy: Variations in lighting, hand positions, and backgrounds can reduce detection accuracy.
- Voice Recognition in Noisy Environments: Background noise may interfere with speech recognition.
- Hardware Limitations: High-performance cameras and microphones are required for accurate detection.

Future Improvements

- AI-Powered Enhancements: Use deep learning models like Vision Transformers (ViTs) for more precise gesture recognition.
- Better Noise Reduction: Advanced speech enhancement techniques can improve voice recognition.
- Integration with Wearable Devices: Smart gloves or AR/VR devices can improve user experience.

TECHNOLOGIES AND TOOLS USED

Component	Technology/Library Used
Gesture Detection	OpenCV, Media Pipe, Tensor Flow
Hand Tracking	Media Pipe Hand Tracking, Open Pose
Gesture Recognition	CNN, SVM, Decision Trees
Voice Processing	Google Speech API, CMU Sphinx, Deep Speech
Speech Recognition	RNN, Transformer-based models
Automation Tools	PyAutoGUI, Windows Speech API, Linux Commands

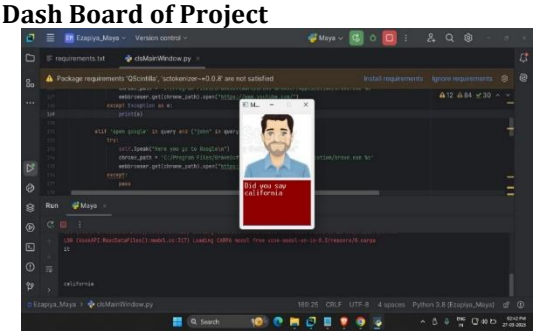
CONCLUSION

The Gesture and Voice-Based Personal Computer Control System represents a significant leap in human-computer interaction (HCI), making computing more natural, intuitive, and accessible. By leveraging computer vision for gesture recognition and speech processing technologies, this system enables users to interact with their computers hands-free, eliminating the need for traditional input devices like keyboards and mice. This system is particularly beneficial for individuals with physical disabilities, professionals in hands-free work environments, and users who seek a more immersive computing experience. The integration of gesture recognition using OpenCV and Media Pipe, along with voice recognition powered by deep learning models, allows for seamless, real-time execution of system commands.

Through a well-defined system architecture, the solution ensures efficient data acquisition, pre-processing, feature extraction, classification, and execution of user commands. The use of machine learning models, speech recognition frameworks, and automation tools like PyAutoGUI further enhances the system's efficiency and usability. Despite its many advantages, the system faces certain challenges, such as gesture recognition accuracy in varying lighting conditions, voice recognition in noisy environments, and hardware limitations. Future improvements could include AI-powered enhancements, better noise reduction techniques, integration with wearable devices (such as smart gloves or AR headsets), and support for additional languages and dialects in speech recognition. In conclusion, the Gesture and Voice-Based Personal Computer Control System has the potential to revolutionize the way users interact with their computers. As advancements in artificial intelligence, deep learning, and sensor technologies continue, the system will evolve to become even more efficient, accurate, and widely adopted, paving the way for a touchless, intelligent, and futuristic computing experience.

RESULT

Dash Board of Project



References

Mitra, S., & Acharya, T. (2007). "Gesture Recognition: A Survey." IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 37(3), 311-324. doi:10.1109/TSMCC.2007.893280

Wu, Y., & Huang, T. S. (1999). "Vision-Based Gesture Recognition: A Review." International Gesture Workshop. Springer, Berlin, Heidelberg.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., ... & Blake, A. (2011). "Real-time human pose recognition in parts from single depth images." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Rabiner, L. R. (1989). "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." Proceedings of the IEEE, 77(2), 257-286. doi:10.1109/5.18626

Koller, O., Zargaran, S., Ney, H., & Bowden, R. (2016). "Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition." British Machine Vision Conference (BMVC).

Google Speech-to-Text API. (n.d.). Retrieved from https://cloud.google.com/speech-to-text

Media Pipe Hand Tracking. (n.d.). Google Research. Retrieved from

https://developers.google.com/mediapipe

OpenCV Library. (n.d.). Retrieved from https://opencv.org

Zhang, Z. (2012). "Microsoft Kinect Sensor and Its Effect." IEEE Multimedia, 19(2), 4-10. doi:10.1109/MMUL.2012.24

PyAutoGUI: Automating the GUI. (n.d.). Retrieved from https://pyautogui.readthedocs.io