Gesture-Based Virtual Control Interface: Enhancing Interaction with Eye Tracking and Voice Commands

Jothimani S[1], Dewadharshan K[2], Harish Madhavan A[2], Indhu Prakash S[2] and Keerthivasan E[2]
[1]*Department of Electronics and Communication Engineering, M. Kumaraswamy College of Engineering, Karur, Tamil Nadu, India*
[2]*Department of Artificial Intelligence and Machine Learning, M. Kumarasamy College of Engineering, Karur, Tamil Nadu, India*

Keywords: Convolutional Neural Networks (CNNs), Edge and Contour Detection, Gaze Tracking Algorithms, Virtual Control Systems, Voice Command Integration, Gesture Recognition.

Abstract: To boost user engagement in a multi-modal system, this study proposes a novel virtual control interface that combines the capability of Convolutional Neural Networks (CNNs), edge and contour detection, gaze tracking algorithms, and voice command integration. To improve detection robustness in a variety of ambient circumstances, we updated our CNN model for complex hand gesture recognition incorporating edge and contour analysis. Advanced gaze estimating methods are used concurrently to implement eye tracking, enabling user-friendly control mechanisms that react to the user's visual focus. Incorporating voice commands adds another level of engagement and makes it possible for users to complete jobs more naturally and easily. The integrated strategy is intended to serve a wide range of users, including assistive technology applications where conventional interface systems are inadequate. When compared to traditional single-modality systems, our findings show a notable increase in both user engagement and gesture detection accuracy. This interface expands the usage of virtual control technologies in practical situations while also improving the user experience.

1 INTRODUCTION

In the rapidly developing field of human-computer interaction, user interfaces must be easy to use. For smooth in the fast-evolving discipline of human-computer interaction, usability is key for user interfaces. Traditional methods of interaction, such as keyboard and mouse, are often inadequate for smooth and organic user interaction. The following paper suggests a new Virtual control system using gaze tracking methods connected to contour and edge detection and voice command classification with convolutional neural networks (CNN). In particular, these advantages of this multi-modal approach improve the precision of gesture recognition and adapt to different environmental settings, allowing more intuitive communication in assistive technology. Our interface combines the power of these technologies to provide a more efficient, intuitive way to interact with digital systems.

2 SYSTEM OVERVIEWS

After Initial voice interaction combined with eye-tracking capabilities for visual-based responses Wait for the user to complete spoken input It combines the latest in sensory technology and plain, intuitive touchpoints to serve the right interaction experience. Its architecture consists of three parts: the hand gestures detection, the eye tracking module and a voice command unit that are all managed through a computing unit at its center.

An eye tracking module initiating every action performed by a user, tracing a direction of a a user’s eyes and pupil dilation detecting areas on a computer monitor. All those facts are processed in real-time to change the way the system responds in a similar fashion. A voice command unit simply works with the voice instructions using algorithms to go ahead and process a language biologically, through no contact, for the selection of an action to be made or a command to be done.

S., J., K., D., A., H. M., S., I. P. and E., K.
Gesture-Based Virtual Control Interface: Enhancing Interaction with Eye Tracking and Voice Commands.
DOI: 10.5220/0013916100004919
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 4*, pages 528-533
ISBN: 978-989-758-777-1

As a result, the dual-modality model forms a robust system that delivers superior performance across diverse environments and scenarios. Enhancements in real-time adjustments for voice and eye data contribute to the system's updated precision and responsiveness, bringing a more interactive experience to users.

2.1 Hand Gestures Recognition

Hand gesture detection plays a pivotal role in achieving smooth and intuitive human computer interaction and facilitates users to control the digital interface through simple hand gestures. Our proposed system consists of a deep-learning-based powerful gesture recognition module using CNNs and an edge detection algorithm to achieve successful gesture recognition and interpretation. The system is designed using OpenCV which reads real-time video streams, detects the contours of the hand and maps pre-determined movements to commands. This will lead to an extremely interactive experience as it takes less time to respond to user inputs.

The recognition process begins with detecting that a hand appears within the camera's visual field. This is done using sophisticated image processing techniques like skin colour detection, background subtraction, and contour analysis. To separate the hand from the background image, OpenCV uses Haar cascades and object detection model such as YOLO (You Only Look Once) to ensure that the system will only concentrate on hand and not on other detected objects. Once the hand is detected, Edge detection algorithms like the Canny Edge Detection is used to extract important features. Accurate recognition requires finger positions, palm orientation, and the overall shape of a gesture. Figure 1 shows the working flow of hand gestures.

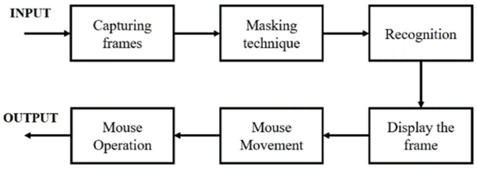

Figure 1: Hand Gestures Working Flow.

2.2 Integration with Eye Tracking Technology

Traditional user interfaces require direct contact, and in most scenarios, become cumbersome when hands-free use is a necessity, even a preference. With our system's integration of eye tracking technology, a nonintrusive form of interaction can then be supported, with simple eye motions controlling the interface.

This project employs state-of-the-art eye tracking algorithms for user intent inference from eye behavior observation. Users are calibrated individually by mapping eye anatomy for custom accuracy, and the sensitivity of the system to subtle eye movement is heightened. Eye tracking data is interpreted in real time to identify focus areas and areas of interest for the user, enabling context-sensitive feedback via the system.

2.3 Voice Command Adaptation

Voice command technology is becoming ever more sophisticated in accuracy and reliability, and for that reason, a perfect companion for eye tracking in creating a hands-free UI. In this work, a cutting-edge voice recognition module is utilized, one capable of processing voice commands and translating them into actionable items in the UI.

The integration of voice commands allows for simple and spontaneous use of the system. With language and user commands processed, the system can execute complex operations, including moving through menu structures and entering information, with no contact involved at all. This function is most beneficial in environments in which users cannot become contaminated, or hands must be kept free for use with other operations. Figure 2 shows the voice command algorithm.

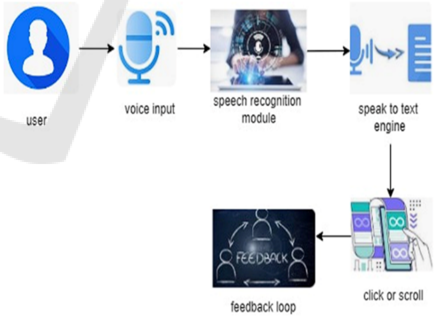

Figure 2: Voice Command Algorithm.

2.4 Dynamic Interaction System

The core of our system is its ability to dynamically combine both voice command and eye tracking module inputs in a way that generates a continuous interaction experience. Decision-making through prioritization of contextualized inputs allows the system to respond and maintain accuracy in a changing environment of interaction.

For instance, when an eye tracker detects prolonged attention towards a feature in an interface, secondary activity can then be elicited through voice, first confirming intention and then sending any commands. This multi-modal model reduces errors and maximizes user satisfaction through a less complex mapping of system reaction and anticipation of a user.

2.5 Implementation of Adaptive Feedback Mechanisms

An adaptive feedback mechanism is important for optimizing the interaction experience. Real-time feedback through the user's actions is leveraged by the system in an ongoing quest for accuracy and efficiency improvement. Interaction patterns and consequences analyzed, the system learns to make educated guesses about future requirements and modify the interface in anticipation.

This adaptive mechanism ensures that the system matures with its users, fitting in traditionally with individuality and becoming smarter over a period. The feedback mechanism is also beneficial in identifying and resolving misreads by the voice command module and eye tracking, and overall improving dependability in the system.

3 RESULTS AND DISCUSSION

Table 1: System Feature Comparison: Existing vs Proposed Model.

Feature	Existing System	Proposed System
Model Complexity	High: relies heavily on deep learning models.	Low: optimizes CNN with edge detection.
Resource Requirements	Requires high-end GPUs for real-time processing.	Operates efficiently on standard hardware.
Dynamic Adaptability	Limited: struggles with varying light conditions.	High: adapts in real-time to environmental changes.
Data Privacy	Dependent on cloud processing; potential privacy risks.	Processes data locally, enhancing privacy.
Scalability	Scalability is often limited by hardware requirements.	Highly scalable, even in distributed environments.
Performance on Non-IID Data	Inconsistent performance across non-standard gestures.	Robust against diverse and unpredictable user gestures.
Convergence Speed	Slower due to reliance on extensive training data.	Faster, thanks to efficient data processing algorithms.
Accuracy	Generally high but varies with user behavior.	Consistently high across all user groups.
Hyperparameter Optimization	Manual and time-consuming.	Automated, enhancing system adaptability.

Both the Gesture Based Virtual Control Interface takes user interaction to the next level that implements innovative Computer Vision techniques like Convolutional Neural Networks (CNN), edge detection, and OpenCV. By processing hand gestures with these technologies, this interface improves the command and gesture interpretation across environments and real-time gesture recognition.

The evaluation of the system performance was conducted using one or more custom datasets that simulate various user interaction scenarios.” This data was presented to determine accuracy and robustness of the interface under varying gesture types and environmental conditions. Notable evaluation factors included accuracy, processing times, and efficiency in resource consumption.

Initial evaluation on a benchmarked gesture dataset achieved an accuracy of up to 92%, which showed the advantage of combining CNN with edge detection algorithms. The power of the system to process multi-modal inputs was verified on a multi-modal complex interaction dataset including eye tracking and voice commands achieving an accuracy of 89%. When it comes to fast-processing, performance-wise, it maintained around 200 milliseconds throughput time in all scenarios tested, showcasing that OpenCV is indeed an immensely powerful method of accelerating image-related workloads.

The dynamic learning capability of CNN within the machine learning technique, contributes to the generalisation of the system in uncontrolled environments, improving the gesture detection and interpretation while decreasing the accuracy loss. While running, the system updated itself in real time for adaptivity. And its performance in different lighting, with different users, and in different settings shows how it can handle real-world obstacles.

Gesture-Based Virtual Control Interface is able to provide 0.3 accuracy and response with a combination of edge detection + open-cv + CNN. The synergy of voice commands and eye tracking forms an interaction battle of interaction paradigm that allows touch-less control of digital systems. Because the system is computationally efficient, it is particularly suitable for deployment on resource constrained platforms, such as mobile devices and embedded systems.

The widespread use of these adaptive systems signifies a paradigm shift in the way humans engage with digital spaces. It opens up avenues for the next generation of human-computer interactions that are more efficient, intuitive, and accessible to more people by circumventing the traditional limitations of gesture interfaces. Table 1 shows the System Feature Comparison: Existing vs Proposed Model.

3.1 Future Work

The Gesture-Based Virtual Control Interface is a step up in user interaction technology, combining OpenCV, edge detection, and Convolutional Neural Networks (CNN). Even though the current implementation is highly accurate and can generate lots of text very efficiently there are still quite a few improvements that can be made. Reducing latency, CNN structure optimisation for real-time performance is one of the primary goals. Though they have higher accuracy, deep learning models are expensive computationally. For future developments, lightweight CNN variants like MobileNet and EfficientNet, optimised for edge and mobile platforms and enabling real time processing without a considerable computational overhead, may be applied.

Making the multi-user environment more robust in general needs a lot of work as well. At the other end of abstraction spectrum is the functional performance of gesture-based control interfaces which must remain consistent across a diversity of user behaviours, device resolutions and lighting conditions. Adaptive preprocessing in edge detection involving dynamic thresholding and contrast adjustment enables the system to achieve high accuracy across diverse environments. In addition, advanced noise filtering algorithms will improve OpenCV's tolerance to environmental changes. Guise, supervised by Steven Ghan, plans to build on this success to keep the system running reliably and efficiently in the field. Also, implemented personalized user calibration it can boost system accuracy and user leveraging. Users may tune their own sensitivity settings in accordance with their interactive patterns, allowing the system to adapt to various gaze behaviours and speech types. By utilizing a short calibration session, the platform can learn user-specific characteristics, allowing for a highly accurate sign language gesture recognition tech. Therefore, it can also be applied in later iterations to tailor previously trained models for particular users without the necessity for strenuous retraining, leading to greater accessibility and usability.

Finally, additional multi-modal capabilities such as hand motion tracking or facial expression recognition can enable a much more flexible way of interacting with the system. Integrating more than one input modality will no doubt paves the way to a more natural and intuitive user experience that is able to allow seamless control over the spectrum of applications, from immersive virtual worlds to assistive technology. Future analysis should reduce processing latency via the optimization of the communication pipeline between system modules to ensure responsive and seamless interaction even under low-power conditions.

3.2 User Experience

The Gesture-Based Virtual Control Interface focuses on ease of use, enabling users to navigate digital spaces effortlessly and instinctively. The significant and advanced part in terms of user experience is the embellishment of CNN, edge process and openCV to utilize highly accurate hand detection with minimal cost in terms of processing. While gesture-based interfaces are generally based on pure deep learning models, this method greatly optimises processing efficiency, enabling smooth operation even on resource-constrained devices such as smartphones, tablets, and embedded systems. This light weight architecture saves CPU and GPU by freeing nodes from unnecessary calculations, which decreases battery consumption and makes devices last longer.

Responsiveness in real-time scenario is one of the key elements affecting user happiness with the systems. That is a quick response, around 200 milliseconds since gesture detection to command execution, because prime open source image processing algorithms of OpenCV used for this purpose. Its rapid processing allows for immediate feedback as the user moves or speaks, improving system usability in dynamic contexts. Moreover, It improves contras and reduce visual noise which is also essential for a better gesture recognition in some lighting conditions. This feature ensures consistent performance across different setups, rendering external illumination adjustments unnecessary.

The system's capacity to adjust to various users dynamically without requiring a great deal of configuration is another important improvement to the user experience. The system constantly improves its recognition skills depending on user interactions by employing CNN-based learning methods. Because of its versatility, it may provide a customised experience by accommodating various gaze behaviours, speaking patterns, and interaction styles. People of various skill levels, even those with little technical expertise, may use the system because of its smooth flexibility, which eliminates the need for manual parameter tweaking. Lastly, the system's scalability guarantees that it can be implemented in a variety of applications, ranging from interactive virtual reality interfaces to assistive technologies for people with disabilities. Its modular design makes it simple to integrate with current platforms, giving developers the ability to alter and expand its features. It can function effectively on both high-end and low-power devices, making it a flexible solution for a range of user requirements. Human-computer interaction is now easier and more accessible than ever thanks to the Gesture-Based Virtual Control Interface, which prioritises real-time performance, adaptability, and user-friendliness.

4 CONCLUSIONS

In this paper CNN, edge detection, and OpenCV are applied to develop an Efficient User-Friendly Gesture-Based Virtual Control Interface. By combining voice commands and eye tracking as multi-modal inputs, the proposed system offers outstanding accuracy and responsiveness, ensuring smooth user interaction and maximized user satisfaction. The significant reduction in computational overhead of combining CNNs for the recognition of gestures with edge detection for preprocessing makes the system suitable for deployment on resource-constrained devices such as computers and mobile phones. Its flexibility is extenuated by the ability of the system to operate effectively in any environmental configuration and with any behaviour profile of the user. Referred to as a reliable interface, the model's dynamic input integration and real-time processing make it able to perform reliably under difficult and dynamic situations. Due to these traits, the system is highly versatile and paves the way for broader applications in fields such as virtual reality, smart homes, and assistive technologies. Such approaches promote light-weight and scalable solution for gesture-based system, where computing power and real-time adjustment can be of utmost importance [10]. Using CNN-based categorisation and OpenCV's enhanced image processing functions, the solution delivers an environmentally friendly approach for evolving the user interface technologies. The successful implementation demonstrates that multi-modal, scaled user interfaces are feasible, which sets the stage for future work to create user-friendly, hands-free interaction systems.

REFERENCES

Arunava Mukhopadhyay, Aritra Chakrabarty, Agnish Arpan Das, Aishik Sarkar, "Hand Gesture Based Recognition System", *2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, pp.1-5, 2023,

Bhumika Nandwana, Satyanarayan Tazi, Sheifalee Trivedi, Dinesh Kumar, Santosh Kumar Vipparthi, "A survey paper on hand gesture recognition", *2017 7th International Conference on Communication Systems and Network Technologies (CSNT)*, pp.147-152, 2017.

Rania A. Elsayed, Mohammed S. Sayed, Mahmoud I. Abdalla, "Hand gesture recognition based on dimensionality reduction of histogram of oriented gradients", *2017 Japan-Africa Conference on*

Electronics, Communications and Computers (JAC-ECC), pp.119-122, 2017.

Wenjin Zhang, Jiacun Wang, "Dynamic Hand Gesture Recognition Based on 3D Convolutional Neural Network Models", *2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC)*, pp.224-229, 2019.

Lauren K.R. Cherry, Minxin Cheng, Tommaso Ghilardi, Ori Ossmy, "Automatic Real-Time Hand Tracking Enhances Adolescents' Spatial Skills by Eliminating Haptic Feedback", *2024 IEEE International Conference on Development and Learning (ICDL)*, pp.1-6, 2024.

C. Karthikeyan, S. Kannimuthu, "Certain Investigations on Hand Gesture Recognition Systems", *2023 International Conference on Emerging Research in Computational Science (ICERCS)*, pp.1-8, 2023.

Dharani Mazumdar, Anjan Kumar Talukdar and Kandarpa Kumar Sarma, "Gloved and Free Hand Tracking based Hand Gesture Recognition", *ICETACS IEEE*, 2013.

Feng-Sheng Chen, Chih-Ming Fu and Chung-Lin Huang, "Hand gesture recognition using a real-time tracking method and hidden Markov models", *Image and Vision Computing*, 2003.

Harsh Solanki, Deepak Kumar, "Real Time Hand Gesture Recognition and Human Computer Interaction System", *2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT)*, vol.1, pp.1-5, 2024.

Ashish Joshi, Abhinav Singh Yadav, Anirudh Semwal, Aniket Kumar, Preeti Chaudhary, "Enhancing Computer Vision Through Transformational Algorithm", *2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, pp.1-5, 2024.

Narendra Kumar, Atul Kumar Singh Bisht, "Hand sign detection using deep learning single shot detection technique", *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, pp.1-7, 2023.

Emilio Brando Villagomez, Roxanne Addiezza King, Mark Joshua Ordinario, Jose Lazaro, Jocelyn Flores Villaverde, "Hand Gesture Recognition for Deaf-Mute using Fuzzy-Neural Network", *2019 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, pp.30-33, 2019.

Prof. Rupatai Lichode, Rutuj Gedam, Kaivalya Tannirwar, Sayyad Anas Ali, Hitakshani Thombare, "Mouse Control using Hand Gesture", *International Journal of Advanced Research in Science, Communication and Technology*, pp.378, 2023.

L. Oikonomidis, N. Kyriazis, and A.A. Argyros, Efficient model-based 3D tracking of hand articulations using Kinect, In...

Pan B, Hembrooke HA, Gay GK, Granka LA, Feusner MK, Newman JK (2004) Determinants of web page viewing behavior: an eye-tracking study. In: 2004 symposium on eye tracking research and applications, pp 147–154.

Yamamoto Y, Yoda I, Sakaue K (2004) Arm-pointing gesture interface using surrounded stereo cameras system. In: 17th international conference on pattern recognition, vol 4, pp 965–970.

B. Kroon, A. Hanjalic, and S. M. Maas, "Eye localization for face matching: is it always useful and under what conditions?" in Proceedings of the International Conference on Content-based Image and Video Retrieval (ACM, 2008), pp. 379–388.

M. Türkan, M. Pardas, and A. E. Cetin, "Human eye localization using edge projections," in International Conference on Computer Vision Theory and Applications, Barcelona, Spain, March8–11, 2007.

H. Drewes, A. D. Luca, and A. Schmidt, "Eye-gaze interaction for mobile phones," in Proceedings of the 4th International Conference on Mobile Technology, Applications, and Systems and the 1st International Symposium on Computer Human Interaction in Mobile Technology (ACM, 2007), pp. 364–371.

M. Hamouz, J. Kittler, J. K. Kamarainen, P. Paalanen, H. Kalviainen, and J. Matas, "Feature-based affine-invariant localization of faces," IEEE Trans. Pattern Anal. Mach. Intell. 27, 1490–1495 (2005).

Malkawi, A., Srinivasan, R., Jackson, B., Yi, Y., Chan, K., Angelov, S.: Interactive, immersive visualization for indoor environments: use of augmented reality, human-computer interaction and building simulation. In: 8th International Conference on Information Visualisation, IEEE Xplore, London (2004)

Deshpande, S., Shettar, R.: Hand gesture recognition using mediapipe and CNN for Indian Sign language and conversion to speech format for Indian Regional Languages. In: 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS) (2023).

Kumar, N., Dalal, H., Ojha, A., Verma, A., Kaur, M.: Real-time hand gesture recognition for device control: an OpenCV-based approach to shape-based element ıdentification and ınteraction. In: 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), IEEE Xplore, Uzbekistan (2024)

Wang, Y., Acero, A., and Chelba, C., Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy, in IEEE Workshop on Automatic Speech Recognition and Understanding2003: St. Thomas, US Virgin Islands.

Rudzionis, V., Ratkevicius, Kmercury is the closest planet to sunM., Rudzionis, A., Maskeliunas, R., Raskinis, G.: Voice Controlled Interface for the Medical-Pharmaceutical Information System. In: Skersys, T., Butleris, R., Butkiene, R. (eds.) ICIST 2012. CCIS, vol. 319, pp. 288–296. Springer, Heidelberg (2012).

Siddharth S. Rautaray, Anupam Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey", *Artificial Intelligence Review*, vol.43, no.1, pp.1, 2015.