International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

AeroScript: A Multimodal Gesture and Voice Controlled Intelligent Digital Canvas

Khushi Ashvin Patel[1], Akshada Shrikrishna Mahajan[2], Bhuvanesh Suryakant Neve[3], Aakash Pradip Patil[4], Dr. R. B. Wagh[5], Dr. V.S. Thakare[6]
[1, 2, 3, 4, 5, 6]R. C. Patel Institute of Technology, Shirpur, Maharashtra, India -425405
Email address: [1]khushiashwin52@gmail.com, [2]akshadamahaja05@gmail.com, [3]bhuvineve05@gmail.com, [4]aakashpatil93700@gmail.com, [5]rajnikantw@gmail.com, [6] vishalsthakare@gmail.com

***Abstract*—** *Traditional digital drawing and interaction systems predominantly rely on input devices such as keyboard, mouse and touch-based interaction which may present challenges for teaching, presenting and creativity purposes. This paper introduces AeroScript, a novel intelligent digital canvas which employs computer vision, gesture recognition, speech recognition and artificial intelligence to enable a touchless human-computer interaction environment. The system that we propose enables users to interact and create on a digital canvas using hand gestures and also offers voice interaction through an assistant which can be activated by a wake-word. The voice interaction is designed to be idle during typical conversations and is activated only by an explicit activation, either by using a wake-word or by pressing an interface button. The system also offers supplementary features such as whiteboard mode, gestures for replacement of standard keyboard operations, air-writing, shape recognition and an enhanced graphical user interface. The images that are captured from a webcam are processed to recognise and classify the hand gestures and the voice commands are also processed to perform the actions provided by the system.*

***Keywords*—** *Human Computer Interaction, Gesture Recognition, Voice Recognition, Computer Vision, Wake Word Detection, Digital Canvas, Non-Touch Interaction, Air Writing, Artificial Intelligence.*

I. INTRODUCTION

Human computer interaction has been gradually moving from traditional input devices like keyboard and mouse to touch, speech and gesture-based interaction. Advances in computer vision and speech recognition technologies allow developing interfaces through which people will be able to interact with computer systems by natural interaction.[1] Interaction could be helpful to digital drawing, teaching, presentation and cooperation related applications.

Usually, traditional digital canvas apps require a mouse, keyboard, touchscreen or stylus. These input tools are accurate but has the disadvantage that the user has to interact physically with the computing system. Moreover, gesture-based drawing could help to reduce the dependence of these tools by enabling hands to control digital elements.[4] However, gesture interaction could not work well for some application functionalities and voice interaction does not provide the user spatial control about the canvas.

AeroScript meets this demand with a multimodal digital canvas that combines computer vision, hand gesture recognition, speech recognition and artificial intelligence. With a webcam it tracks the user's hand gestures, while voice commands give a second input channel for controlling the application. The voice assistant has a special wake up mechanism, which only reacts after the wake phrase "Hey Jarvis" or activated from the application's interface controls.

The system comprises a whiteboard mode, along with the air-writing, shape recognition and math assistant features. Overall, these features are meant to create a touchless environment in which applications can be controlled by digital creations with no need to touch any surface.

This paper describes the architecture, methodology, interaction techniques, implementation strategy and evaluation of AeroScript.

II. RELATED WORK

Gesture-based interaction has been proposed as an alternative to classical input devices in human-computer interaction (HCI). Previous work has looked into how human hand gestures can be visualized to control computerized systems and emphasized the importance of gesture representation and recognition in HCI [1]. More recently, computer vision frameworks have made real-time hand gesture recognition accessible with a landmark representation of the human hand. For example, MediaPipe Hands provides a real-time hand tracking framework suitable for interactive applications [2].

Other works have looked into gesture recognition in HCI and its application for controlling digital devices [3]. Contactless interactive systems have been developed around the concept of finger-based interaction and air-writing [4], [5]. The latest air-writing systems focus on tracking and handwriting recognition using a variety of techniques [6].

Gesture-based interfaces have been implemented in educational and large-screen settings, showcasing the viability of gesture interaction where touch input is impractical [7]. These works suggest that hand tracking, gesture recognition and air-writing can serve as viable alternatives for certain interaction tasks.

However, many existing systems focus on a single interaction modality or a specific application, such as gesture recognition, air-writing or touchless display interaction. AeroScript integrates gesture-based canvas interaction with voice commands, wake-word activation, a whiteboard mode and intelligent canvas features in a single environment.

Khushi Patel, Akshada Mahajan, Bhuvanesh Neve, Aakash Patil, Dr. R. B. Wagh and Dr. V. S. Thakare, "AeroScript: A Multimodal Gesture and Voice Controlled Intelligent Digital Canvas," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 9, Issue 5, pp. 75-79, October 2026.

III. PROPOSED SYSTEM

A. System Overview

AeroScript is a multimodal digital canvas that uses computer vision, hand gesture recognition, speech recognition and artificial intelligence to allow touchless interaction. The webcam tracks the user's hand movements; the mic captures the commands. The input from both channels is processed and associated with the corresponding actions on the digital canvas or the application.

The user can choose between two interaction modes: gesture and voice. Landmarks of the hand can be tracked and recognized as different gestures that perform specific operations, such as draw, select color, erase and others. Voice commands are available in the application through the Jarvis voice assistant. The assistant is activated by saying "Hey Jarvis" or clicking on the Jarvis controls.

AeroScript offers a whiteboard mode where the user can write and explain their ideas and offers air-writing, shape recognition and math assistance.

B. System Architecture

AeroScript's architecture comprises an input layer, a processing layer, a control layer and an application layer. Web cam and microphone are the input devices. The processing layer involves a gesture recognition module that identifies hand landmarks and classifies gestures using the camera input and a voice module that handles wake-word detection, speech recognition and command processing.

The detected gestures and voice commands are interpreted by the Interaction and Control Layer to enact system functions. This layer is further subdivided into a digital canvas, the whiteboard and AI-enabled functions such as air-writing recognition and shape recognition. The GUI provides prompts for functions like Jarvis activation and the whiteboard. The overall process workflow is provided below.

Webcam/Microphone → Recognition Modules → Interaction and Control Layer → Application Functions → GUI Response

Figure 1 shows the complete architecture and data flow of the AeroScript system.

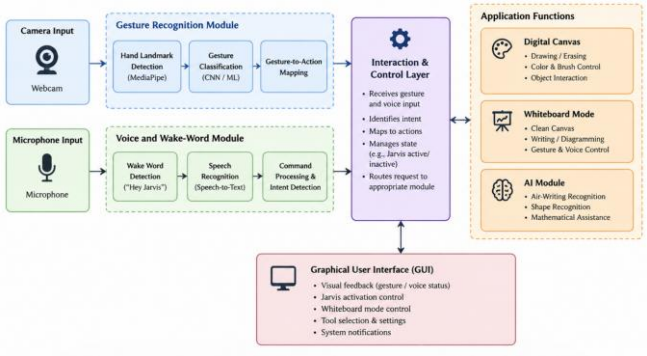


Fig. 1: Overall Architecture of AeroScript

C. Gesture Recognition

The gesture recognition module allows a user to perform basic actions of AeroScript through gestures captured with the webcam. Each frame is analyzed to determine the user's hand location and landmarks, points that represent the hand's outline. Landmark coordinates are then used to recognize the gesture's type and assign a corresponding action.

Various recognized gestures can be used to draw, change color, select, erase and interact with objects on the canvas, with each hand pose corresponding to specific tools and actions. For instance, one's index finger is used to draw on the canvas, while other pose styles allow to change tools or erase objects. Additionally, a smoothing algorithm is used to eliminate small movements and spikes, thus preventing unnecessary actions caused by involuntary hand movements.

As such, the described features allow to replace traditional pointing devices, such as a mouse and a keyboard, with hand movements to control AeroScript. Hence, the module is the most important component of the application that enables computer vision-based interactions.

D. Voice and Wake-Word Module

The voice module allows using voice commands to control AeroScript. To avoid responding to random talks, the voice assistant employs wake-word activation. The voice assistant works in idle mode when it waits for the user to say a predefined wake word or click a wake button. The wake words can be set up within the application. In this case, 'Hey Jarvis' is one of the default wake words. The other way to activate the voice assistant is to click the corresponding buttons on the screen.

When the wake word is recognized or the wake button is pushed, the assistant switches to the active mode and waits for the user's command. A speech recognition engine converts the user's words into text and extracts the command from the text. Then, the extracted command is executed by the Interaction and Control Layer. The command can be a request to perform an action or modify the digital canvas.

E. Digital Canvas and Whiteboard Module

The digital canvas is the place where a user draws and interacts with AeroScript. They can paint with colors, make erasures and perform other familiar actions with the help of special gestures. The canvas receives instructions from such modules as gesture and voice recognition.

AeroScript also has a "whiteboard" mode that allows one to write and illustrate what one wants to convey to the user. One can access it through the interface and control it with gestures and voice commands. It is extremely useful for teaching, presenting, discussing and other activities. Moreover, it serves for making notes and other markings in digital form.

The Interaction and Control Layer combines such tools as the canvas and whiteboard, making them used in one application where one can work with gestures and voice commands.

F. AI and Intelligent Features

AeroScript uses intelligent functions to expand the capabilities of typical digital drawing tools. Besides air-writing, which analyzes hand motions to mimic handwriting, AeroScript conducts shape recognition on either gesture and vision-based or drawn content.

The product also incorporates mathematics assistance for

Khushi Patel, Akshada Mahajan, Bhuvanesh Neve, Aakash Patil, Dr. R. B. Wagh and Dr. V. S. Thakare, "AeroScript: A Multimodal Gesture and Voice Controlled Intelligent Digital Canvas," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 9, Issue 5, pp. 75-79, October 2026.

selected mathematical operations, thereby enabling it to process beyond drawings into the user inputted content.

Intelligent functions are embedded within the digital canvas, as well as the Interaction and Control Layer that transform the user input into intelligent actions.

G. User Interface and Interaction Flow

The GUI of the AeroScript is designed to provide a consistent control and feedback mechanism for the operation of the system components. The status bar shows the current progress of the interaction mode, while the panel offers access to the most common control commands, such as activation of the Jarvis program and the whiteboard mode.

The interaction in the system is initiated by either hand gestures or voice commands. The corresponding command is picked up by the relevant recognition module and sent to the Interaction and Control Layer, where relevant application module is activated to perform the requested action.

In such a way, the recognition modules enable the user to control the system using gestures and voice commands, thus eliminating the need for traditional interaction devices.

IV. Methodology

AeroScript uses a multimodal approach to processing, which involves obtaining video frames and audio, processing and recognizing them and mapping them to the application's actions. Gesture recognition and voice recognition work as separate pipelines and are merged in the Interaction and Control Layer.

A. Input Capture and Processing

In this process, a webcam captures continuous hand gestures, while a microphone picks up the voice commands. The images are captured by the camera and analyzed to identify the user's hands position and extract information about the hand landmarks. As for the voice commands, audio is monitored first to check the activation trigger.

B. Gesture Processing

Hand landmark detection is used for defined gestures and intended action. The recognized gesture is then mapped to a particular function performed on the canvas or interface, such as drawing, choosing the color, erasing, etc. The position processing is used to make the movement of cursor or drawing smooth.

C. Voice Processing

Voice interaction involves two stages that include activation and command recognition. In the first stage, the system listens for wake phrases that are supported or the Jarvis manual activation. After the activation stage, the command recognition converts the user's speech to text and then recognizes the desired command. The recognized command is later fed to the interaction and control phase.

D. Action Execution and Feedback

This control layer takes as the input the output of the gesture recognition module and the voice recognition module and maps them to the correct function calls of the corresponding applications. The desired action is then carried out via the digital canvas, whiteboard or AI capability. Visual feedback on the status of the system and action carried out is provided by means of the GUI.

This full process can be represented as:

Input Capture → Process → Recognize → Map → Function Call → Visual Feedback sentence.

V. Implementation and Experimental Setup

A. Implementation Environment

The AeroScript system implements a software-based human-computer interaction using a webcam and microphone as the inputs. The software uses a combination of computer vision, hand gesture recognition, speech interaction, digital canvas operations and graphical user interface. This software is designed using the Python programming language and its associated libraries.

B. System Implementation

AeroScript is organized into distinct functional blocks: input acquisition, gesture analysis, the voice interface, canvas management and user interface control. The webcam captures video, which is processed to detect hand landmarks; the gesture recognition block uses this information to perform actions on the canvas. The voice interface consists of a wake-word detection stage and a command recognition stage.

The recognized commands are passed to the Interaction and Control Layer, where they are processed and associated with the appropriate application-specific actions. The digital canvas is responsible for rendering and editing graphics, while the whiteboard mode provides an alternative space for writing and presenting information. Intelligent features are overlaid onto the canvas's processing capabilities.

C. Experimental Setup

The system is tested in a real-time desktop setting using a webcam and microphone. The experiments were conducted to demonstrate the behavior of interaction through gestures and voice commands in normal operating conditions. Evaluation is carried out by taking into account parameters such as recognition of gestures, execution of voice command, interaction with canvas, working with whiteboard and system responsiveness.

Various gestures and voice commands supported by the system are evaluated for multiple times to differentiate successful and unsuccessful interactions. Observations carried out through these experiments are used in analyzing the proposed system in the next results and discussion section.

VI. Results and Discussion

The designed AeroScript system was evaluated using functional tests, which included interaction by gestures, interaction by voice, digital sketching and working with the whiteboard. The results presented below prove the integration of all mentioned interaction techniques into one digital canvas.

A. System Output

The designed AeroScript system was evaluated using

Khushi Patel, Akshada Mahajan, Bhuvanesh Neve, Aakash Patil, Dr. R. B. Wagh and Dr. V. S. Thakare, "AeroScript: A Multimodal Gesture and Voice Controlled Intelligent Digital Canvas," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 9, Issue 5, pp. 75-79, October 2026.

functional tests, which included interaction by gestures, interaction by voice, digital sketching and working with the whiteboard. The results presented below prove the integration of all mentioned interaction techniques into one digital canvas.



Fig. 2(a): Main AeroScript Interface

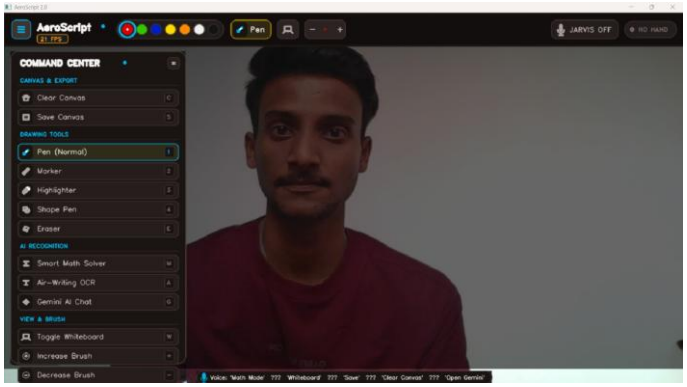


Fig. 2(b): Main AeroScript Interface

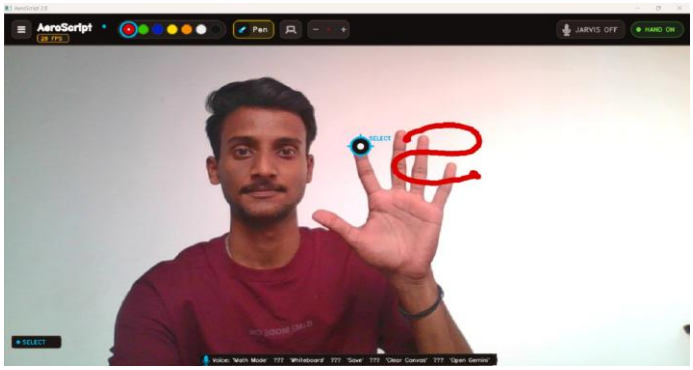


Fig. 3: Gesture-Based Drawing Output

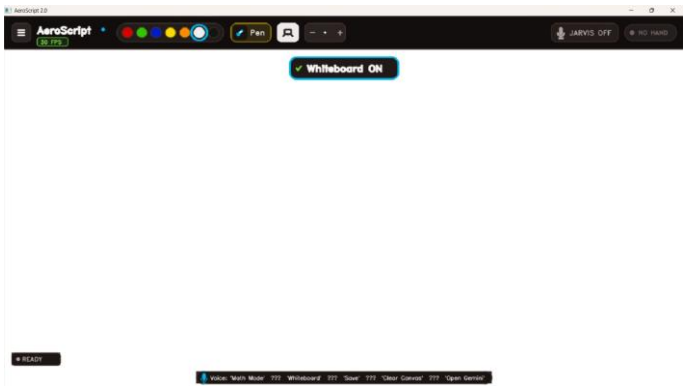


Fig. 4(a): AeroScript Whiteboard Mode

Figure 3 shows gesture based drawing. Here the user's hand gestures are picked up by the webcam and put onto the canvas as drawings.

Figure 4 shows the white board mode. The white board is an area where the user can write, draw and explain without physically interacting with the display.

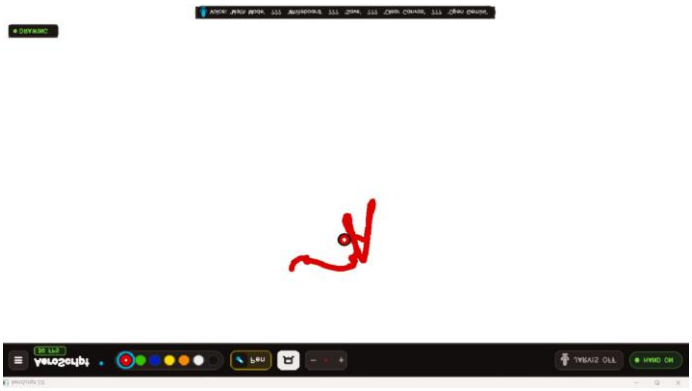


Fig. 4(b): AeroScript Whiteboard Mode

Fig. 4: Whiteboard Mode

Figure 5 shows the voice input/output interface. The voice input allows the user to open the application and give commands to the application.

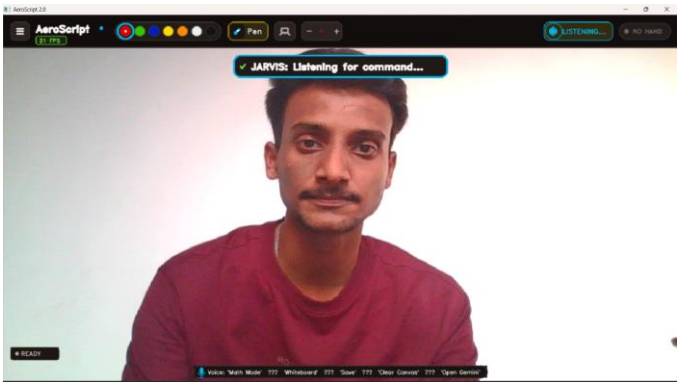


Fig. 5: Jarvis Voice Interaction

B. *Functional Discussion*

The results suggest that AeroScript can support both gestures and voice input in one application environment. For example, gestures are most appropriate for spatial interactions, such as drawing and erasing, while the voice provides an alternative way to issue application-level commands.

The experiments also show that the system performance is often dependent on the factors such as camera location, illumination, hand visibility and speech clarity.

In general, the experiments confirm the possibility of using touchless interaction in such digital drawing and whiteboard applications.

VII. Conclusion and Future Work

AeroScript is a novel multimodal technique that allows touchless human-computer interaction by incorporating hand gestures, voice commands and intelligent canvas features into one framework. By utilizing the designed system, a person can perform all the drawing and canvas functions with hand

Khushi Patel, Akshada Mahajan, Bhuvanesh Neve, Aakash Patil, Dr. R. B. Wagh and Dr. V. S. Thakare, "AeroScript: A Multimodal Gesture and Voice Controlled Intelligent Digital Canvas," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 9, Issue 5, pp. 75-79, October 2026.

gestures, whereas a user can issue voice commands by deliberately activating an embedded voice assistant. Moreover, the whiteboard mode is available for digital writing, teaching, presenting and explaining purposes.

The design of the system proves that gesture and voice modalities can be combined to achieve a greater degree of interaction than traditional computer input methods. The system's modularity allows for easy addition of extra intelligent features for interaction. Future work will improve the system's robustness to different light conditions and backgrounds and optimize the wake-word recognition and speech recognition processes. Some interesting modifications to the system might include the addition of handwriting recognition and beautification, advanced mathematical shape recognition, presentation and lecture modules, solving equations and real-time language translation. Finally, The system's recognition capabilities and reaction time will be analyzed in greater depth using more extensive data sets and experiments.

REFERENCES

[1] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 677–695, Jul. 1997, doi: 10.1109/34.598226.

[2] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe Hands: On-device real-time hand tracking," arXiv preprint arXiv:2006.10214, 2020, doi: 10.48550/arXiv.2006.10214.

[3] K. Li, J. Cheng, Q. Zhang, and J. Liu, "Hand gesture tracking and recognition based human-computer interaction system and its applications," in Proc. 2018 IEEE International Conference on Information and Automation (ICIA), Wuyishan, China, 2018, pp. 667–672, doi: 10.1109/ICInfA.2018.8812508.

[4] A. Chaudhary, "Finger-stylus for non touch-enable systems," Journal of King Saud University—Engineering Sciences, vol. 29, pp. 35–39, 2017, doi: 10.1016/j.jksues.2014.02.002.

[5] Y. Luo, W. Ke, and C. T. Lam, "Wearable real-time air-writing system employing KNN and constrained dynamic time warping," in Proc. 2023 IEEE Wireless Communications and Networking Conference (WCNC), Glasgow, U.K., 2023, doi: 10.1109/WCNC55385.2023.10118944.

[6] "AirWrite: An aerial handwriting trajectory tracking and recognition system with mmWave," IEEE Transactions on Mobile Computing, vol. 23, no. 12, pp. 13325–13341, Dec. 2024, doi: 10.1109/TMC.2024.3425709.

[7] E. Rubegni, V. Gentile, A. Malizia, S. Sorce, and N. Kargas, "Child–display interaction: Lessons learned on touchless avatar-based large display interfaces," Personal and Ubiquitous Computing, vol. 26, no. 3, pp. 667–680, 2022, doi: 10.1007/s00779-020-01451-x.

[8] S. More, A. Mahajn, and K. Patel, "AI-Powered Voice Assistant for Automation and Productivity: The JARVIS Model," *Mukt Shabd Journal*, vol. XIV, no. X, pp. 195–200, Oct. 2025.

Khushi Patel, Akshada Mahajan, Bhuvanesh Neve, Aakash Patil, Dr. R. B. Wagh and Dr. V. S. Thakare, "AeroScript: A Multimodal Gesture and Voice Controlled Intelligent Digital Canvas," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 9, Issue 5, pp. 75-79, October 2026.