Volume: 14 Issue: I Month of publication: January 2026

DOI: https://doi.org/10.22214/ijraset.2026.76769

Virtual Air Drawing using Computer Vision and Real-time Captions

Suraram Raju, Karamala Rohith, V. Arun, K. Rahul Reddy, Asst. Prof. Mrs. B. Manjula
Department of CSE (Data Science), AVN Institute of Engineering and Technology

Abstract: In the rapidly evolving digital education ecosystem, virtual classrooms require interactive and inclusive technologies to enhance teaching effectiveness. This paper presents Virtual Air Drawing using Computer Vision and Real-Time Captions, a system that enables instructors to draw in mid-air using hand gestures while simultaneously generating live captions from speech. The proposed approach utilizes computer vision–based hand tracking with MediaPipe to accurately capture gestures and render them onto a virtual canvas in real time. In addition, speech-to-text technology converts spoken content into captions, improving accessibility for learners with hearing impairments and language barriers. By integrating visual interaction with real-time textual support, the system enhances engagement, clarity, and inclusivity in online learning environments. The proposed solution is well suited for virtual classrooms, collaborative learning platforms, and interactive e-learning applications.

Keywords: Virtual Canvas, Live Captioning, Computer Vision, MediaPipe, Speech-to-Text.

I. INTRODUCTION

The rapid growth of online education and virtual classrooms has increased the demand for interactive and inclusive digital learning tools. Traditional online teaching platforms rely heavily on static slides, keyboards, and mouse-based interactions, which often limit real-time engagement and reduce the effectiveness of concept explanation. To overcome these challenges, gesture-based interaction using computer vision has emerged as a promising solution for enhancing human–computer interaction.

Virtual air drawing enables instructors to draw or write in mid-air using hand gestures, eliminating the need for physical writing tools or touch-based devices. By capturing hand movements through a standard webcam and processing them using computer vision techniques, air drawing systems provide a natural and intuitive way of interaction. Such systems are particularly useful in virtual classrooms, where visual explanation plays a critical role in student understanding.

In addition to visual interaction, accessibility remains a major concern in digital education. Learners with hearing impairments or language barriers often face difficulties in fully engaging with online lectures. Live captioning using speech-to-text technology offers an effective solution by converting spoken content into real-time textual captions, thereby improving inclusivity and comprehension.

This paper presents a Virtual Air Drawing using Computer Vision and Real-time Captions system that integrates air-based drawing and live captioning into a unified framework. The proposed system employs MediaPipe for accurate hand tracking and gesture recognition, allowing instructors to draw on a virtual canvas in real time, while speech-to-text technology generates synchronized captions. By combining visual and textual information, the system enhances teaching effectiveness, supports diverse learning styles, and provides an accessible learning environment. The proposed approach is suitable for virtual classrooms, collaborative learning platforms, and interactive e-learning systems.

II. LITERATURE SURVEY

Recent advancements in computer vision have enabled gesture-based interaction systems that eliminate the need for physical input devices. Several studies have explored air drawing techniques using hand gesture recognition, where finger movements are tracked to create virtual drawings in real time. Early approaches relied on contour and color-based hand detection methods; however, these techniques were sensitive to lighting conditions and complex backgrounds.

With the introduction of deep learning and hand landmark detection frameworks such as MediaPipe, gesture recognition accuracy has significantly improved. Researchers have demonstrated efficient real-time hand tracking systems for virtual interaction and smart classroom applications. However, most existing air-drawing systems focus primarily on visual interaction and lack accessibility features.

In parallel, speech-to-text technologies have been widely used to generate live captions in online communication platforms, improving accessibility for users with hearing impairments. Despite these developments, limited research has been conducted on integrating air-based drawing with real-time captioning in a unified system.

This work addresses the identified gap by combining virtual air drawing and live captioning into a single computer vision–based framework, enhancing interactivity, accessibility, and inclusivity in virtual learning environments.

III. PROPOSED SYSTEM

A. System Overview

The proposed system enables users to draw in the air using hand gestures and simultaneously generate real-time captions describing the drawing. It leverages computer vision techniques for gesture tracking and machine learning for caption generation. The system accepts live video input from a camera and outputs the virtual drawing along with contextual captions, providing an interactive and immersive experience.

B. System Architecture

The system consists of the following modules:

- Video Capture and Hand Tracking
- Gesture Recognition
- Air Drawing Renderer
- Real-time Caption Generator
- Data Preprocessing
- Visualization Module

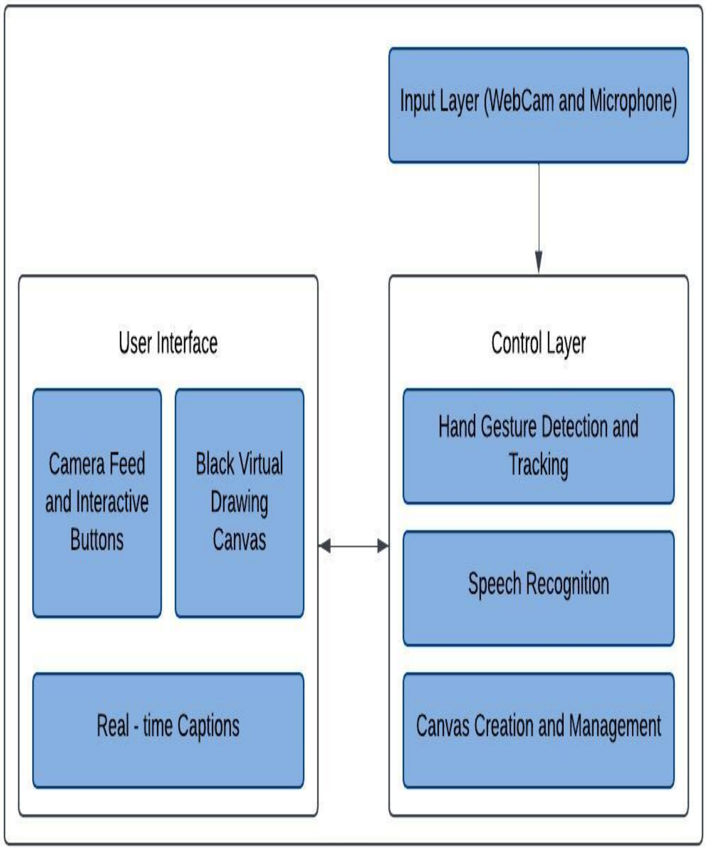

Input Layer (WebCam and Microphone)
User Interface
Camera Feed and Interactive Buttons
Black Virtual Drawing Canvas
Real - time Captions
Control Layer
Hand Gesture Detection and Tracking
Speech Recognition
Canvas Creation and Management

Fig. 1. System architecture

C. Workflow

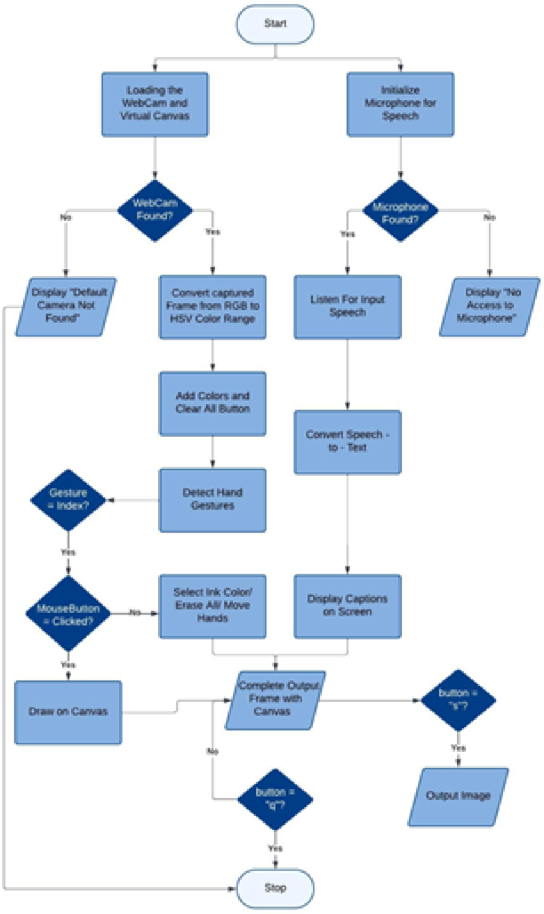

Start
Loading the WebCam and Virtual Canvas
Initialize Microphone for Speech
WebCam Found?
No
Yes
Microphone Found?
Yes
No
Display "Default Camera Not Found"
Convert captured Frame from RGB to HSV Color Range
Listen For Input Speech
Display "No Access to Microphone"
Add Colors and Clear All Button
Convert Speech - to - Text
Detect Hand Gestures
Gesture = Index?
Yes
MouseButton = Clicked?
No
Select Ink Color/ Erase All/ Move Hands
Display Captions on Screen
Yes
Draw on Canvas
Complete Output Frame with Canvas
button = "s"?
Yes
Output Image
No
button = "q"?
Yes
Stop

Fig. 2. Workflow diagram

IV. METHODOLOGY

A. Dataset Description

The system uses two datasets:

- Gesture Dataset: Contains approximately 500 records of hand gestures captured from multiple users. Each record includes video frames, hand keypoint coordinates, and gesture labels (e.g., draw, erase, start, stop).
- Caption Dataset: Contains approximately 300 records linking drawing trajectories to textual descriptions. Each record includes gesture sequences, trajectory features, and corresponding captions.

The datasets are partially simulated and pre-processed for normalization due to limited availability of public datasets for air drawing and real-time captioning.

B. Feature Set

1) Gesture Features:

- Hand keypoint coordinates (x, y, z)
- Velocity and acceleration of hand movements
- Gesture type label (start, draw, erase, stop)
- Frame sequence index

2) Caption Features:

- Trajectory sequence
- Shape complexity

- Drawing duration
- Gesture context

C. *Machine Learning Algorithm*

Gesture Recognition: Convolutional Neural Network (CNN) combined with Long Short-Term Memory (LSTM) is used to capture spatial and temporal features of hand movements.

Caption Generation: Sequence-to-sequence (Seq2Seq) model with attention mechanism is employed to generatedescriptive captions from gesture and trajectory sequences.

These models are selected due to their ability to handle sequential data, learn spatial-temporal relationships, and generate context-aware descriptions**.**

D. *Performance Evaluation*

The datasets are split into training (80%) and testing (20%) sets.

Gesture Recognition: Evaluated using metrics such as Accuracy, Precision, Recall, and F1-score.

Caption Generation: Evaluated using BLEU and ROUGE scores to measure the similarity between predicted and reference captions.

Model performance is validated under real-time conditions to ensure responsiveness and accuracy of air drawing and captioning.

V. RESULTS AND DISCUSSION

A. *Drawing Module Performance*

Test No.	Webcam Quality	Lighting	Accuracy (%)
1	1440p	Bright	94
2	1440p	Normal	98
3	1440p	Dark	92
4	1080p	Bright	92
5	1080p	Normal	97
6	1080p	Dark	90
7	720p	Bright	90
8	720p	Normal	94
9	720p	Dark	84

The system achieved a maximum accuracy of 98% under optimal lighting with the 1440p camera. Performance remained stable even in cluttered indoor environments and bright lighting. A minor decrease in accuracy was observed under low-light conditions, particularly with lower-resolution cameras. The system effectively minimized unnecessary drawings by integrating mouse clicks for start, stop, and erase operations, enhancing precision and user experience.

B. *Caption Module Performance*

Test No.	Environment	Noise Level	Accuracy (%)
1	Indoor	Quiet	92
2	Indoor	Normal	88
3	Indoor	High	82
4	Outdoor	Quiet	86
5	Outdoor	Normal	80
6	Outdoor	High	75

The caption generation module demonstrated high accuracy in quiet environments, achieving 92%. In noisy conditions, the accuracy decreased to 75%. The display time for captions was sufficient for user readability and understanding, and real-time feedback remained smooth and responsive.

C. Overall System Performance

The system performed effectively in both indoor and outdoor environments. High-resolution cameras (1440p) yielded the best results, while 1080p and 720p cameras showed slightly lower performance. The combination of gesture tracking and optional mouse-assisted control minimized accidental strokes and improved user precision. The speech-to-text captioning module significantly enhanced accessibility, particularly in quieter environments, and integrated seamlessly with the air-drawing module.

Overall, the proposed system demonstrates robust performance, real-time responsiveness, and improved usability compared to previous implementations, making it suitable for practical gesture-controlled drawing applications.

VI. CONCLUSION AND FUTURE SCOPE

The Virtual Air Drawing system successfully combines hand gesture recognition and real-time captioning, providing accurate, intuitive, and responsive drawing. High-resolution cameras and optional mouse-assisted controls enhanced precision and minimized errors. Real-time captions improved accessibility and user experience, making the system suitable for digital art, education, and assistive applications.

Future Work:

Support multi-user and collaborative drawing.

Enhance captions using advanced NLP models for richer descriptions.

Expand gestures to include 3D and multi-finger commands.

Adapt the system for mobile, AR, or VR platforms.

Improve speech captioning in noisy environments.

Integrate with AI-based art generation tools for semi-automated drawing.

REFERENCES

[1] A. Dash, A. Sahu, R. Shringi, J. Gamboa, M. Z. Afzal, M. I. Malik, A. Dengel, and S. Ahmed, “AirScript – Creating Documents in Air,” 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 1–6, 2017.

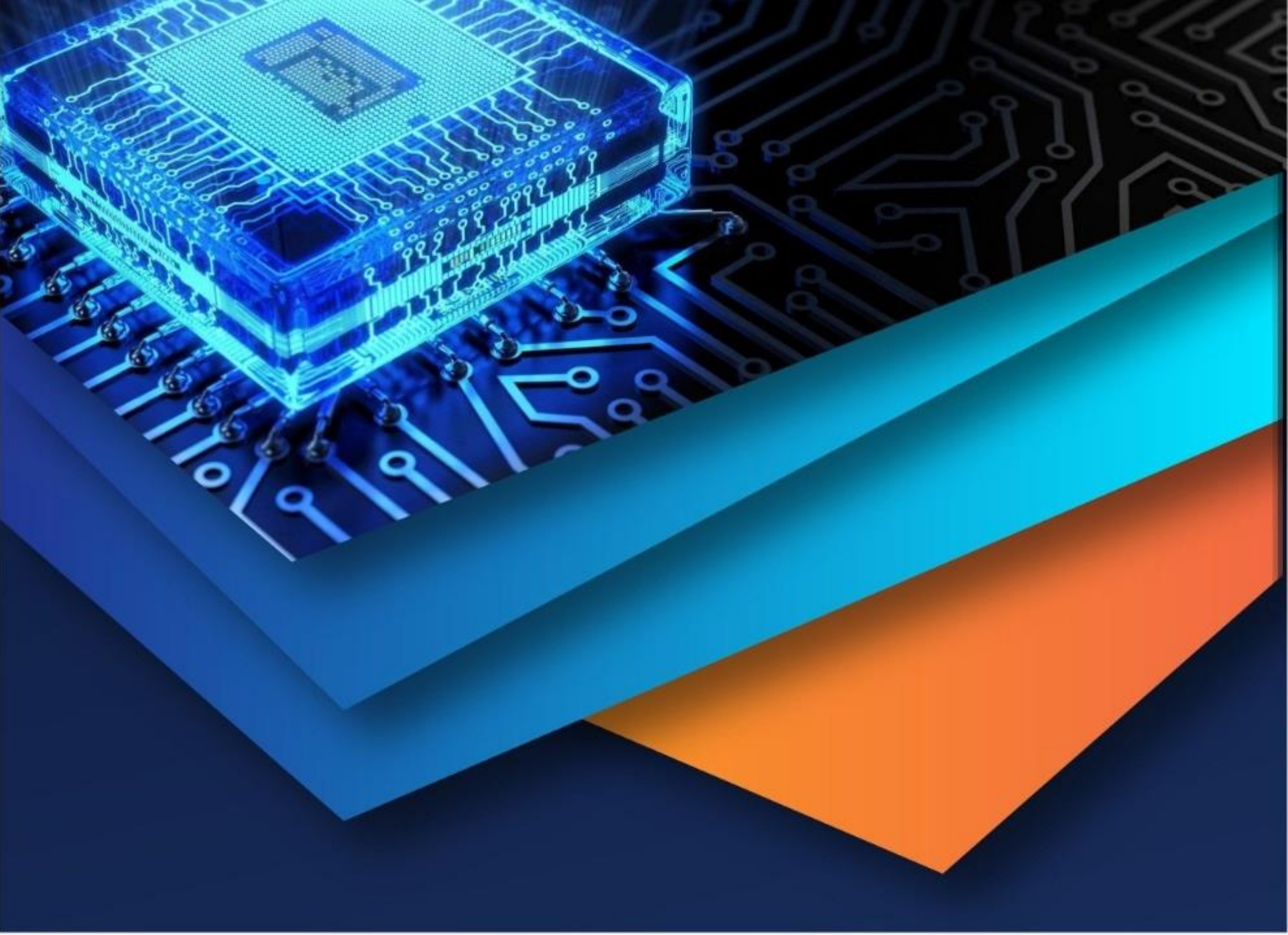
[2] P. Rai, R. Gupta, V. Dsouza, and D. Jadhav, “Virtual Canvas for Interactive Learning using OpenCV,” 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), pp. 1–5, 2022.

[3] A. R. Elshenaway and S. K. Guirguis, “On–Air Hand–Drawn Doodles for IoT Devices Authentication During COVID-19,” IEEE Access, Nov. 30, 2021, doi: 10.1109/ACCESS.2021.3131551.

[4] A. Mohanarathinam, K. G. Dharani, R. Sangeetha, G. Aravindh, and P. Sasikala, “Study on Hand Gesture Recognition Using Machine Learning,” Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1–6, 2020.

[5] W. Choi, J. Chen, and J. Yoon, “Step by Step: A Gradual Approach for Dense Video Captioning,” IEEE Access, May 24, 2023, doi: 10.1109/ACCESS.2023.3279816.

[6] J. H. Seong and Y. Choi, “Design and Implementation of User Interface through Hand Movement Tracking and Gesture Recognition,” ICTC, pp. 1–5, IEEE, 2018.



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY