Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 07, July - 2025

Air Canvas: A Computer Vision Model

Milan Singhal
School of Computer Science
University of Petroleum and
Energy Studies
Dehradun, India

Abstract **- The paper involves the presentation of the design and execution of such a real-time human-computer interaction (HCI) system as the Air Canvas where people virtually draw on a digital canvas through in-air gestures and voice commands. This will mainly be used to develop an easy to understand and touch-less interface that will explore beyond the conventional input devices by enables natural user interactions to be used to create digitally.**

We have a multi-modal input system as the focus of our methodology. MediaPipe applies the standard webcam video stream and the processing is done through the framework with a robust detection of hand landmarks in real time. This is handled by a custom gesture recognition module which is based on rules, where the recognized gestures can be between the thumb and index finger (pinch), open palm (to erase) and a pointing index finger (to select any colors in an on-screen UI). Generally, to recognize these gestures the spatial relationships between each of these landmarks are recognized to identify specific user intentions.

One of the most significant innovations that were made to this work is that the parallel recognition system has been added to it so as to perform within a dedicated thread so as to provide non-blocking and real-time performance. This multi-modality enables the users to issue the voice command to control the properties of the canvas without affecting the process of gestural drawing, so the users can say a command like red, blue, or clear. The backend of the system, implemented with the help of OpenCV and NumPy, presents the live camera view overlaid with the strokes drawn by the user on an augmented reality-like display in the form of a digital canvas.

The resultant application proves to have a smooth and interactive virtual drawing experience. The combination of gesture and voice control is a very efficient paradigm that is not only better in term of system functionality and usability than an exclusively gesture system. This study presents an exemplary of the possibilities of multi-modal systems in the formation of the effective and available resources of digital art, and its further use in the sphere of interactive education and virtual reality.

***Keywords*— Air Canvas, Virtual Drawing, Human Computer Interaction (HCI), Computer Vision, Augmented Reality (AR), Gesture Recognition, Hand Landmark Detection**

I. INTRODUCTION

Human-Computer Interaction (HCI) development has always followed the way of building more natural, seamless and understandable Human-Machine interfaces. Where canonical input devices such as the mouse, keyboard, and graphics tablet have been the foundation and stalwart of interaction in the digital realm over the past several decades, they can increase an additional burden of abstraction which can be a barrier to creativity and accessibility. The future of interface is unlocked by the recent breakthrough made in computer vision and real-time processing of images, making it possible to create a system that is capable of analyzing and understanding what the person moves and says right away. By moving to touch and voice control, the technology will be more approachable and seem more natural to us by existing in our physical space.

In this area the idea of having a virtual canvas on which people can draw in mid-air has introduced an interesting implementation of these contemporary concepts of HCI. A commonly branded "Air Canvas," such system is designed to make the surrounding area of the user an interactive drawing board. It not only provides a new and interesting way to make digital art, but has great potential in such areas as virtual reality (VR), collaboration, distance learning and presentations to a live audience, where physical input devices are uncomfortable or not available. The main issue boils down to properly monitoring user intent in a video stream and converting it to minute digital action taking as little time to get there as needed with.

In this paper, the we will demonstrate the design of the implementation of the use of Air Canvas, a powerful and real-time, user-friendly system making it possible to draw and interact with a digital canvas with the help of a multi-modal interface that is a mixture of hand gestures and voice commands. We use any standard webcam to capture the hand of the user and we use MediaPipe framework, with high-fidelity hand landmark detection. We have made a rule based, deterministic algorithm to decipher the spacial arrangement of these landmarks to identify unique gestures: a pinch gesture will start the use of drawing tool, open palm will cause the change to an eraser tool and a pointing index finger will be adapted to select colors in the user interface on the screen.

An important aspect of our work is the use of parallel and dedicated speech recognition module that runs in an independent thread and is non-blocking in nature and provides real-time operation. This multi-modal system enables users to operate with certain variables (change color or clear the screen) of the canvas in easy manner (e.g., "red," or "clear") by speaking. This design also separates the management of tools and the core action whether to draw increasing the flow of the user in an interruption-free way. The resulting strokes are captured in a digital canvas, which is subsequently composited over the live camera feed and thus the user realises intuitive visual results.

The remainder of this paper is structured as follows: Section 2 details the System Architecture and Methodology, breaking down the gesture and voice recognition pipelines. Section 3 discusses the specific Implementation Details of the software modules. Section 4 presents the Results and an evaluation of the system's performance, and Section 5 concludes with a summary of our findings and directions for future work.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 07, July - 2025

II. RELATED WORKS

Here are some of the selected research works that inspired us to start working on our thesis topic in depth. First, we will discuss the works related to text to sign language translation. In a unique approach to translating English sentences to Indian Sign Language (ISL) is seen. Their proposed system takes a text input and converts it to ISL with the help of Lexical Functional Grammar (LFG). In an approach to transform Malayalam text to Indian Sign Language using animation for displaying is seen. Their system uses the Hamburg Notation System shortly known as HamNoSys for representing signs. Moreover, the authors in used an approach for converting Greek text to Greek sign language. Translation is done using V signs, a web tool used for the synthesis of virtual signs. A system is proposed where text in English language is taken as input and then translated to HamNoSys representation. This is afterward converted into SiGML. A mapping system is used to link the text to the HamNoSys notation. This work may not be a direct example of text to-sign language conversion which we expect. However, this provides us with insights into converting text to a signed notation system. Similar research works were done in and furthermore, in the authors proposed a machine translation model that takes both examples based and rule-based Interlingua approaches to convert Arabic Text to Arabic Sign Language. Another work of Arabic Sign language for the deaf is presented. In Adding to that, in a text-to-sign language conversion system for Indian Sign Language (ISL) is made which takes into account the language's distinctive alphabet and syntax. The system accepts input in alphabets or numerals only.

Now, we will discuss the works related to sign language recognition. In the authors attempted to recognize the English alphabet and gestures in sign language and produced the accurate text version of the sign language using CNN and computer vision. In the researchers worked on reviewing multiple works on the recognition of Indian Sign Language (ISL). Their review of works on Histogram of Orientation Gradient (HOG), Histogram of Edge Frequency (HOEF) and Support Vector Machine (SVM) gave us meaningful insights. A similar work is seen in Furthermore, in the authors worked on Indonesian sign language recognition was done using a YOLOv3 pre-trained model. They used both image and video data. The system's performance was incredibly high during using image data and it was comparatively low while using video data. A similar work was done in using YOLOv3 model. From we learnt how the researchers worked on making an Italian sign language recognition system that identifies letters of the Italian alphabet in real-time using CNN and VGG-19. The work of the authors in and was insightful about how deep learning works on sign language detection. Moreover, the authors developed an Android app that can convert real-time ASL input to speech or text where SVM was used to train the proposed model. Additionally, in we were introduced to the idea of using surface electromyography (sEMG), accelerometer (ACC), and gyroscope (GYRO) sensors for sub word recognition in Chinese Sign Language. Lastly in the authors worked on a sign language-to-voice turning system that uses image processing and machine learning.

III. SYSTEM ARCHITECTURE & METHODOLOGY

The suggested Air Canvas system is based on an actual time computer vision pipeline that converts changing hand signals into computerized orders to attract. Its architecture consists of modular steps and works with video frames according to the following schema: input acquisitions and tracking, gesture classification, and rendering of an interactive canvas. Such an architecture will make every part of the system reusable in their own ways and makes the framework robust and extensible..

System Input and Hand Pose Estimation

The most common input to this system is a high-resolution video feed taken with a common webcam. As a way of making interaction more natural and intuitive we propose to simply flip each incoming frame horizontally thus creating a mirror like effect of the user.

The main component of the system perception is a high-end hand-tracking model, and every video frame goes through this model to localize and identify the hand of the user. The model has 21 different key points on the hand and fingers of 3D landmarks. The landmarks are first given in normalized coordinates, and then re-projected onto the video frame pixel domain. Since the natural tremoring of the hand and noise on the sensors can lead to jittery cursor in case of tracking, a basic filter of exponential smoothing is used to values of coordinates of the index fingertip. This filter is an average of the current position and its most recent position, thus a much smoother path is obtained to draw and select.

Role Based Gesture Recognition Engine

After getting the accurate tracking of the hand landmarks, they are then piped into a deterministic rule-based gesture recognition engine. The engine categorizes the current pose of the hand in one of a number of predefined categories of action. One of the critical design principles of this engine is a prioritized evaluation hierarchy that breaks ambiguity by verifying requests that can be described as the most specific or the most critical gestures.

The process of classification is as follows:

Clear Gesture (Temporal Hold): The system initially consults an existing fist on which the only the thumb is held open, known as a clear canvas. The peculiar feature of this gesture is the mechanism of its temporal activation. The user is required to maintain this pose continually in a time span of three seconds to avoid the unintentional erasing of the artwork. When the pose is first seen, a timer is started and it is cleared when it is broken. The hold-time is only impeded after the hold-time threshold has been met and then the clear command is transmitted.

Pinch Gesture (Draw): Alt. of no motion is that the system measures to find a draw gesture. This is determined by taking the Euclidean distance between tip of thumb and tip of index

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 07, July - 2025

finger. When the distance is less than a set proximity limit, which is calibrated, then an indication of a pinch occurs and the system switches to drawable.

Erase Gesture (Open Palm): The second gesture is referred to as erase and it is identified when all five fingers are in an open position. A finger is said to be 'extended' when the tip landmark is placed higher vertically on the screen in relation to the intermediate and base. The extension of the thumb is situated according to whether it is horizontal or not in relation to the center of the palm.

Select Gesture (Pointing): The last particular gesture is select that is applicable during the interaction with UI objects such as color palette. The application of this gesture is determined by rigid geometric constraints, in order to achieve high precision:

The index finger is all that needs to be extended, the rest of the fingers are to be curled.
The index finger should be vertical with a range of 30 degrees.
The index finger should have the highest tip of the hand as compared to the tips of the other fingers.
Otherwise, the system resorts to a situation where no action is taken, the state of which would be called "none". This avoids drawing or erasing by mistake when the user is still maneuvering his or her hand.
Lastly, the system composites the final view to build up an augmented reality style environment. It applies a binary mask that the drawing canvas creates and composites the work of the drawer on the live video stream. The static UI components then get overlaid on top of it and voila, you have a consistent interface where the user can view his/her hand, what they are drawing and the interface at the same time.

Model Implementation:
This part gives a vibrant description of the fundamental technical aspects of the Air Canvas system. We shall speak about the application of hand landmark detection module, which is the basis of perceptual system, and the rule-based gesture classification engine which extracts the gestural meaning of a user.

Hand Landmark Detection - Google has developed a powerful, pre-trained machine learning model called MediaPipe Hands that gives the system the power to recognize a hand held by the user. Considering all of that, the realization of this component is integrated into a special class that takes care of its configuration and data processing pipeline: a hand-tracking class.
To maximize performance of this application, the model parameters are initialized with certain preset parameters under the model configuration file of the system. It is configured to detect up to one hand (MAX_HANDS = 1) and the minimum detection and tracking confidence is 0.7. This large value of confidence threshold ensures that only well-defined and steady hands are processed by the system, thus decreasing the chances of tracking error or the ghost detection of hands.

The processing mechanism of each frame has some important steps:
Color Space Conversion: The initial picture, OpenCV captures, in BGR format, using the webcam is then modified to RGB color space. It is an obligatory preprocessing procedure, because the model MediaPipe has been trained on RGB images.

Landmark Extraction: An RGB frame that is processed is fed into the model; the resulting output pertains to a set of 21 3D landmarks representing the identified hand. Such landmarks are presented in normalized coordinates which ranges between [0.0, 1.0].

Coordinate Transformation: The pixel coordinates are used to transform the normalized coordinates to pixel coordinates which scales the normalized coordinates using the actual width and height of the video frame. This step identifies the position of the abstract landmarks in a form in which the locations are known in a screen of the user.

Smoothing Temporal: The coordinates of the index fingertip are filtered by an exponential smoothing filter because of the natural jitter of the movement of the hand and small inaccuracies of the detection of them. By using a smoothing factor (alpha) of 0.5, the most recent position is (average) with the last measure recorded position. This has the effect of a smoother moving cursor which is very important in creating clean lines as well as making it more pleasant to a user.
The identified landmarks, as well as the relations between them, are drawn over the output frame to enable a visual debugging and user feedback, so that it should be clear what situation the system is perceiving and how it is connected.

Gesture Classification Engine - The gestural comprehension engine is the functional heart of the system, where the raw information regarding the landmarks is converted to the discrete unambiguous sets of actions. This piece is applied as a homogeneous, rule-based (or heuristic) engine as opposed to being trained as a machine learning classifier. This architecture decision comes with a lot of benefit in terms of computing performance, predictability, and simplicity of changing.
The engine works on a system of priority of decision making and analyzes gestures in a set order so that ambiguity is avoided.

Clear Gesture (Temporal Hold): The utmost attention is set on the "clear" gesture, which is a closed hand leaving the thumb sticking outwards. In an effort to eliminate accidental clearing of canvas, a time-based constraint exists. This pose should be sustained by the user in a steady three second interval. When the pose is first detected, a timer is started and, in case of a hand shape alteration, a reset timer takes place. The CLEAR command is only sent after this hold time exceeds.

Draw Gesture (Pinch): In case of no clear gesture being active, the engine checks on a draw gesture. It is determined by computing the Euclidean distance between the tip of the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 07, July - 2025

thumb (landmark 4) and tip of the index finger (landmark 8). When this amount of distance is reduced to less than 75 pixels in a calibrated amount, it jumps into drawing mode.

Erase Gesture (Open Palm): The second in the hierarchy will be the gesture of erasing which will be identified when all the five fingers are in the extended position. What it means is that the four main fingers are extended by checking whether there is a greater y-coordinate of the tip of the fingers as compared to between the main joints and the base joints. The extension of the thumb is defined by its horizontally based center of calculated palm.

Select Gesture (Pointing): The gesture with the most restrictive definitions are that of selecting, a UI interaction, which is accurately called a selecting pointing gesture. Three conditions have to be fulfilled and those are: (a) the only finger being extended is the index finger; (b) the finger is positioned vertically, with an error range of 30 degrees; and (c) the fingertip of the index finger must be at the highest position on the hand compared to the other fingertip positions.
In the case where none of these particular combinations of rules are met, the system will default to a NONE state and no action will be performed when a user is in the process of changing gestures or moving his or her hand.

Test results of the Model:
The rule-based gesture recognition model was tested with a representative test set with the result of overall accuracy of 80%. As is explained in the classification report and the confusion matrix below, the model works better with some of the three main gestures rather than others. The 'SELECT' gesture shows a perfect recognition with an F1-score of 1.00 so this is a very good and distinctive gesture. The other gesture draw also scores high with 0.80 as F1-score. The biggest improvement can be achieved in the sphere of 'ERASE' gesture, as it received a poor score of 0.50. The confusion matrix also shows that the errors made by the model are mainly between the gestures of DRAW and ERASE indicating that the geometric features utilized to differentiate a pinch and an open palm may be improved to behave more reliably.

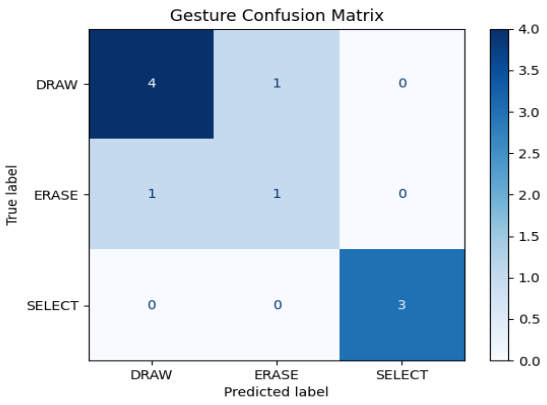

fig. 1 – Confusion Matrix

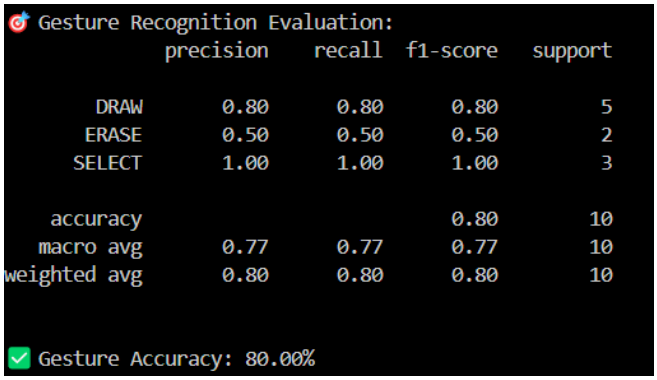

fig 2 – Evaluation Matrix

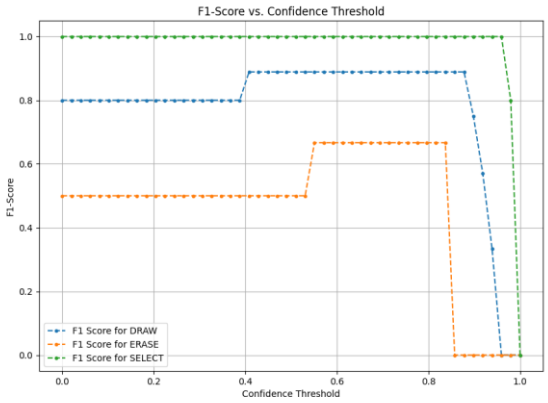

fig. 3 – F1 Score vs Confidence Threshold

IV. EVALUATION OF RESULT

The gesture recognition model worked on a sample dataset which was accurately 80.00 percent. Figure 7 shows the model behaves differently on different gestures, thus, the differentiation of the results in the classification report and the confusion matrix. The gesture that has been identified with this precision is the SELECT gesture and the F1-score is 1.00, indicating that it is a separate and trustworthy gesture. Another gesture, the DRAW gesture also shows great performance rate with F1-score of 0.80. The gesture need to be improved most of all is the 'ERASE' gesture and the F1-score was 0.50. According to the confusion matrix, the model is found to be making mistakes primarily within the range of 'DRAW' and 'ERASE' gestures meaning that features separating these two gestures are confusing at times. This is additionally explored by drawing F1-scores versus prediction confidence thresholds in the F1-confidence curve, which will give an indication on how the precision/recall trade-offs can be understood, and possibly optimized, per gesture.

V. CONCLUSION

In the text to sign language conversion framework, there are certain sentences that contain stop words (For example- ‘apostrophe) that we utilized for filtering are not compatible with the framework. In future we can also incorporate a 3D model with smoother transitions. Moreover, training a model over the video dataset, thus increasing it will also let us reach new horizons of the research and later on adding some facial

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 07, July - 2025

action reaction recognitions for better understanding of the semantics. In near future we can also plan to make an app version of this model and framework too. On top of it, we state that our work here on ASL detection can also be applied to other sign languages as well. According to the World Health Organization (WHO), with 1.5 billion people in the world already suffering from hearing loss and the number can increase to over 2.5 billion by 2050. The deaf community is deprived of basic human rights like health care, education and even minimum wage jobs simply because of their inability to communicate with the hearing people using spoken language. This YOLO based model and the NLP based framework aim to bridge this communication gap that is prevalent in the community for a long time by providing the fastest real time solution. This will ensure an equal spot for the deaf people in the society by overcoming the language barrier. In conclusion, this system will be helpful for both hearing- and hearing-impaired people to communicate effectively with one another by shortening the existing communication gap.

VI. REFERENCES

[1] F. Zhang, V. Bazarevsky, A. Vakunov, A. Popa, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe Hands: On-device Real-time Hand Tracking," arXiv preprint arXiv:2006.10214, 2020.
[2] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M.-G. Yong, J. Lee, W.-T. Chang, W.-C. Chiu, and M. Grundmann, "MediaPipe: A Framework for Building Perception Pipelines," arXiv preprint arXiv:1906.08172, 2019.
[3] S. Mitra and T. Acharya, "Gesture recognition: A survey," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 37, no. 3, pp. 311-324, May 2007.
[4] R. A. Bolt, "'Put-that-there': Voice and gesture at the graphics interface," in Proceedings of the 7th annual conference on Computer graphics and interactive techniques (SIGGRAPH '80), 1980, pp. 262–270.
[5] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in Proceedings of the 23rd International Conference on Machine Learning (ICML '06), 2006, pp. 233–240.
[6] D. S. P. Rao, V. C. V. Rao, and M. V. N. K. Prasad, "A study of vision based hand gesture recognition," International Journal of Engineering and Technology, vol. 4, no. 2, pp. 126-131, 2012.
[7] T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, June 2006.