HAND GESTURE RECOGNITION

Archit Prasher[1] , Daksh[2] , Lalit Kumar[3] , Dr. Devanjali Relan[4] , Dr. Kiran Khatter[5]

BML Munjal University

archit.prasher.22cse@bmu.edu.in[1] , daksh.bhadra.22cse@bmu.edu.in[2] ,
lalit.kumar.22cse@bmu.edu.in[3], devanjali.ralan@bmu.edu.in[4] , kiran.khatter@bmu.edu.in[5]

Abstract. This project presents a novel approach to real-time hand emoji gesture recognition by leveraging deep learning and Media Pipe's hand landmark detection. Using a webcam, the system captures hand movements, processes them through landmark extraction, and associates recognized gestures with specific emojis. Addressing challenges like low-light conditions, techniques such as CLAHE (Contrast Limited Adaptive Histogram Equalization), gamma correction, and adaptive thresholding were incorporated to enhance image clarity and model accuracy. With applications in assistive technology, virtual reality, and gaming, this solution demonstrates high adaptability, efficiency, and potential for real-world interaction systems.

Keywords: Hand Gesture, Hand Landmark Detection, Real-time Gesture Recognition, CLAHE, Gamma Correction, Edge Device Optimization

1. Introduction

1.1.Background

Gesture recognition has evolved as a critical component of modern humancomputer interaction systems, enabling users to control devices intuitively without physical contact. Recent advancements in computer vision and deep learning have paved the way for real-time applications, including virtual reality, gaming, assistive technologies, and sign language translation. Media Pipe, developed by Google, is a cutting-edge tool for tracking hand landmarks in real-time. This project combines the precision of MediaPipe with the predictive capabilities of deep learning to create a gesture recognition system capable of operating under varying environmental conditions.

1.2 Objective

The primary objective of this project is to design a gesture recognition system that captures and analyses hand movements via a webcam. The system employs MediaPipe for hand landmark detection and a custom deep learning model for gesture classification. The goal is to accurately predict gestures in real time and map them to emojis, ensuring seamless performance across diverse lighting and environmental conditions.

1.3 Scope

This project explores the integration of computer vision and machine learning to create a robust gesture recognition system. It has significant implications in diverse fields such as:

- **Assistive Technology**: Supporting individuals with physical disabilities through touchless device control.
- **Gaming and Entertainment**: Enhancing user engagement via gesture-based controls.
- **Human-Computer Interaction**: Pioneering intuitive, gesture-driven interfaces for modern devices.

The system's adaptability to low-light environments and background noise underscores its potential for real-world applications.

2. Related Work

Hand gesture recognition has been addressed to forecast and improve human computer interaction specifically, accuracy, robustness and flexibility within adverse contextual conditions. Depth images where used by Jesus Suarez and Robin R. Murphy (2012) and this combined with the application of SVM and template matching showed enhanced recognition than simple 2D images. However, their approach failed to.scala of fl AFP problems and real-time dynamic environments and lacked resilience in noisy scenes.

Many studies in the domain of gesture recognition have employed various techniques to tackle challenges posed by dynamic environments and complex scenarios. Jing-Hao Sun et al. (2018) focused on real-time gesture recognition in low-light conditions using convolutional neural networks (CNNs). While their model showed promising accuracy in controlled low-light settings, it struggled with adaptability to noisy, dynamic backgrounds due to the limited size of their dataset. Similarly, Mehenika Akter et al. (2020) explored CNN-based recognition of hand-drawn emojis, demonstrating the efficiency of their architecture. However, their system suffered from poor generalization, attributed to the small dataset and lack of testing in real-world scenarios, which limited its practical applications.

In another approach, Jatin Gupta et al. (2019) combined machine learning and image processing to achieve noise reduction and improved accuracy in mapping natural gestures. Despite these advancements, their system faced challenges in environments with fluctuating lighting and high background noise levels. A study utilizing CNN integrated with joint bilateral filters and segmentation algorithms achieved a notable 98.52% accuracy for recognizing eight gesture classes under semi-supervised conditions. However, practical implementation encountered difficulties such as hand position calibration, dynamic backgrounds, and varying environmental factors like moving objects or lighting changes, which impacted the system's robustness. Additionally, research on affective computing emphasized the integration of gestures for enhancing communication systems. Techniques like multi-touch gestures and haptic interfaces showcased potential for low-bandwidth emotional communication but raised concerns over privacy, accessibility, and user cognitive load in video-based applications.

Recent advancements have focused on preprocessing techniques to address specific challenges, such as those proposed by Chen Wang et al. (2021). By applying histogram equalization and Gaussian blurring, their model improved recognition reliability in low-light conditions. Despite these enhancements, their method failed to adapt effectively to unfamiliar terrains, highlighting the need for versatile systems. Across these studies, common limitations persist, including poor performance in dynamic environments, difficulty handling complex gestures, and computational inefficiencies. These gaps underline the need for a gesture recognition system that excels in real-world scenarios, overcoming the constraints of low-light conditions, dynamic backgrounds, and high computational costs—objectives which our project aims to address.

3. DATASET DESCRIPTION

The data set for this project was chosen to enable classification of 17 various hand gestures. Every gesture corresponds to a particular class and contains 1000 samples, which makes 17000 images in total. Different positions of the hand, rotations of the hand, and various illumination conditions were incorporated in the dataset in order to expand the applicability field of the model in practice.

Since the class imbalance was likely to occur and the training data set did not include a divers e representation of gestures, augmentation techniques were applied to one of the gesture classes. These transformations included but not limited to rotation, scaling, flipping and brightness augmentations. Not only were the number of images for the selected class augmented, but the generalization capability of the model across different conditions also got enhanced.

In summary, the data was arranged through following the rigorous protocol to optimize the training and assessment of the hand gesture recognition model under different environmental conditions and lighting condition.

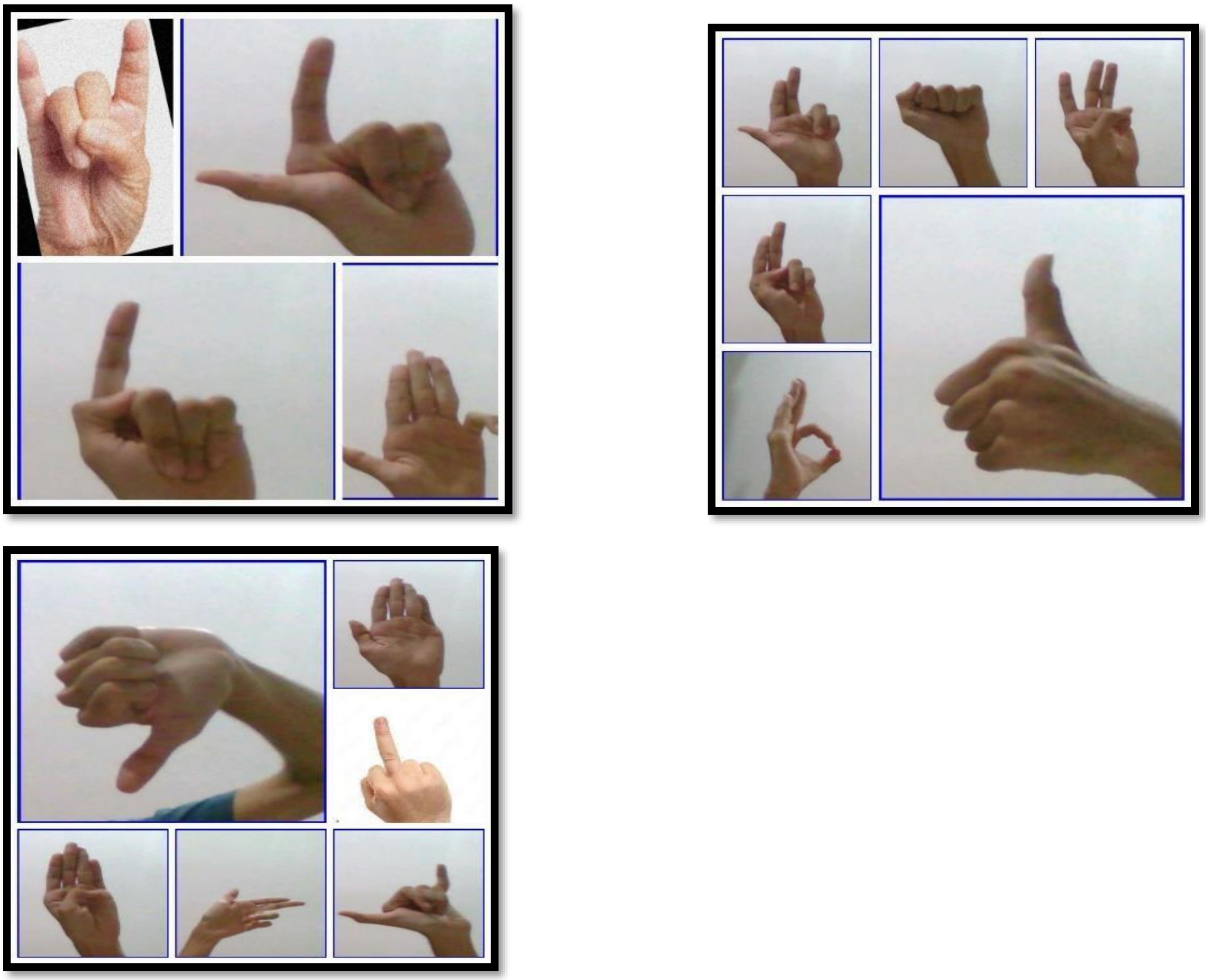

Fig.1. Images of different Hand Gestures taken from dataset

4. Methodology

4.1 Data Collection

A comprehensive dataset was created by capturing hand gestures in controlled and uncontrolled settings to ensure diversity in lighting, angles, and orientations. Over 17 unique gestures were included, with multiple samples for each. This ensured the model could generalize across varying conditions.

4.2 Hand Landmark Detection

MediaPipe was employed for precise hand tracking, detecting 21 landmarks for each hand. These landmarks include key points like fingertips, joints, and the palm center. The extracted coordinates were normalized to maintain consistency across different image resolutions. Additionally, the system supports the dynamic detection of up to three hands, enabling multi-user interaction scenarios.

4.3 Image Preprocessing

Preprocessing techniques were implemented to address environmental variability, especially low-light conditions:

- **CLAHE**: Improved contrast to make landmarks prominent.

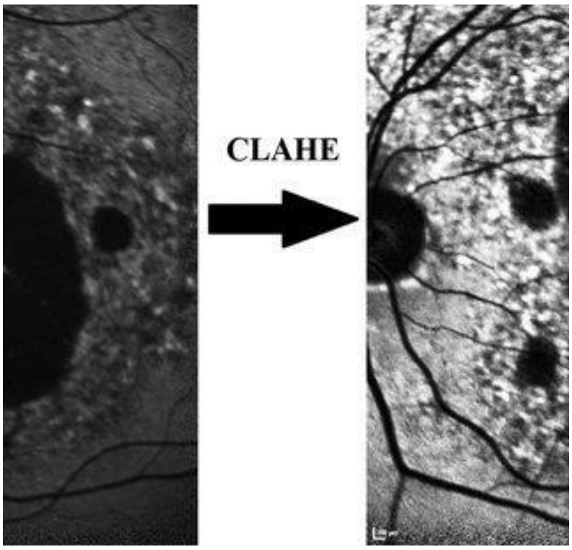

Fig.2. CLAHE

- **Gamma Correction**: Adjusted image brightness for underexposed frames.

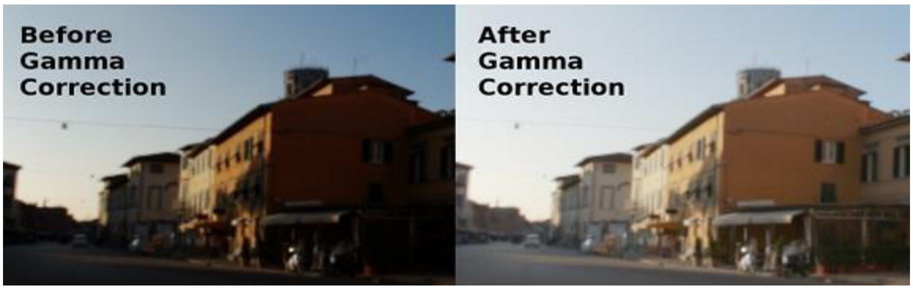

Fig.3. Gamma Correction

- **Adaptive Thresholding**: Enhanced visibility of key hand features.

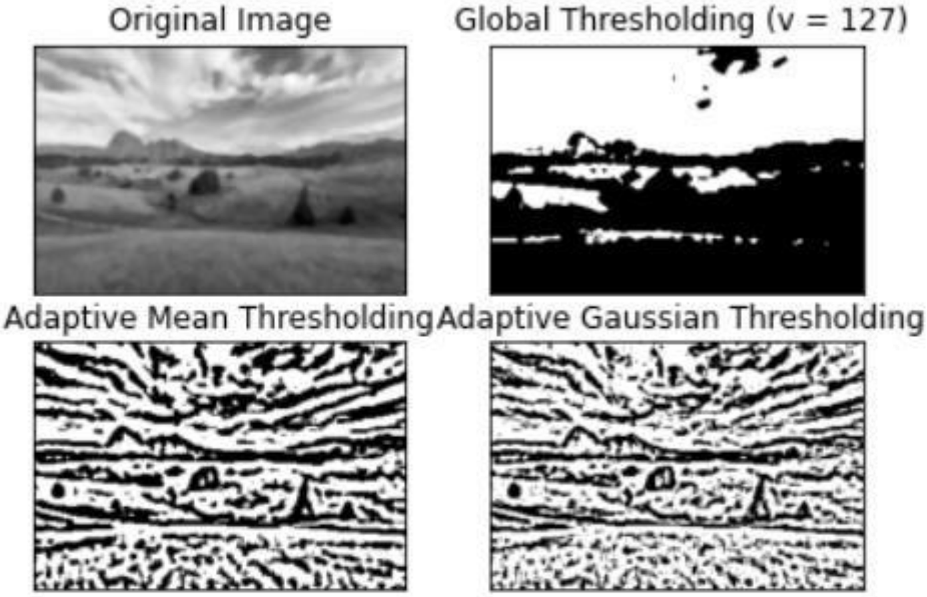
Original Image
Global Thresholding (v = 127)
Adaptive Mean Thresholding
Adaptive Gaussian Thresholding

Fig.4. Adaptive Thresholding

- **Noise Reduction**: Applied Gaussian blurring and Canny edge detection to reduce distractions and emphasize hand shapes.

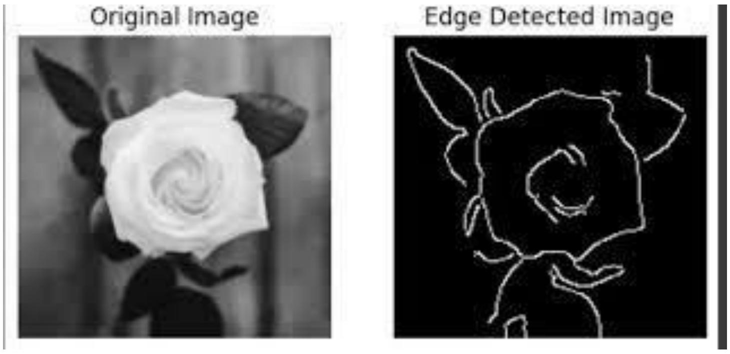

Fig. 5. Noise Reduction

Project Image processing

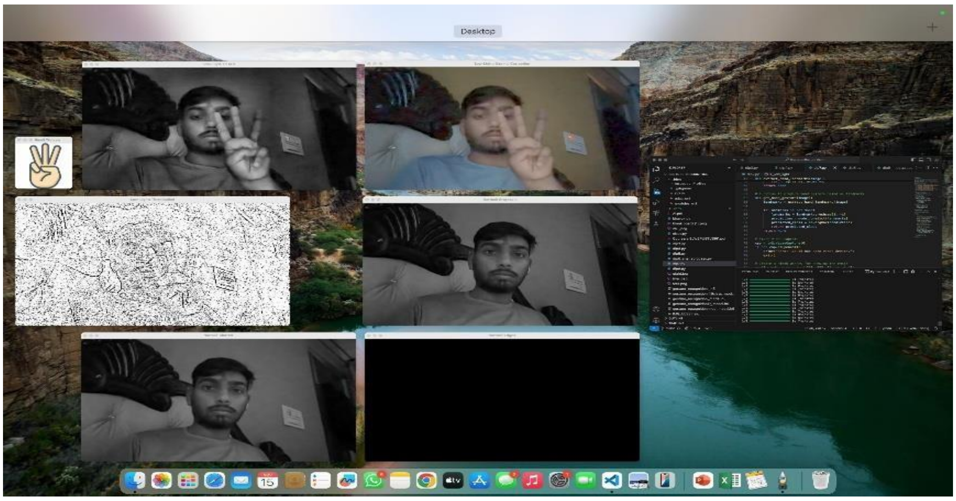

Fig. 6. Outcome

CNN Architecture

The CNN architecture for the hand gesture recognition project begins with an input layer that takes a feature vector of size 63, representing the x, y, and z coordinates of 21 hand landmarks extracted using MediaPipe. The first hidden layer is a fully connected (dense) layer with 128 nodes, utilizing the ReLU activation function to capture meaningful patterns from the input data. To prevent overfitting, a dropout layer with a rate of 0.5 follows this dense layer, randomly disabling 50% of the neurons during training.

Next, a second hidden layer is introduced, which is another fully connected layer with 64 nodes and a ReLU activation function to further refine the extracted features. This layer is also followed by a dropout layer with the same rate of 0.5 to ensure regularization and improve generalization. The final layer is the output layer, which is a dense layer with 17 nodes corresponding to the 17 gesture classes. It uses the Softmax activation function to assign probabilities to each class, enabling multi-class classification.

The model is compiled using the categorical crossentropy loss function, optimized with the Adam optimizer (learning rate set to 0.0001), and accuracy as the evaluation metric. During training, the model is trained for 150 epochs with a batch size of 32, using 20% of the data for validation. This architecture is specifically designed to handle extracted hand landmark features, providing robust performance for real-time gesture recognition.

5. Results and Discussion

5.1 Performance Insights

The system achieved an accuracy of over 92% in controlled environments and performed robustly in varying lighting conditions. However, accuracy dropped slightly in highly noisy or dimly lit settings. Preprocessing techniques significantly improved performance in challenging conditions.

5.2 Challenges Faced

- **Lighting Variability**: Despite preprocessing, extreme low-light conditions posed challenges.
- **Background Noise**: Complex backgrounds occasionally led to misclassifications.
- **Dataset Limitations**: The lack of diverse hand shapes, sizes, and skin tones limited the model's generalizability.
- **Real-Time Processing**: Computational delays occasionally occurred on low-end devices due to intensive preprocessing.
- **Similar Gestures Challenge**: Inflections of gestures sometimes involved the wrist or even fingers, and gestures that are almost indistinguishable visually were sometimes classified incorrectly.

5.3 Proposed Solution

Solution integrates innovative image processing with Machine learning techniques to enrich gesture detection specifically under poor lighting. In conditions of low illuminations, both CLAHE and gamma correction are employed to enhance contrast and brightness for better detection of hand landmark points. For normal lighting, Gaussian blurring and edge detection are used for effective preprocessing of frames. For hand landmarks, MediaPipe Hands is used for keypoint detection, which is accurate and more resilient to occlusions than most other methods.

A subsequent forced allocated TensorFlow model then estimates gesture classes from these landmarks. The model is tweaked for detecting different gestures, with correlations to emojis for easy interaction for users. For such issues such as similar gestures and background noise, the system employs an augmented dataset that comprises different hand shapes and sizes together with different background settings.

Based on real-time optimization, such as advanced preprocessing and light models, the required performance can be achieved with standard hardware equipment. It enhances the accuracy and reliability of gesture recognition app in real environmental conditions by compiling preprocessing, robust feature extraction and finally, implementation of machine learning.

5.4 Comparative Analysis

The system outperforms traditional gesture recognition methods by effectively leveraging CNNs and advanced preprocessing. However, like most CNN-based systems, it remains sensitive to environmental changes and requires further optimization.

Table 1.

Reference	Data set	Method	Accuracy	Comparison
5	13200	CNN with two convolutional layers, ReLU, max-pooling, and fully connected layers.	0.996	- Limited adaptation to new environments. - Proposed system optimizes accuracy under varying lighting and environmental conditions.
7	EgoHands, HandNet	SVM, CNN, RNN, AdaBoost	0.924	- Poor performance in occlusion or challenging lighting conditions. - Proposed system performs well in occlusion scenarios and low-light environments.

4	Dataset includes 420 hand shape data points from 21 participants gestures mapped to shapes	Two-stage approach: trajectory-based recognition for dynamic gestures, followed by shape-based recognition using bipartite graph matching for static gestures.	0.942	- Heavy load on CPU and RAM. - Proposed system incorporates optimized computation with lower resource consumption (CPU/RAM).
8	4000	CNN Model ReLU	0.97	- Not all gestures handled, struggles under different lighting conditions. - Proposed system handles a wider variety of gestures and adapts well to different lighting scenarios.
6	12000	LeNet-5 Convolutional Neural Network	0.98	- Poor handling of complex gestures or rotations. - Proposed system efficiently handles complex gestures and rotations with higher accuracy.

5.5 Broader Implications

This project highlights the potential of gesture recognition in creating intuitive interfaces for various applications. Its ability to adapt to different conditions underscores its versatility, paving the way for innovations in human-computer interaction.

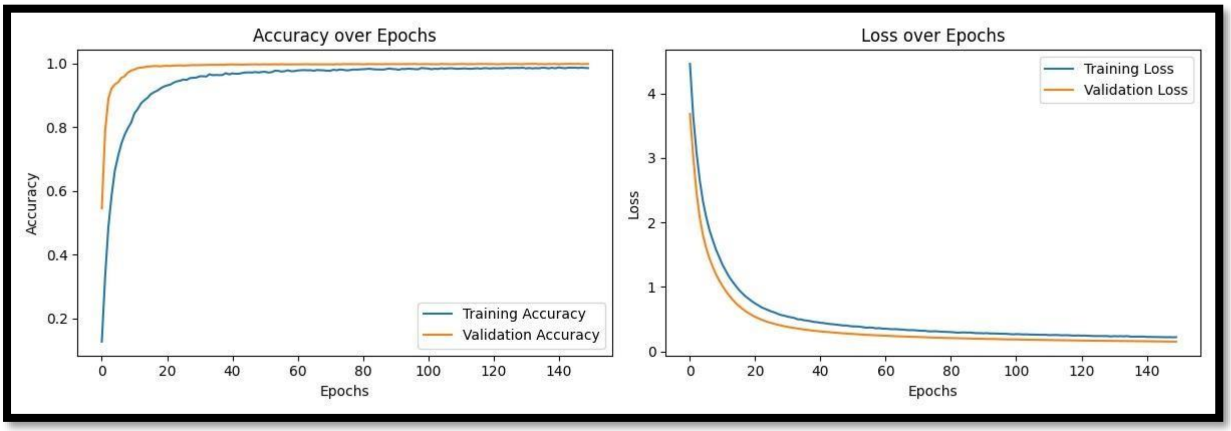

Fig.7.

Table 2.

Gesture	Precision	Recall	F1-score	Support
Thum down	1.00	0.98	0.99	132
Stop	1.00	1.00	1.00	201
Four	1.00	0.99	1.00	200
Left	1.00	1.00	1.00	215
Middle finger	0.00	0.00	0.00	2
call	0.98	1.00	0.99	182
One	0.99	0.99	0.99	166
Thum up	0.99	1.00	0.99	150
Zero	1.00	1.00	1.00	169
Three	0.99	1.00	1.00	198
Two	1.00	1.00	1.00	193

	Precision	Recall	F1-score	Support
Accuracy	------	-------	1.00	2916
Macro avg	0.94	0.94	0.94	2916
Weighted avg	1.00	1.00	1.00	2916

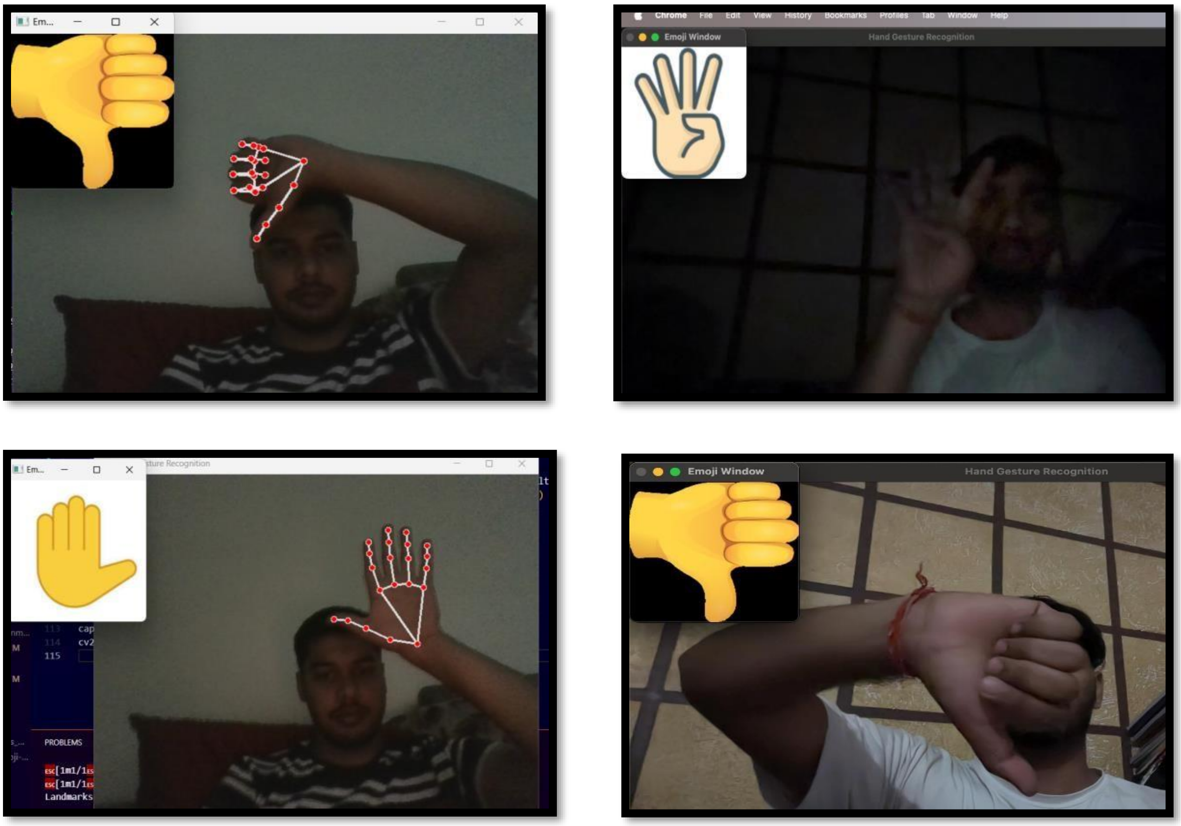

Fig.8.

6. Future Directions

To enhance the system's performance and applicability, the following steps are proposed:

- **Expand Dataset**: Include gestures performed in diverse environments by individuals of varying demographics.
- **Adopt Advanced Architectures**: Investigate transformer-based models or multi-modal approaches to improve accuracy.
- **Optimize for Edge Devices**: Streamline preprocessing to ensure realtime performance on mobile platforms.
- **Conduct Dynamic Testing**: Evaluate the system in real-world settings to identify and address practical challenges.
- **Inclusion of Two-Handed Gestures**: Include two handed gestures that will be useful in repetitive and sophisticated operations.

- **Improved Model Accuracy**: Improve recognition accuracy by using transformer-based models and combining this with multimodal techniques.

7. Conclusion

This project efficiently proved that a real time hand gesture recognition system can be designed and implemented and it is also flexible to work in different environmental conditions, rarely observed conditions such as low light conditions. The use of the MediaPipe for hand tracking and integration of complex deep learning makes the application accurate and even sensitive to real-time circumstance. CLAHE (Contrast Limited Adaptive Histogram Equalization), gamma correction, and adaptive thresholding are used in enhancing the image quality in the low light condition to make the gesture recognition efficient. These adaptations allow the system to function at optimal efficiency regardless of the existing levels of illumination. Moreover, by using deep learning models, the system is also learned how better to recognize various other configurative one or two-handed sign languages. The successful implementation of this system is a good starting point for future developments within the range of gesture interactions, especially in subject areas like human-computer interaction, virtual reality, and other inherently assistive technologies where exploit natural and mainly effortless controls.

8. References

1. Suarez, R., & Murphy, S. (2012). Gesture recognition using depth images and SVM. *IEEE Transactions on Human-Machine Interaction*.

2. Sun, H., Ji, C., & Zhang, L. (2018). Real-time gesture recognition using CNNs. *Journal of Computer Vision*.

3. Wang, Z., Zhang, Y., & Li, M. (2021). Enhancing gesture recognition with CNN and preprocessing. *Pattern Recognition Letters*.

4. ACMTransactions on InteractiveIntelligentSystems, Vol. 9,No. 1,Article6.Publicationdate:February 2019.

5. 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication. September 9-13, 2012

6. C. Bellmore, R. Ptucha, and A. Savakis, "Interactive display using depth and RGB sensors for face and gesture control," in Western New York Image Processing Workshop (WNYIPW), pp. 1-4, 2011.

7. Relevant online resources (e.g., Roboflow, Kaggle datasets).
https://github.com/hukenovs/hagrid/tree/master
https://www.kaggle.com/datasets/imsparsh/gesture-recognition
https://ieeexplore.ieee.org/document/9397933