Real-Time Hand Gesture Recognition Using Mediapipe Landmark Detection

Puccha Poojitha[1], Jaladi Abishek[2], Pedapeniki Durga Prasad[3], Koppada Vinay[4]
[1,2,3,4]Department of Computer Science & Engineering, Avanthi Institute of Engineering and Technology

***Abstract*—Hand gesture recognition is a crucial area in human–computer interaction that enables natural and touchless communication between users and digital systems. Traditional input devices such as keyboards and mice limit intuitive interaction. This project presents a real-time hand gesture recognition system using MediaPipe landmark detection to provide a more efficient interface. The system applies computer vision and machine learning techniques to detect and classify hand gestures from live webcam input. MediaPipe is used to extract 21 key hand landmarks representing joints and fingertips. By analyzing these landmarks, gestures such as open palm, fist, thumbs up, and finger counting are recognized in real time. The system processes video frames, detects the hand, extracts landmark coordinates, and identifies gestures using rule-based or learning-based methods. The implementation is carried out using Python, OpenCV, and MediaPipe libraries. The system achieves high accuracy with minimal latency and can be applied in areas such as virtual control systems, gaming, sign language recognition, robotics, and touchless interfaces.**

***Index Terms*—MediaPipe, Hand Gesture Recognition, Computer Vision, Human–Computer Interaction (HCI), Python, Machine Learning, OpenCV, Landmark Detection, Real-Time Processing, BlazePalm.**

I. INTRODUCTION

Hand gesture recognition is an important task in computer vision and human–computer interaction. It involves detecting and classifying hand movements performed by users. This technology is widely used in virtual reality, gaming, and touchless control systems. Due to variations in hand shapes, positions, and lighting conditions, it is a challenging problem. Machine learning techniques, especially MediaPipe-based models, provide effective solutions. These models can efficiently extract important features from hand landmarks. In this project, real-time webcam input is used for training and testing. The goal is to build an accurate system to recognize different hand gestures. Optimization and normalization techniques improve performance and accuracy. This project highlights the role of computer vision in real-world applications.

II. LITERATURE REVIEW

Hand gesture recognition has been widely studied in computer vision and human–computer interaction. Earlier methods used hardware devices such as data gloves and motion sensors to capture hand movements, but these systems were costly and complex. Later, vision-based approaches using cameras became popular, applying image processing techniques to detect gestures. However, these systems faced challenges such as lighting variations, background noise, and high computational requirements. Recent advancements focus on frameworks like MediaPipe, which enable real-time hand tracking by extracting 21 hand landmarks. Modern approaches also use models like LSTM, GRU, and Transformers to improve accuracy by analyzing gesture sequences. Despite these improvements, issues such as occlusion, environmental conditions, and limited gesture support remain. This review highlights the development of gesture recognition systems and the need for efficient, low-cost, and real-time solutions.

III. PROBLEM STATEMENT

Hand gesture recognition systems aim to provide natural interaction between humans and computers, but existing solutions face several limitations. Earlier systems depend on hardware devices such as data gloves and sensors, which increase cost and reduce accessibility. Vision-based methods using cameras also suffer from challenges like poor lighting conditions, background complexity, and variations in hand shapes and positions. These issues affect the accuracy and reliability of gesture detection. Additionally, many approaches require high computational power and large datasets, making them unsuitable for real-time applications on standard devices. Occlusion and multi-hand interference further reduce performance. Therefore, there is a need to develop a cost-effective and efficient system that can accurately recognize hand gestures in real time using lightweight techniques and minimal hardware requirements. The problem focuses on improving robustness, speed, and usability for practical deployment.

IV. SYSTEM OVERVIEW

The proposed system is designed to recognize hand gestures in real time using MediaPipe and OpenCV. It captures live video input through a webcam and processes each frame to detect the presence of a hand. MediaPipe is used to identify and track 21 key hand landmarks representing joints and fingertips. These landmarks are used to extract coordinate values that describe the position of the hand. The system analyzes these coordinates using rule-based or machine learning methods to classify gestures. The recognized gesture is displayed instantly as output. The system works without additional hardware, making it cost-effective, efficient, and suitable for real-time applications.

V.DATA SET DESCRIPTION

The dataset used in this project combines a public dataset and a custom-collected dataset. The primary dataset is the ASL Alphabet Dataset, which contains labeled images of ASL gestures. To improve real-time performance, additional images were collected using a webcam under varying lighting and background conditions. Extra samples were added for palm (delete), similar gestures, and neutral hand positions. Hand landmarks were extracted using MediaPipe, generating normalized x, y, and z coordinates, resulting in 63 features per sample. This hybrid dataset improves accuracy and robustness in real-world scenarios.

VI. METHODOLOGY

The proposed system follows a structured pipeline for real-time sign language recognition. Initially, image data is collected from both a public dataset and a custom dataset to improve adaptability. Each image is processed using MediaPipe to extract 21 hand landmarks, including spatial depth information. The extracted features are normalized to ensure consistency across varying hand positions. A machine learning model is then trained using these features to classify gestures. During inference, real-time video frames are captured, processed, and passed through the trained model. Prediction smoothing is applied to reduce fluctuations, and the final recognized gesture is displayed as text output.

VII. MODEL IMPLEMENTATION

The model is implemented using a machine learning approach based on extracted hand landmark features. A RandomForest classifier is trained on normalized x, y, and z coordinates obtained from MediaPipe. The model learns to distinguish between different gesture patterns using structured feature vectors. During real-time execution, incoming frames are processed to extract landmarks, which are then fed into the trained model for prediction. A smoothing mechanism is applied to stabilize outputs and ensure consistent gesture recognition.

VIII. RESULT & ANALYSIS

The proposed system achieves high recognition accuracy of approximately 97–98% on the test dataset. The integration of depth-based features and a hybrid dataset significantly improve the distinction between similar gestures.

Real-time performance remains stable, with prediction consistency above 95% due to smoothing techniques. The palm gesture used for deletion attains an accuracy of around 94–96%. The system performs reliably under normal lighting conditions, with minimal errors during rapid motion. Overall, the model demonstrates strong performance and robustness, making it suitable for practical real-time sign language recognition applications.

IX. SYSTEM INTERFACE/ IMPLEMENTAION

The system interface is designed to provide a simple and interactive user experience for real-time gesture recognition. A live webcam feed is displayed using OpenCV, where detected hand landmarks are visually overlaid for user reference. The current predicted gesture is shown on the screen along with the dynamically generated text output. A dedicated display area is used to present the constructed sentence clearly. The implementation ensures smooth interaction by incorporating prediction stabilization and gesture locking mechanisms. The interface responds instantly to user gestures, making it intuitive and efficient for continuous real-time communication and testing purposes.

X. APPLICATIONS AND ADVANTAGES

The proposed system has wide applications in assistive communication, enabling interaction between individuals with hearing or speech impairments and others. It can be used in educational tools, public service systems, and human-computer interaction interfaces. The system offers advantages such as real-time performance, high accuracy, and ease of use without requiring specialized hardware. Its modular design allows scalability and integration into mobile or web platforms. Additionally, the use of depth-based features and custom data enhances robustness, making the system adaptable to varying environments and practical for real-world applications.

XI. FUTURE WORK

The current system can be further enhanced by incorporating deep learning models such as Convolutional Neural Networks and recurrent architectures to improve recognition of complex and dynamic gestures. Integration with mobile platforms using TensorFlow Lite can make the system more accessible. Adding text-to-speech functionality would enable audible communication of recognized gestures. Future work may also include multi-hand detection and support for complete word-level recognition instead of individual alphabets. Expanding the dataset with more diverse samples and environmental conditions can further improve robustness. Additionally, optimizing the system for low-power devices can enhance its usability in real-world applications.