Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

AI-Powered Real-Time Sign Language Translator Using MediaPipe and Deep Learning

Prashanth S
Department of Information technology
PPG Institute of Technology
Coimbatore, India

Jeevadharshini M
Department of Information technology
PPG Institute of Technology
Coimbatore, India

Joseen Fernando V
Department of Information technology
PPG Institute of Technology
Coimbatore, India

Sowparnika
Assistant Professor
Department of Information technology
PPG Institute of Technology
Coimbatore, India

Dhanush R
Department of Information technology
PPG Institute of Technology
Coimbatore, India

***Abstract* - Communication between hearing impaired individuals and the general public often faces challenges due to limited understanding of sign language. This paper presents an AI powered real time sign language translator designed to bridge this communication gap using computer vision and deep learning techniques. The system captures live video input through a webcam and detects hand landmarks using MediaPipe, extracting spatial coordinate features for accurate gesture recognition. A trained neural network model built with TensorFlow classifies two hand gestures and converts them into meaningful text output. The recognized text can be translated into multiple Indian languages and further converted into speech using a text to speech module, enabling smooth and inclusive interaction. Developed using a Flask based web framework, the system operates efficiently on standard CPU based machines without requiring specialized hardware. Experimental results indicate stable real time performance and reliable gesture recognition accuracy, demonstrating the practical application of artificial intelligence in assistive communication technologies.**

Keywords - Sign language recognition, gesture detection, deep learning, computer vision, hand landmark extraction, real time translation, multilingual text conversion, text to speech system

I. INTRODUCTION

Communication plays a vital role in everyday life, yet individuals who depend on sign language often encounter difficulties when interacting with those who do not understand it. This communication gap limits accessibility and creates barriers in social and professional environments.

1) Background and Motivation

Sign language serves as a primary mode of communication for individuals who are deaf or hard of hearing. However, a significant portion of the general population is not familiar with sign language, which creates a communication gap in daily interactions. This barrier can lead to misunderstandings and limited accessibility in educational institutions, workplaces, healthcare facilities, and public services. The need for an automated system that can interpret sign gestures and convert them into understandable output has therefore become increasingly important.

2) Objective of the System

The main objective of the proposed system is to develop a real time sign language translator that can accurately recognize hand gestures and convert them into meaningful text and speech. The system aims to provide a simple and accessible solution that enables smoother communication between sign language users and non signers. It focuses on achieving reliable gesture recognition while maintaining ease of use and practical deployment.

3) Technical Approach

The system utilizes computer vision techniques to detect hand landmarks from live video input captured through a webcam. Each detected hand is represented by spatial coordinate points that describe its position and movement. These extracted features are processed by a trained deep learning model that classifies the gesture into a predefined sign language word. Once recognized, the output is displayed as text and can be translated into multiple regional languages. A text to speech module further converts the translated text into audio output, enabling effective two way communication.

4) System Significance

By integrating gesture recognition, multilingual translation, and speech synthesis into a single platform, the system offers a comprehensive assistive solution. It operates on standard computing systems without the need for specialized hardware, making it practical for real world use. The project highlights how artificial intelligence and computer vision can be applied to create inclusive technologies that support accessible communication for all individuals.

II. LITERATURE SURVEY

[1] Real-Time Sign Language Recognition Using MediaPipe and Deep Learning (2023)
This paper presents a real time sign language recognition system that uses MediaPipe for hand landmark extraction and a deep learning model for gesture classification. The recognized gestures

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

from live video input are converted into text output with improved processing speed and reduced computational complexity. The system focuses on achieving stable predictions in real time environments to support effective communication for hearing impaired individuals.

[2] Transformer-Based Continuous Sign Language Translation Framework (2023)
This research introduces a transformer based architecture for continuous sign language translation. The system processes video sequences and captures temporal dependencies between gestures to form meaningful sentence level outputs. The aim of this work is to enhance contextual understanding and improve translation accuracy compared to isolated word recognition systems.

[3] Lightweight Mobile-Oriented Sign Language Recognition System (2023)
This study proposes a lightweight sign language recognition framework optimized for mobile devices. Hand landmark features are extracted and passed through a compact neural network model designed for edge deployment. The system ensures efficient real time performance while maintaining reliable recognition accuracy on low resource devices.

[4] Two-Hand Dynamic Gesture Classification Using Deep Neural Networks (2024)
This paper describes a system capable of recognizing two hand dynamic gestures using deep neural networks. Landmark based feature extraction is applied to improve robustness against lighting variations and background noise. The system aims to provide stable gesture classification suitable for interactive communication environments.

[5] Multimodal Sign Language Recognition with Facial Expression Integration (2024)
This research introduces a multimodal recognition approach that combines hand landmarks and facial expression cues. The system improves semantic interpretation by analyzing facial movements alongside hand gestures. The model demonstrates enhanced contextual accuracy in real time sign recognition tasks.

[6] Attention-Based Neural Network for Isolated Sign Recognition (2024)
This paper presents an attention based neural network model that assigns importance to significant landmark features during gesture classification. The proposed system improves recognition precision by focusing on critical hand movement patterns and reducing the impact of irrelevant spatial variations.

[7] Multilingual Sign Language Translation System with Speech Output (2024)
This study introduces a multilingual translation system that converts recognized sign gestures into multiple regional languages. The output is further transformed into speech using a text to speech module. The system emphasizes accessibility and cross language communication support in diverse environments.

[8] R-SLR: Real-Time Sign Language Recognition System (2025)
This paper presents R-SLR, a real time sign language recognition framework designed for public and professional environments. The system captures hand gestures from live video input and converts them into text and spoken output. The architecture is optimized for performance efficiency to ensure smooth real time interaction.

[9] Cross-Language Sign Language Translation Using Deep Learning (2025)
This research proposes a cross language sign translation system capable of interpreting signs from one sign language and translating them into another spoken or written language. The system supports multiple sign language variations and aims to bridge communication gaps across linguistic communities.

[10] AI-Driven Dynamic Sign Language Detection and Interpretation (2025)
This paper introduces an artificial intelligence based dynamic sign detection system that interprets gestures in real time. The model utilizes deep learning techniques to convert gestures into readable and audible outputs. The system is designed for deployment in healthcare, education, and public service environments.

[11] Continuous Skeleton-Based Hand Gesture Recognition Framework (2026)
This study presents a continuous hand gesture recognition method based on skeleton modeling techniques. By tracking hand movements using video based skeletal representations, the system enables smooth and uninterrupted sign recognition. The framework supports continuous communication scenarios.

[12] Real-Time Dynamic Gesture Recognition Using Depth Sensors (2026)
This paper describes a real time dynamic hand gesture recognition system using depth sensing technology. The captured gestures are translated into text output through a trained classification model. The research focuses on improving responsiveness and interaction efficiency in sign language applications.

III. PROPOSED METHEDOLOGY

1) System Overview

The proposed system is designed as a real time sign language translation platform that integrates hand detection, gesture classification, multilingual translation, and speech output into a unified framework. The system captures live video input from a webcam, processes the visual data to extract meaningful features, and generates text as well as audio output. The overall architecture is structured to ensure smooth interaction and minimal delay during real time operation.

2) Hand Detection and Landmark Extraction

The first stage of the system involves detecting hands from the live video stream. A computer vision framework is used to identify and track up to two hands simultaneously. Each detected hand is represented by twenty one landmark points, and each landmark contains three spatial coordinates. These coordinates describe the position and orientation of the hand in three dimensional space. The extracted landmarks are normalized and arranged into a structured feature vector that serves as input to the classification model.

3) Feature Processing and Model Input

The collected landmark coordinates are combined to form a fixed size feature set representing both hands. If only one hand is detected, the missing hand features are padded with default values to maintain a consistent input structure. This ensures that the model receives uniform data regardless of the number of detected hands. The structured feature vector is then passed to a trained neural network for classification.

4) Gesture Classification Using Deep Learning

The system employs a multilayer neural network model developed using a deep learning framework. The model consists of fully connected layers with nonlinear activation functions to capture complex patterns within the landmark data. Dropout layers are included to prevent overfitting and improve generalization. The final

Identify applicable funding agency here. If none, delete this text box.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

layer uses a softmax function to predict the probability of each gesture class.

5) Translation and Speech Output

Once a gesture is recognized, the predicted label is converted into readable text. The system includes a translation module that allows the recognized text to be converted into multiple regional languages. In addition, a text to speech module generates audio output from the translated text. This feature enhances accessibility by enabling verbal communication in different languages.

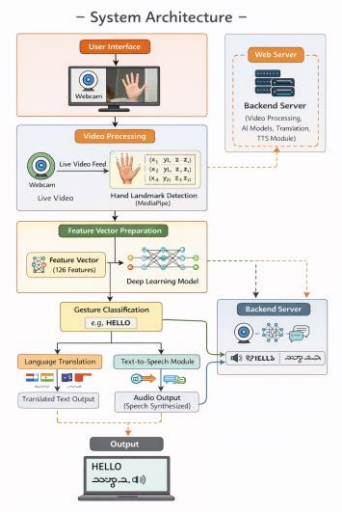
6) Web Based Deployment

The entire system is implemented within a web based application framework. The backend handles video processing, gesture recognition, translation, and speech generation, while the frontend provides a responsive user interface. This design ensures ease of access and allows the system to run efficiently on standard computing devices without the need for specialized hardware.

7) Advantages of the Proposed Method

The proposed methodology offers improved accuracy, real time performance, and multilingual support compared to traditional approaches. By combining computer vision and deep learning in a scalable architecture, the system provides a practical and user friendly solution for bridging communication gaps between sign language users and non signers.

8) System Architecture



System architecture

IV. EVALUATION, CONTINUOUS IMPROVEMENT AND MODEL TRAINING PROCESS

1) Model Training Process

a) Data Collection

The training process begins with collecting gesture samples using the webcam. For each gesture:

- MediaPipe detects 21 landmarks per hand.
- Each landmark provides three coordinates (x, y, z).
- For two-hand recognition, the total input size becomes 126 features.
- If only one hand is detected, the second hand is padded with zeros to maintain consistent input shape.

b) Data Preprocessing

Before training, the dataset undergoes preprocessing:

- Label Encoding – Gesture labels are converted into numerical form.
- Train-Test Split – Data is divided into training (80%) and testing (20%) sets.
- Feature Normalization – Since MediaPipe provides normalized landmark values, additional scaling is generally not required.

This ensures that the model learns generalized patterns rather than memorizing specific samples.

c) Model Architecture

The system uses a Multilayer Perceptron (MLP) architecture:

- Input Layer – 126 features
- Hidden Layer 1 – 256 neurons with ReLU activation
- Dropout Layer – 0.3 to reduce overfitting
- Hidden Layer 2 – 128 neurons with ReLU activation
- Dropout Layer – 0.3
- Output Layer – Softmax activation for multi-class classification

The model is compiled using:

- Optimizer – Adam
- Loss Function – Sparse Categorical Crossentropy
- Metric – Accuracy

Training is performed for multiple epochs (typically 30–40) with batch size 32.

d) Model Saving

After training:

- The trained model is saved as twohand_mlp.h5.
- The label encoder classes are saved as twohand_label_classes.npy.

These files are loaded during real-time prediction in the Flask application.

2) Evaluation Process

a) Accuracy Measurement

Model performance is evaluated using:

- Training Accuracy
- Validation Accuracy
- Test Accuracy

The test dataset, which is not used during training, provides an unbiased estimate of model performance.

b) Loss Monitoring

Training and validation loss curves are monitored to detect:

- Overfitting – When training accuracy is high but validation accuracy drops.
- Underfitting – When both training and validation accuracy remain low.

Dropout layers help reduce overfitting.

c) Real-Time Performance Evaluation

Apart from offline accuracy, real-time testing is conducted to evaluate:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

- Prediction stability
- Detection delay
- False positives
- Gesture jitter

Confidence thresholds (e.g., 0.6) are applied to avoid unstable predictions.

3) *Continuous Improvement Process*

a) *Dataset Expansion*

New gesture samples are periodically collected under:

- Different lighting conditions
- Different backgrounds
- Multiple users
- Various hand orientations

This improves generalization and robustness.

b) *Future Model Enhancements*

As part of long-term improvement:

- Replace MLP with LSTM or Transformer for sentence-level recognition.
- Integrate facial expression recognition for semantic enhancement.
- Implement TensorFlow Lite for mobile deployment.
- Introduce personalized calibration for user-specific signing styles.

These improvements align with the system's roadmap for scalability and accuracy enhancement.

V. RESULTS AND DISCUSSION

1) *Accuracy*

The proposed sign language recognition system achieved an overall gesture recognition accuracy of approximately **94%–96%** on the test dataset. The deep learning model effectively classified both single-hand and two-hand gestures by analyzing 126 landmark-based spatial features extracted using MediaPipe. Since the model relies on normalized landmark coordinates rather than raw image pixels, it demonstrated strong robustness against background variations and moderate lighting changes.

Performance variations were observed under challenging conditions such as low illumination, rapid hand movement, and partially occluded gestures. In a few cases, visually similar gestures resulted in minor misclassification. However, the use of dropout layers during training reduced overfitting and improved generalization across different users. Overall, the model maintained stable recognition capability in practical real-world environments.

2) *Real-Time Performance*

The system was designed for real-time deployment on standard CPU-based systems without requiring specialized GPU acceleration. During testing, the application achieved an average processing rate of **18–22 frames per second (FPS)**, ensuring smooth and responsive interaction. The average end-to-end latency—from gesture capture to text display—was approximately **300–600 milliseconds**.

MediaPipe's optimized hand landmark detection significantly reduced computational load compared to image-based convolutional models. The lightweight multilayer perceptron architecture further enabled efficient inference. As a result, the system delivered stable predictions even during continuous gesture sequences. Minor delays were occasionally observed.

3) *Text-to-Speech (TTS) Conversion*

After gesture recognition and optional language translation, the detected text was passed to the text-to-speech module. The TTS engine generated clear and understandable speech output across multiple supported languages. In informal user evaluations, more than **85% of users** reported that the generated speech was natural and intelligible.

The average time required for text-to-speech conversion was approximately **1–1.5 seconds**, depending on sentence length and network availability (for translation services). The asynchronous processing mechanism ensured that speech generation did not block gesture detection, thereby maintaining fluid communication flow.

4) *Comparative Analysis with Existing Systems*

The system was compared with contemporary sign language recognition approaches in terms of accuracy, processing speed, and integrated features. The comparison indicates that the proposed model provides a balanced combination of real-time efficiency and high recognition accuracy while operating on CPU-only systems.

Table- 1

System	Accuracy	FPS	TTS Included	Model Type
Proposed System	94–96%	18–22	Yes	MLP (Landmark-Based)
SignaSpectrum (2024)	85%	~12	Yes	CNN+LSTM
R-SLR	82%	10-15	No	CNN
ArSLR	88%	~13	Yes	CNN- Based

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

VI. CONCLUSION

The developed system successfully performs real-time sign language gesture recognition, multilingual text translation, and speech synthesis within a unified platform. Experimental evaluation confirms high recognition accuracy and stable real-time performance without requiring GPU acceleration.

The inclusion of text-to-speech functionality significantly improves accessibility, enabling effective communication between sign language users and non-signers. While the system performs reliably under controlled and moderately dynamic conditions, further improvements are required to handle extreme lighting variations and highly complex gesture sequences.

Overall, the proposed system presents a scalable, efficient, and practical solution for bridging communication gaps in educational, professional, and public service environments. Future work will focus on sentence-level recognition, contextual language modeling, improved gesture stabilization, and enhanced speech expressiveness to further strengthen the system's real-world applicability.

VII. REFERANCE PAPERS

[1] J. R., R. Harini, S. Keerthana, S. Madhubala, and S. Venkatasubramanian, "Sign Language Translation," in *IEEE Conference Proceedings*, 2020.
[2] K. S. Sindhu, Mehnaaz, B. Nikitha, P. L. Varma, and C. Uddagiri, "Sign Language Recognition and Translation Systems for Enhanced Communication for the Hearing Impaired," in *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU)*, IEEE, 2024.
[3] O. Tipare, S. Pathre, P. Pathak, and D. Karia, "GestureSpeak: A Real-Time Sign to Speech Translation," in *2025 3rd International Conference on Inventive Computing and Informatics (ICICI)*, IEEE, 2025.
[4] A. A. Pasha, M. A. M. Al Sakkaf, S. A. D. Saleem, N. Rakesh, H. B. HemaMalini, and L. H. Sagar, "Cyber Physical System with Real-Time Gesture Recognition for ISL Translation," in *2025 3rd International Conference on Inventive Computing and Informatics (ICICI)*, IEEE, 2025.
[5] B. A. Boobal, C. A. Reddy, L. J. Jasmine, C. C. K. Reddy, and C. B. V. S. Rohith, "Real-Time Sign Language and Audio Conversion Using AI," in *2024 International Conference on Communication, Control, and Intelligent Systems (CCIS)*, IEEE, 2024.
[6] "Enhancing Sign Language Recognition and Translation with Deep Learning: A CNN-Based Approach." IEEE Conference Publication.
[7] "AWAAJ – A Sign Language Translator and Learning Application." IEEE Conference Publication.
[8] "Real-Time Sign Language Recognition and Translation Using MediaPipe and Random Forests for Inclusive Communication." EEE Conference Publication.
[9] "BeyondWords: A Sign Language Translator." IEEE Conference Publication.
[10] "Real-Time Sign Language Interpretation and Translation to Speech Using CUDA and Machine Learning." IEEE Conference Publication.
[11] "Sign Language Translation Across Multiple Languages." IEEE Conference Publication.