© Volume 2, Issue 10, October 2026 | JATIR
e-ISSN: 2839-5786, p-ISSN: 3139-8464

Real Time Sign Language Recognition and Translation Using Mediapipe Ann and Lstm

[1]Shivani shitole, [2]Sushant Bhosale, [3]Akshata Gone, [4]Laxmi Gujar
[5]Omkar Ghanawat
[1]*Assistant Professor. Information Technology Parvatibai Genba Moze College, Pune, India*
[2,3,4,5]*Student, Information Technology Parvatibai Genba Moze College, Pune, India*

***Abstract*—This paper presents a real time sign language translation system that converts hand gestures into text and speech. The system uses a webcam, OpenCV, and MediaPipe for hand landmark detection. An Artificial Neural Network recognizes static gestures, while a Long Short Term Memory network recognizes dynamic gestures. Recognized signs are combined into sentences and converted into speech using text to speech. The proposed system provides a low cost and contactless approach for assisting communication between sign language users and non-signers.**

***Index Terms*—Sign language, gesture recognition, MediaPipe, artificial neural network, long short-term memory, computer vision, text to speech.**

I. INTRODUCTION

I. Sign language is an important communication method for deaf and hard of hearing individuals. However, communication becomes difficult when people who do not understand sign language interact with sign language users. This limitation can affect communication in educational institutions, workplaces, hospitals, public services, and other everyday situations.
II. Recent developments in computer vision and artificial intelligence have made it possible to recognize hand gestures using ordinary cameras. A vision-based sign language recognition system can reduce the dependency on specialized wearable devices and provide a convenient method for communication assistance.
III. The proposed Sign Language Translator is a real time system that recognizes hand gestures and converts them into readable text and audible speech. A webcam captures the user's hand gestures, while OpenCV is used for video capture and processing. MediaPipe is used to detect hands and extract their landmark coordinates.
IV. The proposed system uses two different neural network models according to the type of gesture. An Artificial Neural Network (ANN) is used for recognizing static gestures such as alphabets and digits. A Long Short-Term Memory (LSTM) network is used for recognizing dynamic gestures where the meaning depends on hand movement over a sequence of frames.
V. The recognized predictions are passed through confidence checking and prediction stabilization before being added to a sentence buffer. Instead of converting every individual

gesture immediately into speech, the system constructs a sentence and sends the completed text to a text to speech module.

The main objective of the system is to provide an accessible, contactless, and relatively low-cost communication assistance solution using commonly available hardware such as a webcam and computer.

II. PROBLEM STATEMENT

Sign language provides an effective communication method for deaf and hard of hearing individuals, but most people do not understand sign language. This creates communication barriers in educational institutions, workplaces, hospitals, public offices, and social environments.

Existing systems may focus mainly on static gestures or require specialized sensors and wearable devices. Systems based on complete image processing can also require significant computational resources.

Therefore, there is a need for a real time system that can recognize both static and dynamic hand gestures using an ordinary webcam and translate them into text and speech.

III. MOTIVATION

The motivation behind the proposed system is to reduce communication barriers between sign language users and people who do not understand sign language.

Human interpreters can provide effective communication, but they may not always be immediately available. A computer vision-based system can provide an additional communication interface without requiring gloves, sensors, or other wearable equipment.

MediaPipe provides efficient hand landmark detection, while machine learning models can classify the extracted landmark information. Combining ANN and LSTM models provides an approach for handling both static hand configurations and dynamic hand movements.

The project therefore aims to develop an accessible and practical system that can perform sign recognition in real time using commonly available computing hardware.

IV. AIM AND OBJECTIVES

Aim: - To develop a real time sign language recognition and translation system that recognizes static and dynamic hand gestures and converts them into text and speech using MediaPipe, ANN, and LSTM.

Objective: -

1. To capture hand gestures in real time using a webcam and OpenCV.
2. To detect hands and extract landmark coordinates using MediaPipe.
3. To prepare and preprocess datasets for static and dynamic gestures.
4. To develop an ANN model for recognizing static gestures such as alphabets and digits.
5. To develop an LSTM model for recognizing dynamic gestures using sequences of hand movements.
6. To convert recognized gestures into readable text.
7. To construct sentences using a prediction buffer and stability mechanism.
8. To convert the generated text into audible speech using a text to speech module.

9. To develop a low-cost system that does not require wearable sensors.

V. SCOPE OF THE PROJECT

The scope of the proposed Sign Language Translator is to develop a real time computer vision based system capable of recognizing a predefined vocabulary of sign language gestures.
The system will cover:

1. Video Capture – Capturing live hand gestures through a webcam.
2. Hand Detection – Detecting hands using MediaPipe.
3. Landmark Extraction – Extracting numerical hand landmark coordinates.
4. Static Gesture Recognition – Recognizing alphabets and digits using ANN.
5. Dynamic Gesture Recognition – Recognizing movement-based gestures using LSTM.
6. Text Generation – Converting recognized gestures into characters and words.
7. Sentence Construction – Combining recognized predictions into a sentence.
8. Speech Generation – Converting the completed sentence into audible speech.
9. Prediction Validation – Applying confidence thresholds and stability checks to reduce incorrect or repeated predictions.

The system will initially support a predefined set of gestures. Expansion of the vocabulary can be performed by collecting additional training data and retraining the models.

VI. EXISTING SYSTEM AND ITS LIMITATION

Existing System
Existing sign language recognition systems use different techniques including sensor-based gloves, image-based recognition, machine learning, and deep learning.
Sensor based systems use flex sensors, accelerometers, gyroscopes, or similar devices attached to the hand. Image based systems use cameras to capture hand gestures and apply computer vision or deep learning techniques.
Some systems focus mainly on static alphabets and numbers, while others attempt to recognize dynamic signs using video sequences.

Limitations

- Sensor based systems require additional wearable hardware.
- Wearable devices may be expensive or inconvenient for regular communication.
- Many systems focus mainly on static gestures.
- Dynamic gestures require analysis of movement over multiple frames.
- Recognition performance can be affected by lighting conditions.
- Hand orientation and occlusion can reduce recognition accuracy.
- Similar looking gestures may be incorrectly classified.
- Complete image based deep learning models can require greater computational resources.
- Many systems do not provide sentence level speech generation.

VII. PROPOSED SYSTEM

The proposed Sign Language Translator is a real time vision-based system that recognizes static and dynamic hand gestures and translates them into text and speech.

The system uses a standard webcam as the input device. OpenCV captures and processes the video frames, while MediaPipe detects the user's hand and extracts landmark coordinates.
For static gestures, landmark features from individual frames are provided to an Artificial Neural Network. The ANN classifies gestures such as alphabets and numerical digits.
For dynamic gestures, landmark information is collected over multiple consecutive frames. These sequences are provided to a Long Short Term Memory network, which learns temporal patterns associated with movement based gestures.
The outputs of both models are passed through a prediction processing stage. Confidence thresholding and prediction stability are used to reduce uncertain predictions. Duplicate prevention is also applied so that a gesture is not repeatedly added to the sentence.
The validated predictions are stored in a sentence buffer. Once an end of sentence condition is detected, the generated text is passed to the text to speech module. The resulting speech provides an audible representation of the recognized sign language.

VIII. PROPOSED METHODOLOGY

The proposed methodology consists of the following stages:
Data Collection – Static and dynamic sign language samples are collected using a webcam.

Hand Landmark Detection – MediaPipe detects the hand and extracts landmark coordinates from each video frame.

Data Preprocessing – Landmark coordinates are normalized and organized into suitable feature representations.

Static Gesture Training – Individual frame landmark features are used to train an ANN model for static gestures.

Dynamic Gesture Training – Consecutive landmark frames are grouped into sequences and used to train an LSTM model.
Real Time Detection – During execution, webcam frames are continuously processed to detect hand landmarks.

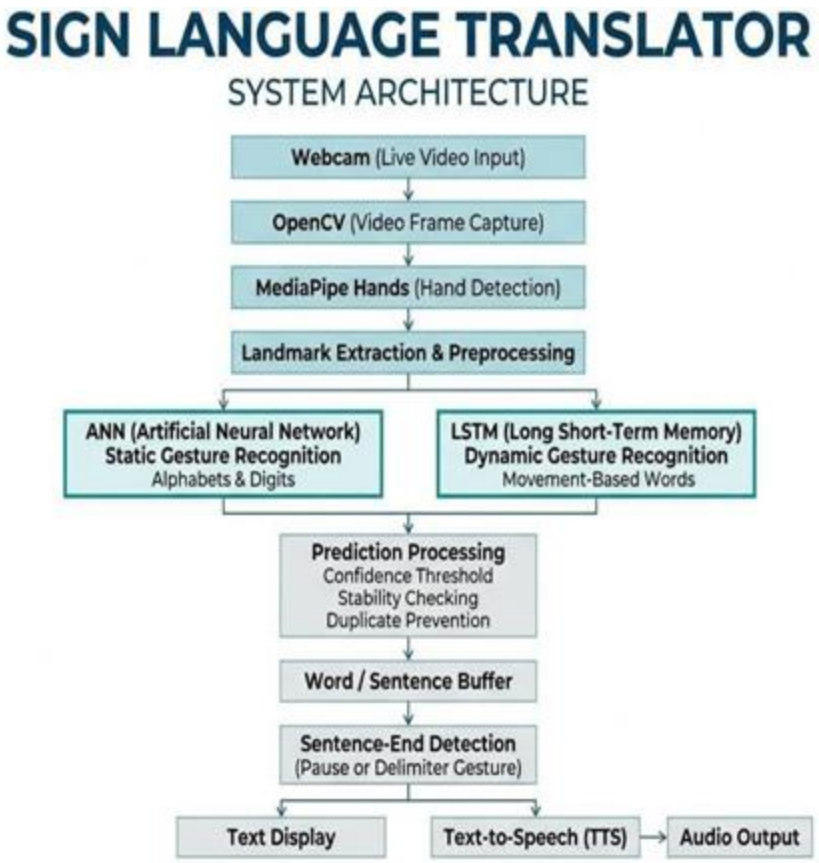
Gesture Classification – The appropriate ANN or LSTM model predicts the corresponding gesture.

Prediction Validation – Confidence thresholds and temporal stability checks are applied to improve prediction reliability.

Sentence Construction – Valid predictions are added to a sentence buffer while duplicate predictions are controlled.

Text to Speech – The completed sentence is converted into audible speech using a text to speech module.

IX. SYSTEM ARCHITECTURE



X. MODULE DESCRIPTION

Module	Responsibility
Video Capture Module	Captures real time video frames from the webcam using OpenCV.
Hand Detection and Landmark Extraction	Detects hands using MediaPipe and extracts hand landmark coordinates.
Static Gesture Recognition	Uses ANN to classify static gestures such as alphabets and digits.
Dynamic Gesture Recognition	Uses LSTM to recognize gestures based on hand movement sequences.
Prediction Processing	Applies confidence thresholding, stability checking, and duplicate prevention.
Sentence Construction	Combines recognized characters and words into a sentence buffer.

XI. HARDWARE AND SOFTWARE

Hardware

- Processor: Intel Core i5 / AMD Ryzen 5 or equivalent.
- RAM: Minimum 8 GB; 16 GB recommended for model training.
- Storage: Minimum 10 GB free space.

- Webcam: Built-in or external HD webcam with minimum 720p resolution.
- Display: Standard laptop or desktop display.

Software
- Operating System: Windows 10/11 or Linux.
- Programming Language: Python 3.10 or compatible version.
- Development Environment: Visual Studio Code / Jupyter Notebook / Google Colab.
- OpenCV: Real time video capture and image processing.
- MediaPipe: Hand detection and landmark extraction.
- TensorFlow/Keras: ANN and LSTM model development.
- NumPy: Numerical and array operations.

Pandas: Dataset organization and preprocessing.
- Scikit-learn: Data preprocessing and model evaluation.
- Matplotlib/Seaborn: Result visualization.
- Text to Speech Library: pyttsx3 or another suitable TTS library

XII. DATABASE/UML DESIGN

The proposed system primarily uses machine learning datasets rather than a traditional relational database.
The static gesture dataset contains landmark features extracted from individual frames along with corresponding gesture labels.
The dynamic gesture dataset contains sequences of landmark coordinates collected across multiple consecutive frames along with their corresponding gesture labels.
The major data structures include:
- Static gesture samples
- Dynamic gesture sequences
- Gesture labels
- ANN model
- LSTM model
- Prediction history
- Sentence buffer
- The system can be represented using UML diagrams such as:
- Use Case Diagram
- Activity Diagram
- Sequence Diagram
- Class Diagram
- Data Flow Diagram

XIII. INNOVATION/NOVELTY

The proposed system provides the following features:
- Hybrid ANN and LSTM architecture for static and dynamic gesture recognition.
- MediaPipe based landmark extraction instead of processing complete images directly.
- Real time gesture recognition using a standard webcam.

- Recognition of both static and movement-based gestures.
- Sentence level text generation instead of isolated gesture output.
- Confidence checking and prediction stabilization.
- Duplicate prevention for continuous gesture recognition.
- Text to speech conversion for audible communication.
- Low cost and contactless operation without wearable sensors.

The main novelty of the project is the integration of static and dynamic gesture recognition into a single real time translation pipeline using landmark based features.

XIV. EXPECTED OUTCOMES

The expected outcomes of the proposed system are:

- A functional real time sign language recognition system.
- Recognition of predefined static gestures such as alphabets and digits.
- Recognition of selected dynamic gestures using LSTM.
- Real time conversion of recognized gestures into text.
- Sentence construction using recognized signs.
- Conversion of completed sentences into speech.
- Reduced dependency on specialized wearable hardware.
- Efficient landmark based processing suitable for ordinary computer systems.
- The final recognition accuracy, precision, recall, F1 score, and response latency will be measured experimentally after model training and testing.

XV. APPLICATIONS

The proposed Sign Language Translator can be used in:

- Educational Institutions – To assist communication between sign language users, teachers, and students.
- Hospitals and Clinics – To support basic communication between patients and medical staff.
- Government Offices – To assist communication at information and service counters.
- Banks and Customer Service Centres – To provide an additional communication interface.
- Railway Stations and Airports – To assist communication at enquiry and assistance counters.
- Public Places – To support basic communication between sign language users and non signers.
- Research and Education – To demonstrate applications of computer vision and artificial intelligence.

XVI. ADVANTAGES AND LIMITATION

- Advantages
- Provides real time sign recognition.
- Supports both static and dynamic gestures.
- Uses a standard webcam instead of wearable sensors.
- MediaPipe landmark features reduce the amount of data processed by the recognition models.
- ANN is suitable for static gesture classification.
- LSTM can model temporal movement patterns.
- Provides both text and speech output.

- Sentence buffering improves the usability of speech generation.
- Relatively low cost and contactless system.

Limitations

- The system can recognize only gestures included in its trained vocabulary.
- Recognition accuracy can be affected by lighting conditions.
- Camera quality and hand distance can affect landmark detection.
- Hand occlusion can result in incorrect predictions.
- Similar looking gestures may be misclassified.
- Dynamic gestures may be affected by differences in movement speed and signing style.

The system may not represent all linguistic aspects of natural sign language.
Additional training data is required to expand the vocabulary.

XVII. FUTURE SCOPE

The proposed system can be enhanced in several ways:

- Expand the vocabulary of supported signs and words.
- Add facial expression and body-pose recognition.
- Improve dynamic gesture recognition using advanced models such as Transformers.
- Support continuous sign language sentence recognition.
- Develop a mobile application for real-time translation.
- Support multiple regional and international sign languages.
- Improve accuracy under different lighting and background conditions.

XVIII. LITERATURE SURVEY

- Sign language recognition has been studied using computer vision, machine learning, deep learning, wearable sensors, and skeletal landmark representations.
- MediaPipe provides an efficient framework for real time perception and hand landmark tracking. MediaPipe Hands is particularly useful for this project because it provides hand landmark coordinates that can be used as numerical input features for machine learning models.
- Artificial Neural Networks have been widely used for classification tasks involving numerical feature vectors. In the proposed system, ANN is used to classify static hand configurations obtained from individual video frames.
- However, static classification models do not adequately represent gestures where the meaning depends on movement over time. Long Short Term Memory networks are designed to process sequential information and can therefore be used to recognize dynamic hand gestures from consecutive landmark frames.
- Previous sign language recognition approaches also show that recognition performance can be affected by lighting, hand orientation, occlusion, gesture variation, and dataset size. The proposed system addresses some of these challenges through landmark based processing, prediction stabilization, and a hybrid ANN-LSTM architecture.
- The research gap identified is the need for an accessible real time system that combines static and dynamic gesture recognition with sentence level text and speech output while operating with commonly available hardware.

REFERENCE

[1]. C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. Yong, J. Lee, W. Chang, W. Hua, M. Grundmann, and J. A. Rehg, "MediaPipe: A framework for building perception pipelines," arXiv preprint arXiv:1906.08172, 2019.
[2]. F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C. L. Chang, and M. Grundmann, "MediaPipe Hands: On-device real-time hand tracking," arXiv preprint arXiv:2006.10214, 2020.
[3]. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[4]. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A system for large-scale machine learning," in Proc. 12th USENIX Conf. Operating Systems Design and Implementation, 2016, pp. 265–283.
[5]. G. Bradski, "The OpenCV library," Dr. Dobb's Journal of Software Tools, 2000.
[6]. F. Chollet, "Keras," 2015. [Online].