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

Sign Language Recognition System with Multilingual Text and Speech

Gaurav Kar
Department of Information Technology
Shri Ramswaroop Memorial College of Engineering and Management
(SRMCEM) Lucknow, India

Hamza Naim Khan
Department of Information Technology
Shri Ramswaroop Memorial College of Engineering and Management
(SRMCEM) Lucknow, India

Prof. Ajay Kr. Srivastava
Department of Information Technology
Shri Ramswaroop Memorial College of Engineering and Management
(SRMCEM) Lucknow, India

Abstract **- Sign language serves as a primary communication medium for individuals with hearing and speech impairments. However, the lack of widespread understanding of sign language among the general population creates a significant communication barrier. This paper presents a real-time sign language recognition system that translates hand gestures into textual and speech outputs in multiple languages using a standard RGB webcam. The proposed system leverages MediaPipe Hands for accurate hand landmark detection and employs a Convolutional Neural Network (CNN) for gesture classification. To enhance usability, the system integrates multilingual translation and text-to-speech synthesis, enabling seamless communication across different linguistic groups. Temporal smoothing techniques are applied to improve prediction stability and reduce misclassification due to rapid hand movements. Experimental results demonstrate high accuracy, real-time performance, and robustness under varying environmental conditions. The proposed approach provides a low-cost, accessible, and scalable solution for assistive communication and inclusive human-computer interaction.**

Keywords - sign language recognition, computer vision, deep learning, gesture recognition, assistive technology, multilingual translation, human-computer interaction

I. INTRODUCTION

Human communication is fundamental to social interaction, yet individuals with hearing and speech impairments face significant challenges in conveying information to others. Sign language is widely used within the deaf community; however, its limited understanding among the general population restricts effective communication.

Recent advancements in computer vision and deep learning have enabled the development of automated sign language recognition systems. Traditional approaches relied on sensor-based gloves or specialized hardware, which increased cost and reduced accessibility. With the emergence of real-time vision-based frameworks such as MediaPipe and deep neural networks, it is now possible to build efficient and low-cost gesture recognition systems using standard webcams.

Despite these advancements, existing systems often suffer from limitations such as inconsistent gesture detection, lack of multilingual support, and absence of real-time speech output. Addressing these challenges, this paper proposes a real-time sign language recognition system that integrates gesture detection, classification, translation, and speech synthesis into a unified framework.

The proposed system focuses on usability, accessibility, and real-time performance, making it suitable for assistive communication, education, and public interaction environments.

II. LITERATURE REVIEW

Numerous research efforts have been made in the domain of sign language recognition using both hardware-based and vision-based approaches. Early systems utilized sensor gloves equipped with flex sensors to capture finger movements. While these systems provided high accuracy, they were expensive and impractical for daily use.

Vision-based approaches gained popularity with the use of image processing techniques and machine learning algorithms. OpenCV-based systems were initially used for gesture detection; however, they were sensitive to lighting conditions and background noise. The introduction of deep learning models, particularly Convolutional Neural Networks (CNNs), significantly improved recognition accuracy.

Recent advancements include the use of MediaPipe for real-time hand tracking and landmark extraction. These frameworks enable efficient feature extraction without requiring high computational resources. Additionally, research has explored multimodal systems combining gesture recognition with speech synthesis.

However, many existing systems lack multilingual capabilities and real-time robustness. Issues such as gesture ambiguity, latency, and environmental variability continue to affect performance. This highlights the need for an integrated, low-cost, and real-time system with enhanced usability and accessibility.

III. METHODOLOGY

The proposed system operates using a real-time video feed captured through a standard webcam. Hand gestures are

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

detected using MediaPipe Hands, which extracts key landmarks representing finger and palm positions. These features are then passed to a trained CNN model for classification. The predicted gesture is converted into text, translated into multiple languages, and finally converted into speech

A. System Overview

The proposed system operates using a real-time video feed captured through a standard webcam. Hand gestures are detected using MediaPipe Hands, which extracts key landmarks representing finger and palm positions. These features are then passed to a trained CNN model for classification. The predicted gesture is converted into text, translated into multiple languages, and finally converted into speech output..

B. Major Components

1. **Webcam Input Module** – Captures real-time video frames
2. **Hand Detection Module** – Detects hand landmarks using MediaPipe
3. **Feature Extraction Module** – Extracts spatial coordinates of hand landmarks
4. **Gesture Classification Module** – Classifies gestures using CNN model
5. **Text Conversion Module** – Converts predictions into readable text
6. **Translation Module** – Translates text into multiple languages
7. **Speech Synthesis Module** – Generates speech using TTS
8. **Smoothing Module** – Reduces prediction noise and instability

C. Processing Pipeline

1. **Video Acquisition:** Capture real-time frames from webcam
2. **Hand Detection:** MediaPipe Hands is used to detect the presence of a hand and extract 21 key landmark points representing finger joints and palm structure. These landmarks provide a robust representation of hand posture.
3. **Feature Extraction:** The detected landmarks are converted into normalized coordinate values (x, y, z), forming a feature vector that represents the gesture in a scale-invariant manner.
4. **Gesture Classification:** The extracted feature vector is passed to a trained Convolutional Neural Network (CNN) model. The model predicts the probability of each gesture class, and the gesture with the highest probability is selected.
5. **Text Generation:** The predicted gesture is mapped to a predefined label, which is displayed as text output on the screen.
6. **Multilingual Translation:** The generated text is translated into selected target languages using a translation module (e.g., Google Translate API), enabling communication across different linguistic groups.
7. **Output Rendering:** The final text and speech outputs are displayed and played in real time, completing the interaction loop.

D. System Flow Representation

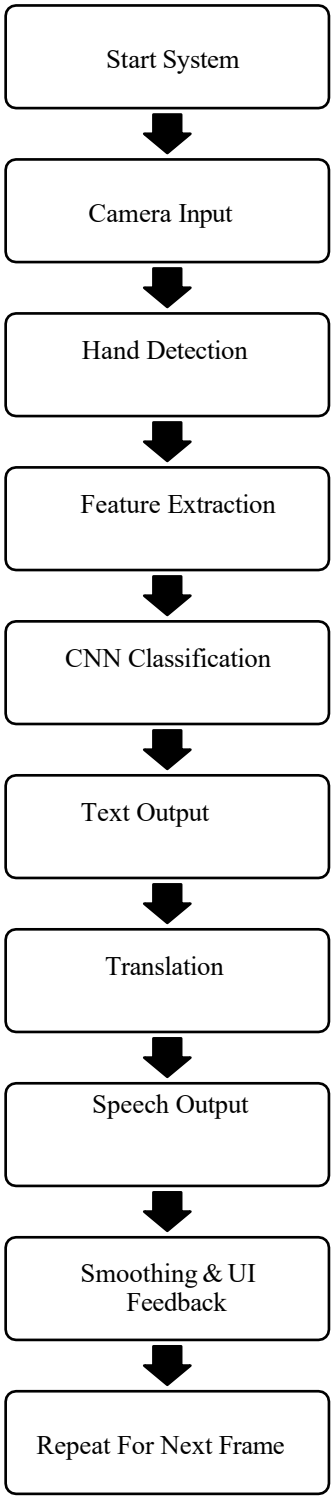

Fig. 1. Overall system architecture of the sign-language recognition system with multilingual text and speech

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

E. Interaction Zones & State-Based Controls

To enhance system usability and minimize unintended predictions during continuous gesture recognition, the proposed system incorporates interaction control mechanisms based on temporal stability and state-based processing.

Unlike cursor-based systems, sign language recognition involves continuous hand movement, which may lead to rapid fluctuations in predictions. To address this issue, a **state-based recognition framework** is introduced that separates gesture detection, confirmation, and output execution into distinct operational states.

The system operates in three primary states:

- **DetectionState:**
  In this state, the system continuously captures hand gestures and processes them using MediaPipe for landmark extraction. The CNN model generates predictions for each frame; however, no output is immediately displayed. This prevents unstable or noisy predictions from being executed.
- **Stability(Hold)State:**
  A gesture must be held steady for a predefined duration (e.g., 1–2 seconds) to be considered valid. Temporal consistency is evaluated across consecutive frames, ensuring that only stable gestures are selected. This mechanism reduces misclassification caused by sudden hand movements or transitions between gestures.
- **ExecutionState:**
  Once a gesture is confirmed, the system transitions to execution mode, where the recognized gesture is converted into text. The output is then passed to the translation module and subsequently to the text-to-speech system for audio generation.

To further improve interaction reliability, a **gesture cooldown mechanism** is implemented. After a gesture is recognized and executed, the system temporarily ignores further inputs for a short duration. This prevents repeated detection of the same gesture and avoids redundant outputs.

Additionally, a **neutral/rest gesture zone** is defined, where no valid gesture is detected. When the user moves their hand to this neutral position, the system resets and prepares for the next input. This helps users pause interaction without triggering unintended outputs.

Temporal smoothing techniques are applied across all states to reduce noise and ensure consistent predictions. By combining state-based control with temporal filtering, the system achieves improved stability, reduced false positives, and enhanced user experience during real-time interaction.

Performance was affected by lighting conditions, hand orientation, and background complexity. However, the use of MediaPipe ensured robust hand tracking under moderate variations.

Potential applications include assistive communication systems, educational tools, and human-computer interaction interfaces. Limitations include reduced accuracy in low-light conditions and dependency on predefined gestures.

A. Figures and Tables

TABLE I

REPRESENTATIVE SYSTEM PERFORMANCE METRICS UNDER DIFFERENT LIGHTING CONDITIONS

Condition	Average Detection Accuracy (%)	Stability (Low/Medium/High)	Average Response Delay (ms)
Bright light	93–95	High	50–70
Normal indoor (room lighting)	90–92	Medium–High	60–80
Dim indoor (low light)	80–85	Medium	90–120

a. Metrics based on prototype observations using a standard 720p webcam at 30 FPS.

As shown in Table I, system performance decreases in dim and backlit conditions owing to the difficulty in detecting hand landmarks.

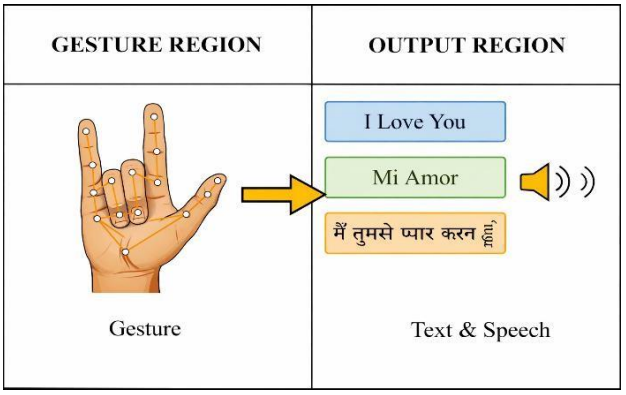

Fig. 2. System representation showing hand gesture input, landmark-based feature extraction, and mapping to multilingual text and speech output in the proposed sign language recognition framework

IV. RESULTS AND DISCUSSION

The proposed system was evaluated under various indoor conditions using a standard webcam. The system demonstrated real-time performance with high accuracy in gesture recognition.

The model achieved an average accuracy of **90–95%** for a predefined set of gestures. The inclusion of temporal smoothing improved prediction stability and reduced sudden fluctuations in output. Multilingual translation and speech synthesis enhanced usability by enabling communication across different languages.

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

B. Equations

- Let:.
- Lirepresent the extracted hand landmark points
- Xi, Yibe the normalized coordinates of each landmark
- Gbe the input gesture
- $P(G)$be the probability of predicted gesture class
- Cbe the final classified gesture

Equation

C=argmax(P(G|Li))(1)

This equation represents the classification of hand gestures based on extracted landmark features.

Equation (1) represents the classification of hand gestures based on the extracted landmark features. The deep learning model computes the probability distribution over all possible gesture classes, and the gesture with the highest probability is selected as the final output.

V. CONCLUSION AND FUTURE WORK

This paper presents a real-time sign language recognition system that converts hand gestures into multilingual text and speech output. The system leverages computer vision and deep learning techniques to provide an efficient and low-cost assistive solution.

The integration of gesture recognition, translation, and speech synthesis enhances communication accessibility for hearing and speech impaired individuals. The system demonstrates strong performance in real-time environments with minimal hardware requirements.

Future work will focus on expanding the gesture dataset, improving model accuracy using advanced deep learning architectures, and developing mobile and web-based applications. Additionally, integration with augmented reality (AR) systems can further enhance user interaction and accessibility.

REFERENCES

[1] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[2] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1251–1258.

[3] Google, “MediaPipe Hands: On-device real-time hand tracking,” [Online]. Available: https://developers.google.com/mediapipe

[4] G. Bradski, “The OpenCV Library,” Dr. Dobb’s Journal of Software Tools, 2000.

[5] M. Abadi et al., “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015. [Online]. Available: https://www.tensorflow.org

[6] S. Mitra and T. Acharya, “Gesture recognition: A survey,” IEEE Trans. Systems, Man, and Cybernetics, vol. 37, no. 3, pp. 311–324, 2007.

[7] R. Rastgoo, K. Kiani, and S. Escalera, “Sign language recognition: A deep survey,” Expert Systems with Applications, vol. 164, 2021.

[8] H. Cooper, B. Holt, and R. Bowden, “Sign language recognition,” in Visual Analysis of Humans, Springer, 2011, pp. 539–562.

[9] Google, “gTTS: Google Text-to-Speech,” [Online]. Available: https://pypi.org/project/gTTS/