Sign Language Translation Via Computer Vision Technique

S.Hemavarthini[1], S.Aarthi[2], P.Narmatha[3],Mrs.B.Sangeetha[4]

[1’2’3] Student, Department of CSE , Kings College of Engineering, Punalkulam, Near Thanjavur.

[4] Assistant Professor ,Department of CSE, Kings College of Engineering.

Abstract

Communication barriers between deaf and hard-of-hearing individuals and the general population remain a significant challenge, particularly in regions where Indian Sign Language (ISL) is not widely understood. This paper presents a real-time Indian Sign Language translation system based on computer vision and deep learning techniques. The proposed system captures live hand, face, and body movements using a standard webcam and processes these visual cues to recognize commonly used ISL gestures. MediaPipe Holistic is employed for keypoint extraction, and a Long Short-Term Memory (LSTM) network is used to classify dynamic gesture sequences. Recognized gestures are translated into readable text or audible speech to facilitate seamless interaction with hearing individuals. The system is designed to be lightweight and efficient, ensuring real-time performance on affordable consumer hardware while addressing challenges such as lighting variations, background noise, and gesture diversity. The proposed solution aims to promote accessibility, inclusivity, and independent communication in everyday environments.

Keywords

Indian Sign Language, Computer Vision, MediaPipe, LSTM, Gesture Recognition, Real-Time Translation, Assistive Technology

1. Introduction

Sign language serves as a primary mode of communication for millions of deaf and hard-of-hearing individuals worldwide. In India, Indian Sign Language (ISL) is widely used; however, a lack of awareness and trained interpreters creates communication barriers in education, healthcare, public services, and emergency situations. Most hearing individuals are unfamiliar with ISL, making everyday interactions difficult for the deaf community.

Existing sign language translation systems often suffer from limitations such as high computational requirements, lack of real-time performance, high cost, or poor accuracy under real-world conditions. Many solutions focus on offline processing or require specialized hardware, making them impractical for daily use.

This project addresses these challenges by proposing a real-time ISL translation system that captures live gestures using a webcam, processes them efficiently using computer vision and deep learning, and converts recognized gestures into text or speech. The system emphasizes low latency, affordability, and real-world usability, making it suitable for practical deployment

2. Literature Review

Several studies have explored sign language recognition using computer vision and deep learning techniques. Joshi et al. investigated deep learning approaches for semantic sign language translation using CNNs, RNNs, and LSTMs; however, their work primarily focuses on semantic representation and lacks real-time ISL implementation. Gupta and Singh utilized MediaPipe Holistic for multilingual sign language recognition, but their approach emphasized keypoint detection rather than comprehensive dynamic gesture classification.

Tekin and Yalcin proposed hand gesture recognition using MediaPipe and CNNs under varying conditions, though their dataset was limited and focused mainly on static gestures. Mukherjee and Kar introduced attention-based deep learning models for continuous sign language recognition, achieving high accuracy but at the cost of computational complexity unsuitable for low-resource systems.

Islam et al. demonstrated real-time American Sign Language translation using deep recurrent neural networks, but their approach did not address ISL-specific variations. Graph Convolution Networks have also been explored for pose-based sign recognition; however, their complexity limits real-time deployment on affordable hardware.

Overall, existing research highlights the effectiveness of deep learning for gesture recognition but reveals gaps in real-time, low-cost, ISL-focused systems. This work aims to bridge these gaps by combining MediaPipe-based keypoint extraction with LSTM-based temporal modeling for efficient real-time ISL translation.

3. Problem Statement

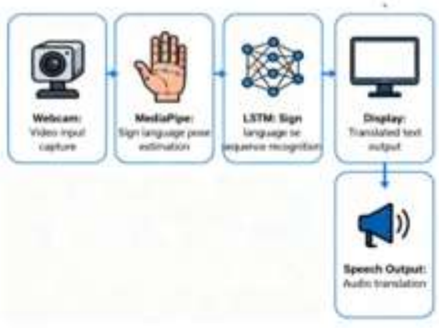
Although significant progress has been made in assistive technologies, real-time translation of Indian Sign Language (ISL) remains a challenging problem. Many existing sign language recognition systems are computationally expensive, require specialized hardware, or fail to operate reliably in real-world environments. Variations in lighting conditions, background clutter, camera quality, and individual gesture styles often degrade system performance. Additionally, several approaches focus on offline processing or static gestures, limiting their applicability for continuous, real-time communication.

Affordability and accessibility are also major concerns, as high-end hardware requirements restrict widespread adoption. Therefore, there is a strong need for an efficient, low-latency, and cost-effective ISL translation system that can accurately recognize dynamic gestures in real time and translate them into understandable text or speech for everyday communication scenarios.

4. Proposed System

The proposed system is a real-time Indian Sign Language translation framework that leverages computer vision and deep learning techniques to enable seamless communication between deaf or hard-of-hearing individuals and the hearing population. The system captures live video input through a standard webcam and processes continuous hand, face, and body movements to recognize ISL gestures.

A modular architecture is adopted to ensure scalability, flexibility, and ease of integration. The system extracts meaningful gesture features from video frames, classifies dynamic gestures using a sequence-based deep learning model, and converts recognized gestures into readable text or audible speech. The design emphasizes low computational overhead and minimal latency, making it suitable for deployment on affordable consumer-grade hardware.



5.System Architecture Modules

The proposed Indian Sign Language translation system is designed using a modular architecture to ensure real-time performance, scalability, and ease of integration. Each module performs a specific function and communicates efficiently with adjacent modules.

1. Data Acquisition Module

This module is responsible for capturing live video input of the user performing Indian Sign Language gestures.

- Uses a standard webcam to capture continuous video frames.
- Records hand movements, facial expressions, and body posture required for accurate gesture recognition.
- Supports real-time streaming to enable uninterrupted gesture input.
- Acts as the primary interface between the user and the system.

---

2. Preprocessing Module

The preprocessing module prepares raw video data for feature extraction and model inference.

- Converts captured video into sequential frames.
- Normalizes keypoint coordinates to ensure scale and position consistency.
- Removes noise caused by lighting variation, background clutter, and camera movement.
- Handles missing or occluded keypoints to maintain sequence continuity.
- Organizes frames into fixed-length temporal sequences suitable for LSTM input.

---

3. Feature Extraction Module

This module extracts meaningful gesture-related features from the preprocessed video frames.

- Utilizes **MediaPipe Holistic** to detect hand landmarks, facial landmarks, and body pose keypoints.
- Converts visual gesture information into numerical representations.
- Captures spatial relationships and motion patterns of gestures.
- Reduces dependency on raw pixel data, improving processing speed and accuracy.

---

4. Model Training and Classification Module

This module performs gesture learning and classification using deep learning techniques.

- Employs **Long Short-Term Memory (LSTM)** networks to model temporal dependencies in gesture sequences.
- Learns dynamic gesture patterns rather than static hand positions.
- Trained on labeled ISL gesture datasets for supervised learning.
- Continuously classifies incoming gesture sequences during real-time execution.
- Optimized for low-latency inference on consumer-grade hardware.

---

5. Prediction and Output Module

The prediction module translates recognized gestures into meaningful output for users.

- Converts classified gestures into readable text format.
- Integrates text-to-speech synthesis to generate audible output.
- Displays output on the frontend interface in real time.
- Ensures minimal delay between gesture recognition and output generation.
- Enables effective communication between sign language users and non-signers.

---

6. System Control and Integration Module*(Optional but reviewer-friendly)*

This module manages coordination between all system components.

- Handles data flow between frontend and backend services.
- Manages model loading, inference triggers, and system synchronization.
- Supports scalability and future feature integration.
- Ensures smooth and stable system execution.

6. Algorithms and Models Used

1. Frame Acquisition

Let the live video stream be:

$V=\{F1,F2,F3,...,Ft\}$

Where
FtF_tFt = frame captured at time t.

2. Preprocessing Function

Each frame is preprocessed using function P(·)

$Ft'=P(Ft)$

Where preprocessing includes:

$P(Ft)=Normalize(Resize(RGB(Ft)))$

3. Hand Landmark Extraction

Apply landmark detection function ϕ(·)\phi(\cdot)ϕ(·):

Lt=ϕ(Ft′)L_t = \phi(F'_t)Lt=ϕ(Ft′)

Where

$Lt=\{(x1,y1,z1),(x2,y2,z2),...,(x21,y21,z21)\}$

(21 hand keypoints)

4. Feature Vector Formation

Flatten landmark coordinates into feature vector:

$Xt=Flatten(Lt)$ $Xt\in R63$

(21 × 3 coordinates)

5.Temporal Sequence Construction

For dynamic signs, construct sequence of length N:

$Seq=\{Xt-N+1,...,Xt\}$

$Seq\in RN\times63$

6.Classification Model

Let trained model be f(·)f(\cdot)f(·)

$Y=f(Seq)$

Where:

$Y=\{y1,y2,...,yk\}$

k = number of sign classes

7. Softmax Probability

$P(Si)=\sum j=1keyjeyi$

8. Final Prediction

$S*=argmaxP(Si)$

9. Translation Mapping

$T=Map(S*)$

Where
TTT = Output translated text

Final System Equation (Compact Form)

$T=Map(argmaxSoftmax(f(\{\phi(P(Ft))\})))$

- **MediaPipe Holistic:** Extracts real-time hand, face, and body pose keypoints with high accuracy and low computational cost, suitable for real-time ISL recognition.
- **LSTM Network:** Classifies dynamic ISL gestures by learning temporal dependencies from sequential keypoint data.

- **Feature Normalization:** Scales input features to maintain consistency and improve model stability.
- **Noise Filtering:** Reduces variations caused by lighting and background disturbances.
- **Data Augmentation:** Enhances dataset diversity and improves generalization across users.
- **Lightweight Model Design:** Ensures low latency and smooth real-time performance on consumer-grade hardware.

7. Implementation Details

Hardware Requirements

- Processor: Intel® Core™ i9-14900K (3.20 GHz)
- RAM: 16 GB
- Storage: 1 TB

Software Requirements

- Frontend: HTML, CSS
- Backend: Python
- Framework: Flask

The backend, developed using Python and Flask, manages video frame processing, feature extraction, model inference, and output generation. RESTful communication enables smooth data flow between components. The frontend displays recognized text output and supports speech synthesis for audio feedback, providing an intuitive user experience.

8. Testing and Validation

The system is evaluated using multiple performance metrics to ensure reliability and real-time usability. **Accuracy** measures the percentage of correctly recognized gestures compared to labeled ground truth data. **Precision** evaluates the correctness of predicted gestures, reducing false positives.

Response time is measured as the delay between gesture input and output generation, ensuring minimal latency. **Throughput** assesses the number of gestures processed per second, confirming system efficiency during continuous use.

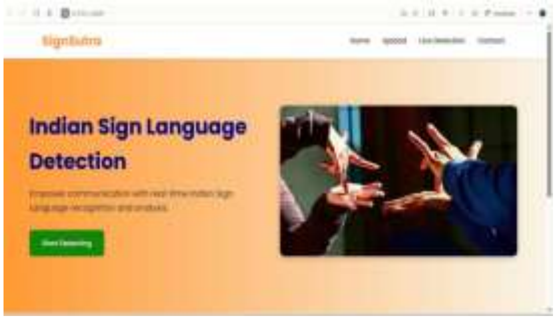
Testing is conducted under varying lighting conditions, backgrounds, and user styles to evaluate robustness. Performance analysis is used to fine-tune model parameters and validate system stability in real-world environments.

9. Results and Discussion

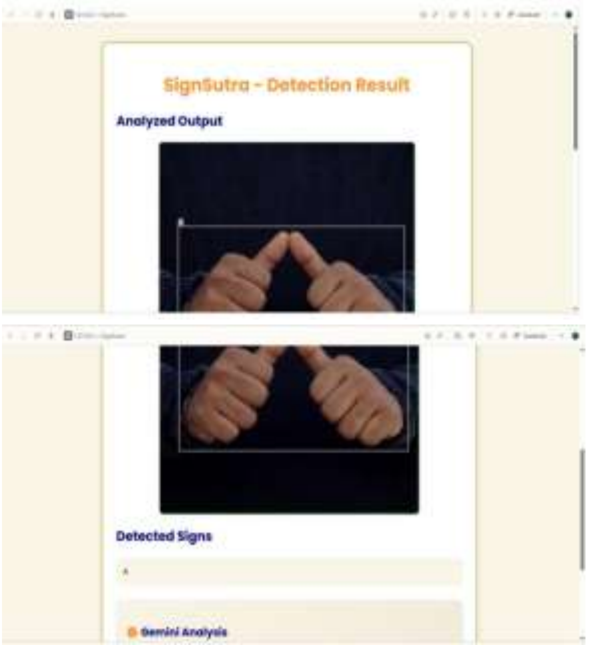
Experimental results demonstrate that the proposed system achieves reliable real-time ISL translation with low latency and satisfactory accuracy. The integration of MediaPipe Holistic with LSTM-based sequence modeling enables effective recognition of dynamic gestures while maintaining computational efficiency.

The system performs consistently across different environments and user variations, highlighting its practicality for real-world deployment. Compared to traditional approaches, the proposed solution offers improved responsiveness, affordability, and adaptability, making it suitable for applications in education, healthcare, and public communication.

10.IMPLEMENTATION



Indian Sign Language Detection



10. Future Work

Future enhancements include expanding the ISL gesture vocabulary to support more complex and continuous expressions. Multilingual support can be incorporated to translate gestures into multiple spoken languages. Personalization features may be introduced to adapt the system to individual gesture styles, improving accuracy.

Integration with mobile devices and cloud-based learning can enable continuous model updates and scalability. Additionally, advanced deep learning architectures such as Transformer-based models may be explored to improve recognition of overlapping or complex gesture.

11. Conclusion

This paper presents a real-time Indian Sign Language translation system using computer vision and deep learning. By capturing live gestures and translating them into text or speech, the system significantly reduces communication barriers for deaf and hard-of-hearing individuals. The lightweight and low-latency design ensures practicality on affordable hardware, promoting inclusivity and accessibility in everyday communication. The proposed solution demonstrates the potential of assistive technologies to empower the deaf community and foster seamless interaction with the hearing population.

References

[1] S. K. Dwivedi and A. S. Kushwaha, “Vision-based Indian Sign Language recognition using deep convolution neural network,” *IEEE Access*, vol. 8, pp. 131471–131479,2020.

[2] A. M. Kumar and P. S. Varma, “Real-time gesture recognition using MediaPipe and LSTM for sign language translation,” *Proc. Int. Conf. SmartTech*, pp. 211–216, 2021.

[3] R. R. Mukherjee and S. Kar, “Continuous sign language recognition using attention-based deep learning,” *IEEE Access*, vol. 9, pp. 117890–117901, 2021.

[4] F. Tekin and H. A. Yalcin, “Human hand gesture recognition using MediaPipe framework,” *Proc. Int. Conf. Signal Processing*, pp. 143–148, 2022.

[5] M. M. Islam et al., “Real-time American Sign Language translation using deep recurrent neural networks,” *IEEE Transactions on Multimedia*, vol. 23, no. 5, pp. 1573–1585, 2021.