A Multilingual Two-Way Sign Language Communication System

Amrita Reddy K[1], Subhi Agarwal[2], Akhter Zeeshan Siddiqua[3]

[1]Department of Artificial Intelligence and Data Science, Stanley College of Engineering and Technology for Women, India
Osmania University, Hyderabad, India
Email: *ishamrita[at]proton.me*

[2]Department of Artificial Intelligence and Data Science, Stanley College of Engineering and Technology for Women, India
Osmania University, Hyderabad, India
Email: *subhiagarwal14[at]gmail.com*

[3]Department of Artificial Intelligence and Data Science, Stanley College of Engineering and Technology for Women, India
Osmania University, Hyderabad, India
Email: *akhterzeeshaansiddiqua[at]gmail.com*

Abstract: ***In social settings such as workplaces, schools, or hospitals where various languages are spoken, interaction between people who are deaf and those who aren't familiar with sign language poses a real challenge. As a result, this communication barrier leads to exclusion and limited opportunities. Communication via sign language is the basic means for the deaf and hard-of-hearing; yet, these aren't widely understood and vary from one language or region to another, leading to ambiguities in areas that require collaboration, learning, and even service provision. To tackle this issue, our project aims to develop a real-time two-way multilingual sign language communication system that recognizes signing and translates it into text or spoken words in multiple languages while at the same time converting the text responses back into sign language. The system captures hand movements using a webcam and along with computer vision techniques using MediaPipe and gesture classification based on deep learning. The gestures that are detected are then translated to text and can also be output in speech in the user's selected language, which is done using natural language processing.***

Keywords: real-time, two-way multilingual sign language communication system, computer vision, MediaPipe, gesture classification, deep learning, natural language processing.

1. Introduction

Communication between hearing-impaired individuals and those unacquainted with sign language is a problem that has prevailed in social environments. Existing tools largely offer one-way communication with limited vocabulary, leaving individuals dependent on sign language interpreters. This project addresses this gap with the help of a two-way system, which is built for multilingual communication via sign language. Using a webcam to capture real-time hand gestures, the system processes them using MediaPipe for landmark detection and classifies it using gesture classification via deep learning. Once recognised, the output is provided as both text and speech, in the user's chosen language. The text responses can also be converted back to sign language via video clips or animation, allowing seamless communication.

2. Literature Review

This review tells us about the projects, theories, and technologies in sign language recognition and communication across multiple languages. But what's currently available is Hernandez et al., and also Gupta & Singh (2022), who experimented with a 3D CNN + Optical Flow for gesture recognition. Worked well inside but failed miserably outside. Lee & Park (2023) used encoders and decoders for sign language outputs. Kumar & Rao (2023) used MediaPipe and Bi -LSTM for Indian Sign Language, but only managed to get single words for translations. Sharma and group (2024) used CNN- LSTM with MediaPipe for signs in different languages. Patel and Wong (2024) used Vision Transformers for translation. Patel and Wong (2024) went all-in with Vision Transformers for cross-lingual translation. Rodriguez et al. (2025) went all-in with building a massive foundation model on millions of sign videos. This really made a difference as Chen and Kumar (2025), made the accuracy leaderboard with a Sparse Vision Transformer. However, there are still some issues that are present. Most systems need expensive hardware. Most systems struggle with new signers. And getting full sentences, especially in Indian Sign Language (ISL), and still translating smoothly has posed a challenge. That's where we come in and finally bridge these gaps.

3. Methodology

The idea behind this process is quitesimple. We wanted to make something that helps bridge the communication gap between someone that's hard-of- hearing and someone that's unfamiliar with sign language, talk to each other without needing any special device or someone to interpret for them. If a Deaf person wants to talk they can sign in front of the webcam, the camera captures their signs and uses a tool called MediaPipe to find 21 important points on each hand to understand the shape and position of the sign. These points are then processed by a computer program made to understand what the sign means. When the program is able to understand which sign it is, it displays the text that correlates with the sign and provides an audio output in the language of the other person's choice. For example, if a Deaf person signs "hello" the other person will see the word on the screen and be able to hear it being said out loud.

For basic statements, it works the other way around as well. If someone wants to display a video of a sign language gesture, they may type what they want to say into the system. The text is then translated if required and it shows a video of the sign being signed from the set of videos existing in the computer. The deaf person can see the signed video on thescreen, understand what the other person is saying.

4. Proposed System

The proposed system allows both parties to interact in their preferred language because interpreters normally use gestures and writing to communicate. The concept requires a Deaf person to sign in front of the camera after which, the system detects the gesture through its webcam and uses computer vision techniques to transform it into text or an audio output in the selected language. The system interprets text responses from users to produce sign language through pre-recorded videos. The system combines MediaPipe hand landmark detection with Gesture Recognition, Machine Learning to create an efficient and user-friendly solution that operates effectively in environments where people face communication difficulties. The project aims to establish a platform which enables future development of 3D avatars and expanded vocabulary and advanced deep-learning models to create more natural communication methods for people with disabilities.

5. Requirements

1) **Functional Requirements**

- Real-time Sign Recognition: The system requires webcam-based hand movement recognition which needs to operate with sign detection functionality.
- Hand Landmark Extraction: MediaPipe to pinpoint hand landmarks with great precision for the purpose of gesture recognition.
- Gesture Recognition: The machine learning model will classify observable gestures into specific word categories.
- Sign-to-Text and Sign-to-Speech Output: The system will display all recognized signs as text which users can choose in their preferred language.
- Text-to-Sign Translation: The user's text input should output the relevant sign language video clip.
- Multilingual Support: The system provides text and speech output in three languages.
- User Interface: A neat user-friendly interface which enables users to interact with the system,

2) **Non-Functional Requirements**

- Real-Time Performance: Quick processing so that the communication is natural and dialogue-like.
- Accuracy: Gesture Recognition must reach sufficient accuracy levels to prevent any potential user misunderstandings which could occur through the system's performance.
- Scalability: The system's architecture should be designed to enable capacity expansion for supporting additional signs, languages, and 3D avatar characters.
- Usability: The system needs to be user-friendly so that users who possess little to no technical knowledge can use it with ease.
- Portability: The system should be operational on common devices which include laptops and webcams without needing any expensive specialized hardware.
- Maintainability: The code and architecture should allow for easy updates or enhancements which support the addition of new features.

3) **Hardware Requirements**

- Processor: Intel Core i5/Apple M1 or above to ensure smooth execution of real-time gesture recognition and machine learning tasks.
- RAM: Minimum 8 GB RAM to support efficient multitasking and fast application performance.
- Storage: Minimum 256 GB HDD or SSD to store the application, datasets, trained models, and sign language videos.
- Webcam: An HD webcam is required to capture hand gestures accurately for real-time sign recognition.
- Microphone: A microphone is used to receive voice input for speech-based interactions.
- Speakers: Speakers are required to provide clear sign-to-speech audio output to the user.

4) **Software Requirements**

- Operating System: Windows, Linux, or macOS is required to run and develop the application.
- Programming Language: Python is used as the primary programming language due to its rich AI and computer vision libraries.
- MediaPipe: MediaPipe is used to detect and track hand landmarks with high accuracy.
- OpenCV: Used for image processing, webcam access, and real-time video analysis.
- TensorFlow/Scikit-learn: Frameworks used to develop, train, and deploy gesture recognition models.
- gTTS or pyttsx3: To convert recognized text into speech for effective user communication.
- Google Translate API: The API enables multilingual translation of recognized text and speech.
- IDE/Development Tools: Jupyter Notebook, Visual Studio Code (VS Code), is used for coding, debugging, and testing the application.
- Additional Python Libraries: Libraries such as NumPy and Pandas are used for data manipulation, numerical computation, and preprocessing

6. System Architecture

The sign language system captures a real-time video at the user's end. The backend has an inference engine which operates on CNN. It processes incoming data, decodes the signs, and turns them into text or other languages or vice versa, as per the users' needs. For data management, communication, and storage, the system has object storage, a messaging system, and a database as supporting services. At the end of the process, the users receive the output in video, text, or audio form. The Sign Language System is a scalable, end-to-end pipeline constructed to start from the client's device and passes through several backend services. The recording is done automatically when a video is captured by a camera, and the local processing provides the main hand landmarks which saves the bandwidth and ensures the user privacy at the same time. The processed features are then sent to a gateway where user authentication, rate limiting, and feature toggling are applied, thus providing access that is both secure and reliable. Machine Learning inference services are deciphering the sign language while carrying out translation and text-to-sign services simultaneously. In the end, the system provides multiple choices of outputs: video, text, or audio, thereby facilitating communication in sign language through different channels as per user requirements.

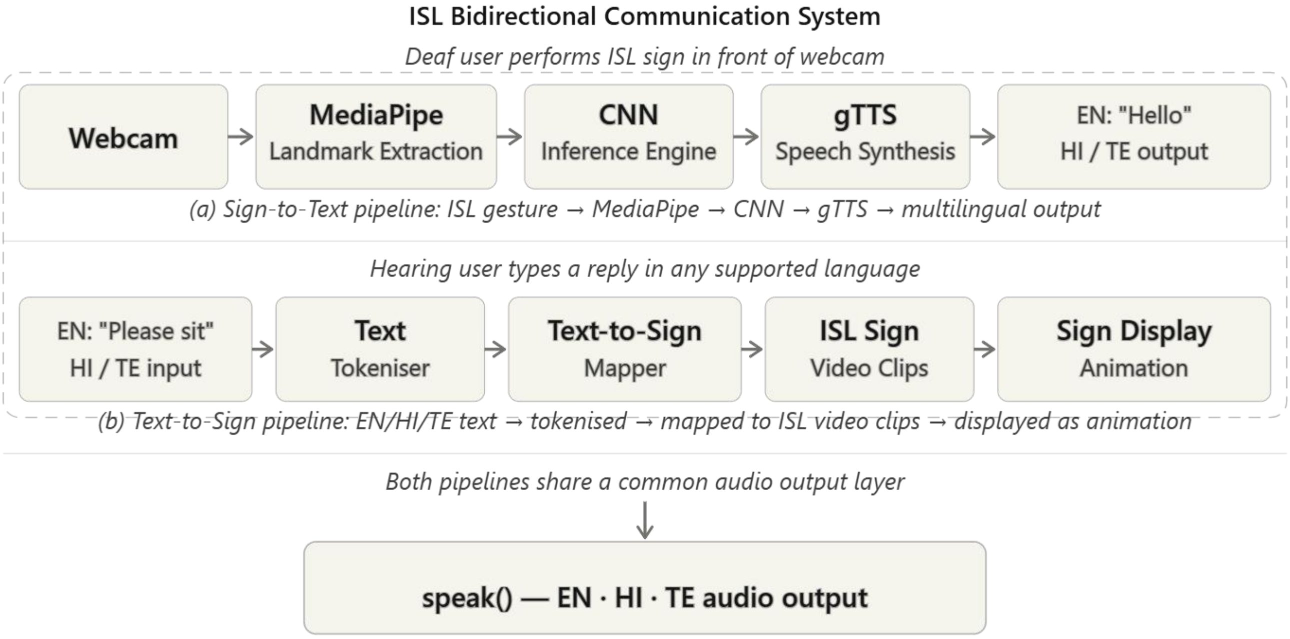

Figure 1: System Architecture of the Model

7. Modules

- Data Acquisition: This module is responsible for collecting input data required by the system. It captures real-time video streams or images using a camera. The acquired data serves as the raw input for further processing and analysis.
- Hand Detection: The Hand Detection module identifies and localizes human hands from the live video stream or images captured by the camera. This module outputs the position and region of the hand, enabling accurate tracking for further gesture analysis.
- Gesture Recognition: This module interprets the captured hand movements to recognize specific sign language gestures. It extracts features such as hand landmarks, finger orientation, and motion patterns, which are then classified using machine learning.
- Speech Conversion: The speech conversion module converts spoken language into text using Automatic Speech Recognition (ASR).
- Translation (Sign-to-Text & Text-to-Sign): This module acts as the core translation engine. It translates recognized gestures into meaningful text or displays sign language representations for corresponding input text (uses pre-recorded sign videos).
- User Interface: The user interface provides an interactive platform for users to communicate with the system. It displays video input, recognized signs, translated text, and generated sign outputs. The UI is designed to be user-friendly, accessible, and responsive for users.

8. Discussion

The experimental evaluation of the proposed Multilingual Two-Way Sign Language Communication System demonstrates that landmark-based gesture recognition combined with multilingual translation can effectively support real-time bi-directional communication. The system maintained consistent performance throughout standard lighting conditions by achieving frame rates between 25 and 30 fps and detecting pre-defined static gestures with accuracy between 85 percent and 9 percent. The results show that MediaPipe-based hand landmark extraction reaches economical processing power while delivering enough spatial information for accurate classification results. The landmark-based feature representation decreases processing requirements, which allows for faster calculations compared to deep image-based convolutional methods that need to analyze complete frames. The system achieves standard hardware deployment because it does not need GPUs or specialized sensors, which enhances system accessibility and decreases operational expenses. The accuracy experiences a decline when the system encounters difficult situations that include inadequate lighting, movement blur, object blockage, and intricate background patterns. The system needs more powerful pre-processing methods and model designs to achieve greater stability against these restrictions. The system's multilingual translation feature enables its users to operate successfully in different language-based environments. The system provides more than basic sign-to-text functionality by converting detected gestures into various languages which enables inclusive communication across educational institutions, healthcare facilities, and professional workplaces. The system sets itself apart from standard sign recognition systems through its ability to establish two-way communication through its text-to-sign feature. The current system shows positive outcomes but it needs to enhance its capabilities for handling static gestures which require users to stay in one position. The field of continuous sign recognition at the sentence level encounters major difficulties because of two factors which include the need for temporal dependencies and the need for contextual information. The proposed architecture achieves an optimal balance between three key factors which are accuracy and computational efficiency and usability according to the discussion. The combination of Computer Vision and Machine Learning with multi-lingual processing creates a research base which allows further development of assistive communication technologies.

9. Conclusion

We have developed a Multilingual Two-Way Sign Language Communication System which enables hearing-impaired individuals to communicate with people who do not know sign language. The system uses computer vision and machine learning and multilingual translation methods to achieve live two-way communication through its sign-to-text and text-to-sign translation system. The system uses a reverse pipeline which converts text into visual sign language representations, thus enabling users to communicate with each other in both directions. The experimental results showed that the system operated continuously in real-time while achieving frame rates between 25 and 30 FPS and detecting gestures with an accuracy that ranged from 85% to 90% under normal operational conditions. The system demonstration achieved successful frontend-backend operation through standard hardware use, which demonstrated its ability to function without specialized sensors or advanced computing equipment. The proposed framework provides a scalable and economical and immediate solution for multilingual sign language interpretation which improves assistive communication technologies. The system demonstration showed that standard hardware could support system deployment without needing special sensors or advanced computing systems. The ability to translate multiple languages makes the system more accessible to various environments. The results show that landmark-based gesture recognition together with language processing modules enable effective communication for hearing-impaired people by improving accessibility. This solution enables and encourages the creation of communication-free zones between people, thus assisting the involvement of communities. We believe the project not only has practical applications but also feeds into the research of the assistive technology domain and opens doors to even more advanced sign language solutions in the coming future.

References

[1] Y. Chen and R. Kumar, "Ultra-real-time multilingual sign translation using sparse vision transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 2, 2025.

[2] M. Rodriguez, A. Lopez, and S. Chen, "Universal sign language foundation model (SignFM)," in *Proc. 38th Conf. Neural Information Processing Systems (NeurIPS)*, 2025.

[3] Anonymous, "ISL continuous sentence translator using NVIDIA DeepStream," *Pattern Recognition Letters*, Elsevier, 2025.

[4] M. Arib, H. Benali, and K. Saeed, "SignFormer-GCN: Continuous sign language translation using spatio-temporal graph convolutional networks," *PLOS ONE*, vol. 20, no. 3, 2025.

[5] A. Singhal, R. Mehra, and P. Verma, "Real-time Indian sign language detection for translation," in *Proc. IEEE RAIT Conf.*, 2025.

[6] Anonymous, "Lightweight sign language recognition using improved YOLOv5s," *Computers*, MDPI, vol. 14, no. 1, 2025.

[7] S. Vani, "Multimodal Indian sign language recognition and translation," *International Research Journal of Computer Science*, 2025.

[8] R. Singh, A. Malhotra, and S. Gupta, "Real-time small object detection in video streams," *Image and Vision Computing*, vol. 138, 2024.

[9] P. Sharma, N. Joshi, and R. Iyer, "Multi-lingual sign language recognition system using machine learning," *Multimedia Tools and Applications*, vol. 83, no. 5, 2024.

[10] K. Patel and L. Wong, "Cross-lingual sign language translation using vision transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 9, 2024.

[11] J. Lee and S. Park, "Transformer-based continuous sign language translation," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, 2023.

[12] A. Kumar and S. Rao, "Indian sign language recognition using MediaPipe and deep learning," *Pattern Recognition Letters*, vol. 168, 2023.

[13] D. Hernandez, M. Xu, and L. Chen, "SignBERT: Pre-trained transformer for sign language understanding," *Neural Computing and Applications*, vol. 34, no. 18, 2022.

[14] R. Gupta and P. Singh, "Real-time American sign language dynamic gesture recognition," *Neural Computing and Applications*, vol. 34, no. 7, 2022.

[15] T. Nguyen and H. Tran, "End-to-end sign language to speech translation using deep learning," *Computer Speech & Language*, vol. 68, 2021.

[16] Y. Chen, L. Zhao, and Q. Wang, "Sign-to-text translation using pose graph networks," *IEEE Access*, vol. 9, 2021.

[17] A. Mehta and S. Das, "Robust Indian sign language recognition using capsule networks," *Expert Systems with Applications*, vol. 176, 2021.

[18] Anonymous, "Spatio-temporal graph convolutional networks for sign language recognition," *IEEE Transactions on Multimedia*, vol. 23, 2021.

[19] O. Koller and J. Cooper, "Real-time American sign language recognition using YOLOv3 and LSTM," *Pattern Recognition Letters*, vol. 147, 2021.

[20] R. Zhao, Y. Wang, and L. Chen, "Spatio-temporal graph convolutional networks for sign language recognition," *IEEE Transactions on Multimedia*, vol. 23, 2021.

[21] M. Rahman, S. Ahmed, and T. Islam, "Deep residual Bi-GRU for sign language translation," *Image and Vision Computing*, vol. 109, 2021.

[22] Y. Liu and J. Xiao, "Hybrid CNN–HMM model for continuous sign language recognition," *Neurocomputing*, vol. 396, 2020.

[23] K. Tanaka and S. Mori, "RGB–depth fusion for robust sign language recognition," *Sensors*, vol. 20, no. 14, 2020.

[24] M. Rahman, S. Ahmed, and T. Islam, "Deep residual networks for sign language video recognition," *Image and Vision Computing*, vol. 102, 2020.

[25] K. Tanaka and S. Mori, "Fusion-based sign language recognition using skeleton and RGB data," *Sensors*, vol. 20, no. 9, 2020.
[26] N. Carion et al., "End-to-end object detection with transformers," in *Proc. European Conf. Computer Vision (ECCV)*, 2020.
[27] W. Liu et al., "SSD: Single shot multibox detector," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, 2020.
[28] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, 2020.
[29] T.-Y. Lin et al., "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, 2020.
[30] K. He et al., "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, 2020.