Inclusive Communication: Leveraging AI for Sign Language Translation and Real-Time Audio Transcription

Shwethashree G C[1], Aditya T[2], Neha Dayanand[3], Rishin S[4], Adi Sharan A[5]

[1]*Shwethashree G C, Dept. of Computer Science and Engineering, JSS Science and Technology University,*
[2]*Aditya T, Dept. of Computer Science and Engineering, JSS Science and Technology University,*
[3]*Neha Dayanand, Dept. of Computer Science and Engineering, JSS Science and Technology University,*
[4]*Rishin S, Dept. of Computer Science and Engineering, JSS Science and Technology University,*
[5]*Adi Sharan A, Dept. of Computer Science and Engineering, JSS Science and Technology University*

Abstract - Humans communicate through both natural language and body language, including gestures, facial expressions, and lip movements. While understanding spoken language is essential, recognizing sign language is equally important, especially for individuals with hearing impairments. Deaf individuals often struggle to communicate with those unfamiliar with sign language, making real-time translation systems invaluable. This paper proposes a real-time meeting platform that recognizes Indian Sign Language (ISL) gestures and converts them into text and speech, enabling smooth interaction between deaf and hearing individuals. The system uses image processing, computer vision, and deep learning—specifically Long Short-Term Memory (LSTM) networks—to analyze hand gestures from a live video stream. LSTM models effectively capture temporal patterns in gesture sequences, enhancing recognition accuracy. The identified gestures are then translated into text and synthesized into speech. This system aims to bridge communication gaps and improve accessibility for the hearing impaired.

Key words: Indian Sign Language (ISL), Gesture Recognition, LSTM, Real-Time Translation, Accessibility, Communication, Deep Learning, Speech Synthesis, Computer Vision, Virtual Meetings.

1. INTRODUCTION

Communication is an essential aspect of human interaction, and while spoken language is the primary mode of communication, non-verbal methods such as gestures, facial expressions, and body movements also play a crucial role. Sign language is a structured visual language that relies on hand movements, facial expressions, and body posture to convey meaning, making it the primary mode of communication for individuals with hearing impairments. Unlike spoken languages, sign languages differ across regions and cultures, with Indian Sign Language (ISL) being distinct from American Sign Language (ASL) or British Sign Language (BSL).

In India, the lack of widespread awareness and formal education in ISL creates barriers for the deaf community, making communication with the hearing population challenging. Existing solutions, such as human interpreters or text-based communication, are not always readily available, leading to social and professional isolation for individuals with hearing disabilities. To address this challenge, we propose a real-time meeting platform that recognizes Indian Sign Language (ISL) gestures and converts them into both text and speech, enabling seamless interaction between deaf and hearing individuals.

Our system leverages image processing, computer vision, and deep learning techniques to accurately detect and interpret hand gestures from a real-time video feed. By capturing and analyzing the shape, position, and movement of hands, as well as facial expressions, the system translates ISL gestures into text and synthesized speech. This innovation facilitates seamless communication in both virtual and physical settings, eliminating the need for a human interpreter.

Apart from enhancing accessibility, this technology can be applied to educational settings, enabling individuals to learn and practice ISL, as well as to human-computer interaction scenarios, such as gesture-based control systems. By bridging the communication gap between the hearing and deaf communities, it promotes inclusivity, improves accessibility, and empowers individuals with hearing impairments in various personal, professional, and social contexts.

2. LITERATURE SURVEY

The development of AI-driven sign language translation and real-time audio transcription has seen significant progress in recent years. Several studies have explored different approaches to Indian Sign Language (ISL) recognition, focusing on gesture classification, real-time processing, and audio integration.

Abraham et al. [1] proposed a wearable glove-based system using LSTM networks to translate ISL signs into speech. While

achieving high accuracy (98%), the system's reliance on specialized hardware limits its scalability and practicality in real-world scenarios. Similarly, Poornima et al. [2] introduced a recognition method using global features and Hu moments for ISL alphanumeric gestures. Despite achieving a 98% accuracy rate, the model struggles with dynamic gestures, which are essential for conversational signing.

Goel et al. [3] utilized MediaPipe and LSTM models for real-time ISL translation, effectively mapping hand gestures to text and speech. However, challenges remain in processing sequential gestures efficiently, with latency affecting real-time usability. Thayiparampil et al. [4] developed a TensorFlow Object Detection-based system, demonstrating high accuracy for static ISL alphabets. However, the model's limited dataset (500 images) restricts its effectiveness for dynamic gesture recognition in diverse environments.

To improve dataset diversity, Poornima et al. [5] introduced the ISL2022 dataset, covering alphanumeric, word-level, and sentence-level gestures. While this dataset enhances model training, handling complex gestures remains a challenge.

These studies highlight the advancements in ISL recognition while underscoring existing challenges in real-time performance, dataset diversity, and scalability. Our research aims to bridge these gaps by developing a real-time AI-powered meeting platform that integrates deep learning for gesture recognition, NLP for text generation, and speech synthesis for improved accessibility and communication.

3. SYSTEM ARCHITECTURE

The proposed Indian Sign Language (ISL) recognition system follows a modular approach, comprising three essential phases: Pre-processing, Classification, and Speech Synthesis. These phases work together to ensure real-time ISL translation into text and speech.

3.1 Preprocessing Phase:

The preprocessing phase is crucial for ensuring high-quality gesture recognition. It involves:

1) Real-Time Video Capture:
   a) A standard webcam captures the user's hand movements continuously at a frame rate of 30 FPS (frames per second).
   b) The camera input is processed to identify the region of interest (i.e., the user's hand and fingers).
2) Image Enhancement Techniques:
   a) Background Subtraction: Removes unnecessary elements to focus on hand gestures.
   b) Grayscale Conversion: Converts the RGB image to grayscale, reducing computational complexity.
   c) Histogram Equalization: Enhances contrast for better visibility of hand shapes and key points.
3) Hand Tracking and Key-Point Detection:
   a) MediaPipe Hand Tracking: Uses 21 key points to detect hand landmarks accurately.

3.2 Classification Phase:

The classification phase involves using deep learning models to recognize hand gestures and convert them into meaningful text representations.

1) Feature Extraction:
   a) The detected 21 hand landmarks (from MediaPipe) are mapped into a structured format.
   b) Bounding boxes of hand movements provide spatial data for tracking gestures in real time.
2) Deep Learning Model:
   a) A Long Short-Term Memory (LSTM) network is used for gesture sequence analysis.
   b) The LSTM model is trained on Indian Sign Language datasets, enabling it to recognize dynamic hand gestures over time.
3) Text Mapping:
   a) Once a gesture is classified, it is mapped to predefined text representations stored in the system.
   b) The system supports word-level and sentence-level recognition for improved communication.

3.3 Speech Synthesis Phase:

The final phase converts the recognized text into audible speech using text-to-speech (TTS) conversion techniques.

1) Text-to-Speech (TTS) Conversion:
   a) Google Text-to-Speech (gTTS) is used for online speech synthesis with natural-sounding voices.
   b) pyttsx3 is used as an offline alternative, allowing speech conversion without internet dependency.
2) Real-Time Audio Feedback:
   a) The system generates and plays audio output corresponding to the recognized ISL gesture.
   b) Users can adjust voice pitch, speed, and volume for better accessibility.

4. METHODOLOGY

4.1 Data Preprocessing:

1) The dataset consists of 10,000+ ISL gesture images, including both static and dynamic signs.
2) Roboflow Annotation Tool was used to label the dataset, ensuring proper class distribution.

4.2 Preprocessing Techniques Applied:

1) Grayscale conversion to reduce computational load.

2) Contrast enhancement and noise reduction to improve image clarity.
3) Resizing and normalization to standardize input dimensions.
4) Data augmentation (rotation, flipping, brightness adjustments) to increase dataset diversity.

4.3 Model Training:

1) The model was trained using a Long Short-Term Memory (LSTM) neural network for gesture recognition.
2) The MediaPipe framework extracted 21 hand key points.
3) Training was performed on TensorFlow/PyTorch with the following hyperparameters:
   a) Batch Size: 16
   b) Learning Rate: 0.001
   c) Epochs: 50
   d) Optimizer: Adam
4) The model was validated on a test dataset, achieving a validation accuracy of 99.8%

4.4 Real-Time Gesture Recognition:

1) The system captures live video input using a standard webcam.
2) MediaPipe tracks hand key points, while LSTM classifies the gesture sequence.
3) The recognized gesture is mapped to a predefined text representation.

This Methodology ensures an efficient, real-time ISL recognition pipeline that can be deployed for various assistive applications.

5. DEVELOPMENT ENVIRONMENT

The ISL recognition system is built using advanced software tools and machine learning frameworks for efficient performance.

5.1 Programming Language & Frameworks:

1) Python: Used for backend processing, gesture classification, and speech synthesis.
2) React.js: Used to develop a user-friendly frontend for displaying recognized text.

5.2 Computer Vision Libraries:

1) OpenCV: Performs image processing, such as cropping, thresholding, and filtering.
2) MediaPipe: Detects and tracks hand key points for gesture classification.

5.3 Machine Learning & Deep Learning:

1) TensorFlow & PyTorch: Train and implement the LSTM model for gesture recognition.
2) NumPy & Pandas: Handle data processing, including gesture dataset transformations.

5.4 Backend & API Management:

3) FastAPI: Handles API calls for gesture classification, text-to-speech conversion, and real-time processing.
1) SQLite/PostgreSQL: Stores user interaction data for improving model accuracy over time.

This robust development environment ensures fast processing, real-time interaction, and scalable deployment across multiple platforms.

6. CHALLENGES AND LIMITATIONS

Despite the promising results, the system faces several challenges and limitations that affect its performance in real-world scenarios.

6.1 Dataset Challenges:

1) Unlabeled Data: Initial dataset required manual annotation, increasing preprocessing time.
2) Limited Variability: Some signs have fewer samples, leading to an imbalance in gesture classification.
3) Background Noise: Variations in lighting and complex backgrounds reduce recognition accuracy.

6.2 Model Performance Limitations:

1) Dynamic Gesture Recognition: The LSTM model struggles with long-sequence gestures due to temporal inconsistencies.
2) Real-Time Processing: Achieving low latency (below 50ms) remains a challenge for high-speed applications.
3) Gesture Overlap: Some gestures have similar hand shapes, leading to potential misclassification.

6.3 Environmental Factors:

1) Lighting Conditions: Low-light environments reduce hand visibility, affecting recognition accuracy.
2) Hand Variations: Differences in signer hand shape, size, and skin tone impact model generalization.
3) Occlusions: Partial hand visibility (e.g., due to clothing or objects) can lead to incorrect predictions.

6.4 Deployment & Scalability:

1) Hardware Limitations: Real-time processing on low-end devices remains computationally expensive.
2) Internet Dependency: The Google TTS model requires an internet connection for speech synthesis.

3) Sign Language Diversity: The model currently supports a limited vocabulary, requiring further dataset expansion.

7. EXPERIMENTAL RESULTS

7.1 Performance Evaluation:

The ISL recognition system was tested in various real-world scenarios to evaluate its accuracy, speed, and reliability.

1) Key Performance Metrics

a) Validation Accuracy: 99.8%

b) Processing Speed: 30 FPS

Fig -1: Training vs Validation Accuracy

Fig -2: Training vs Validation Loss

7.2 Real-World Testing Scenarios:

1) Controlled Environment: Achieved high accuracy in well-lit conditions.

2) Low-Light Conditions: Accuracy slightly dropped to 88% due to limited hand visibility.

3) Multiple Signers: The system successfully identified gestures from different users with a 90% recognition rate.

4) Virtual Meeting Integration: The system was successfully integrated with Zoom, Google Meet, and Microsoft Teams.

These results confirm that the system effectively recognizes ISL gestures in real-world applications while maintaining high accuracy and real-time performance.

8. CONCLUSION

This paper presents a real-time AI-powered ISL recognition system that bridges the communication gap between sign language users and non-signers. By integrating computer vision and deep learning techniques, the system accurately translates ISL gestures into text and speech.

The system achieves:

1) High accuracy (99.8%) for static gestures.
2) Fast processing speed (30 FPS) for real-time interaction.

This project has the potential to enhance accessibility for the deaf and speech-impaired community, improving communication in education, workplaces, and social interactions.

9. FUTURE SCOPE

To further improve the ISL recognition system, several advancements are planned:

9.1 Dataset Expansion

1) Include regional sign variations for better accuracy.
2) Expand the dataset to cover full-body sign language gestures.

9.2 Improved Gesture Recognition

1) Implement transformer-based architectures (e.g., ViT, GPT-4 Vision) for better dynamic gesture classification.
2) Integrate motion-based sensors for enhanced gesture tracking.

9.3 Mobile & Edge Deployment

1) Optimize the model for mobile applications and embedded devices (e.g., Raspberry Pi, Jetson Nano).
2) Reduce model size using TensorFlow Lite & ONNX for low-power devices.

9.4 Integration with Assistive Devices

1) Connect the system to hearing aids and augmented reality (AR) glasses for better accessibility.
2) Develop a smart AI assistant that automatically detects and translates sign language in real-time conversations.

By implementing these advancements, the system can become a comprehensive, real-time communication tool for the speech-impaired community, empowering them to interact seamlessly in a digitally connected world.

ACKNOWLEDGMENT

The authors want to dedicate this work by presenting thanks the people and bodies below for providing great encouragement and contributions during this study:

Project Mentor: Shwethashree G C is kindly acknowledged as she guided them constantly by giving good feedback on her part and encouraged her with various aspects throughout this course of work. We extend our gratitude to JSS Science and Technology University, with special thanks to the Computer Science Department, for giving us the necessary resources and infrastructure to conduct this research. Thank you, faculty, and staff, for your support.

REFERENCES

1. Abraham, E., Nayak, A., & Iqbal, A. (2024). Real-time translation of Indian sign language using LSTM.
2. P. B. V., R. R., Srinath, & Rashmi. (2024). Recognition of Indian sign language alphanumeric gestures based on global features.
3. Goel, P., Sharma, A., Goel, V., & Jain, V. (2024). Real-time sign language to text and speech translation using LSTM.
4. Thayiparampil, J. P., K. K., Binu, B. B., & Viju, S. (2023). Real-time sign language translation using TensorFlow object detection.
5. P. B. V., R. R., Srinath, & Rashmi. (2024). ISL2022: A novel dataset creation on Indian sign language.
6. Kaur, S., Sharma, R. K., & Gupta, A. (2023). Deep learning-based sign language recognition: A comparative study. *IEEE Access, 9*, 112345–112358.
7. Hossain, M., Chakraborty, T., & Roy, P. (2024). Real-time gesture recognition using MediaPipe and LSTM for sign language translation. *International Journal of Computer Vision and AI, 15*(4), 567–582.
8. Kumar, R., & Singh, P. (2023). Sign language recognition using CNN and transfer learning: A case study on Indian sign language. *Journal of AI Research, 31*(2), 98–113.
9. Verma, A., Tiwari, S., & Ramesh, M. (2024). Enhancing real-time sign language recognition through transformer networks. In *Proceedings of the IEEE International Conference on AI & Human-Computer Interaction*.
10. Patel, L., & Bose, D. (2023). Speech and gesture integration for real-time sign language communication. *International Journal of Human-Computer Interaction, 40*(1), 55–72.