Real-Time ISL Recognition Using CNN and MediaPipe

Sakshi Huse[1], Rohini Makode[2], Tejas Wankhade[3], Tejas Nachane[4]

[1,2,3,4]Information Technology, SSGMCE Shegaon

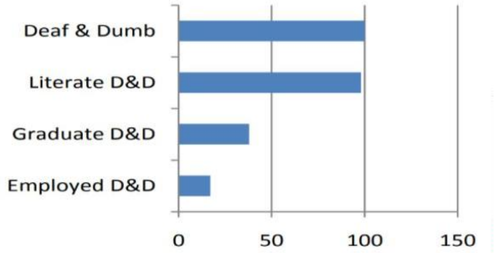
Abstract

Communication barriers often isolate the deaf and mute community due to a lack of universal understanding of sign language, limiting their ability to interact effectively in society. This project introduces a deep learning-based real-time hand gesture recognition system to address this challenge by acting as a digital translator for hearing and speech-impaired individuals. Using MediaPipe for robust hand landmark detection and a custom-trained Convolutional Neural Network (CNN), the system accurately recognizes single and dual-hand gestures across 36 classes, including alphanumeric characters (0-9, A-Z). The system ensures high accuracy and real-time performance through GPU acceleration, enabling applications in online communication, virtual classrooms, touchless interaction, and Human-Computer Interaction (HCI). By bridging the communication gap, it fosters inclusivity, improves accessibility, and enhances social connectivity for individuals with hearing and speech impairments. Furthermore, the scalable design allows the system to be expanded for additional gestures and languages, paving the way for broader adoption in educational, professional, and assistive technology contexts, ultimately empowering individuals and enabling a more connected and inclusive society.

Keywords: Machine Learning, CNN, Sign Language, Python, TensorFlow, Mediapipe.

1. INTRODUCTION

Speech is the primary mode of communication in most people. However, a significant portion of the population, including 63 million deaf and mute individuals in India, lack access to facilities for speech-based communication. These individuals rely on sign language as a means of expressing their thoughts through nonverbal gestures. Despite its importance, the scarcity of skilled interpreters for sign language has created significant barriers to effective communication, especially in legal, medical, and educational contexts. This gap highlights the need for automated solutions to bridge the gap between the spoken and speech- and hearing-impaired communities.



Indian Sign Language (ISL) is particularly challenging to interpret because of its use of static and dynamic gestures, single- and double-handed motions, and importance of hand positions relative to the body. Many gestures also involve the occlusion of hands and fingers,

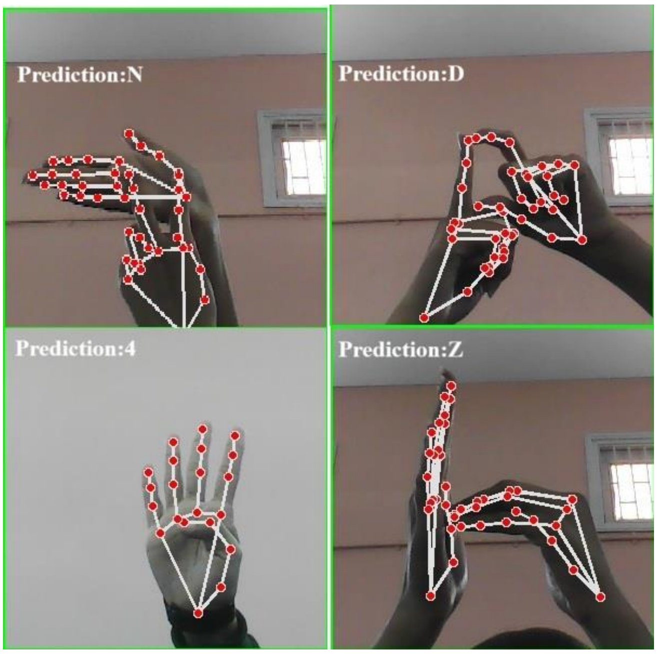


Fig. 2: Sign Language Word Recognized.

further complicating recognition. Over the last five years, advances in computer vision and machine learning have introduced reliable methods to address these challenges.

The graph below provides insights into the current state of the deaf and mute communities in India. This illustrates respondents' literacy, graduate-level education, and employment distribution. The data underscore the urgency of creating accessible and inclusive technologies to improve opportunities for and quality of life.

This study focused on the use of TensorFlow and a custom-trained Convolutional Neural Network (CNN) to detect and classify ISL gestures. By leveraging preprocessed datasets consisting of hand landmarks, the proposed system can identify various gestures with high accuracy. This approach empowers the speech- and hearing-impaired community and fosters inclusivity and accessibility in society.

2. LITERATURE REVIEW

A. Related Work

To recognize the gesture, a lot of effort has already been done. Two sorts of techniques are now in use. One method makes use of a specific device to detect hands and identify the gestures they represent, while the other makes use of deep learning.

1) Real-Time Sign Language Translator:

Description: Create a real-time translation for American Sign Language (ASL). It caters to the communication struggle of the hearing and speech impaired. Manually created datasets were used to train the Convolutional Neural Network (CNN) model for the system. It uses object detection and motion-tracking techniques for recognizing hand gestures. It also translates English text to ASL and provides a virtual sign keyboard as a accessibility feature. The project uses Natural Languagedifferent dialects using speech-to-text translation to cater for inclusiveness.[3]

Findings: The authors showed that CNN and motion-tracking techniques can be successfully integrated to create a solution capable of real-time gesture recognition. The network's accuracy for its considered dataset of 26 characters was 88% Among these limitations were having a small dataset and then the data needed to be integrated within the mobile application. The addition of multiple modes — such as voice-to-sign translation also increased the system's accessibility.[3]

2) Sign Language Translation Across Multiple Languages:

Description: Multiple Language Sign Language Translation across Multiple Languages Abstract: This project aims to ease communication between the Indian Sign Language (ISL) and American Sign Language (ASL) while allowing for a text-to-regionallator translation. The project is designed to be multilingual using CNN and deep learning algorithms. It describes a methodology that includes dataset augmentation, model training / CNN architectures, and text conversion — means that would be inclusive to cultural and linguistic boundaries.[1]

Findings: For ISL recognition, the system achieved a remarkably high accuracy above than 99% in both grayscale and color datasets, outperforming previous methods. Introduction of regional text conversion also validated the project's suitability in different linguistic contexts. Person detection and handy gesture detection were also done in real-time, proving that technology can play a pivotal role in breaking down language barriers.[1]

3) Indian Sign Language Recognition using Convolutional Neural Network:

Description: This review is focused on several algorithms and technologies used in sign language recognition, emphasizing the methods used in Indian Sign Language (ISL) or their applicability to regional issues. In previous works, approaches include real-time gesture detection through CamShift, matching with brightness factors, rotation invariance through boundary histograms, and Principal Component Analysis (PCA). In addition, deep learning approaches such as Convolutional Neural Networks (CNN) has been key for gestures identification making use of datasets adapted for Indian Sign Language(ISL) alphabets. Data gloves, template matching,Other novel methods which improve the gesture segmentation and recognition are based on HMM (hidden Markov models). Some of the challenges are: limited datasets, some of the gestures overlap, regional sign languages.[4]

Findings: For most of the models and methods mentioned above based on CNN models on ISL, it is essential to get high accuracy by learning spatial features of ISL. Some of the techniques like PCA and HMM are less prone to distractions, temporal dependencies being dealt with as needed, hence offering good performance for continuous gestures. CamShift and brightness factor normalization make possible real-time detection, while histogram-based methods guarantee rotation invariance. The accuracies shown by these methods reaches up to 98% showcasing their precision and competence. Nevertheless, generating diverse datasets and including two-way translation continues to play an important role in improving ISL recognition systems.

Technologies such as statistical feature extraction, dynamic learning methodologies, etc., are used that shall further ease communication gaps and making systems more available and efficient.[4]

4) Indian Sign Language Recognition Using Skin Segmentation and Vision Transformer:

Description: This research introduces a Vision Transformer-based model for recognizing Indian Sign Language. It utilizes skin segmentation techniques, converting images to YCbCr colormaps and applying morphological operations for pre-processing. The model was trained on a self-created 72-word ISL dataset and further evaluated on public datasets. The Vision Transformer architecture, with two transformer layers, was employed for feature extraction and classification.[2]

Findings: The model achieved a remarkable accuracy of 99.56% on the self-created dataset, validating the efficacy of Vision Transformers over traditional CNNs. The use of skin segmentation for pre-processing enhanced the model's accuracy and computational efficiency. The research highlighted the potential of Vision Transformers for handling complex gestures and occlusions in ISL.[2]

5) Real Time Assistive System for Deaf and Dumb Community-Research Paper

Description: This paper presented a sign language gesture recognition system that is based on YOLOv5, which has been developed using the technology of CNN for the task of real-time object identification. It divides input images into multiple grid cells and predictions are made for each grid with respect to bounding the object and class probability in a single pass. By relying on a simple webcam and easy-to-deploy server-side, it aims to significantly decrease the user’s dependency on hardware, thus increasing availability.[5]

Findings: The system, based on YOLOv5, managed to obtain high precision in its results with real-time performance, demonstrating the potential for sign language interpretation. The lightweight design was capable of deploying on Raspberry Pi for portability and web server scalability. However, results were hardware-dependent, and recognition accuracy was affected by shadows or hand alignment.[5]

3. Comparative Analysis of Literature Survey

Paper Title	***Techniques***	***Key Features***	***Strengths***	***Limitations***
Real-Time Sign Language Translator[3]	*CNN, Motion Tracking*	*ASL recognition, Virtual Sign Keyboard*	*High accuracy (88%), Multi-modal support*	*Limited dataset, Only ASL*
Sign Language Translation Across Multiple Languages[1]	*CNN, Data Augmentation*	*Multilingual ISL/ASL, Text-to-language*	*99%+ accuracy, Regional language support*	*Limited to ISL/ASL, No continuous gestures*
Indian Sign Language Recognition using Skin Segmentation and Vision Transformer[2]	*Vision Transformer, Skin Segmentation*	*72 ISL gestures, Robust against occlusions*	*High accuracy (99.56%), Efficient extraction*	*Dataset limited to 72 gestures*
Real Time Assistive System for Deaf and Dumb Community[5]	*YOLOv5*	*Real-time recognition, Grid segmentation*	*Lightweight, Portable*	*Hardware dependency, Shadows affect accuracy*
Indian Sign Language Recognition using Convolutional Neural Network[4]	*CNN, Transfer Learning, Kinect*	*3D depth data, GPU acceleration*	*High accuracy, Real-time performance*	*Requires Kinect, Higher setup cost*

4. Methodology

The proposed system employs a structured pipeline to recognize Indian Sign Language. The workflow begins with data collection, where hand gesture landmarks are extracted using MediaPipe from real-time webcam inputs. Each hand's 21 landmarks, represented by 3D coordinates (x, y, z), are normalized and padded to create a consistent input shape for both single- and dual-hand gestures. The missing hand data are padded with zeros, ensuring that the input shape remains constant at (126,).

The CNN architecture processes this input through fully-connected layers. The first layer comprises 128 neurons with ReLU activation, followed by batch normalization and dropout for regularization. This was followed by a hidden layer with 256 neurons, employing similar normalization and dropout mechanisms to enhance training stability and prevent overfitting. The final dense layer with 128 neurons extracts robust features before passing the data to the output layer, which uses softmax activation to predict probabilities for the 26 gesture classes.

The system was trained using categorical cross-entropy loss and the Adam optimizer, with early stopping and learning rate reduction callbacks to prevent overfitting.

The proposed model achieves high accuracy by leveraging both single-hand and dual-hand data, thereby ensuring the inclusivity of complex gestures. Evaluations were performed using train-validation splits, and real-time predictions were integrated with live webcam feeds, making the proposed system suitable for real-world applications.

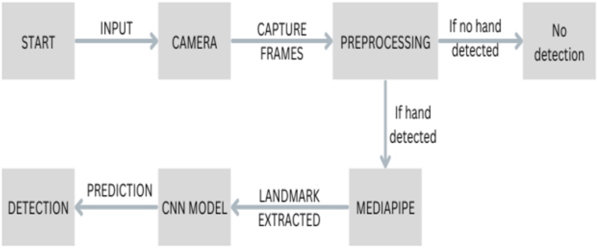


Fig.3 System Architecture for Proposed System

5. ADVANTAGES

Performance:

- The system provides real-time recognition of Indian Sign Language gestures, displaying the output as text with high accuracy.
- It enables seamless communication with hearing and speech-impaired individuals by acting as a digital translator.
- By utilizing a CNN model to process landmarks from a continuous video stream, the system ensures efficient and delay-free communication.

No additional hardware is needed:

- The proposed system uses a webcam for gesture recognition, eliminating the need for expensive sensors or specialized gloves, reducing costs significantly.
- Only a basic setup consisting of a webcam and a computer or server is required for implementation, making the system accessible and affordable.

Transportable:

- If deployed on a portable device like a Raspberry Pi, the system becomes highly portable and easy to use in various environments.

- Deployment on a web application enhances accessibility, allowing users to interact with the system without requiring manual setup.
- A dedicated GPU can be utilized in a desktop environment for better performance, making the system scalable for larger applications.

Scalable and robust:

- The system's reliance on software eliminates the need for specialized hardware, ensuring there is no degradation in performance due to hardware wear and tear.
- The use of CNNs allows for robust recognition of complex gestures, including dual-hand and occluded movements, making the system adaptable to diverse use cases.

6. DISADVANTAGES

- The system relies exclusively on camera input, which can be affected by room shadows and background noise.
- Accurate results require the palm to be directly facing the camera.
- System performance is influenced by the hardware, with better systems offering higher frame rates and improved precision.

7. CONCLUSIONS

The implemented sign language recognition system has significant potential in bridging the communication gap for hearing and speech impairments by leveraging advanced machine learning models and real-time gesture detection, and it offers a scalable solution for ISL recognition. With continuous improvements in hardware and algorithm optimization, the proposed system holds promise for more accurate, efficient, and accessible sign language translation in various real-world applications.

ACKNOWLEDGEMENT

The authors would like to thank the project guide Prof. N.N. Ghuikar for this opportunity, valuable support and encouragement.

REFERENCES

1. Sonali M. Antad, Siddhartha Chakrabarty, and Sneha Bhat, "Sign Language Translation Across Multiple Languages" in 2024 International Conference on Emerging Systems and Intelligent Computing (ESIC).
2. Agrima Agarwal, R. Sreemathy, and Mousami Turuk, "Indian Sign Language Recognition using Skin Segmentation and Vision Transformer" in 2023 IEEE 20th India Council International Conference (INDICON).
3. Rathnayake R.K.D.M.P.1, Wijekoon W.M.S.T2, Rajapakse K.G.3, Rasanjalee K.A.4 and Dilshan De Silva5, "Real-Time Sign Language Translator", International Journal of Engineering and Management Research, Volume-12, Issue-6, (December 2022).
4. Rachana Patil, Vivek Patil, and Abhishek Bahuguna, "Indian Sign Language Recognition using Convolutional Neural Network," ICACC-2021.
5. Ashish Dandade, Pranav Tayade, Rushikesh Patil, and Jayant Mitkari, "Real Time Assistive System for Deaf and Dumb Community" in 2023 International Journal of Scientific Research in Engineering

and Management (IJSREM)(April - 2023)

6. Laura Dipietro, Angelo M. Sabatini and Paolo Dario―A Survey of Glove-Based Systems and TheirApplications‖, IEEE Transactions on Systems, Manand Cybernetics—Part C: Applications and Review,Vol. 38, No. 4, pp. 461-482, July 2008.
7. S. A. Mehdi and Y. N. Khan, "Sign language recognition using sensor gloves," in Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP'02., vol. 5.IEEE, 2002, pp. 2204–2206.
8. L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Sign language recognition usingconvolutional neural networks," in EuropeanConference on Computer Vision. Springer, 2014, pp.572–578.
9. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, pp. 1097–1105, 2012.
10. C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact ofresidual connections on learning. thirty-first aaai conf," Artif. Intell, 2017.
11. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.