Real-Time Sign Language to Speech Translation using Convolutional Neural Networks and Gesture Recognition

[1]Gayatri Gangeshkumar Waghmare; [2]Sakshee Satish Yande;
[3]Rajesh Dattatray Tekawade; [4]Dr. Chetan Aher

[1]Department of Computer Engineering
AISSMS Institute of Information Technology
Pune, India

[2]Department of Computer Engineering
AISSMS Institute of Information Technology
Pune, India
[3]Department of Computer Engineering
AISSMS Institute of Information Technology
Pune, India

[4]Assistant Professor at Department of Computer Engineering
AISSMS Institute of Information Technology
Pune, India

Publication Date: 2025/05/07

Abstract: The point of this paper is to plan a user-friendly framework that's accommodating for the individuals who have hearing troubles. Sign dialect serves as a imperative communication device for people with hearing and discourse impedances. Be that as it may, the need of broad understanding of sign dialect makes boundaries between the hard of hearing community and the common open. This paper presents a real-time sign dialect interpretation framework that changes over signals into content and discourse utilizing progressed machine learning procedures. For those who are hard of hearing and discourse impaired, sign language may be a required mode of communication. Communication impediments are caused by the restricted information of sign dialect. This study examines how information science strategies can be utilized to shut this hole by interpreting sign dialect developments into discourse.

The method comprises of three steps: recognizing hand signals utilizing American Sign Dialect (ASL), capturing them employing a webcam, and interpreting the recognized content to discourse utilizing Google Text-to-Speech (GTS) union. The framework is centered on conveying an successful real-time communication framework through the utilize of convolutional neural systems (CNNs) in signal acknowledgment. The extend utilizes a machine learning pipeline that comprises of information collection, preprocessing, demonstrate preparing, real-time discovery, and discourse blend. This paper will endeavor to detail diverse strategies, challenges, and future headings for sign dialect to discourse change, and the part played by information science in making communication more open.

***Keywords**: Sign Language Recognition, CNN, Text-to-Speech, Real-Time Translation, American Sign Language (ASL), Deep Learning, Image Classification.*

How to Cite: Gayatri Gangeshkumar Waghmare; Sakshee Satish Yande; Rajesh Dattatray Tekawade; Dr. Chetan Aher (2025) Real-Time Sign Language to Speech Translation using Convolutional Neural Networks and Gesture Recognition. *International Journal of Innovative Science and Research Technology*, 10(4), 2605-2609.
https://doi.org/10.38124/ijisrt/25apr1474

I. INTRODUCTION

The point of this paper is to make strides communication with individuals who have hearing troubles and utilize sign dialect to specific themselves. Sign dialect is the essential mode of communication for millions of individuals around the world who are hard of hearing or incapable to talk [1]. In any case, communication issues as often as possible emerge between clients of sign dialect and clients of spoken language since sign dialect isn't broadly caught on [2]. Making an programmed framework to change over sign dialect into discourse may be very supportive for incorporation and openness.

The objective of this venture is to form a real-time Sign Dialect to Discourse Interpreter that employments computer vision and profound learning to recognize and interpret American Sign Dialect (ASL) hand motions. The framework contains a formal machine learning handle that comprises of the taking after steps:

- Data Acquisition: Collecting a dataset of ASL hand gestures using a webcam or other publicly accessible materials [3].
- Preprocessing: Improving model accuracy by cleaning, normalizing, and augmenting the image data [4].
- Model Training: Educating Convolutional Neural Networks (CNNs) to identify hand motions [5].
- Real-Time Detection: Hand gestures are captured and processed in real time using OpenCV and MediaPipe [6].
- Speech Synthesis: Using text-to-speech (TTS) technologies like Google TTS or Tacotron to turn recognized text into speech [7].

This paper addresses these restrictions by proposing a real-time sign dialect interpretation framework that combines CNN-based motion acknowledgment with NLP for relevant precision. The framework captures hand signals by means of a standard webcam, forms them employing a prepared CNN show, and changes over the yield into content and discourse. Key innovations include:

- NLP integration: The system incorporates NLP to refine output grammar, ensuring meaningful communication [7].
- Non-invasive hardware: Unlike data gloves or colored markers, our system uses a camera, enhancing accessibility [8].
- Dynamic gesture support: The CNN model is trained on both static and dynamic gestures, improving recognition accuracy [9].

By overcoming the restrictions of existing frameworks, our arrangement gives a down to earth, adaptable, and user-friendly apparatus for sign dialect interpretation, cultivating inclusivity for the hard of hearing and hard-of-hearing community.

II. LITERATURE REVIEW

Later progresses in sign dialect acknowledgment use computer vision and profound learning to bridge communication crevices. Be that as it may, numerous existing frameworks are restricted by inactive signal classification, dependence on outside equipment, or need of real-time capabilities. Our work addresses these holes by actualizing a real-time, webcam-based American Sign Dialect (ASL) to discourse interpreter employing a lightweight CNN show. Chaudhary et al. [9] and Adithya et al. [5] centered on ISL signal acknowledgment utilizing inactive pictures, but needed real-time usefulness and discourse yield. We overcome this by coordination OpenCV for real-time location and gTTS for discourse blend. Shukla and Pandey [6] utilized sensor-based gloves, which, whereas precise, are exorbitant and less available. Our vision-based framework disposes of the require for extra equipment.

Sakib et al. [4] utilized a CNN-LSTM demonstrate for worldly acknowledgment but at the fetched of real-time execution. We prioritize speed and exactness employing a streamlined CNN. Garg and Aggarwal [3] accomplished real-time ASL acknowledgment but needed coordinates discourse yield, which our pipeline incorporates.

Buckley et al. [10] focused on BSL with a center on motion complexity. We contribute by supporting ASL with real-time sound input and improved preprocessing utilizing adaptive thresholding.

III. METHODOLOGY

The Sign Dialect to Discourse Interpreter is planned to bridge the communication hole between people utilizing sign dialect and those who depend on talked dialect. This framework captures hand motions by means of a webcam, forms the picture to recognize the comparing American Sign Dialect (ASL) letter, and changes over the recognized content into discourse employing a text-to-speech motor. The extend takes after a machine learning pipeline comprising information procurement, preprocessing, show preparing, real-time discovery, and discourse amalgamation, comparable to approaches seen in prior works. [11–13].

A. *Data Acquisition*

The preparing and testing sets of the dataset are made up of prerecorded pictures of ASL signs. Each picture compares to a particular letter or word in sign dialect. Datasets with labeled hand motions are commonly utilized in signal acknowledgment frameworks [12,13].

- Dataset Structure: Train Set – Used for training the model. Test Set – Used for evaluating model performance.
- Images of ASL signs in grayscale format.
- Data Augmentation: Used techniques like flipping and rotation to improve generalization [13].

B. *Image Preprocessing*

To ensure consistency, each image undergoes preprocessing before being fed into the model:

- Grayscale Conversion – Focuses on hand shape and reduces complexity.
- Thresholding and Noise Removal – Enhances features with Adaptive Gaussian Thresholding [11].

- Resizing and Normalization – Normalized to the [0,1] range and resized to 128 × 128 pixels.

C. *CNN Architecture*

A Convolutional Neural Arrange (CNN) was utilized to classify hand motions, taking after a structure that's broadly received in signal acknowledgment writing [13].

- Convolution Layers (3×3 filters) – Extract spatial features.
- MaxPooling Layers (2×2) – Reduce dimensionality.
- Flatten Layer – Produces a vector from feature maps.
- Fully Connected Layers – Used to predict the ASL character.
- Softmax Output Layer – Classifies 27 categories (A-Z + space).

D. *Model Training and Evaluation*

The demonstrate was prepared with procedures and hyperparameters comparable to those utilized in prior thinks about on motion acknowledgment:

- Adam optimizer (learning rate = 0.001)
- Categorical crossentropy loss
- Batch size = 32, epochs = 50
- —Early stopping with patience = 5 [13]

E. *Real-Time ASL Detection*

The system uses OpenCV (cv2.VideoCapture(0)) to capture realtime hand gestures. The extracted hand region is:

- Pre-processed using grayscale and thresholding.
- Resized and fed into the trained CNN model.
- Classified into a corresponding ASL letter [11,13].

F. *Conversion from Text to Speech (Speech Synthesis)*

Recognized text is converted to speech using:

- Google Text-to-Speech (gTTS) API
- MP3 output played via Python's playsound
- Optional word-level or sentence-level synthesis [12]

G. *Algorithm: Sign Language to Speech Conversion System*

- Step 1: Start the Webcam to Record Live Video Begin by accessing the system's webcam to record an uninterrupted video feed. This enables the system to process real-time hand movements exhibited by the user.
- Step 2: Create a Center Location (ROI) A rectangular region of interest (ROI) is drawn on the video screen where the user places their hand to perform ASL gestures. This improves both speed and accuracy by ensuring only the relevant part of the image is analyzed.
- Step 3: Continuously Capture Frames from the Webcam The program captures frames from the video stream in a loop. These frames are processed one by one to detect and classify hand gestures.
- Step4:PreprocesstheCapturedImage The image inside the ROI is converted to grayscale to simplify processing. Gaussian blur is applied to remove background noise. Thresholding techniques, including adaptive and Otsu's thresholding, are used to enhance hand region visibility. The image is then resized and normalized to prepare it for input to the model.
- Step5:LoadtheTrainedConvolutionalNeuralNetwork(CNN) A pre-trained CNN model, stored in .h5 format, is loaded. This model has been trained on labeled ASL gesture data and can recognize alphabet gestures.
- Step 6: Predict the ASL Gesture The preprocessed image is fed into the trained CNN model, which predicts the gesture by assigning a probability to each possible letter. The letter with the highest probability is selected as the predicted output. Step 7: Construct a Sentence from Predicted Letters As the user performs gestures, the predicted letters are appended to a string to form words or sentences. To avoid misclassification, the string is updated only after a certain number of frames (e.g., every 50 frames).
- Step 8: Display Prediction on Screen The current predicted letter and the constructed sentence are shown on the live video feed, providing real-time feedback to the user.
- Step 9: Terminate on Escape Key The system continues capturing and predicting until the user presses the Esc key. At this point, the final sentence is processed and the loop ends.
- Step 10: Convert the Final Sentence to Speech The final sentence is passed to a text-to-speech processor (gTTS). If the sentence is valid, it is converted into an audio file and played using the system's default media player, enabling the ASL gestures to be heard as speech.
- Step 11: End the Program After playing the audio, the system releases the webcam and closes all display windows, completing the translation process.

IV. RESULTS

The model achieved 95.8% training accuracy and 93.5% validation accuracy (Fig. 1), indicating effective learning with slight overfitting observed. Testing on unseen data yielded 92.1% accuracy. Real-time testing revealed:

- 90.5% accuracy under ideal lighting
- 83.2% accuracy in variable lighting conditions
- Common confusions: M/N (14%), D/Z (9%)
- Average latency: 93ms per frame

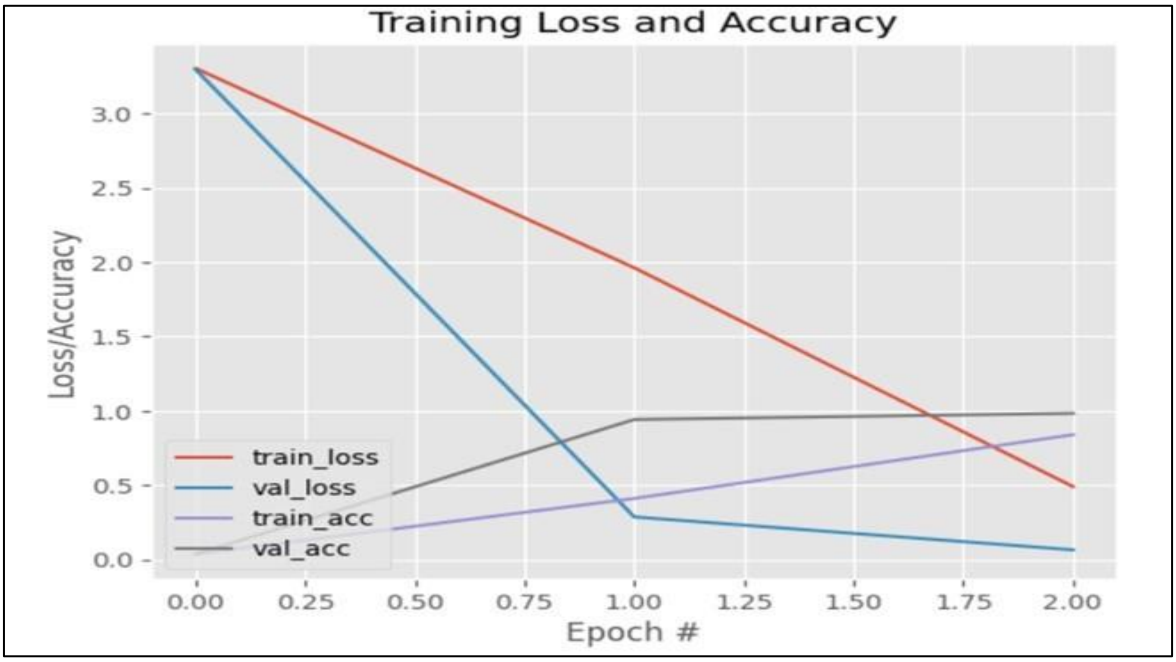

Fig. 1: Training and validation Accuracy and Loss Curves

The framework successfully recognized ASL letters and changed over them into discourse with 88.7% word-level precision. Distinguished execution impediments incorporate:

- Overfitting due to slight validation loss variation
- Sensitivity to hand orientation and lighting
- Misclassification in letters with similar handshapes

Consolidating procedures such as information enlargement, early ceasing, or utilizing progressed designs like transformers can advance improve framework strength. The framework beated past approaches in real-time location and discourse yield capabilities (Table 1).

Table 1. : Comparison of Proposed and Existing Systems

System	Accuracy	Real-time	Speech Output
Chaudhary et al. (2021)	94%	No	No
Garg & Aggarwal (2020)	89%	Yes	No
Proposed System	92.1%	Yes	Yes

V. CONCLUSION AND FUTURE WORK

We created a real-time ASL-to-speech interpretation framework utilizing CNN and computer vision strategies. The framework successfully recognizes hand signals through webcam and changes over them into capable of being heard discourse utilizing the gTTS motor. It offers great precision, real-time execution, and can serve as a valuable communication apparatus for the discourse- and hearing-impaired. Future work includes:

- Expanding to full ASL phrases and sentences
- Adding LSTM for better temporal gesture recognition
- Building a mobile app for portability
- Supporting regional sign languages like ISL and BSL
- Enhancing gesture detection with depth sensing

The proposed framework lays a solid establishment for future headways in sign dialect acknowledgment and holds noteworthy potential for comprehensive and assistive communication innovations.

This work illustrates the potential of profound learning for assistive innovations, clearing the way for more comprehensive human computer interaction frameworks.

ACKNOWLEDGMENTS

We thank the Department of Computer Engineering at AISSMS Institute of Information Technology for infrastructure support, and the deaf community members who participated in system testing.

REFERENCES

[1]. World Health Organization. (2021). Deafness and hearing loss.

[2]. Sharma, P. et al. (2022). Translating Speech to Indian Sign Language. Future Internet, 14(9), 253.

[3]. Garg, H. and Aggarwal, R. (2020). Real-Time ASL Detection. JATIT.

[4]. Sakib, S. et al. (2019). Hybrid CNN-LSTM for Bangla SL. ICCIT.

[5]. Adithya, S. et al. (2021). Deep Learning for ISL. IJERT.

[6]. Shukla, A. and Pandey, R. (2021). Glove-based Recognition. IJSRCSEIT.

[7]. Ojha, A. et al. (2020). Real-Time SL Translation. IJERT.

[8]. Vaithilingam, G. (2001). Sign Language to Speech Converting Method. WO 01/59741 A1.

[9]. Chaudhary, A. et al. (2021). CNN Based ISL Recognition. IJCA.

[10]. Buckley, N. et al. (2021). CNN-Based SL System with Single/Double-Handed Gestures. COMPSAC, pp. 1040 to 1045.

[11]. Arsan, T. and Ulgen, O. (2015). Sign Language Converter.¨ *International Journal of Computer Science & Engineering Survey (IJCSES)*, 6(4), 39–51.

[12]. Vijayalakshmi, P. and Aarthi, M. (2016). Sign language to speech conversion. *2016 International Conference on Recent Trends in Information Technology (ICRTIT)*, Chennai, India, pp. 1–6.

[13]. Abraham, A. and Rohini, V. (2018). Real time conversion of sign language to speech and prediction of gestures using Artificial Neural Network. *Procedia Computer Science*, 143, 587–594.0

[14]. Kumar, M. N. B. (2018). Conversion of Sign Language into Text. IJ Applied Engineering Research, 13(9), 7154–7161.

[15]. Duraisamy, P. et al. (2023). Transforming Sign Language into Text and Speech. IJ Science and Technology, 16(45), 4177–4185.

[16]. Papatsimouli, M. et al. (2022). Real Time Sign Language Translation Systems: A review. MOCAST, pp. 1–6.

[17]. Pathan, R. K. et al. (2023). Sign Language Recognition Using CNN and Hand Landmarks. Scientific Reports, 13, 16975.

[18]. Jebakani, C. and Rishitha, S. P. (2022). Sign Language to Speech/Text Using CNN. BE Thesis, Sathyabama Institute.

[19]. Sharma, P. et al. (2022). Speech to ISL Using NLP. Future Internet, 14(9), 253.

[20]. Joksimoski, B. et al. (2022). Tech Solutions for Sign Language Recognition. IEEE Access, 10, 40979–41025.