American Sign Language Translation into Text and Speech

Pranjali Kolawale[1], Pranav Deore[2], Dr. Disha Gabhane[3]
[1,2] *Student, MIT-ADT University, Pune, Maharashtra, India*
[3]*Assistant Professor, MIT-ADT University, Pune Maharashtra, India*

***Abstract*—Sign language is one of the communication ways for the people who have hearing issues. Common people find it a challenge to communicate with deaf people, that's where this system is useful. This project presents the development of real-time American sign language translation into text and voice. This System captures the hand gestures using a webcam and processes using the Mediapipe and OpenCv Python libraries. These 3D landmark coordinates are then structured into a consistent format and passed to a trained 2D Convolutional Neural Network (CNN) for classification. The model accurately predicts ASL Gestures. The predicted output is displayed as text, with optional conversion to voice, enabling more accessible and seamless communication across hearing boundaries.**

Keywords—ASL, CNN, Landmark.

I. INTRODUCTION

Communication is very important for humans, yet millions of individuals rely on sign language, face barriers in both personal and professional environments. Sign language is the visual method of communication that based on hand gestures, facial expressions, and body movements to convey message [1]. American Sign Language is one of the way for communication, but lack of understanding of sign language in common people results in feeling left out, miscommunication and having fewer chances in educational, professional and social context [3].

With the rapid evolution of artificial intelligence and computer vision technologies, there is opportunity to bridge this gap. There are some tools that attempt to do this, but many of them are too expensive, lack of contextual understanding or slower real-time performance [3].

The proposed solution uses different technologies to understand and translate sign language. It includes gesture recognition, computer vision to track hand movements. After recognizing a sign, the system gives text interpretation of the sign which can be converted into speech using text-to-speech technology, allowing smooth communication [2]. The main challenge in the software is capturing hand gestures.

This research presents an AI-driven system designed to translate ASL gestures into real-time text and spoken language. This solution is further enhanced with features such as live interpretation during virtual sessions and proposed wearable smart glasses or AR Systems or through room Cameras for offline, mobile usage.

This study adds new insights to the field where assistive technology meets artificial intelligence. It focuses on making communication smoother for the deaf people and aims to create equal opportunities for them in professional and social life. By solving technical, ethical, and real-world problems of real-time translation, this project hopes to support ASL users and improve how we design inclusive communication tools for the modern digital world.

II. LITERATURE SURVEY

In this section, we examine previous existing research related to the translating sign language into text using computer vision and machine learning techniques. Each study is reviewed for its approach, strengths, and limitations, offering a thorough understanding of techniques and identifying gaps that the proposed system aims to address.

Author and Year	Methodology	Key Contributions
Cem Keskin, Furkan Kıraç, Yunus Emre Kara, and Lale Akarun (2012)	Uses multi-layered RDF framework combining Global Expert Network (GEN) and Local Expert Network (LEN) for more accurate and generalized pose estimation.	Introduces a multi-layered Randomized Decision Forest (RDF) framework for real-time hand pose estimation and shape classification using depth sensors.
W. Tao, M. C. Leu, and Z. Yin (2018)	Uses CNN with multi view data and fusion of results for improved accuracy.	Proposes a CNN-based ASL alphabet recognition system with multi view augmentation and inference fusion.
Premkumar Duraisamy, A. Abinayasrijanani2, M Amrit Candida2, P Dinesh Babu (2023)	Used MediaPipe for hand landmark detection and CNN for classification	Developed real-time sign language recognition system with good accuracy using webcam input
Akash Kamble, Jitendra Musale, Rahul Chalavade, Rahul Dalvi, Shrikar Shriyal (2023)	Utilizes MediaPipe for key point detection, data pre-processing, labelling, and trains an LSTM neural network for gesture recognition.	Proposes a system that converts Indian Sign Language (ISL) gestures into text using computer vision and deep learning techniques.
Proposed System	Hand landmark extraction using MediaPipe + 2D CNN on coordinate-based features	Real-time, cost-effective sign recognition system using webcam and spatial landmarks without extra sensors

Table. 1

To make our model more effective in future, we can work on dynamic signs, for that we need to work on combining different methodology to get better predictions.

III. DATASETS

To build the sign language translation system, we used a publicly available dataset containing hand gesture images for American Sign Language letters and other signs. The dataset has more than 150,000 labeled images, covering all English alphabets along with several non-letter signs [4]. These labels help the model learn both alphabetic and non-alphabetic gestures, which are important for accurate sign language translation [1], [5].

A. Data Collection

- Training Data: The training dataset contains labeled images of hand gestures into different categories: the 26 letters of the English alphabet, “space”, “delete”, etc. Each image has a hand gesture and has been captured under different light conditions and backgrounds.
- Testing Data: The testing dataset contains additional labeled images of hand gestures. These images are used to evaluate model performance [6].

B. Data Preprocessing

- Hand Landmark Detection: The input image is preprocessed using MediaPipe hand tracking model. Then extraction of hand landmark is done. There are 21 landmarks each having 3 coordinates x, y, and z. But we are only using x and y. If no hand landmark is detected in the image then the image gets skipped during data loading.
- Normalization: The extracted landmark data is normalized to a common scale. This is done by dividing each value by the largest value among landmarks. It helps the model to learn better and work more efficiently.
- Reshaping: The hand landmarks are arranged into a fixed format which is suitable for Convolutional Neural Network. The data was reshaped into a 21×3×1 four-dimensional array, where the dimensions encode hand landmarks (21), their spatial coordinates (x, y, z), and grayscale channel (1) [6].

C. Labeling

Each image is labeled according to gestures. Firstly Letters ‘A’ to ‘Z’ are assigned numerical labels from 0 to 25 based on alphabetical order after which different signs are labeled from number 26. The labels are one hot-encoded during preprocessing and converting them into suitable for multi-class-classification [4].

D. Challenges in Dataset

- Different Hand Shapes and Sizes: Different hand sizes can affect the landmark detection. This difficulty we try to reduce using a large dataset.
- Far from Camera: In some of the images the hand is at a far distance from the camera which resulted in skipping that image while training.
- Hand Visibility Issue: In some of the images the hand is not clearly visible due to different light conditions.

IV. ARCHITECTURE

This system starts by capturing hand gestures from a webcam. Then MediaPipe extracts 21 landmarks from the frame. These landmarks are preprocessed. The processed data is fed into a 2D CNN trained model to classify and predict the output. The predicted output can be converted into speech. This is the complete flow of the system.[7]

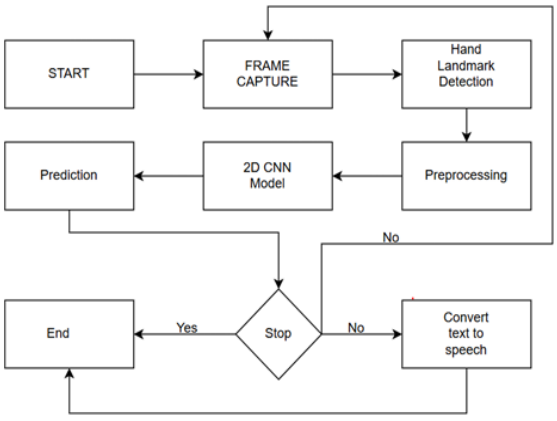

Figure.1

V. METHODOLOGY

The proposed system is designed to classify hand gestures. The system is divided into 5 different stages such as image capture, preprocessing, model

classification, evaluation, text to speech translation [1].

1. Image Capture: In this module image is captured, which contains the hand gestures, which will be processed for gesture classification
    - Data Collection: The system uses publicly available datasets, which have alphabets, different signs. Each image represents specific ASL gesture.
    - Hand Detection and Processing: MediaPipe Hands model is used to detect hand landmarks on captured images. It provides a set of 21 key landmarks for each hand, in a 3D coordinate space. The landmarks are then returned as a 3D list of coordinates (x, y, z) for each hand detected.
2. Preprocessing Module: It focuses on preparing the extracted hand landmarks for input into the classification model.
    - Landmark Extraction: In this, 21 landmarks are extracted from each captured image in a 3D coordinate form.
    - Normalization: In this, data is normalized to a range of 0 to 1 by dividing each coordinate by the maximum value found in the dataset. This process ensures the model receives standard input, which improves efficiency and model accuracy.
    - Reshaping: After that, landmark data is reshaped into a 4D array with the shape (samples, 21, 3, 1).
    - One-Hot Encoding: Each gesture has a unique number. To help a model understand these labels better, they are converted into one-hot encoded vectors. Basically, each gesture turned into a list of 0s and 1s. This helps the model to guess the correct gesture by giving a list of probabilities, and the gesture with the highest value is chosen as prediction.
3. Model Training and Classification: model training and classification is core of the mode, where a Convolutional Neural Network (CNN) is defined, trained and validated to recognize ASL gestures [2].

    1.CNN Architecture
    - Conv2D layers: The first two convolutional layers (64 filters with a kernel size of (2, 2) and 128 filters with the same kernel size) use the ReLU activation function. These layers are responsible for learning the spatial features of the hand landmarks [2].
    - MaxPooling2D layers: After each Conv2D layer, a MaxPooling2D layer with a (2,1) window reduces feature map width by half while keeping height unchanged. This preserves the 3D structure (x, y, z) of hand landmarks and emphasizes spatial relationships between joints. The approach improves efficiency, reduces overfitting, and maintains key features [2].
    - Flatten and Dense layers: The model is flattened into a 1D vector. This representation passed through two fully connected Dense layers with 128 and 64 units, respectively. Each layer uses ReLU activation. To reduce the risk of overfitting, a dropout mechanism with a rate of 0.3 is applied
    - Output layer: The final layer of the network consists of 29 neurons. The Softmax activation function is used to convert the outputs into a probability distribution, where the highest probability corresponds to the predicted gesture [2].

2.Model Compilation: The model is built using the Adam optimizer and uses categorical cross-entropy to measure loss. Accuracy is used as the metric to evaluate how well the model performs during both training and validation.

3.Model Training: The model undergoes training over 20 epochs using batches of 32 samples each, with the processed training dataset. Its performance is checked on the test set during training to see how well it generalizes. The model learns to recognize ASL gestures by reducing the categorical cross-entropy loss [9].

4. Model Evaluation

    After training, the model is evaluated on the test dataset to determine its classification accuracy. This step allows for assessing the model's ability to generalize to new, unseen data. The performance metrics include accuracy and loss on the test data [2].
5. Text to Speech Module: In this module, using google Text-To-Speech (gTTS) library we are converting our recognized text into spoken audio output.

VI. RESULTS

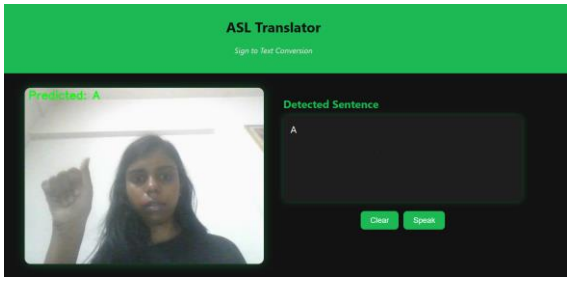

Figure. 2

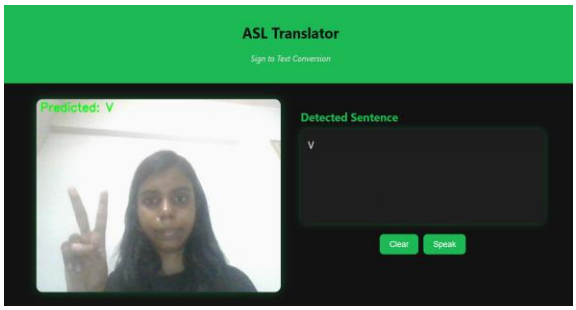

Figure. 3

VII. CONCLUSION

This research successfully demonstrates an ASL recognition system using hand landmark detection and a 2D CNN model. The process involves capturing hand gestures, extracting and preprocessing landmark data, and classifying gestures into ASL alphabets and commands.

The use of MediaPipe for hand tracking and a lightweight CNN architecture ensures accurate and efficient recognition. The model achieves reliable results and is saved for future deployment, offering potential for real-time applications to assist the hearing and speech impaired. Future improvements could include dynamic gesture recognition and integration into real-time communication tools.

REFERENCES

[1] R. Harini, R. Janani, S. Keerthana, S. Madhubala and S. Venkatasubramanian, "Sign Language Translation," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 883-886, doi: 10.1109/ICACCS48705.2020.9074370. keywords: {Gesture recognition;Assistive technology;Feature extraction;Training;Computer science;Cameras;Computer vision;Machine learning;Gesture recognition;Sign language;Human Computer Interaction;Computer Vision},

[2] P. Duraisamy, A. Abinayasrijanani, M. A. Candida, and P. D. Babu, “Transforming Sign Language into Text and Speech through Deep Learning Technologies,” *Indian Journal Of Science And Technology*, vol. 16, no. 45, pp. 4177–4185, Dec. 2023, doi: 10.17485/IJST/v16i45.2583.

[3] M. S. CHANDRAGANDHI, A. RAJ R, M. SHAMIL ML, A. S, and P. PT, “REAL TIME TRANSLATION OF SIGN LANGUAGE TO SPEECH AND TEXT,” *IARJSET*, vol. 8, no. 4, Apr. 2021, doi: 10.17148/iarjset.2021.8412.

[4] W. Tao, M. C. Leu, and Z. Yin, “American Sign Language alphabet recognition using Convolutional Neural Networks with multiview augmentation and inference fusion,” *Engineering Applications of Artificial Intelligence*, vol. 76, 2018, doi: 10.1016/j.engappai.2018.09.006.

[5] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun, “LNCS 7577 - Hand Pose Estimation and Hand Shape Classification Using Multi-layered Randomized Decision Forests,” 2012.

[6] A. Kamble, J. Musale, R. Chalavade, R. Dalvi, and S. Shriyal, “Issue 3 www.jetir.org (ISSN-2349-5162),” JETIR, 2023. [Online]. Available: www.jetir.orgd291

[7] M. Kumar, “Conversion of Sign Language into Text,” 2018. [Online]. Available: http://www.ripublication.com

[8] M. Prabhakar, P. Hundekar, S. B. Deepthi P, S. Tiwari, and V. M. S, “SIGN LANGUAGE CONVERSION TO TEXT AND SPEECH,” JETIR, 2022. [Online]. Available: www.jetir.org

[9] A. Dagur, K. Singh, P. S. Mehra, and D. K. Shukla, “Intelligent Computing and Communication Techniques – Volume 2”, CRC Press, 2025.

[10] A. Banafa, “Artificial Intelligence in Action: Real-World Applications and Innovations”, River Publishers, 2025.