

Deep Learning-Based Real-Time Sign Language Recognition with Text and Speech Output

K.Krupa Sagari¹, K.Pavani², A.Supriya³

#1 Assistant Professor & Head of Department of MCA, SRK Institute of Technology, Vijayawada.

#2 Assistant Professor in the Department of MCA, SRK Institute of Technology, Vijayawada

#3 Student in the Department of MCA, SRK Institute of Technology, Vijayawada

Abstract: This paper presents a real-time sign language recognition system that translates hand gestures into text and speech using deep learning techniques. The system utilizes MediaPipe for efficient hand tracking and a Convolutional Neural Network (CNN) for accurate gesture classification. The recognized gestures are converted into meaningful text and further synthesized into speech for effective communication. This approach provides a cost-effective and user-friendly solution to bridge the communication gap between speech-impaired individuals and non-sign language users.

Index terms - — *Sign Language Recognition, MediaPipe, Convolutional Neural Network (CNN), Hand Gesture Detection, Deep Learning, Text-to-Speech (TTS), Computer Vision, Real-Time System, Human-Computer Interaction (HCI)*

1. INTRODUCTION

Sign language is an essential communication medium for speech-impaired individuals, where hand gestures and movements are used to represent alphabets, words, and expressions. Despite its importance, most

people are not familiar with sign language, which creates a major communication barrier between deaf-mute individuals and the general public. This gap often leads to difficulties in everyday interactions, limiting accessibility and social inclusion.

With the rapid growth of Artificial Intelligence (AI) and Computer Vision, automated sign language recognition systems have gained significant attention. These systems aim to interpret hand gestures using cameras and machine learning techniques, reducing the dependency on human interpreters. However, many traditional approaches rely on additional hardware such as gloves or sensors, and often face challenges like poor accuracy, sensitivity to lighting conditions, and lack of real-time performance.

To overcome these limitations, this paper proposes a real-time sign language recognition and speech conversion system using MediaPipe and deep learning techniques. The system captures hand gestures through a webcam and extracts hand landmark features using MediaPipe. These features are then processed using a Convolutional Neural

Network (CNN) to accurately classify gestures into corresponding text.

Furthermore, the recognized text is converted into speech using a Text-to-Speech (TTS) module, enabling effective communication with non-sign language users. The proposed system is cost-effective, efficient, and easy to use, providing a practical solution for real-time interaction. Overall, this work contributes to improving accessibility and supports inclusive human-computer interaction by bridging the communication gap between sign language users and society.

2. LITERATURE SURVEY

a) A Real-Time Hand Gesture Recognition Method

Vision-based hand involvement is faster and more natural than using a pen, mouse, or keyboard. A reliable technique for real-time hand gesture detection was presented in this paper. In order to overcome aspect ratio constraints, most learning-based hand gesture systems depend on scale-space feature detection for gesture identification. A gesture is the first step in tracking and detecting hands. The fingers may be distinguished by their color and movement. The experiments show that our photo browsing navigation method is successful.

b) Sign Language Recognition Using Convolutional Neural Networks:

The hearing majority finds it challenging to interact with the Deaf. This communication barrier is threatened by recent advancements in automated sign language detection. Convolutional neural networks (CNNs), graphics

processing unit (GPU) acceleration, and Microsoft Kinect are used in this identification system. By automating the process, CNNs remove the need for complex features to be created by hand. We may easily recognize twenty Italian gestures. With a cross-validation accuracy of 91.7%, the prediction model may generalize to individuals and circumstances that were not experienced during training. With a mean Jaccard Index of 0.789, our model is competitive in the human gesture detection domain (ChaLearn 2014).

c) Hand gesture recognition using machine learning algorithms:

The identification of gestures is the primary focus of modern technologies. This study presents mathematical methods for computer-assisted gesture recognition. The restricted methods of computer interface include keyboards, mouse, touch screens, etc. These devices differ in their ability to include more adaptable computer technologies. Gesture detection is crucial for user-friendly interfaces. Gestures can come from anywhere on the body, however they are most frequently made with the hands and face. Instead of touching gadgets, you may operate them using gestures. This study demonstrates learned hand motions for scrolling and page turning.

d) Hand Gesture Recognition for Sign Language Using 3DCNN:

These reasons have led to an increase in the adoption of vision-based apps, touchless control on ubiquitous devices, and automated hand motion detection. Since hand gesture identification is essential to the

understanding of sign language, a good system should incorporate both spatial and temporal features. Establishing discriminative spatiotemporal descriptors for hand gesture sequences is difficult. In this study, we introduced a novel method for hand gesture detection that makes use of deep convolutional neural networks. Transfer learning was used to address the lack of tagged hand gesture datasets. In all, forty, twenty-three, and ten criteria were used to assess color video gestures. Identification rates for the three datasets are 98.12%, 100%, and 76.67% when signer-dependent mode is used. For each of the three datasets, the signer-independent mode identification rates were 84.38%, 34.9%, and 70.0%.

e) Hand Gesture Recognition in Complex Background Based on Convolutional Pose Machine and Fuzzy Gaussian Mixture Models:

Because hand gestures are a natural way for people to communicate with computers, they are widely used in human-computer interaction. It remains difficult because to complex real-world backgrounds, light variations, and occlusion. This paper proposes a two-stage approach for hand motion detection to address these issues. Regardless of the backdrop's complexity, the first step is to build a hand posture estimate that uses the convolutional pose machine to find the hand's keypoints. In the second stage, gestures are classified using computed hand keypoints, and nongesture patterns are eliminated using fuzzy Gaussian mixed models (FGMMs). Numerous tests have shown the method's efficacy, resilience, and user satisfaction in real-world situations.

3. METHODOLOGY

i) Proposed Work:

The proposed system focuses on developing a real-time sign language recognition and translation framework that enables seamless communication between speech-impaired individuals and non-sign language users. The system captures live video input using a webcam and processes the frames using MediaPipe for accurate hand detection and landmark extraction. These extracted features are then fed into a Convolutional Neural Network (CNN) model, which classifies the hand gestures into corresponding alphabets, words, or phrases with high accuracy.

Once the gesture is recognized, the system converts it into text using a predefined mapping mechanism. The generated text is further processed by a Text-to-Speech (TTS) module to produce natural and clear speech output. This end-to-end pipeline ensures real-time performance, eliminates the need for additional hardware, and provides a cost-effective, user-friendly solution for communication. The proposed system enhances accessibility, supports efficient human-computer interaction, and works effectively under varying environmental conditions.

ii) System Architecture:

The proposed system architecture is designed as a real-time processing pipeline that converts hand gestures into text and speech. Initially, the system captures live video input through a webcam, which continuously streams frames for processing. These frames are passed to the MediaPipe framework, where hand detection and landmark extraction are performed. MediaPipe identifies 21 key points of the hand, enabling accurate tracking of finger positions

and movements regardless of background or lighting conditions.

As illustrated in the system architecture diagram on page 24, the extracted landmark features are then forwarded to a Convolutional Neural Network (CNN) model for gesture classification. The predicted gesture is mapped to its corresponding text representation, forming meaningful words or sentences. Finally, a Text-to-Speech (TTS) module converts the generated text into natural speech output. This structured flow—from image acquisition to speech generation—ensures real-time performance, high accuracy, and seamless communication between sign language users and non-signers.

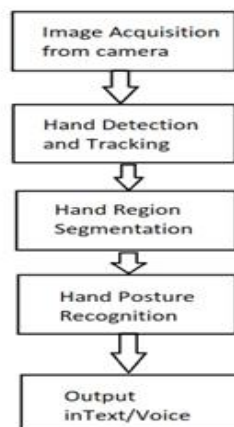


Fig1 proposed architecture

iii) Modules:

1. Data Acquisition

This module captures real-time hand gesture images using a webcam. It collects continuous video frames, which serve as input for further processing and ensures proper visibility of hand movements.

2. Data Preprocessing

In this stage, the captured images are processed using techniques like grayscale conversion, noise reduction, thresholding, and Region of Interest (ROI) extraction. This improves image quality and prepares the data for accurate feature extraction.

3. Feature Extraction (MediaPipe)

This module uses MediaPipe to detect and track hand landmarks. It extracts 21 key points from the hand, representing finger joints and positions, which are essential for gesture recognition.

4. Model Training (CNN)

A Convolutional Neural Network (CNN) is trained using preprocessed gesture images. The model learns spatial patterns and features associated with different sign language gestures for accurate classification.

5. Gesture Prediction

In real-time operation, the trained CNN model predicts the gesture based on extracted features. It identifies the corresponding alphabet, word, or phrase with high accuracy.

6. Text Conversion

The predicted gestures are mapped to their respective textual representations using a predefined mapping system. Sequential gestures are combined to form meaningful words or sentences.

7. Speech Synthesis (TTS)

This module converts the generated text into speech using a Text-to-Speech engine. It produces clear and natural audio output, enabling effective communication with non-sign language users.

8. User Interface

A web-based interface allows users to interact with the system easily. It displays recognized text and

plays the generated speech output, ensuring accessibility across different devices.

iv) ALGORITHMS:

1. MediaPipe Hand Tracking Algorithm

The MediaPipe algorithm is used for real-time hand detection and tracking. It identifies and extracts 21 key landmark points of the hand, including fingertips and joints, from each video frame. These landmarks provide a structured representation of hand gestures, enabling accurate tracking even under varying lighting conditions and complex backgrounds.

2. Convolutional Neural Network (CNN) Algorithm

The CNN algorithm is employed for gesture classification. It processes the extracted hand landmark features and learns spatial patterns associated with different gestures through multiple convolutional and pooling layers. The trained model classifies gestures into alphabets, words, or phrases with high accuracy, making it suitable for image-based recognition tasks.

3. Text Mapping Algorithm

This algorithm converts the predicted gesture labels into corresponding textual representations using a predefined mapping database. It ensures that each recognized gesture is correctly translated into letters, words, or sentences. Sequential gesture recognition is also handled to form meaningful text output.

4. Text-to-Speech (TTS) Algorithm

The TTS algorithm converts the generated text into natural and understandable speech. It processes the text into phonetic representations and synthesizes audio output using deep learning-based speech models. This enables effective communication by providing clear voice output for non-sign language users.

5. Deep Learning Model Optimization Algorithm

This algorithm focuses on improving the performance and accuracy of the system. It involves techniques such as dataset augmentation, hyperparameter tuning, and model training optimization. These enhancements ensure faster processing, reduced latency, and better recognition accuracy in real-time scenarios.

4. EXPERIMENTAL RESULTS

The proposed system was evaluated using real-time webcam input along with a dataset of hand gesture images representing alphabets from A to Z. Each class was trained using approximately 180 preprocessed images, and the Convolutional Neural Network (CNN) model was trained for multiple epochs to achieve optimal performance. The integration of MediaPipe for hand landmark extraction significantly improved feature detection, enabling accurate recognition even under varying lighting conditions and complex backgrounds.

The experimental results demonstrate that the system achieves high accuracy in gesture classification with minimal latency, making it suitable for real-time applications. The end-to-end pipeline—from gesture capture to text and speech output—operates efficiently, providing quick and reliable responses. The system successfully converts recognized gestures into meaningful text and clear speech, ensuring effective communication between sign language users and non-signers in practical environments.

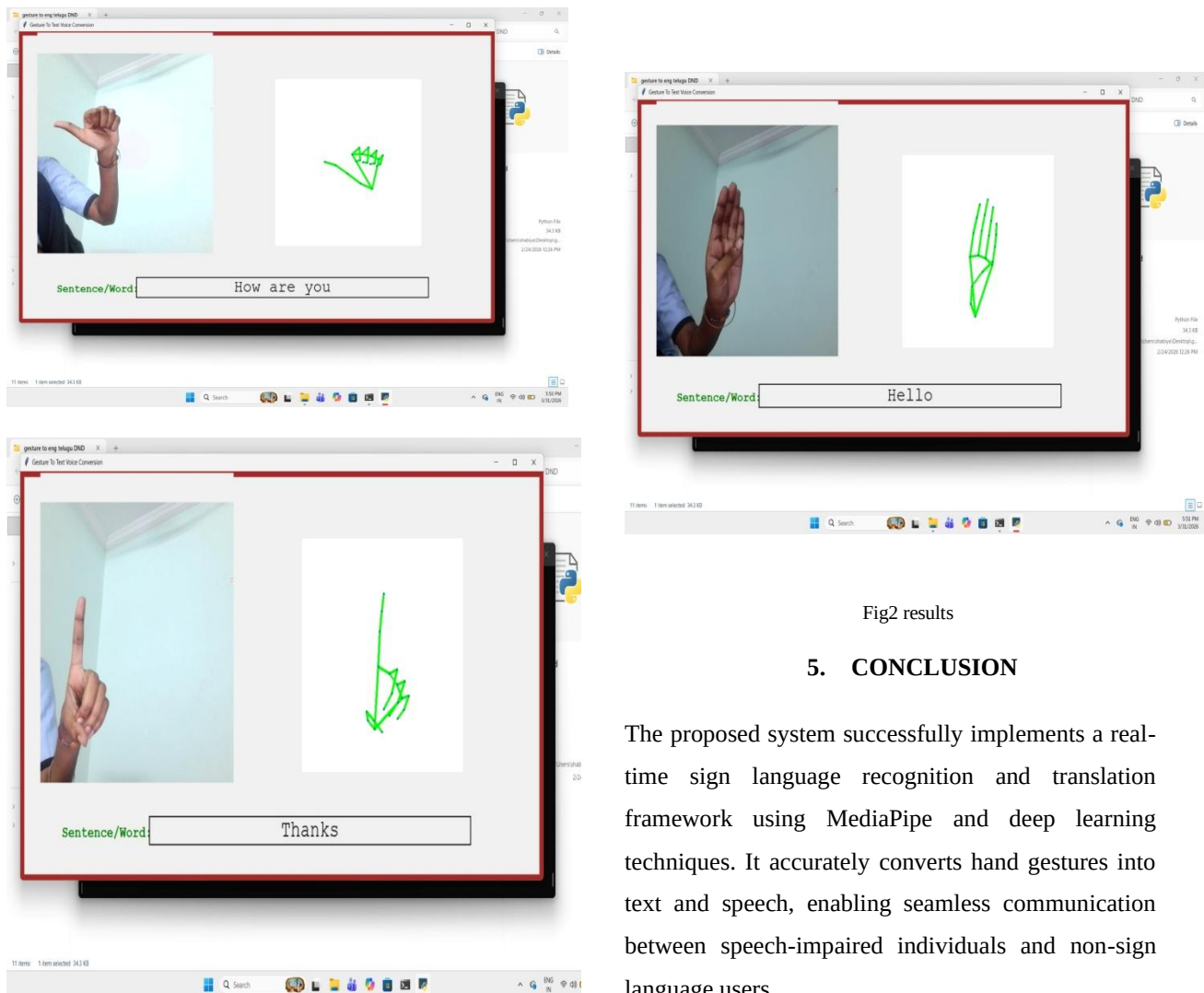


Fig2 results

5. CONCLUSION

The proposed system successfully implements a real-time sign language recognition and translation framework using MediaPipe and deep learning techniques. It accurately converts hand gestures into text and speech, enabling seamless communication between speech-impaired individuals and non-sign language users.

The system is efficient, cost-effective, and user-friendly, eliminating the need for additional hardware while maintaining high accuracy and real-time performance. It contributes significantly to improving accessibility and promoting inclusive human-computer interaction.

6. FUTURE SCOPE

- Extend the system to support dynamic gestures and full sentence recognition instead of only alphabet-level detection.
- Integrate advanced deep learning models such as LSTM, Transformer, or hybrid

CNN-RNN for improved temporal understanding.

- Expand support for multiple sign languages (e.g., ASL, ISL) to increase usability across regions.
- Develop a mobile and edge-based application for real-time deployment on smartphones and wearable devices.
- Improve robustness by handling complex backgrounds, occlusions, and low-light conditions more effectively.
- Incorporate voice customization and multilingual Text-to-Speech (TTS) for better user experience.
- Enable cloud integration and dataset expansion to continuously improve model accuracy and scalability.

REFERENCES

- [1] N. A. Ahmad, "A Globally Convergent Stochastic Pairwise Conjugate Gradient-Based Algorithm for Adaptive Filtering," in *IEEE Signal Processing Letters*, vol. 15, pp. 914-917, 2008, doi: 10.1109/LSP.2008.2005437.
- [2] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," in *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311-324, May 2007, doi: 10.1109/TSMCC.2007.893280.
- [3] N. H. Dardas and N. D. Georganas, "Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques," in *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 11, pp. 3592-3607, Nov. 2011, doi: 10.1109/TIM.2011.2161140.
- [4] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree and J. Kautz, "Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4207-4215, doi: 10.1109/CVPR.2016.456.
- [5] P. Molchanov, S. Gupta, K. Kim and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2015, pp. 1-7, doi: 10.1109/CVPRW.2015.7301342.
- [6] Ming-Hsuan Yang, N. Ahuja and M. Tabb, "Extraction of 2D motion trajectories and its application to hand gesture recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1061-1074, Aug. 2002, doi: 10.1109/TPAMI.2002.1023803.
- [7] A. S. Ghotkar, R. Khatal, S. Khupase, S. Asati and M. Hadap, "Hand

gesture recognition for Indian Sign Language,"
2012 International

Conference on Computer Communication and
Informatics, 2012, pp.

1-4, doi: 10.1109/ICCCI.2012.6158807



Ms.A.Supriya is an MCA Student in the Department of Computer Application at SRK Institute Of Technology, Enikepadu, Vijayawada, NTR District. She has Completed Degree in B.Sc(computers) Viveka arts and science college. Her area of interest are DBMS and Machine Learning with Python.

Author Profiles



Ms.K.Krupa Sagari working as an Assistant Professor in the Department of MCA at SRK Institute of Technology, Vijayawada. She has completed her MCA and M.Sc. (Mathematics). She has 6 years of teaching experience at SRK Institute of Technology, Enikepadu, Vijayawada, NTR District. Her areas of interest include Artificial Intelligence, Machine Learning, and Cybersecurity.



Mrs. K. Pavani is working as an Assistant and Head of Department of MCA, in SRK Institute of technology in Vijayawada. She completed her MCA and M.Tech in Computer Science. She has 10 years of teaching experience in SRK Institute of technology, Enikepadu, Vijayawada, NTR District. Her areas of interest include AI and ML, etc.