2025 International Conference on Emerging Systems and Intelligent Computing (ESIC) | 979-8-3315-2210-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/ESIC64052.2025.10962728

Real-Time Sign Language to Speech Converter Using OpenCV and MediaPipe

Anvisha Pathak
Dept. of Information Technology
Fr. Conceicao Rodrigues Institute of Technology
Navi Mumbai, India
anvishapathak4@gmail.com

Niraj Patil
Dept. of Information Technology
Fr. Conceicao Rodrigues Institute of Technology
Navi Mumbai, India
nirajpatil849@gmail.com

Prashant Padhy
Dept. of Information Technology
Fr. Conceicao Rodrigues Institute of Technology
Navi Mumbai, India
padhyprashant21@gmail.com

Adarsh Jadhav
Dept. of Information Technology
Fr. Conceicao Rodrigues Institute of Technology
Navi Mumbai, India
jadhavadarsh1710@gmail.com

Smita Rukhande
Dept. of Computer Engineering
Fr. Conceicao Rodrigues Institute of Technology
Navi Mumbai, India
smita.rukhande@fcrit.ac.in

Lakshmi Gadhikar
Dept. of Information Technology
Fr. Conceicao Rodrigues Institute of Technology
Navi Mumbai, India
lakshmi.gadhikar@fcrit.ac.in

***Abstract*—Communication barriers between deaf individuals and hearing people can lead to several significant problems such as social integration, reduced access to information and reduced opportunities thereby affecting quality of life of deaf individuals. Use of sign language for communication serves as the primary means to address these challenges. However, understanding of the sign language is very difficult for hearing individuals. To address this issue, we present a Real-Time Sign Language to text and Speech Converter to facilitate effective communication between deaf and hearing individuals. This paper presents a novel approach to sign language recognition using OpenCV and MediaPipe. The main aim of the paper is to develop real time computer vision that can translate sign language gestures into text in real-time. We use image processing techniques to segment and analyze sign gestures, OpenCV to record and process these gestures, MediaPipe for hand tracking and gesture prediction and a text-to-speech engine to translate the text into speech for auditory communication. Thus, this study offers a comprehensive sign language to speech converter system that has the potential to significantly improve communication between the deaf community and the hearing people.**

Keywords— Sign Language Recognition, Gesture Recognition System, Real-Time Hand Tracking, OpenCV, MediaPipe, Deaf Communication Tools, Text to Speech Conversion, Computer Vision

I. INTRODUCTION

Sign language is a natural and incredibly expressive form of communication for people who are deaf or hard of hearing. Effective communication with the deaf community is a big challenge, regrettably, many people without hearing impairments might not even attempt to learn sign language. To address this communication gap, our proposed system seeks to minimize the differences between individuals with and without hearing impairments. This can be accomplished by having the system be programmed to convert any sign language gesture into a standard text format.

The system attempts to make communication easier between people who don't understand sign language and those who use it frequently. It uses technology to translate sign language gestures into text that is simple to read by recognizing and interpreting them. The introduction emphasizes the advancements in science and engineering, particularly in improving human-computer interaction for the hearing impaired. By training computers to translate sign language into text, the system caters to both static and dynamic gestures captured through a webcam. The challenge lies in accurately capturing and filtering these gestures from diverse backgrounds, addressed through background subtraction techniques.

The motivation for developing this system stems from the struggles faced by individuals with hearing impairments in effectively communicating with the broader society due to limited recognition of their sign language gestures. Statistical data from sources such as the Indian census and the National Association of the Deaf underscores the significant number of people affected by hearing impairments, highlighting the necessity for a communication channel that can translate sign language into text for those unfamiliar with it.

The system addresses the difficulty faced by sign language users in communicating, eliminating the barrier between them and individuals unfamiliar with these gestures. The objective is to establish seamless communication without relying on specific background colors, gloves, or sensors. The system innovatively stores image pixel values in CSV files, reducing memory requirements and enhancing prediction accuracy. Comprising four core modules—Image Capturing [1], Preprocessing [2], Classification [3], and Prediction [4]. The system's primary goal is to convert sign language gestures into readable text, promoting smoother interaction between individuals with hearing impairments and the general population.

II. RELATED WORK

The purpose of the paper [7] is to examine the performance of various techniques for converting sign language to text/speech format. An Android application that can translate real-time American Sign Language (ASL) signs into text or speech is developed after analysis. The work that is suggested in the paper [8] helps those who are blind or have hearing or speech impairments communicate with others. The suggested model creates a platform that is easy to use by offering a speech/text output for a sign input. majority of people are not familiar with their sign language because they are a minority. Consequently, American Sign Language

979-8-3315-2210-0/25/$31.00 ©2025 IEEE

(ASL) is converted to text and speech output by the suggested system. To identify ASL hand gestures, this study uses convolutional neural networks (CNN) to extract effective hand features. The focus of the paper [9] is on the different strategies that have been put forth to translate Indian sign language into audio signals. The Sign Language Recognition (SLR) system classifies input into two categories: isolated and continuous sign language models. In interpersonal communication, language is used both verbally and nonverbally. Nonverbal communication is limited to people with speech and hearing impairments and is uncommon in the general population. Developing an automated language translation model that can easily translate sign language into text or audio is one method to bridge the communication gap between verbal and nonverbal communicators. Despite extensive research in this field, no reliable, affordable system exists that can effectively translate signs into speech. A gesture-based communication to text/speech model that permits a two-way exchange without the need for an interpreter is presented in the paper [10]. According to the paper, Deaf-Mute individuals encounter various difficulties even in carrying out every-day, basic tasks. Using sign language to communicate with others is one of the many challenges. The goal of the proposed "Sign Language Translator" system is to close this communication gap by enabling Deaf-mute individuals to communicate with non-ASL users. The system translates speech to American Sign Language using an RNN (Recurrent Neural Network) with LSTM (Recurrent Neural Network - Long Short-Term Memory) neural network trained by a Connectionist Temporal Classification (CTC) neural network. The purpose of the paper [11] is to assess the efficacy of various techniques for translating sign language into text and speech formats. Following analysis, an Android application is created that can translate real-time American Sign Language (ASL) signs into text or speech.

Prior work on sign language recognition has contributed significantly to the advancements in the field, but they typically lack effectiveness in real-time gesture prediction and effective interaction between the deaf and hearing communities. Most of the prior work concentrates either on static image recognition or require additional hardware, like gloves or a depth sensor, thereby limiting its accessibility. Moreover, the recognition process sometimes has problems with environmental noise or inconsistent lighting conditions. Our proposed system obliterates these drawbacks through the use of OpenCV for sophisticated image processing, MediaPipe for more accurate hand-tracking and noise reduction techniques for gesture recognition under more trustworthy conditions. Thus, it enables real-time, accurate translations without the need for specialized hardware making it much more scalable and user-friendly.

III. Methodology

This section presents tools and methodology used for implementation.

A. Tools used for implementation

Different tools used for implementation of system are OpenCV, MediaPipe and Flask.

OpenCV [5]:

OpenCV provides a wide array of tools, algorithms, and functions for real-time computer vision applications. OpenCV is widely used in industries and research for tasks like image and video analysis, object detection and recognition, facial recognition, gesture recognition, and more. Its functionalities include image processing, feature detection, machine learning support, and it offers a comprehensive set of libraries that enable developers to perform complex tasks in the field of computer vision. The library is designed to be efficient and optimized, allowing for real-time applications in various domains, including robotics, healthcare, security, and augmented reality.

MediaPipe [12]:

MediaPipe is a framework developed by Google that aims to simplify the integration of machine learning models, into media processing workflows. It provides a range of solutions for processing media data with a specific focus on video and camera inputs. By offering built models and processing modules MediaPipe enables developers to create applications in areas such as augmented reality, computer vision and gesture recognition. The key objective behind this open-source framework is to facilitate the integration of machine learning capabilities into media workflows. It offers developers access to tools including trained models and components for real time visual data processing. One of the key features of MediaPipe is its modular design, which allows developers to easily integrate pre-trained models and components into their projects.

Flask [13]:

Flask is a lightweight and versatile web application framework for Python. It is designed to build web applications in a simple and straightforward manner, providing the essentials to create powerful, web-based software. Flask offers flexibility and simplicity, allowing developers to create web applications swiftly with a minimalistic approach. It is known for its modular design, which means it is easy to extend with additional libraries, making it highly customizable based on specific project requirements. Flask is considered ideal for small to medium-sized projects due to its simplicity, but it's also robust enough to handle more complex web applications. It includes features for URL routing, template rendering, and request handling. Additionally, Flask is complemented by a vast array of extensions, making it capable of handling various tasks and functionalities, such as user authentication, database integration, and more, depending on the specific needs of the project.

B. Methodology for Implementation

The primary goal of the system is to facilitate communication between individuals using sign language and those who may face challenges in understanding sign language. Unlike traditional systems that rely on extensive databases storing numerous copies of alphabets and gestures, the proposed system adopts an innovative approach by utilizing the latest hand tracking algorithm. Typically, sign language translation systems involve large datasets containing various representations of each alphabet or letter in the form of gestures, leading to significant storage consumption. In contrast, the proposed system employs a hand tracking algorithm to track and interpret hand gestures captured by a camera. This algorithm serves as the core mechanism for recognizing and translating sign language.

The hand tracking algorithm operates by continuously analysing the movements and configurations of the hands captured in the camera feed. Instead of relying on a pre-existing database with multiple copies of each alphabet, the system dynamically interprets gestures in real-time, applying the algorithm associated with a specific alphabet. This eliminates the need for storing extensive datasets, reducing storage requirements, and making the system more efficient. The logic embedded in the algorithm corresponds to the unique characteristics of each sign language alphabet. As a user displays a sign language gesture in front of the camera, the hand tracking algorithm processes the information, identifies the relevant alphabet or letter, and generates the corresponding output. This output is then displayed, providing a real-time translation of the sign language gesture into a format that is easily understandable for individuals who may not be familiar with sign language.

By utilizing a hand tracking algorithm, the system not only minimizes storage demands but also enhances the efficiency and accuracy of sign language translation. This approach represents a modern and dynamic solution, aligning with the latest advancements in technology to bridge the communication gap between sign language users and those who may face difficulties in comprehending sign language gestures.

Fig 1 visually depicts a comprehensive array of hand gestures comprising the American Sign Language (ASL). ASL is a complex and nuanced visual language, primarily relying on hand movements, facial expressions, and body postures to convey meanings and concepts. Each gesture within ASL holds significance, representing letters, words, phrases, or even abstract ideas. The gestures showcased in Fig 1 encompass a diverse range of hand configurations and movements, each assigned specific meanings within the grammar and lexicon of ASL. Through an intricate combination of handshapes, orientations, and motions, ASL users can express a wide spectrum of linguistic and contextual information, forming the basis for communication within the deaf and hard-of-hearing community. This visual representation serves as a fundamental reference for understanding the foundational elements of ASL and provides a starting point for the development of technologies aimed at interpreting and translating these gestures into spoken language, fostering enhanced communication between ASL users and individuals reliant on spoken communication.

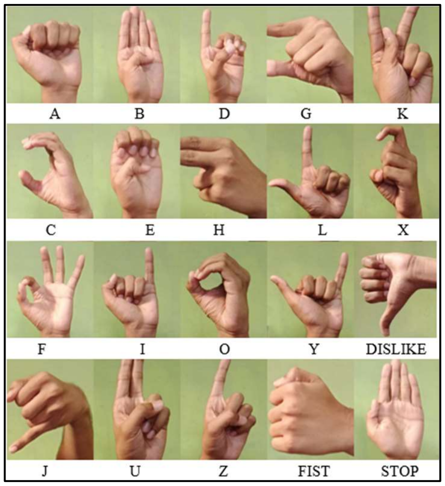


Fig. 1. Hand Gestures

The process for developing a sign language to speech converter involved the following steps:

- Create a List of American Sign Gestures: Define a comprehensive list of American Sign Language gestures, representing various letters, numbers, or words.
- Import OpenCV & MediaPipe Libraries: Utilize the OpenCV and MediaPipe libraries in your programming environment. These libraries provide essential tools for image processing, computer vision, and hand tracking.
- Apply Hand Tracking Algorithm using Hand Class: Implement a hand tracking algorithm using the Hand class from the MediaPipe library. This algorithm tracks the movement and position of the user's hand in real-time through the webcam.
- Create an Array with Fingertips and Thumb: Establish an array to store the coordinates of fingertips and the thumb detected by the hand tracking algorithm. This array forms the basis for analyzing the hand gestures.
- Write Logic for Each Sign using If-Else Statements: Employ if-else statements to create logic for each sign in the list. Based on the coordinates and directions of the fingertips and thumb, define specific conditions that correspond to each sign gesture.
- Print the Name of the Sign Gesture: When a matching sign gesture is identified through the logic, print the corresponding name of the sign gesture. This step ensures that the system accurately recognizes and interprets the user's sign language input.
- Apply API to Convert Text into Speech: Integrate a Text-to-Speech (TTS) API to convert the identified sign language gesture into spoken words. This API transforms the printed name of the sign gesture into audible speech, enabling communication for individuals with hearing impairments.

The combination of hand tracking, logical conditions, and TTS API integration enables the system to interpret sign language gestures in real-time and provide spoken output. This comprehensive approach ensures that users can effectively communicate using sign language while the system translates their gestures into audible speech, fostering inclusivity and accessibility.

The user interaction process with the sign language to speech converter involves the following sequential steps:

- User Performs Hand Gestures: The interaction begins with the user making hand gestures in front of the web camera. These gestures correspond to American Sign Language letters, numbers, or words.
- Image Capture for Image Processing: The system captures images of the user's hand gestures through the webcam. This step is crucial for obtaining visual data that will undergo image processing.
- Image Processing: Employ image processing techniques to analyze and interpret the captured images of hand gestures. This involves using computer vision algorithms to identify key features such as fingertips, thumb, and the overall hand configuration.
- Hand Gesture Classification: After image processing, the system classifies the hand gesture based on the identified features. Each gesture is associated with a specific set of criteria and coordinates, allowing the system to recognize and categorize the sign language input accurately.
- Prediction of Sign Language Meaning: Utilize the classified hand gesture information to predict the intended meaning of the sign language input. This involves mapping the recognized gesture to a pre-defined list of American Sign Language letters, numbers, or words.
- Conversion to Text Format: Once the system predicts the sign language meaning, it converts this interpretation into text format. This step is essential for creating a readable representation of the user's sign language input.
- Text-to-Speech Conversion: Integrate a Text-to-Speech (TTS) mechanism to convert the generated text into spoken words. This ensures that the system can audibly communicate the interpreted meaning of the user's sign language input.

Working of the system is shown in Fig 2.

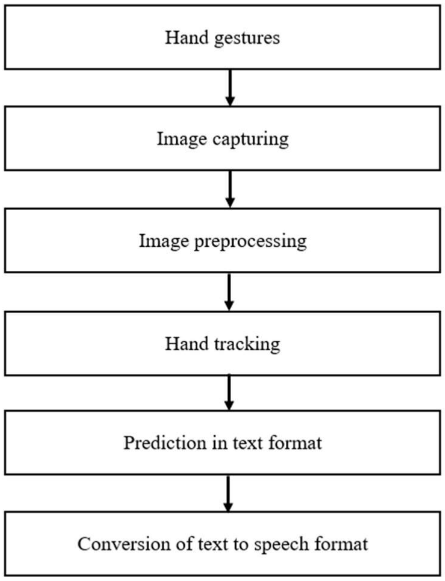

Hand gestures
Image capturing
Image preprocessing
Hand tracking
Prediction in text format
Conversion of text to speech format

Fig 2 - Working of the system

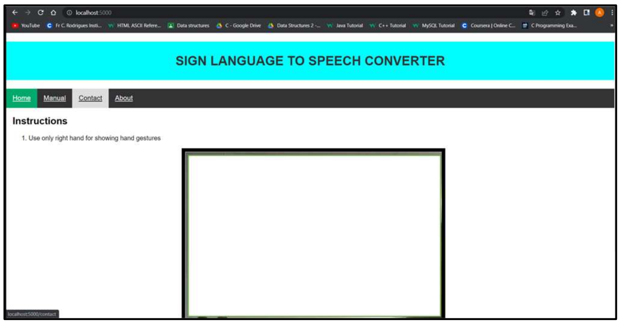
Modules of the system are explained below:

- Image Capturing Module: This module is responsible for recording the gestures using a webcam or any other camera source. It captures real-time images of hand movements for further processing.
- Preprocessing Module: Once the images are captured, they are resized, converted into grayscale, and subjected to noise reduction. This preprocessing enhances the accuracy of the gesture recognition system.
- Hand Tracking Module: MediaPipe's hand tracking algorithm is used to detect and track hand movements in real time. It identifies key points on the hand to form a skeleton model that helps predict gestures.
- Gesture Recognition Module: Based on the tracked hand skeleton, gestures are compared against a pre-trained machine learning model to classify them into specific signs.
- Text Conversion Module: The recognized gestures are then converted into corresponding text, displayed on the screen for the user.
- Speech Output Module: After converting the gestures into text, a text-to-speech engine translates the text into speech for auditory communication.

IV. Result

The GUI of the system is shown in Fig 3. The graphical user interface (GUI) of the system is designed with three main sections: Home, Contact, and About, each serving specific functionalities.

- The home section features the option to activate the web camera, allowing users to record and interpret their hand gestures in real-time.
- Users can switch on the camera to engage with the system, capturing their sign language gestures for translation.
- The Home page is the interactive space where the primary functionality of translating sign language into text occurs.
- The Contact page provides information about the development team behind the system.
- Users can find contact details for reaching out to the team, which could include email addresses, phone numbers, or other relevant communication channels.
- This section aims to establish a connection between users and the development team, facilitating communication and support.
- The About page offers insights into the motivation behind the system and details about the development team.
- It may explain the inspiration or necessity that led to the creation of the system, emphasizing the problem it addresses.
- Users can gain a deeper understanding of the system's purpose, objectives, and the vision of the development team.
- This section might also showcase the team members, their expertise, and their roles in creating the system.



SIGN LANGUAGE TO SPEECH CONVERTER

Instructions

1. Use only right hand for showing hand gestures

Fig 3 – GUI of the system

Fig 4 depicts various examples illustrating the sign language to speech converter's capability to accurately interpret diverse hand gestures. the system successfully demonstrated real-time translation of sign language gestures into text and speech. The home interface allows users to interact with the system seamlessly, while the contact and about sections provide additional information. Throughout the test cases, the system consistently captured and processed hand gestures, converting them into text and speech with minimal lag. In various test examples, the system translated gestures such as "Stop," "Like," and "Dislike" into their corresponding text outputs. These examples illustrate the robustness of the hand tracking and gesture recognition algorithms, which performed well even in varying lighting conditions. The text-to-speech conversion produced clear and understandable speech, bridging the communication gap effectively.

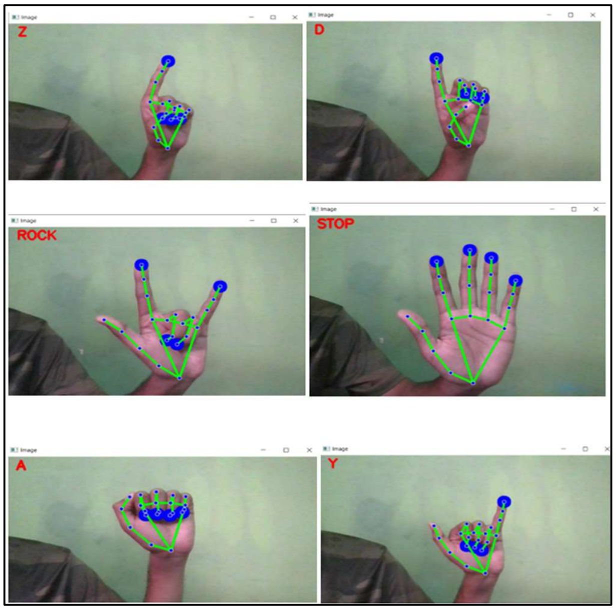


Fig 4 - Hand gesture recognition images

V. Conclusion

The development of a sign language to speech converter holds paramount significance in its commitment to aiding individuals with hearing disabilities, ensuring social relevance, and championing equal rights. This system facilitates effective communication, breaking down barriers that may impede interaction for those with hearing impairments. Its user-friendly design and intuitive interface cater to users with varying technical expertise, fostering a seamless and accessible communication experience. The converter's cost-effectiveness ensures widespread accessibility, empowering individuals without the need for expensive technology. By promoting inclusivity and bridging communication gaps, the system emerges as a valuable tool for enhancing the lives of individuals with hearing disabilities, fostering equal opportunities, and contributing to a more understanding and supportive society.

Acknowledgment

I sincerely thank Smita Rukhande and Lakshmi Gadhikar for their invaluable guidance and support throughout this research. I am also grateful to my colleagues for their collaboration and encouragement, and to my family and loved ones for their unwavering support and understanding during this journey.

References

[1] Aung San Oo, Min Yan Naing, Itthisek Nilkhamhang, Thida Than "Experiment on Real-Time Image Processing in the Controlling of Mecanum Wheel Robotic Car", Published in: 2019 First International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP)

[2] Gerardo Orellana, Belen Arias, Marcos Orellana, Victor Saquicela, Fernando Baculima, Nelson Piedra "A Study on the Impact of Pre-Processing Techniques in Spanish and English Text Classification over Short and Large Text Documents", Published in: 2018 International Conference on Information Systems and Computer Science (INCISCOS)

[3] Chao Ma, Shuo Xu, Xianyong Yi, Linyi Li, Chenglong Yu "Research on Image Classification Method Based on DCNN", Published in: 2020 International Conference on Computer Engineering and Application (ICCEA)

[4] Xu Dazhan, Wu Xiaoyu, Sun Guoquan "Image Memorability Prediction Based on Machine Learning", Published in: 2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology (CCET)

[5] Maliha Khan, Sudeshna Chakraborty, Rani Astya, Shaveta Khepra "Face Detection and Recognition Using OpenCV", Published in: 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)

[6] Chenyu Liu, Zhijun Li, Chengyao Zhang, Yufei Yan, Rui Zhang "An Improved Hand Tracking Algorithm for Chinese Sign Language Recognition", Published in: 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)

[7] Kohsheen Tiku , Jayshree Maloo , Aishwarya Ramesh and Indra R. "Real-time Conversion of Sign Language to Text and Speech", Published in: 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)

[8] C.Uma Bharathi , G. Ragavi , K. Karthika . "Signtalk: Sign Language to Text and Speech Conversion", Published in: 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)

[9] C. U. Om Kumar , K. P. K. Devan , P. Renukadevi , V Balaji , Adarsh Srinivas , R. Krithiga . "Real Time Detection and Conversion of Gestures to Text and Speech to Sign System", Published in: 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)

[10] K Amrutha , P Prabu . "A Comparative Study on Indian Sign Language Representation", Published in: 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)

[11] Vibhu Gupta , Mansi Jain , Garima Aggarwal . " Sign Language to Text for Deaf and Dumb ", Published in: 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)

[12] Chidvika Gunda, Manohar Maddelabanda, Hariharan Shanmugasundaram "Free Hand Text Displaying Through Hand Gestures Using MediaPipe", Published in: 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)

[13] Claudia Audia Trianti; Budhi Kristianto; Hendry "Integration of Flask and Python on The Face Recognition Based Attendance System", Published in: 2021 2nd International Conference on Innovative and Creative Information Technology (ICITech)