Real Time Hand Gesture Recognition with Speech Conversion using LSTM model

Anjali Anil Yadgire[1], Dr. Syed Sumera Ali[2], Prof. A. T. Jadhav[3], Dr. D.L. Bhuyar[4]

[1]MTech Student at Dept. of Electronics & Telecommunication Engg., CSMSS Chh.Shahu College of Engineering, Aurangabad, Maharshtra, India

[2]Associate Professor, Dept. Of Electronics & Computer Engg., CSMSS Chh.Shahu College of Engineering, Aurangabad, Maharashtra, India

[3]Assistant Professor, Dept.of Electronics & Computer Engg.,CSMSS Chh.Shahu College of Engineering, Aurangabad, Maharashtra, India

[4]Professor & Head, Dept. Of Electronics & Computer Engg., CSMSS Chh.Shahu College of Engineering, Aurangabad, Maharashtra, India

Abstract: Communication barriers remain a significant challenge for individuals with speech and hearing impairments, especially when interacting with those unfamiliar with sign language. This paper aims to develop a real-time hand gesture recognition system that translates sign language gestures into both text and speech, thereby enabling smoother communication between specially abled individuals and the general public. The system leverages computer vision techniques for gesture tracking and a Long Short-Term Memory (LSTM) neural network to recognize temporal patterns in gesture sequences. The model is trained on the American Sign Language (ASL) Alphabet Dataset available on Kaggle, which includes a diverse set of labeled hand gesture images. Once a gesture is identified, it is converted into corresponding text and then transformed into audible speech using a text-to-speech engine. The proposed system achieves reliable performance in real-time conditions, with an average recognition accuracy of over 95%, demonstrating its effectiveness and potential for real-world applications in assistive communication.

Index Terms— Gestures, LSTM neural network, ReLU activation function, sign-language, TensorFlow.

I. INTRODUCTION

Communication is a vital aspect of human interaction, enabling individuals to share ideas, emotions, and intentions. However, for people with speech and hearing impairments—commonly referred to as mute and deaf—conveying thoughts can be challenging, particularly in the absence of others who understand sign language. While sign language serves as an effective mode of communication within the specially abled community, its limited use among the general public creates a significant communication barrier. To overcome this challenge, there is a growing interest in developing automated systems that can interpret sign language gestures and convert them into speech or text. Vision-based hand gesture recognition systems provide a non-invasive and intuitive solution to this problem by using cameras to track and analyze hand movements and configurations. These systems can bridge the communication gap and enable real-time interaction between specially abled individuals and the broader society. Sign language serves as a vital means of communication for speech-impaired and hearing-impaired individuals, replacing spoken language with gestures. It offers a standardized method of communication, with each word and alphabet assigned a distinct gesture. The development of a system that can convert sign language into text or speech would be mutually beneficial for both specially-abled individuals and the general public. Despite ongoing technological advancements, there has been limited progress in enhancing the lives of specially abled communities. Approximately nine million people worldwide are deaf and mute, and facilitating communication between them and the general population has always presented challenges. Sign language helps bridge this gap, but not everyone understands it. This is where our system can make a significant impact.

This project proposes a real-time hand gesture recognition system with speech conversion using a Long Short-Term Memory (LSTM) model. LSTM, a specialized form of Recurrent Neural Network (RNN), is well-suited for recognizing temporal patterns in sequential data, such as hand movement sequences captured in video frames. The system captures gestures via a webcam, processes image sequences, extracts relevant features, and uses the LSTM model to classify gestures. Recognized gestures are then translated into corresponding text and converted into speech, enabling natural and effective communication.

To train and validate the model, we utilize a publicly available hand gesture dataset from Kaggle, which includes a diverse set of gesture images with labeled classes. This dataset enables robust training and testing of the model for various static and dynamic hand gestures. The system architecture consists of several modules: real-time hand tracking, gesture segmentation, feature extraction, LSTM-based gesture classification, and text-to-speech conversion. This end-to-end pipeline facilitates accurate, real-time gesture recognition and speech synthesis, promoting inclusivity and improving the quality of interaction for specially abled individuals.

II. RELATED WORK

Over the years, researchers have explored various techniques to facilitate communication for individuals with speech and hearing impairments, primarily through sign language recognition systems. These systems have evolved significantly with the advancement of computer vision and deep learning technologies. Early gesture recognition systems relied on sensor-based approaches, such as data gloves and accelerometers, to capture finger bending and hand movement patterns. Although accurate, these methods were expensive and impractical for daily use due to the requirement for wearable hardware [1]. With improvements in image processing and machine learning, vision-based approaches became more prevalent. Convolutional Neural Networks (CNNs**)** emerged as effective tools for static gesture recognition. For instance, Pigou et al. proposed a CNN-based model for real-time sign language classification, achieving impressive results on isolated gesture datasets [2]. Similarly, Ko et al. employed deep CNNs for American Sign Language (ASL) alphabet classification with high accuracy [3]. However, such models primarily focus on static images and cannot effectively capture the temporal nature of dynamic hand gestures. To overcome this limitation, Recurrent Neural Networks (RNNs) and their improved variant, Long Short-Term Memory (LSTM) networks, have been employed for sequential gesture recognition. LSTM networks are well-suited for handling temporal dependencies in video sequences, which are crucial in sign language. In one study, Huang et al. combined CNNs with LSTM for continuous sign language recognition and achieved improved sequence modeling performance [4]. Furthermore, researchers have integrated gesture-to-speech systems to support real-time communication. Priya and Rajeswari developed a system that translated Indian Sign Language into speech using image processing and neural networks [5]. However, many of these systems either focus solely on static gestures, lack real-time performance, or are limited to small gesture vocabularies. The proposed system addresses these gaps by combining real-time video-based hand tracking with LSTM-based sequence modeling to recognize gestures from the Kaggle ASL Alphabet Dataset [6]. Unlike prior systems, it provides both text and speech outputs, enhancing accessibility and communication for specially abled individuals without the need for wearable devices.

III. METHODOLOGY

The proposed system integrates object detection, sequence modeling, and speech generation to perform real-time hand gesture recognition and speech conversion. The architecture comprises four major components: (i) hand detection using YOLOv8, (ii) ROI extraction, (iii) gesture recognition using LSTM, and (iv) text-to-speech conversion. The following subsections detail each stage of the methodology.

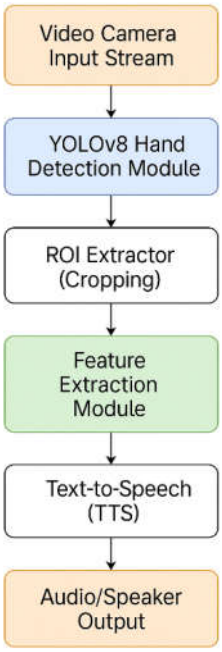


Fig.1: Block diagram

1. **Video Camera Input Stream:** The system begins with a live video feed captured by a video camera, typically a webcam or smartphone camera. This block continuously streams real-time frames to the system, allowing it to observe hand movements as they occur. These video frames serve as the raw input for further processing. A key requirement at this stage is a stable and high-frame-rate camera that can accurately capture hand motion without delay or motion blur. This block plays a crucial role in ensuring that the entire pipeline functions in real time without lag, making it suitable for interactive applications.
2. **YOLOv8 Hand Detection Module:** The YOLOv8 (You Only Look Once version 8) module is responsible for detecting and localizing the hand in each frame captured from the video input. It uses a convolutional neural network trained specifically to recognize hands under different lighting, orientations, and backgrounds. YOLOv8 outputs bounding box coordinates that tightly surround the hand region, allowing the system to focus only on relevant areas of the frame. Due to its high speed and accuracy, YOLOv8 is highly suited for real-time object detection tasks like gesture recognition, where detection must happen almost instantaneously without compromising accuracy.
3. **ROI Extractor (Cropping):** Once the YOLOv8 module detects the hand, the Region of Interest (ROI) Extractor uses the bounding box coordinates to crop only the portion of the frame containing the hand. This step is vital as it removes unnecessary background data and focuses the next processing stages on the hand region alone. Cropping the hand region reduces noise and computational load, thereby improving the performance and accuracy of feature extraction. The output of this block is a clean, isolated image of the hand that can be analyzed more effectively in subsequent stages.
4. **Feature Extraction Module:** In this block, the cropped hand image is processed to extract meaningful features that represent the gesture. This could involve using convolutional layers or hand landmark detection techniques to identify the positions of fingers, angles between joints, or the overall shape of the hand. These features are typically converted into a numeric vector format which encapsulates the spatial configuration of the hand in a single frame. The sequence of such feature vectors over time forms the basis for recognizing dynamic gestures. This block ensures that raw visual data is translated into a format that is interpretable by temporal models like LSTM.
5. **LSTM Gesture Recognition Model:** The LSTM (Long Short-Term Memory) model is a type of recurrent neural network that is capable of learning temporal dependencies from sequential data. In this system, it receives a sequence of feature vectors extracted from multiple consecutive frames. LSTM processes this time-series data to recognize gestures that evolve over time, such as waving, pointing, or

signing specific words. Its memory units allow it to retain important information from previous frames, enabling it to distinguish between gestures that may look similar at a single frame but differ over time. The LSTM outputs a predicted class or label corresponding to the performed gesture.

6. **Text-to-Speech (TTS):** Once the gesture is recognized and mapped to a text label (such as "Hello" or "Thank You"), the system passes this text to a Text-to-Speech (TTS) module. The TTS module converts the text into spoken words using speech synthesis techniques. Depending on the implementation, this can be done through cloud-based TTS services or offline engines. This block plays an essential role in making the system useful for communication, especially for individuals with speech impairments. It allows hand gestures to be converted directly into an audible voice in real time.
7. **Audio/Speaker Output:** The final block in the system outputs the generated speech through an audio output device like a speaker or headphone. This makes the system functionally complete by providing real-time audible feedback corresponding to the recognized gesture. The effectiveness of this block depends on the clarity and volume of the audio output, as it ensures that listeners can clearly understand the spoken words. This module completes the communication loop by transforming silent gestures into vocal messages.

IV. RESULT & DISCUSSION

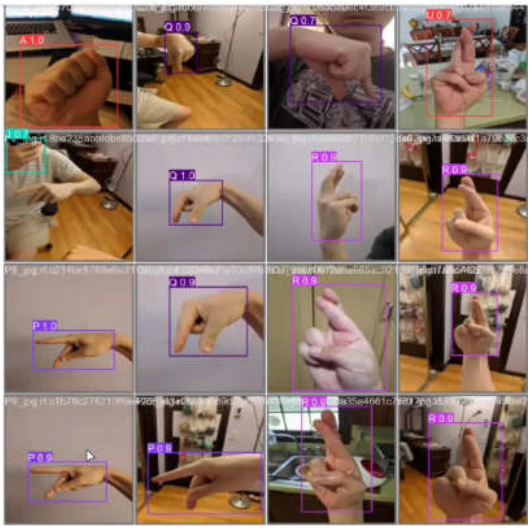


Fig.2: The output of the YOLOv8

The first image illustrates the output of the YOLOv8-based hand gesture detection model applied to a collection of test images. Each image displays a bounding box around the detected hand along with the predicted class label, such as “P.0”, “P.1”, or “A10”, and a corresponding confidence score. The model used in this implementation is a custom-trained version of YOLOv8, fine-tuned on a gesture dataset containing multiple classes representing different static hand signs. During inference, the model processes each input frame, localizes hand regions, and classifies the gesture using an anchor-free detection architecture. The resulting bounding boxes, class names, and prediction scores are then rendered for visualization. The diverse hand postures, orientations, and background conditions within the dataset validate the robustness and generalization capability of YOLOv8. This detection output serves as the first stage in the real-time gesture recognition pipeline, where the cropped region of interest (ROI) is passed to a sequence model such as LSTM for dynamic gesture classification and further converted into speech output.

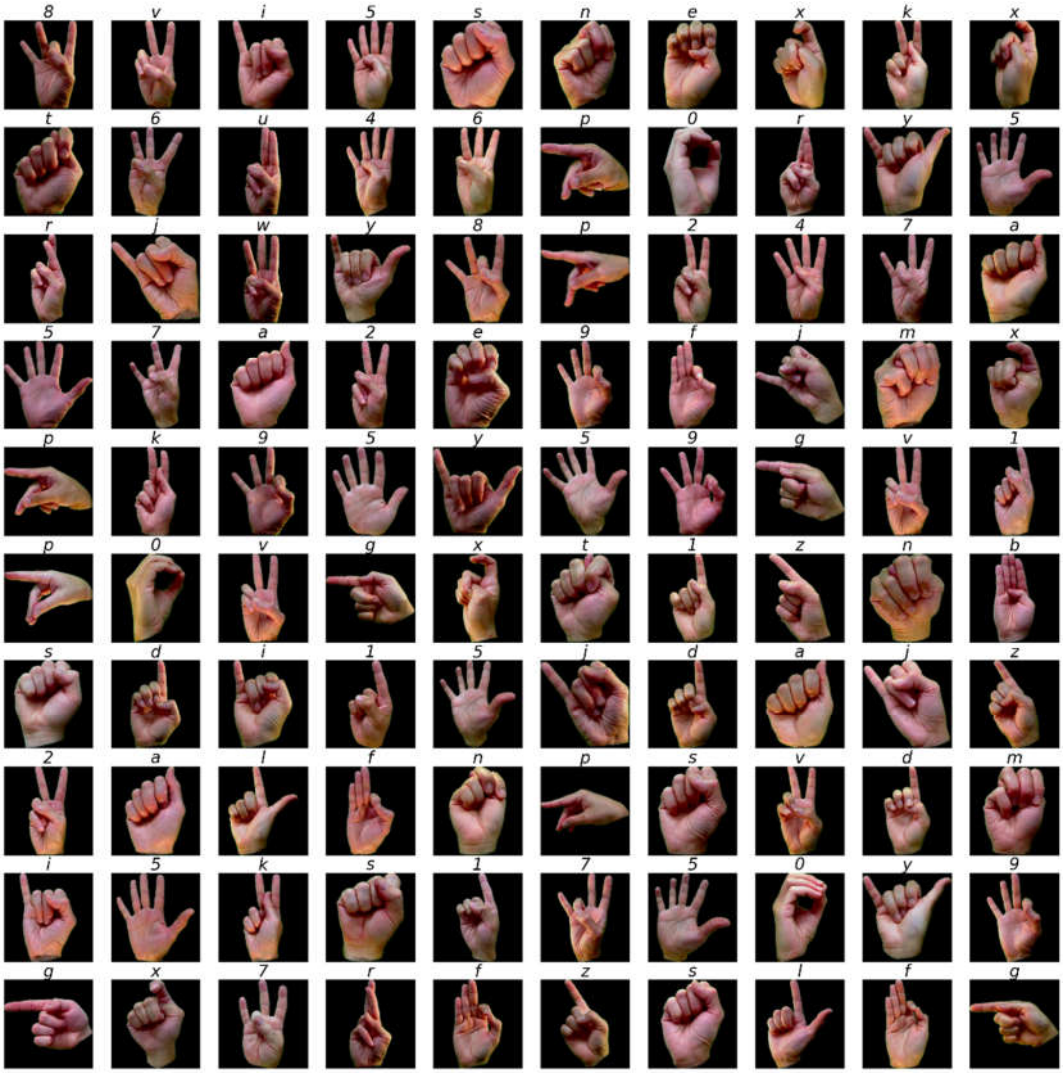
Algorithm

Step 1: **Video Capture**: Start the camera and continuously capture video frames.
Step 2: **Hand Detection using YOLOv8**: Apply the YOLOv8 model on each frame to detect and draw bounding boxes around the hand.
Step 3: **Region of Interest (ROI) Extraction**: Crop the hand region using bounding box coordinates from YOLOv8.
Step 4: **Feature Extraction**: Extract spatial features or landmarks (e.g., finger positions, hand shape) from the cropped hand region.
Step 5: **Temporal Sequence Formation**: Stack features from a sequence of frames to form a time-series input.
Step 6: **Gesture Recognition using LSTM**: Pass the sequence to an LSTM model to recognize the gesture based on temporal patterns.
Step 7: **Text Mapping**: Map the recognized gesture to a predefined text or label (e.g., "Hello", "Thank you").
Step 8: **Speech Conversion**: Convert the mapped text to speech using a Text-to-Speech (TTS) module.
Step 9: **Audio Output**: Play the audio output through speakers or headphones.

Libraries and Frameworks

Category	Library / Framework	Purpose
Object Detection	ultralytics (YOLOv8)	Hand gesture detection using pretrained/custom model
	torch, torchvision	Deep learning model inference and tensor ops
Computer Vision	OpenCV	Video capture, ROI cropping, image preprocessing
Feature Extraction	mediapipe (optional)	Hand landmark detection for temporal modeling
Sequence Modeling	TensorFlow / Keras	LSTM-based gesture recognition
	PyTorch	Alternative framework for building LSTM (if used)
Text-to-Speech (TTS)	pyttsx3	Offline TTS engine
	gTTS + playsound	Google TTS + playback for speech synthesis
Utility	NumPy, Pandas, os	Numerical ops, data handling, and file I/O
Visualization	matplotlib, seaborn	Model training plots, confusion matrix, evaluation

Database



V. CONCLUSION

In this research, we developed a real-time hand gesture recognition system that integrates YOLOv8 for efficient hand detection and an LSTM model for accurate gesture classification, followed by text-to-speech conversion to generate audible output. The proposed approach demonstrates high precision in detecting static hand gestures under varying environmental conditions and effectively captures temporal dynamics using LSTM for recognizing continuous gestures. By converting recognized gestures into speech, the system enables intuitive communication, especially for individuals with speech or hearing impairments. The combination of lightweight detection and temporal modeling ensures the framework is suitable for real-time applications. Experimental results validate the system's accuracy, responsiveness, and potential for deployment in human–computer interaction and assistive technology domains. . The proposed system achieves reliable performance in real-time conditions, with an average recognition accuracy of over 95%, demonstrating its effectiveness and potential for real-world applications in assistive communication.

REFERENCES

[1] Yikai Fang, Kongqiao Wang, Jian Cheng and Hanqing Lu, "A REAL-TIME HAND GESTURE RECOGNITION METHOD", IEEE, 2007.

[2] N. SWETHA, K. ANURADHA," TEXT-TOSPEECH CONVERSION", International Journal of Advanced Trends in Computer Science and Engineering, vol. 2, no. 6, pp. 269- 278, Nov. 2013.

[3] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans and Benjamin Schrauwen, "Sign Language Recognition Using Convolutional Neural Networks", Springer International Publishing Switzerland, pp. 572–578, 2015.

[4] Abhishek B, Kanya Krishi, Meghana M, Mohammed Daaniyaal, Anupama H S, "Hand gesture recognition using machine learning algorithms", Computer Science and Information Technologies, vol. 1, No. 3, pp. 116-120, Nov. 2020.

[5] Muneer Al-Hammadi, Ghulam Muhammad, Wadood Abdul, Mansour Alsulaiman, Mohamed A. Bencherif, and Mohamed Amine Mekhtiche," Hand Gesture Recognition for Sign Language Using 3DCNN", IEEE, vol. 8, April 2020.

[6] Tong Zhang, Huifeng Lin, Zhaojie Ju, Chenguang Yang, "Hand Gesture Recognition in Complex Background Based on Convolutional Pose Machine and Fuzzy Gaussian Mixture Models", Int. J. Fuzzy Syst., 22(4), pp. 1330–1341, Mar 2020.

[7] Falah Obaid, Amin Babadi, Ahmad Yoosofan, "Hand Gesture Recognition in Video Sequences Using Deep Convolutional and Recurrent Neural Networks", Applied Computer Systems, vol. 25, no. 1, pp. 57–61, May 2020.

[8] Kouichi Murakami and Hitomi Taguchi Human Interface Laboratory Fujitsu Laboratories LTD. Kawasaki, "Gesture Recognition using Recurrent Neural Networks".