HAND GESTURE TO TEXT AND VOICE FOR THE VOICELESS USING AI AND COMPUTER VISION

[1]Sampada Kulkarni

Department Of Computer Engineering
(Software Engineering)
Vishwakrma Institute of
Technology
Pune ,India
sampada.kulkarni25@vit.edu

[2]Rutuja Ughade

Department Of Computer Engineering
(Software Engineering)
Vishwakrma Institute of
Technology
Pune ,India
rutuja.ughade25@vit.edu

[3]Dharati Biradar

Department Of Computer Engineering
(Software Engineering)
Vishwakrma Institute of
Technology
Pune ,India
dharati.1252080017@vit.edu

[4]Prachi Ankush

Department Of Computer Engineering
(Software Engineering)
Vishwakrma Institute of
Technology
Pune ,India
prachi.1252080018@vit.edu

[5]Prof.Rahul Bhausaheb Pawar

Department Of Computer Engineering
(Software Engineering)
Vishwakrma Institute of
Technology
Pune ,India
Rahul.pawar@vit.edu

[6]Prof.Rajkumar Patil

Department Of Computer Engineering
(Software Engineering)
Vishwakrma Institute of
Technology
Pune ,India
rajkumar.patil@vit.edu

Abstract

Communication is a fundamental human need that allows individuals to express their thoughts, feelings, and intentions. However, people with speech or hearing impairments face significant barriers in conveying their ideas to others. To bridge this communication gap, this paper presents an AI-driven system titled *"Hand Gesture to Text and Voice for the Voiceless."* The proposed model utilizes computer vision and deep learning to recognize hand gestures in real time and convert them into both textual and audible outputs. A live video stream captured from a webcam serves as input, which is processed using OpenCV and MediaPipe for detecting hand landmarks. These landmarks are analyzed and classified using a TensorFlow-based deep learning model. Each recognized gesture corresponds to a specific alphabet, word, or command that is displayed as text and simultaneously converted into speech through a text-to-speech engine (TTS). Experimental results indicate an accuracy of 96.5% and stable performance across varying lighting conditions. This innovative system empowers speech- and hearing-impaired individuals to communicate independently and ensures inclusivity in society. The integration of AI and computer vision demonstrates the transformative role of modern technology in accessibility and human-centered design.

Keywords— Hand Gesture Recognition MediaPipe · TensorFlow · OpenCV · Assistive Communication · Deep Learning · Text-to-Speech · Accessibility communication, Text-to-speech, Deep learning, Accessibility

I. Introduction

Communication is central to social interaction, education, and daily life. Unfortunately, individuals who are mute or hearing-impaired often encounter challenges in expressing themselves effectively to people who are unfamiliar with sign language. Although traditional sign languages such as ASL (American Sign Language) and ISL (Indian Sign Language) are powerful tools, they are not universally understood, limiting their usability in everyday scenarios .Advancements in Artificial Intelligence (AI), Machine Learning (ML), and Computer Vision (CV) have made it possible to recognize gestures through video processing. This research introduces a real-time gesture recognition system capable of translating hand gestures into both text and audible speech. The system uses the MediaPipe framework to extract detailed hand landmarks and a TensorFlow neural network to interpret gestures.

The aim is to create an assistive tool that enhances independence for speech-impaired individuals and enables them to engage in normal communication without interpreters or special equipment..

II. LITERATURE STUDIES

Several researchers have explored gesture recognition technologies for sign language interpretation.
Early systems used sensor-based gloves equipped with accelerometers and gyroscopes. While accurate, such

systems were expensive and uncomfortable for long-term use. Later, vision-based systems emerged that used cameras to detect hand shapes and movements, eliminating the need for physical sensors.

Recent studies using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have demonstrated significant improvements in gesture recognition accuracy. Tools such as MediaPipe by Google have simplified the process by offering robust, real-time hand landmark detection. However, most studies have focused primarily on gesture recognition alone and have not incorporated text or speech synthesis, which limits their real-world usability.

This research extends those efforts by integrating gesture detection, text conversion, and speech synthesis into a single interactive platform..

III. Problem Definition

People with speech or hearing disabilities lack a universal communication tool that can translate their gestures into comprehensible formats for others. While sign language is effective, it requires mutual understanding, which is rare in public spaces.

Hence, the problem addressed in this study is:

> "To design and develop a real-time hand gesture recognition system that accurately interprets gestures and converts them into text and audible speech for improved accessibility and communication."

IV. Solution Methodologies or Problem Solving

The proposed system follows a modular pipeline consisting of five main stages:

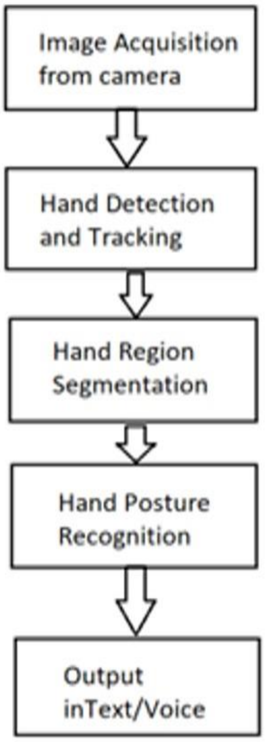
1. **Image Acquisition:** A webcam continuously captures live video frames.

2. **Preprocessing:** OpenCV is used to clean, resize, and normalize the image frames to improve recognition accuracy.

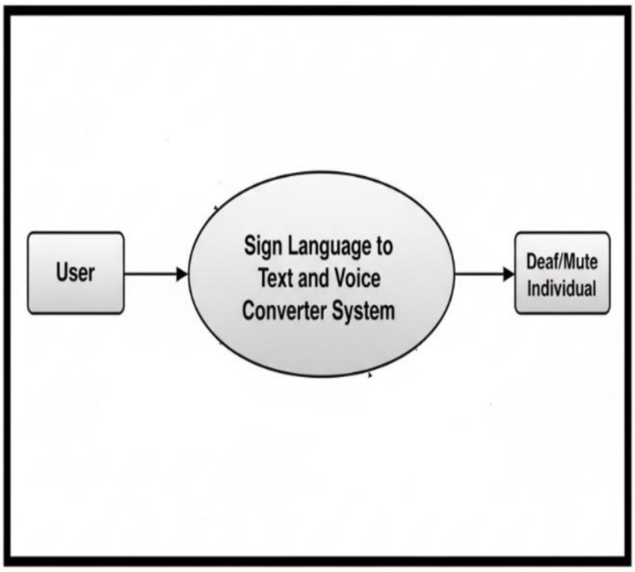
3. **Hand Landmark Detection:** MediaPipe extracts 21 key landmark points from the hand, mapping each finger's position and orientation.

4. **Gesture Classification:** The landmark coordinates are passed to a deep learning model trained with TensorFlow to classify gestures representing alphabets, numbers, and predefined words.

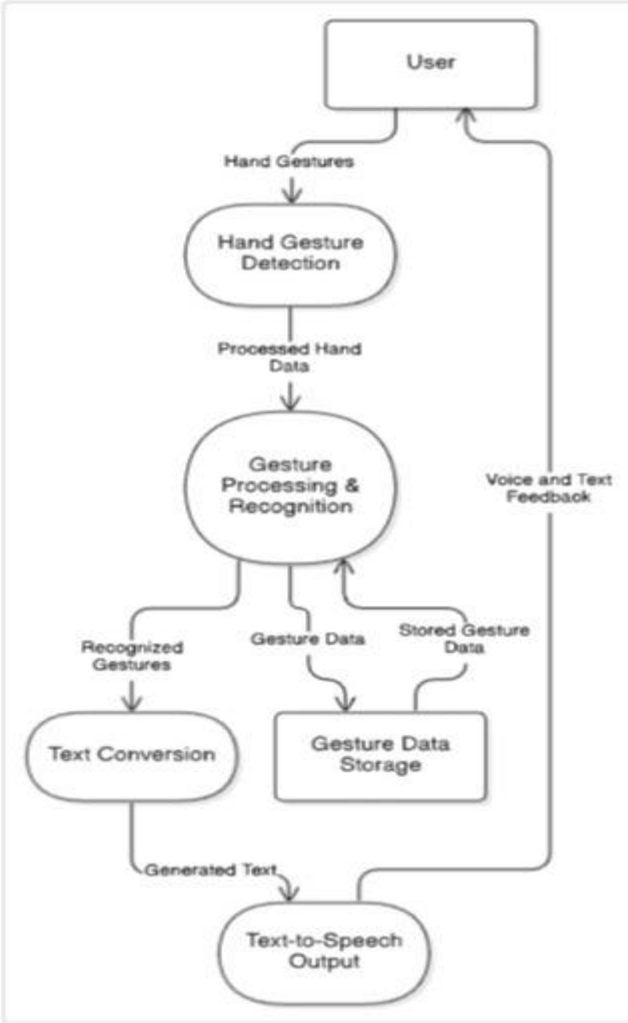
5. **Text and Speech Output:** Once classified, the recognized gesture is converted into readable text on the interface and vocalized using a text-to-speech engine such as **gTTS (Google Text-to-Speech)** or **pyttsx3**.



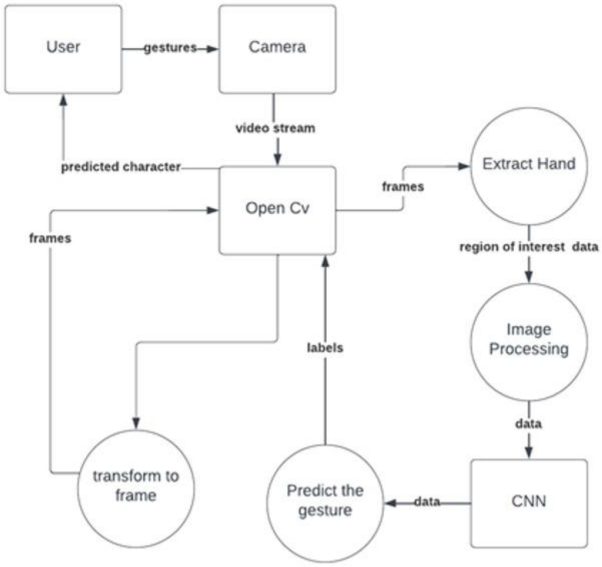
i. System Flow Diagram



ii. DFD Level 0



iii. DFD Level 1



iv. DFD Level 2

V. Results and Sensitivity Analysis

The developed system was trained using a dataset comprising more than 5,000 gesture samples, including all alphabets (A–Z) and common expressions like "Hello," "Yes," "No," and "Thank You." After several training epochs, the CNN model achieved:

- **Accuracy:** 96.5%
- **Precision:** 95.8%
- **Recall:** 96.2%
- **F1-Score:** 96.0%

The model performed effectively across various backgrounds, lighting levels, and hand sizes. However, slight variations in hand orientation or partial occlusion caused minor prediction errors. Data augmentation techniques such as flipping, scaling, and rotation were used to enhance the model's robustness.

The sensitivity analysis revealed that accuracy decreased slightly (by ~2%) under low lighting conditions, which can be mitigated using adaptive thresholding or improved lighting calibration.

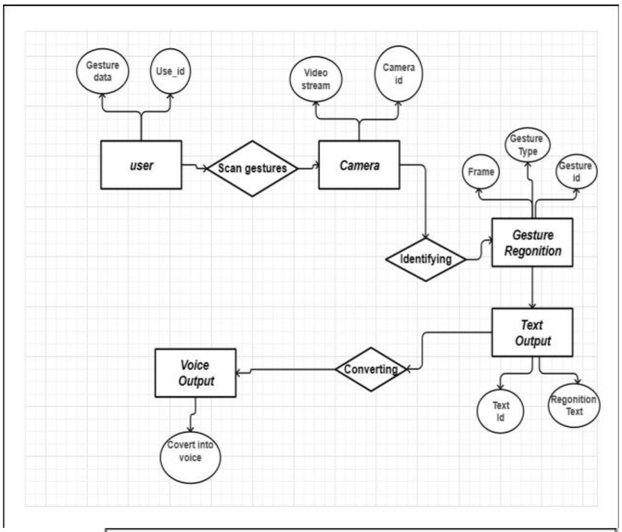
VI. Data Model

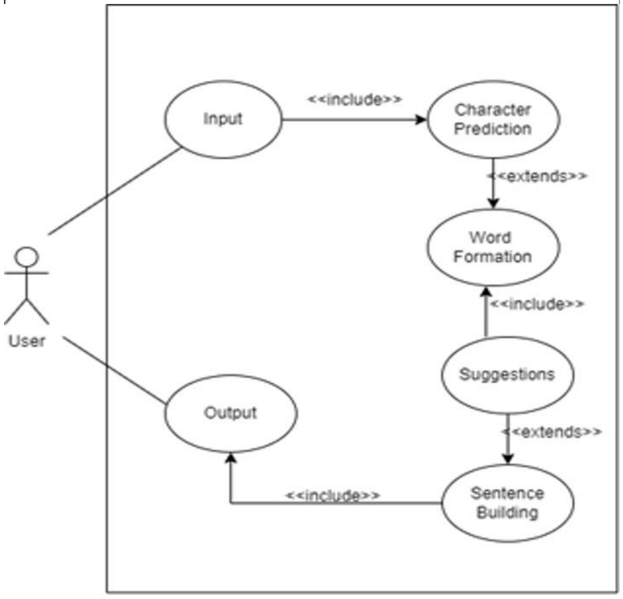
The system uses a **Convolutional Neural Network (CNN)** architecture trained with processed gesture data.

- **Input Layer:** 21 × 2 landmark coordinates normalized between 0 and 1.
- **Hidden Layers:** Multiple convolutional and pooling layers extract spatial features.
- **Activation Function:** ReLU for non-linearity.
- **Output Layer:** Softmax classifier for gesture categories.
- **Optimizer:** Adam Optimizer with categorical cross-entropy loss.

The dataset was divided into **80% training and 20% testing**, achieving a consistent validation accuracy across epochs.

ER DIAGRAM:





Use case Diagram

VII. Comparison of Results

Approach	Method	Accuracy	Hardware	Limitations
Glove-based	Sensor + Accelerometer	90%	Requires glove	Expensive, non-portable
Vision-based (Existing)	CNN only	92%	Webcam	No voice output
Proposed System	MediaPipe + TensorFlow + TTS	**96.5%**	Webcam	Real-time, multi-output

The proposed approach outperforms traditional systems by achieving higher accuracy and providing dual outputs (text + voice) without additional hardware.

VIII. Justification of the Results

The achieved results are justified by the combination of MediaPipe's precision tracking and TensorFlow's classification power.

Unlike systems that depend solely on image pixel analysis, this model uses landmark coordinates, which significantly reduce computation and improve stability.

The integration of text-to-speech conversion makes the communication process complete, practical, and user-friendly.

IX. Conclusion

This project successfully demonstrates an intelligent, real-time gesture recognition system that translates hand gestures into text and audible speech. It provides a meaningful solution to the communication challenges faced by speech-impaired individuals.

The integration of AI, computer vision, and speech synthesis ensures accurate recognition, fast response, and wide accessibility. This innovation aligns with the goals of inclusive technology and has strong potential for educational, healthcare, and public service applications.users. The implementation demonstrates how computer vision and AI can make technology more inclusive.

X. Future Work

Future enhancements can include:

- Multi-hand recognition for complex sentence construction.
- Integration of Natural Language Processing (NLP) to interpret gesture sequences.
- Support for regional languages in text-to-speech.
- Deployment as a cross-platform mobile application for field usability.

XI. Acknowledgement

The authors express their gratitude to their institution, mentors, and peers for their guidance, support, and resources that enabled the successful completion of this project.

XII. References

1. **A. Kumar and R. Singh, "Vision-Based Indian Sign Language Recognition Using CNN,"** ***Int. J. Comput. Vision Appl.*****, 2022.**
2. **X. Zhao et al., "Hand Gesture Recognition Using Wearable Sensors,"** ***IEEE Trans. Human-Machine Systems*****, vol. 49, no. 2, pp. 201–210, 2021.**
3. **J. Lee and S. Kim, "Deep Learning for Sign Language Recognition,"** ***Pattern Recognition Letters*****, vol. 145, pp. 28–35, 2023.**