IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Gesture Based Text and Voice Output System

Kunika Patil, Sakshi Ladake, Vaishnavi Naphade, Shraddha Nirgude

Dr. Praveen Blessington Thummalakunta

[1]Student, Information Technology, Zeal College of Engineering and Research, Pune, Maharashtra, India

[2]Student, Information Technology, Zeal College of Engineering and Research, Pune, Maharashtra, India

[3]Student, Information Technology, Zeal College of Engineering and Research, Pune, Maharashtra, India

[4]Student, Information Technology, Zeal College of Engineering and Research, Pune, Maharashtra, India

[5]Prof., Information Technology, Zeal College of Engineering and Research, Pune, Maharashtra, India

Abstract

The increasing demand for inclusive communication solutions has led to significant advancements in hand sign gesture recognition systems. By combining state-of-the-art deep learning and computer vision technologies with Dense Neural Networks (DNNs), this work provides a thorough foundation for translating hand sign gestures into speech and text. By integrating MediaPipe for real-time hand tracking and landmark identification with OpenCV, the proposed approach guarantees precise gesture recognition. TensorFlow and Keras are used to create and train a powerful DNN that can accurately recognize gestures. Google Text-to-Speech (GTTS) is then used to convert the motions into natural-sounding speech. A user-friendly graphical user interface that promotes accessibility and seamless interaction was developed using the Tkinter (Tk) toolkit. To extract hand landmarks, the approach preprocesses input video streams using MediaPipe and OpenCV. These landmarks are normalized before being added to the DNN model for classification. The resulting text output is shown on the interface and concurrently converted into speech using GTTS. Experiments reveal that the system can accurately identify both static and dynamic motions in a range of lighting and background conditions. This research contributes to bridging the communication gap for those with speech and hearing impairments by providing a portable, real-time, and cost-effective option for sign language interpretation.

Keywords:- Deep Learning, Dense Neural Network (DNN), MediaPipe, Google Text-to-Speech(gTTS)

1. Introduction

Effective communication is essential to human connection because it enables people to effectively convey their objectives, feelings, and thoughts. People with speech or hearing impairments, on the other hand, may find it very challenging to engage in everyday conversations, which can result in social isolation and limited access to essential services. These individuals mostly communicate through sign language, however there is a big communication gap because most people do not comprehend sign language. To bridge this gap, cutting-edge technical tools that can quickly convert hand sign gestures into speech and text are required.

The proposed framework uses OpenCV and MediaPipe for real-time hand tracking and landmark detection, which guarantees accurate gesture identification even in dynamic environments. These landmarks are processed and classified using a Dense Neural Network (DNN) model built using TensorFlow and Keras, which produces very high gesture recognition accuracy. To enhance accessibility and provide users with a more inclusive communication experience, the recognized gestures are translated into text and then into speech using Google Text-to-Speech (GTTS).

Additionally, a graphical user interface (GUI) developed using the Tkinter (Tk) toolkit is integrated into the system to enable seamless user interaction with the application.

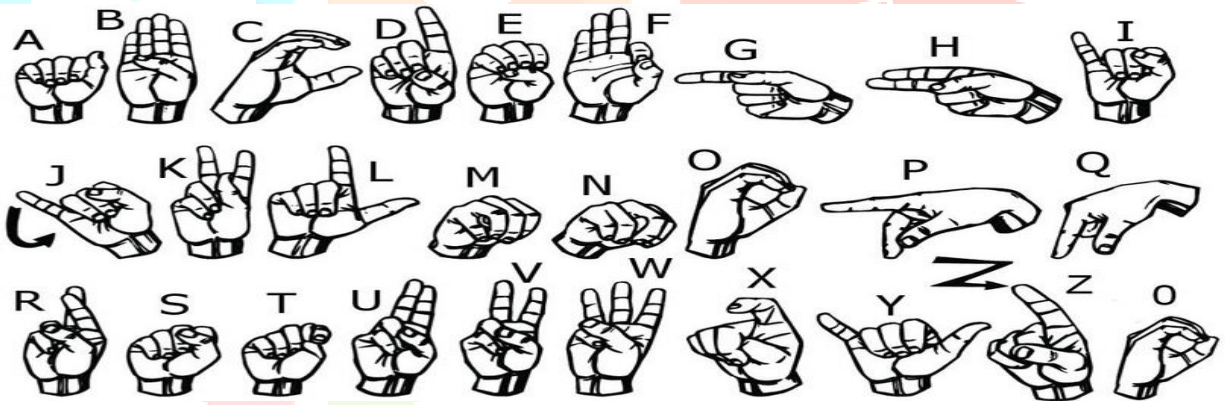


Fig 1.1 Hand Gesture Signs

The biggest challenge in converting hand gestures into speech and writing is achieving high accuracy. Human gestures vary widely due to individual differences, ambient factors, and dynamic motions. This variation, together with potential noise and occlusions, makes it difficult for systems to reliably interpret and translate hand motions into meaningful output. Furthermore, the need for large, accurately labeled datasets makes it difficult to develop robust machine learning models, and real-time processing requires efficient algorithms to handle the continuous flow of data. To overcome these challenges, better methods for gathering and processing data, more advanced real-time processing capabilities, and advancements in machine learning techniques are also required.

The goal of the proposed project is to develop a system that can convert hand gestures into spoken and written language in order to enhance communication for those who have speech or hearing problems. This research will explore state-of-the-art gesture recognition systems that employ computer vision and deep learning approaches to accurately identify and interpret hand gestures in real-time. By integrating speech synthesis methods with natural language processing (NLP) for text production, the system would easily convert gestures into meaningful text and audio outputs. This innovative approach aims to remove barriers to communication and provide an inclusive, user-friendly tool that may be applied in industries including education, healthcare, and public services. The study will advance assistive technology and human-computer interaction by addressing problems like contextual interpretation and real-time

To bridge the communication gap between the deaf and hard-of-hearing community and non-sign language users, research on hand sign gesture translation to speech and text is being conducted. Through the creation of more inclusive settings in a range of settings, the improvement of communication through real-time interaction, and the development of assistive technology such as learning aids and real-time translation devices, this field of study seeks to increase accessibility. This study's significance is summarized in the following aims, which include facilitating smooth communication in a range of contexts, such as public services, healthcare, and education. Facilitate real-time communication between individuals who do not understand sign language and others who have hearing problems to create more inclusive environments. In emergency situations, it is useful and easy to use.

2. Literature Review:

The application of deep learning models—specifically, Convolutional Neural Networks (CNNs)—has shown significant progress in a number of fields, such as medical image processing, natural language processing, and visual object recognition, showcasing their versatility and effectiveness in handling complex data [1].

A novel model called Global Average Pooling Residual Network (G-ResNet) has been introduced for the classification of brain tumor images using the ResNet34 architecture, and it has a 95.00% classification accuracy. By using a global average pooling layer to reduce parameters and prevent overfitting, this model shows gains in medical picture classification [2].

In the field of remote sensing, a method for scene-level classification of high spatial resolution photographs has been created by extracting both low-level and depth properties using a residual learning network (ResNet). This technique has demonstrated its applicability in real-world scenarios with an amazing 95.71% classification accuracy with a limited number of training samples [3].
Sign language recognition has witnessed significant progress, and the research has examined a variety of methods, including vision-based and data glove-based systems. These techniques demonstrate how important accurate hand gesture recognition is to effective communication, particularly for the hard-of-hearing community [4] [5].

Numerous studies have examined Hidden Markov Models' (HMMs') ability to handle data variations, with applications in statistical modeling. Understanding the characteristics of various HMM types, such as ergodic and left-to-right models, is crucial for appreciating their effectiveness in various contexts [4]. In deep learning and computer vision applications, the combination of TensorFlow and OpenCV has been highlighted; TensorFlow is renowned for its ability to manage large neural networks, while OpenCV is commended for its efficiency in processing large datasets in real time. [1].

Improving communication requires the ability to recognize and interpret hand gestures in sign language. The need for systems that can reliably translate sign language to text is highlighted in the literature [5].

3. Motivation:

Millions of people with speech or hearing impairments struggle every day to adequately express themselves, despite the fact that communication is a basic human need. Despite their effectiveness, traditional nonverbal communication techniques like sign language frequently necessitate the presence of an interpreter or prior knowledge from both parties, which restricts accessibility in real-world encounters. There is increasing potential to close this communication gap with automated gesture identification and translation systems thanks to developments in machine learning. The objective of this project is to convert hand motions into text and speech by utilizing deep learning, more especially Dense Neural Networks (DNNs). The technology can correctly identify gestures in real time after training the model on a dataset of 28 distinct hand signs. Upon detection, the system initially translates the gesture into legible text before generating an audio output in response to a predetermined 'Enter' motion. The output's twin modalities—speech and text—improve inclusion and practicality. This study presents a viable way to increase accessibility and promote more inclusive human-computer interaction by enabling touchless, intuitive communication through gesture recognition. It advances the larger goal of AI-enhanced assistive devices that empower people and remove communication barriers.

4. Proposed System Design:

The block diagram that follows describes a gesture recognition system that uses a Dense Neural Network (DNN) to convert hand gestures into speech and text. The procedure starts with gesture input recorded by the laptop's webcam, which captures hand movements in real time. To increase the input's clarity and quality, the collected image is preprocessed using techniques like background noise reduction and image enhancement. These procedures guarantee that the accuracy of recognition is not hampered by extraneous background components. After preprocessing, the algorithm extracts important hand points, including the fingertips and joints, by performing landmark detection using methods like MediaPipe. These landmarks provide a structured depiction of the hand's location and shape, which is crucial for successful gesture classification.

A trained Dense Neural Network (DNN), which has been trained on 28 distinct hand motions, is then fed the collected landmark data. Using the landmark features, the model determines which particular gesture is being displayed and outputs the relevant text label. The system activates a text-to-speech module to translate the

recognized gesture into audible speech if it detects the designated "Enter" gesture. This enables real-time hand sign communication with both visual and aural input.

This approach provides a smooth transition between written language, spoken communication, and gestures in an effort to help those with speech or hearing problems.

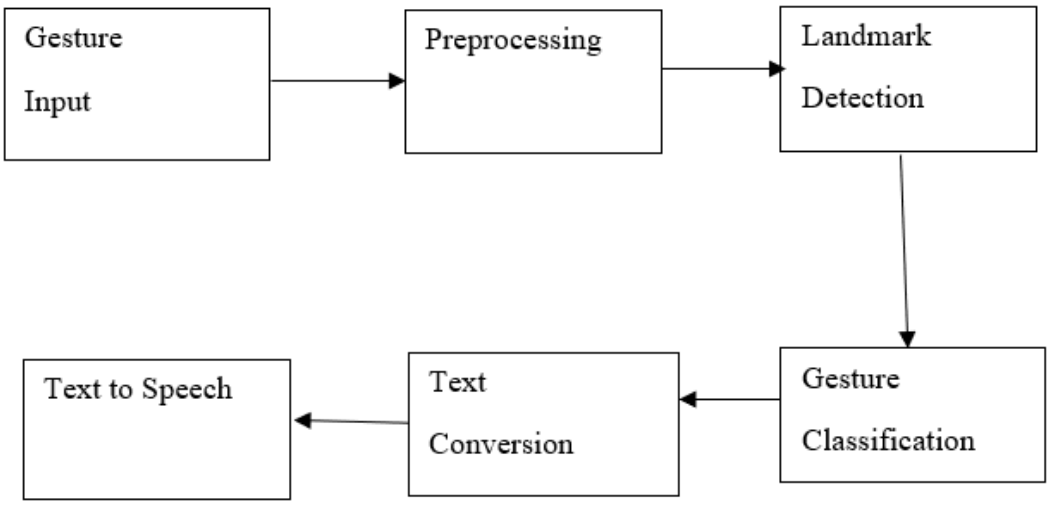
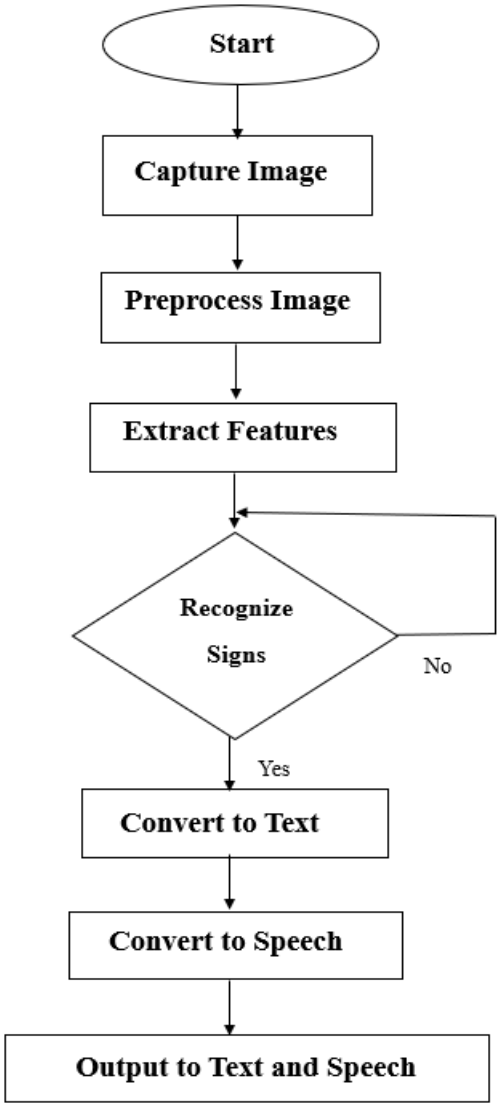


Fig 4.1 Proposed System Flow

❖ **Flowchart**

The process of applying deep learning to translate hand gestures into text and speech is shown in the flowchart above. First, an input gesture is captured by the system's webcam. This gesture is used to distinguish particular signs from a predetermined collection of hand gestures and is the main means of communication. After the gesture is recorded, it goes through a preprocessing step that includes normalization, frame refining, and background noise reduction to make sure that only the hand's pertinent features are preserved. This stage increases the accuracy of gesture recognition and its clarity.

After preprocessing, the system proceeds to the feature extraction stage, where key visual characteristics of the gesture are identified. A landmark identification module receives these features and uses them to identify important hand traits including the palm position, joints, and fingertips. In order to precisely trace the gesture's shape and motion, certain landmarks are necessary.



4.2 Flowchart

Following landmark detection, the system moves on to classifying gestures. The appropriate gesture is submitted for text conversion, where the associated gesture is translated into its text representation, if it is identified. The text is then transformed into audio output by a text-to-speech engine after the user makes a certain "Enter" gesture or a gesture trigger.

The system offers the opportunity to fix a mistake in gesture detection by using the backspace gesture, which removes the preceding entry. Particularly in situations involving real-time communication, this corrective process guarantees that the system stays flexible and easy to use.

The suggested system functions as an assistive technology solution to improve communication between people with speech or hearing impairments by combining gesture recognition, error correction, and multimodal output (text and audio).

5. Methodology:

In this study, we use deep learning techniques, machine learning models, and third-party libraries to create a real-time system that converts hand gestures into speech and text. Each of the system's many parts is essential to the translation and recognition processes. We go into further detail about these elements below.

5.1 Confusion Matrix:

This technique evaluates the performance of the hand gesture recognition model. It makes it easier to see how the model can distinguish between different hand motions. A confusion matrix shows the proportion of accurate (true positive) and inaccurate (false positive and false negative) predictions for each gesture class. This matrix is necessary to comprehend the model's precision, recall, and overall accuracy. For example, the confusion matrix can be used to identify if a misclassified gesture is due to input noise or the model's inability to distinguish between similar movements. Accuracy, precision, and recall—performance metrics derived from the confusion matrix—are used to improve the model.

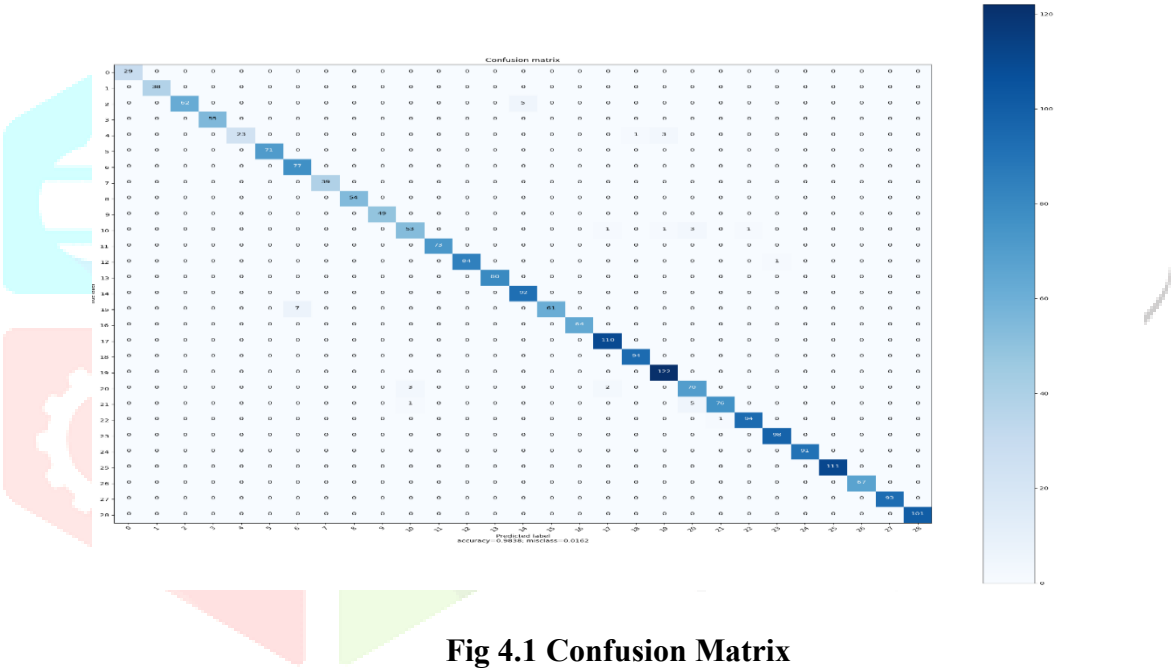


Fig 4.1 Confusion Matrix

5.2 Neural Network:

A neural network serves as the foundation for our gesture recognition model. Examples of input features that the neural network learns to convert to a corresponding output (gesture classification) include key points from hand movements. Each layer of the network processes input data and contains neurons that adjust their weights to lower prediction error during training. Neural networks are perfect for this purpose because they can recognize complex patterns from large datasets, such as the dynamic fluctuations in hand movements. The network is trained using labeled gesture datasets to learn the relationships between the observed input (video frames or gesture images) and the corresponding class labels (specific motions). by adjusting the neural network's weights.

5.3 Dense Neural Network (DNN):

In a dense neural network (DNN), every neuron in one layer is connected to every other neuron in the layer underneath it. Because of its wide connection, the DNN can identify hand motions with very minor modifications, allowing it to identify intricate and detailed patterns in the data. A DNN can handle these challenges in gesture recognition with ease.A set of input variables, such as hand form, location, and motion, are used to represent each gesture. The DNN learns to generate the most likely gesture class based on the input features. Once trained, the DNN can be utilized in real-time systems to classify gestures as they are captured by the camera.

5.4 Google Libraries (gTTS):

The system uses Google libraries such as gTTS (Google Text-to-Speech) to convert the detected text into audible speech. Once the system detects a gesture and links it to a text label (such "hello" or "thank you"), gTTS is utilized to convert this text into spoken words. This enables users, particularly those with hearing impairments, to get real-time spoken input from the system. The gTTS library is easy to use and provides high-quality, natural-sounding voice output in several languages. Other Google libraries, including MediaPipe, are also used for hand tracking and gesture recognition. MediaPipe employs advanced machine learning models to recognize and track hand landmarks (such finger positions) in real time to guarantee that the system accurately recognizes hand gestures.

5.5 Used Dataset:

A pixel value dataset for American Sign Language (ASL) gesture identification is composed of numerical representations of hand gestures rather than raw images and is mostly used for landmark detection. Each dataset entry contains pixel intensity values that help identify key hand landmarks such the fingertips, knuckles, and wrist. Tools like Google's deep learning-based MediaPipe Hands framework, which extracts precise hand key points from pixel data, are frequently used for real-time hand tracking and landmark detection. Additionally, OpenCV, an open-source computer vision toolkit, is used to enrich the dataset through preprocessing tasks like edge detection, contour extraction, and background subtraction.

6. Result and Conclusion

Our study investigates how well a gesture recognition-based system can translate hand sign movements into appropriate text and vocal outputs. The technology improves communication for people with speech or hearing difficulties by accurately recognizing hand motions and translating them into legible text and intelligible speech. It provides a user-friendly interface for real-time interaction and consistently recognizes static hand motions. The overall performance of the system, however, may be impacted by issues like identifying dynamic movements and gesture similarities.

In conclusion, by facilitating the smooth conversion of hand signals into text and audio forms, gesture recognition technology offers a viable answer for sign language interpretation. The technology has a lot of promise for helpful applications in the real world, particularly when it comes to closing communication gaps.

Future research should concentrate on enhancing the system's ability to recognize dynamic gestures and adapting it to a variety of settings. By tackling these issues, gesture-to-text-and-speech systems will become more reliable, inclusive, and effective in everyday communication situations.

7. References

[1] "Conversion of Sign Language to Text" by Akash Kamble1, Jitendra Musale2, Rahul Chalavade3, Rahul Dalvi4, and Shrikar Shriyal, vol. 11, May 2023, International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN 2321-9653

[2] "Sign language recognition using the fusion of image and hand landmarks through multi-headed convolutional neural network," Munmun Biswas, Suraiya Yasmin, Mayeen Uddin Khandaker, Mohammad Salman, and Ahmed A. F. Yasmin, 2023

[3]The International Journal of Academic Information Systems Research (IJAISR) ISSN 2643-9026 Volume 6, August 2022; Tanseem N., Abu-Jamie, Prof. Dr. Samy S., and Abu-Naser, "Classification of Sign-Language Using Deep Learning by ResNet"

[4] "Real Time Sign Language Recognition Using Deep Learning," International Research Journal of Engineering and Technology (IRJET) ISSN 2395-0072, Volume 09, April 2022, Sanket Bankar, Tushar Kadam, Vedant Korhale, and Mrs. A. A. Kulkarni

[5] "Real Time Sign Language Detection" by Aman Pathak, Avinash Kumar, Priyam, Priyanshu Gupta, and Gunjan Chugh, International Journal for Modern Trends in Science and Technology, ISSN 2455-3778, December 31, 2021

[6] Mahesh Kumar N B, "Translation of Sign Language into Text," International Journal of Applied Engineering Research, Volume 13, Issue 9, ISSN 0973-4562 (2018)

[7]The paper "Real-time Indian Sign Language (ISL) Recognition" by Kartik Shenoy, Tejas Dastane, Varun Rao, and Devendra Vyavaharkar was published in July 2018 by IEEE, 43488.

[8] P Surekha, Niharika Vitta, Pranavi Duggirala, Venkata Surya Saranya Ambadipudi, "Hand Gesture Detection and Conversion to Speech and Text", IEEE Xplore, May 2022. Available: https://ieeexplore.ieee.org/document/9743064.

[9] K. Manikandan, Ayush Patidar, Pallav Walia, Aneek Barman Roy, "Hand Gesture Detection and Conversion to Speech and Text", ResearchGate, December 2018. Available: https://www.researchgate.net/publication/329305443_Hand_Gesture_Detection_and_Conversion_to_Speech_and_Text.

[10] P. Gupta and R. Verma, "Scientific Exploration of Hand Gesture Recognition to Text", IEEE Xplore, August 2020. Available: https://ieeexplore.ieee.org/document/9155652/.

[11] Sheng-Tzong Cheng 1, Chih-Wei Hsu 1, Jian-Pan Li , "Combined Hand Gesture — Speech Model for Human Action Recognition", PMC, January 2014. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3892835/.

[12] R.Priyakantha, N.M.Sai Krishnab, Radha Abburi, "Hand Gesture Recognition and Voice Conversion for Speech Impaired", ResearchGate, January 2020. Available: https://www.researchgate.net/publication/347242673_Hand_Gesture_Recognition_and_Voice_Conversion_for_Speech_Impaired.

[13] Harale, Avinash D., Karande, Kailash J., "Literature review on dynamic hand gesture recognition", AIP Conference Proceedings, October 2022. Available: https://ui.adsabs.harvard.edu/abs/2022AIPC.2494c0005H/abstract.