real-time conversion of sign language to text and speech, and vice-versa

Harshavardhan Latkar, Sohan Wasker, Aman Vashistha, Atharva Kanse

1, 2, 3, 4Department of Computer Engineering, 1, 2, 3, 4Dhole Patil College of Engineering, Wagholi, Pune - 412207

Abstract - By translating sign language into text and speech and vice versa, the research seeks to develop a real-time system that will close communication gaps between hearing and hearing-impaired people. The system promises accurate and efficient translation by utilizing cutting-edge technologies such as image processing and convolutional neural networks (CNN). CNN is essential for identifying sign language gestures from video inputs because it has been extensively trained on a variety of datasets. Image processing techniques ensure robustness in varying lighting and backgrounds by reducing noise and extracting features, thereby improving system performance. Simulating real gestures for precise capture and enhanced user experience is a unique feature. Because of the system's real-time operation, users of spoken and sign language can communicate instantly with minimal latency. By combining these technologies, it may be possible to improve inclusivity in social and professional contexts and dissolve barriers to communication. This invention adds a dynamic and expressive layer to communication in addition to accurately recognizing a large variety of sign language gestures.

Index Terms - Translation, Sign language, Real-time system, Convolutional Neural Networks (CNN), Image processing, Inclusivity

I. INTRODUCTION

A vital component of human interaction, communication shapes our ability to relate to, comprehend, and collaborate with one another. However, for millions worldwide, this exchange of ideas and emotions can be a significant challenge due to physical disabilities like hearing impairment. Statistics from the World Health Organization (WHO) reveal that 466 million individuals, including 34 million teenagers, currently experience some form of hearing loss [4]. By 2050, these numbers are expected to surpass 900 million, with many of those affected residing in low- and middle-income countries.

To bridge this communication gap, sign languages have emerged as pivotal tools for the hearing impaired to convey their thoughts and feelings. These visual languages, such as American Sign Language (ASL), serve as a primary means of communication [2][7]. A vital component of human interaction, communication shapes our ability to relate to, comprehend, and collaborate with one another. However, for millions worldwide, this exchange of ideas and emotions can be a significant challenge due to physical disabilities like hearing impairment. Statistics from the World Health Organization (WHO) reveal that 466 million individuals, including 34 million teenagers, currently experience some form of hearing loss [4]. By 2050, these numbers are expected to surpass 900 million, with many of those affected residing in low- and middle-income countries.

To bridge this communication gap, sign languages have emerged as pivotal tools for the hearing impaired to convey their thoughts and feelings. These visual languages, such as American Sign Language (ASL), Indian Sign Language (ISL), and others worldwide, provide a means for the deaf and hard of hearing to interact effectively with each other and the broader community [1][2][4]. From the past to the present, these sign languages have evolved, creating rich vocabularies and grammatical structures unique to each language. [6]

For instance, in India, where approximately 6.3% of the population faces significant hearing disabilities, Indian Sign Language (ISL) serves as a primary means of communication [2][7]. However, despite its importance, many individuals within this community lack proficiency in ISL due to various factors such as limited access to ISL interpreters and tools, and insufficient research on ISL. [2]. This underscores the critical need for technological solutions that facilitate better communication for the hearing impaired, especially in countries like India where millions are affected. [9]

Advances in machine learning (ML) and computer vision present promising avenues for improving communication accessibility [1][5][10]. Researchers have explored various techniques, from deep learning models like Convolutional Neural Networks (CNN) for hand gesture recognition [1][3]. These technologies aim to create systems capable of accurately detecting and interpreting sign language gestures, offering real-time translation into text or speech [10].

Moreover, the development of applications like vision-based sign language translation for smartphones provides cost-effective and accessible solutions [12]. By capturing hand gestures through cameras, these applications convert sign language into text or speech, empowering deaf individuals to communicate more effectively in various settings, from online conferences to healthcare interactions [14].

However, challenges persist, particularly in the complexity of sign languages themselves [2][6]. Different regions have their own sign languages, each with unique semantics and gestures. This diversity poses obstacles to creating universal sign language recognition (SLR) systems. Researchers are addressing these challenges, exploring methods to train SLR systems with diverse sign language data and grammatical structures [2][11].
In the pursuit of effective communication solutions for the hearing impaired, it is crucial to consider not only the technological advancements but also the linguistic intricacies of sign languages [1][8][13]. As these technologies continue to evolve, they hold the potential to significantly improve the lives of millions by fostering inclusivity and accessibility in communication. Through ongoing research and development, we move closer to a future where communication barriers are minimized, enabling a more connected and inclusive society for all.[15] Several studies have highlighted the importance of sign language in empowering the deaf community, especially in emergency situations. One study discusses the development and organization of American Sign Language (ASL) and introduces a hand gesture recognition system using YOLOv3, showing promising accuracy in identifying ASL gestures. From this paper, we glean insights into the recognition of gestures using the YOLO model. [1] In India, another study focuses on Indian Sign Language (ISL), a fully-fledged natural language, utilizing Unity3D, Microsoft Kinect, and an Android app for real-time translation, achieving 91% accuracy. From this paper, we grasp the utilization of facial expressions or body postures to convey the desired message and emotion. [2] Furthermore, a study we learned about the utilization of ANN and MATLAB to achieve accuracy ranging from 92 to 100%. [7] Addressing challenges such as sudden hand movements, an ML-based Sign Language Recognition (SLR), from this study, we learned that KNN is used for classification, yielding a 65% accuracy rate.[4] Lastly a research explores real-time ASL conversion to text and speech using vision-based methods in an Android application, offering a cost-effective solution for enhanced communication accessibility. From this study, we comprehended the use of ASL to convert sign language into text/speech.[5] These advancements demonstrate a concerted effort to enhance communication and accessibility for the deaf and mute communities globally.

II. LITERATURE SURVEY

Several studies have highlighted the importance of sign language in empowering the deaf community, especially in emergency situations. One study discusses the development and organization of American Sign Language (ASL) and introduces a hand gesture recognition system using YOLOv3, showing promising accuracy in identifying ASL gestures. From this paper, we glean insights into the recognition of gestures using the YOLO model. [1] In India, another study focuses on Indian Sign Language (ISL), a fully-fledged natural language, utilizing Unity3D, Microsoft Kinect, and an Android app for real-time translation, achieving 91% accuracy. From this paper, we grasp the utilization of facial expressions or body postures to convey the desired message and emotion. [2] Furthermore, a study we learned about the utilization of ANN and MATLAB to achieve accuracy ranging from 92 to 100%. [7] Addressing challenges such as sudden hand movements, an ML-based Sign Language Recognition (SLR), from this study, we learned that KNN is used for classification, yielding a 65% accuracy rate.[4] Lastly a research explores real-time ASL conversion to text and speech using vision-based methods in an Android application, offering a cost-effective solution for enhanced communication accessibility. From this study, we comprehended the use of ASL to convert sign language into text/speech.[5] These advancements demonstrate a concerted effort to enhance communication and accessibility for the deaf and mute communities globally.

III. PROPOSED SYSTEM

The paper presents an innovative solution that integrates Convolutional Neural Networks (CNNs), image processing, Natural Language Processing (NLP), and animation gesture recognition to redefine real-time communication between sign language users and those dependent on text and speech. It delves into the training process of CNNs on diverse datasets, showcasing their pivotal role in accurately interpreting a broad spectrum of sign language gestures in various scenarios. The advanced image processing techniques are explored for their contribution to refining video frames, eliminating noise, and enhancing overall precision in gesture recognition. It is highlighted how NLP can be integrated as a linguistic bridge, facilitating inclusive real-time interactions by translating recognized gestures into coherent text and spoken language with ease. Animation gesture recognition is given a lot of attention, explaining how it can be used to mimic and identify real gestures in order to provide more expressive and immersive communication.

The paper concludes by synthesizing findings across these components, highlighting the holistic and technologically sophisticated nature of the proposed system, which not only contributes to academic understanding but also holds practical implications for improving communication accessibility and inclusivity. Overall, the research represents a significant advancement, pushing the boundaries of gesture recognition and communication technologies in a unified and impactful manner.

A. SYSTEM ARCHITECTURE

The system architecture is designed with a sophisticated integration of key components, namely Convolutional Neural Networks (CNNs), image processing, Natural Language Processing (NLP), and animation gesture recognition. The process involves a seamless flow of operations to ensure efficient real-time communication.
1. **Image Processing**: The system begins by carefully processing the video frames of sign language gestures that are input. This crucial step uses sophisticated image processing methods to clean up the video frames, carefully removing any extraneous noise as it extracts the most important information from the image data. The system hopes to greatly improve the quality and clarity of the input data through this process. Through image processing, the input is refined so that the system's later stages can function with clear and pertinent visual data. This is an important step because it lays the groundwork for later stages of accurate sign language gesture recognition and interpretation.

2. **Convolutional Neural Networks**: Convolutional Neural Networks (CNNs) play a pivotal role in sign language recognition by accurately identifying and interpreting a wide range of gestures. Trained extensively on diverse sign language data, CNNs receive refined information post-image processing to translate visual representations into understandable formats for the system. Their training phase ensures precise identification of gestures, crucial for comprehending sign language expressions. Gesture detection, facilitated by CNNs, is vital for aiding communication for individuals with hearing and speech impairments. This study concentrates on a CNN-based approach for real-time gesture detection, boasting an impressive average accuracy rate of 90%.

3. **Natural Language Processing (NLP)**: The system moves on to the Natural Language Processing (NLP) stage after receiving the CNNs' output, which shows the gestures in sign language that have been identified. Here, the gestures that have been identified are carefully examined and transformed into logical textual representations. Grammar and syntax are just two of the elements that NLP algorithms take into account when parsing the visual language of signs into structured and meaningful text. This conversion procedure is essential because it fills in the gaps between written text and visual sign language expressions. The system makes sure that the meaning expressed through sign language is precisely recorded and represented in a way that is easily understood by translating the recognized gestures into text.

4. **Speech-to-Text Conversion**: Using a Speech-to-Text conversion module like Whisper AI, which enables us to precisely and accurately detect speech and translate it into text. In addition to translating multiple non-English languages into English, Whisper AI can transcribe speech in both English and a number of other languages. The model served as the foundation for a single speech recognition model.

B. **METHODOLOGY**

The workflow of the system goes as follows, the hand gesture performed in front of webcam is detected and further converted to text and later to speech and vice-versa.
The technical workflow of the system goes as follows, the system is trained by feeding it with different hand gestures at various positions and simultaneously the coordinates are generated in background. Later in prediction mode the frames are captured and the Convolution Neural Network (CNN) locates the hands. And the Google’s MediaPipe identifies the coordinates of hands, further it is provided to the Feed Forward Neural Network for identifying Hand Sign. The text generated is converted to speech using Text- To-speech python API as output.

For vice-versa workflow, the OpenAI’s Whisper API detects the speech and generates the .wav audio file, then it is converted to text. Later the text is mapped to Sign Hand gesture and the output is generated.

**TENSORFLOW**

The best part of using TensorFlow library is that it is an open-source Library with lots of pre designed models, useful in Machine Learning and especially Deep Learning. For understanding the conceptual use of TensorFlow is required to understand the meaning of two terms, where the Tensor here is considered as N-Dimensional Array and Flow refers to graph of operations. Every mathematical computation in TensorFlow is considered as graph of operations where Nodes in the Graph are Operations and Edges are nothing but tensors.

Any mathematical computation is written in form of data flow diagram in Python Frontend or C++ or Java, as in our case Python is used. Then, TensorFlow Execution Engine comes into picture and makes it deployable on any of the hardware of Embedded System let it be CPU or Android or IOS, TensorFlow is a Machine learning framework that comprises of uses the dataset to train Deep learning models and helps in prediction and also improvise future results.

The biggest advantage of using TensorFlow is its feature of providing Abstraction, that is the developer does not need to work on every small aspect of designing the model as it is managed by the library itself, thus giving the developer the freedom to focus on logic building, which was clearly explained in [16].

TensorFlow in our system helps as in training the model using the provided dataset. TensorFlow object recognition algorithms help us classify and identify different hand gestures when combined with use of OpenCV. TensorFlow can assist in the classification and identification of hand gestures in real time by analyzing thousands of images. It enables the creation of a model that can assist in the identification of 3D images and their classification using 2D images from the feed dataset. TensorFlow has the capacity to process more data and identify more patterns.

**OPENCV**

An open-source computer vision library is called OpenCV. Since all of the training and classification processes are now prepared to be carried out, the system's eye was required to take real-time pictures of hand gestures, which could subsequently be sent for identification and classification. For image processing visualization, OpenCV infuses Deep Learning models with intelligence. Here, images are divided into two channels: RGB and greyscale. This means that after an image is taken with OpenCV, it first converts to the greyscale channel so that it can be processed morphologically, as demonstrated in [17]. The NumPy Library is used by OpenCV to compute images numerically as pixel matrices.

**MEDIA PIPE**

An open-source framework called Media Pipe can be used to create machine learning pipelines that are multi-modal, such as audio and video. It provides a range of pre-built parts and instruments for building intricate machine learning models and real-time media data processing.

You can identify hand gestures in real time with the Media Pipe Gesture Recognizer task, which also displays the landmarks of the hands that were detected. With this task, you can identify particular hand gestures made by the user and use the features of the application that go along with those gestures.

This task can accept either static data or a continuous stream and uses a machine learning (ML) model to operate on image data. Hand landmarks in world coordinates, image coordinates, handedness (left/right hand), and the hand gesture categories of multiple hands are the outputs of the task.

**CONVOLUTIONAL NEURAL NETWORK**

A convolutional Neural Network is nothing but a Deep Learning algorithm that is capable of assigning biases and weights to different objects in an Image and on basis of the same it can differentiate one image from another. It consists of processing different layers of Image Classification and it is designed with means of representing functioning of Neurons in Human Brain as explained in [18].

Even if the most minimalist pixelated image is considered, it still needs 4x4 matrix and required to consider the same image in different channels of colour formats like RGB, Greyscale, HSV, etc so it is very difficult to process thousands of images in high rates of pixels for instance 1020x1980 pixels. Here comes the need of Convolutional Neural Network that convolutes every image into its basic reduced form of matrix which can be differentiable at the same time. These increases the Accuracy and Speed and also reducing the processing of Classifier model. The convolutional layer is also supported with Pooling layer to decrease the processing need of classifier model. It also convolutes the matrix but on basis of dominant features. Pooling is majorly of two types; MAX Pooling and AVG Pooling, this is clearly explained in [19].

**PYTHON TEXT AND SPEECH APIs**

The Python text-to-speech library that used is very simple and easy to use. It makes use of modules like pyttsx3 and engine.io which let us change different properties like rate and intervals of text to speech conversion and outflow. The Python speech-to-text library by
which practicing makes use of speech recognition module. It let us adjust the ambient noise and also helps in recording the audio in form of mp4 files.

**EVALUATION**

When it comes to model compilation and evaluation the focus is on the accuracy of the correct predictions. The accuracy is calculated as the division of the sum of the true-positive instances (TP) and the true-negative instances (TN) through the total population (TOTAL):

\[
\text{Accuracy} = \frac{TP + TN}{TOTAL} \tag{1}
\]

Additionally, to show deeper insights into the classification errors the confusion matrix is used. Supported by this visualization we present and discuss the precision (2) and the true-positive-rate (TPR) also called Sensitivity (SN) or Recall in this context. The precision is calculated by the division of the TP through the positive predictions (POS PRED). The TPR is calculated as all TP divided through the number of actually positive instances (3).

\[
\text{Precision} = \frac{TP}{POS PRED} \tag{2}
\]

\[
\text{Sensitivity(TPR)} = \frac{TP}{ACTUAL POS} \tag{3}
\]

**HAND GESTURE DATASET**

The dataset is generated by training model with the dynamic hand gestures at various positions, and simultaneously the coordinates of the hand are generated in background in csv file.

**IV. DATA FLOW DIAGRAMS**

A Data Flow Diagram (DFD) illustrates the processes that alter data, data stores, and external entities as it moves through a system. A Level 0 Data Flow Diagram (DFD) provides an overview of the system, showing the main processes, data flows between them, and external entities interacting with the system. Here the DFD level 0 tells us about how the user can convert sign language to text/speech and vice versa, as shown in fig 2.
V. RESULTS

The demonstration of project is given using output images.

In prediction mode the hand gesture is detected and simultaneously converted to text and speech. Figure 6 illustrates how the model identified the hand gesture and accurately reported the result as dislike.

![Fig 6. Sign Language Recognition](image)

As illustrated in Figure 7, the model correctly identified the hand gesture and reported an okay result.

![Fig 7. Sign Language Recognition](image)

Figure 8 illustrates how the model identified the hand gesture and accurately reported victory as the outcome.

![Fig 8. Sign Language Recognition](image)
The speech is recorded and the transcript is generated, as shown in fig 9.

![Speech to Text Conversion](image)

**Fig 9. Speech to Text Conversion**

The transcript generated from previous output is converted into sign language as shown in fig 10.

![Text To Sign Language](image)

**Fig 10. Text To Sign Language**

The Performance evaluation is derived using Confusion matrix, as shown in fig 11. A clear visual depiction of a classification model's performance is called a confusion matrix. An instant evaluation of the model's accuracy, precision, recall, and other metrics is made possible by the way it presents the number of true positives, true negatives, false positives, and false negatives. The average classification accuracy of the Sign Language is, macro average for precision is 0.97, recall is 0.97, f1-score is 0.997 & support is 1324; And the weighted average for precision is 0.97, recall is 0.97, f1-score is 0.997 & support is 1324 respectively.
VI. CONCLUSION AND FUTURE SCOPE
This innovative project utilizes complex algorithms and computer vision to convert sign language into speech and text, and vice versa, offering significant benefits for individuals with hearing impairments. Despite challenges like dialect disparities and complex gestures, the technology holds immense potential for enhancing communication, inclusivity, and relationships. It can facilitate real-time interaction, inclusive education, and emergency support, removing barriers to seamless communication. As machine learning and AI progress, the accuracy of sign language recognition is likely to improve, paving the way for its integration into mobile applications, emergency situations, and educational settings. Moreover, advancements in identifying sign language dialects may facilitate international communication among users worldwide, ultimately leading towards a more inclusive and connected society.

VII. REFERENCE


[14] Indian Sign Language text generation from English/Hindi Text, Pawan Kumar, Savita Khatri et al. International Journal of Recent Research Aspects (IJRRA) ISSN: 2349-7688, Special Issue: Conscientious and Unimpeachable Technologies 2016, pp. 30-33

