International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 04 | Apr 2025 www.irjet.net p-ISSN: 2395-0072

Sign Language Vision to Text Using Deep Learning

Vinayak Suryavanshi[1], Ritika Parab[2], Abhijeet Yadav[3], Nishek Sharma[4], Prof.Vrushali Thakur[5]

[1,2,3,4]*Student, Dept of Computer Engineering, MGM's College of Engineering and Technology, Kamothe, Navi Mumbai, India*

[5]Prof. Vrushali Thakur: Professor, *Department of Computer Engineering, MGM's College of Engineering and Technology, Kamothe, Navi Mumbai, India*

Abstract - *For deaf and hard-of-hearing people, sign language is an essential medium of communication. However, since sign language is not universally understood, it poses a range of problems in social as well as professional contexts. A vision-based deep learning approach using sign language gestures to provide text translation is proposed in this paper, providing better access and more people. In this system Convolutional Neural Networks (CNNs) are used for gesture recognition; and Long Short-Term Memory (LSTM) networks are employed in the temporal sequence learning process. For improved recognition accuracy, the model incorporates attention mechanisms and uses a hybrid feature extraction method.*

A wide variety of sign languages are used to create the dataset, which covers enough general conditions and a range of scenarios. One must observe that using a variety of representations significantly improves the likelihood of effective generalization. The suggested approach is tested using both standard datasets and custom-collected sign motions. By displaying mistake rates, it produces promising accuracy results. Data augmentation and subs are two adaptive learning strategies that address issues including hand occlusions, signer variability, and various sign language patterns.

Key Words: **Sign Language Recognition, Deep Learning, CNN, LSTM, Computer Vision, Gesture Translation, Real-Time Processing, Accessibility, Sequence Learning, Human-Computer Interaction.**

1.INTRODUCTION

Sign language is a rich, expressive visual language that is utilized by millions of deaf and hard-of-hearing people across the globe. Sign languages differ from spoken or written languages in that they are dependent on the accurate combination of hand movement, facial expression, and body movement to communicate meaning. Each nation or region might use its own type of sign language, e.g., American Sign Language (ASL), British Sign Language (BSL), and Indian Sign Language (ISL), each with different grammar and vocabulary [2].

Traditional methods for sign language recognition have used sensor-based gloves, accelerometers, or depth sensors, which are good but sometimes costly, invasive, and impractical for broad use. Some of the pioneering work by Starner and Pentland [2] proved real-time sign language recognition with Hidden Markov Models (HMMs) feasible, setting the stage for current recognition systems. Our system utilizes CNNs to extract visual features and LSTM and Transformer models for recognition of temporal sequences. The solution also exploits the use of MediaPipe to achieve real-time landmark detection as well as the use of OpenCV for preprocessing images and image segmentation. Our objective is to design a low-cost, easy-to-use, high-accuracy system that works within real environments that have unpredictable lighting conditions and backgrounds. This research advances the area of assistive technologies by offering a scalable and versatile system which can be instantiated for various sign languages, is deployable on multiple devices, and can ultimately accommodate features such as sentence recognition and voice synthesis.

Fig 1 Sign Gestures

1.1Objectives

The primary objective of this study is to create a real-time, vision-based sign language to text translation system based on deep learning. The specific objectives are as follows –

1.Real-TimeTranslation:
To create a system that can recognize and translate sign language gestures into text in real time with the use of standard camera input without requiring any extra sensors or gloves, hence making it scalable and cost-effective [2].

2. User-Friendly Interface:
To incorporate a simple and interactive GUI with the help of tools such as Python Tkinter, so that end users—non-technical users as well—can interact with the system in an intuitive manner.

3. Robustness to Environmental Variations:
To make the system work reliably under different lighting conditions and complex backgrounds, employing data augmentation and preprocessing techniques [4].

4. Support for Continuous Gesture Recognition:
In order to move beyond stand-alone sign identification and process continuous strings of signs to create complete words or sentences, more natural patterns of communication [1].

5. Adaptability Across Languages and Users:
In order to make the model transferrable to a variety of users and sign languages through transfer learning and multi-language data sets so that in the future it can be extended to ASL, ISL, BSL, etc. [5].

6. Assistive Application Development:
With a focus on being placed on the mobile and embedded markets as a help tool to aid the deaf and hard-of-hearing user community in private and public modes of communication.

1.2 SCOPE

The Main motive of these project is to help the disabled people which has problem to interact with other people. Our aim is to provide real-time sign language-to-text translation through a camera and deep learning algorithms without dependence on any additional sensors or wearable technologies. This is not only more affordable, but also provides greater comfort to the user and scalability to the system [2].

This system specializes in detecting static and dynamic hand movements of Indian Sign Language (ISL) but is meant to be generalizable to other sign languages like ASL and BSL using transfer learning and multilingual training data sets [5]. The system comprises pre-processing, feature extraction, gesture classification, and text generation modules that are capable of execution on typical desktop or embedded platforms, e.g., laptops, Raspberry Pi, or Android-based phones. The GUI enables users to view real-time translation and provides suggestions for the predicted words, making it more useful in educational, personal, and public environments.

1.3 PROJECT MODULES

1.3.1. Data Acquisition

1.3.2. Data pre-processing and Feature extraction

1.3.3. Gesture Classification

1.3.4 Text and Speech Translation

1.3.5 Feature Extraction Module

1.3.6 Text Generation and Language Processing Module

2. Problem Statement

Although a common and very expressive means of communication for the deaf and hard-of-hearing population, sign language is still not easily accessible to the general public. This leaves a communication gap between signers and non-signers in schools, public services, healthcare, and daily life. The absence of technology which may affect to those who use sign language, frequently face problem in lack of communication with the society [2].

Conventional methods of sign language recognition use wearable sensors, gloves, or depth cameras, which are either too costly, invasive, or not very mobile. Although these systems have shown high accuracy, their practical use is limited by hardware dependencies and environmental sensitivity [3].

Latest developments in computer vision and deep learning hold promising alternatives in the form of camera-based, non-invasive approaches. But issues still remain for capturing real-time performance, processing ongoing gestures, conformity to different sign languages, and sustaining accuracy in diverse surroundings and illumination conditions [1][4].
Thus, a vision-based, low-cost, real-time sign language recognition system with the ability to work effectively on commonly available devices like webcams or mobile phone cameras is certainly the need of the hour. The system should be able to recognize both static and dynamic hand signs and transcribe them into correct textual output and be a useful communication aid for the deaf and hard-of-hearing community [5].

3. LIMITATIONS OF EXISTING SYSTEMS

Despite major breakthroughs in sign language recognition technology, current systems still present many limitations that limit their applicability in real-world settings. A key limitation is the need for specialized hardware like sensor-based gloves, accelerometers, or depth cameras, which are typically costly, intrusive, and not feasible for everyday use by the masses (Starner & Pentland, 1997). Though such systems can be precise in a controlled setting, they are not scalable for general use. Additionally, most previous models

only concern themselves with static gesture recognition and cannot accurately translate dynamic, continuous sign sequences, which are needed to express complete sentences and thoughts (Pu et al., 2019; Huang et al., 2015).

Another relevant constraint is limited generalizability across sign languages and signers. Models learned on one corpus or language, e.g., ASL, tend to perform suboptimally when subjected to other regional dialects like ISL or BSL. This is worsened by the inability of previous systems to have powerful transfer learning methods (Koller et al., 2020). Furthermore, vision-based recognition systems tend to be extremely sensitive to environmental conditions like low lighting, cluttered backgrounds, or partial occlusions, all of which lead to decreased accuracy in real-world applications (Pigou et al., 2014). Furthermore, most of these systems lack Natural Language Processing (NLP) modules, leading to grammatically incorrect or contextually unnatural translations when translating gestures into text (Pu et al., 2019).

Finally, certain deep learning-based systems, while precise, are computationally intensive and hence not deployable on mobile or embedded systems without extensive optimization (Huang et al., 2015). These constraints as a whole point towards the necessity of a real-time, low-cost, and hardware-independent sign language recognition system that is flexible, scalable, and user-friendly for static and dynamic gestures.

4. LITERATURE SURVEYS

Sign language recognition has been a most common language which has been talk for about decades, and the systems are tended to emphasize hardware solutions and rule-based algorithms. One of the earliest seminal contributions in this area was presented by Starner and Pentland (1997), who presented a real-time recognition system for American Sign Language (ASL) based on Hidden Markov Models (HMMs) and wearable computing. Although effective, their model utilized head-mounted cameras and was confined to isolated sign recognition (Starner & Pentland, 1997). With the emergence of deep learning, researchers started discovering vision-based solutions based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Pigou et al. (2014) showed how CNNs are used to recognize video-based sign gestures in a non-intrusive manner, making CNNs a strong means of extracting spatial features from visual inputs. These systems, however, tended to be restricted to static or short temporal gestures.

More recent developments came with Pu et al. (2019), who introduced the Iterative Alignment Network (IAN), a model intended to align input video frames and gesture sequences better. Their method has achieved the state of the art on continuous sign language benchmarks like RWTH-PHOENIX-Weather. More recently, investigators have tackled the limitations of language dependency and signer variation. Koller et al. (2020) suggested employing transfer learning to allow models learned on a single sign language to generalize across others. The study is indicative of the work in creating scalable and multilingual recognition systems that are more universal in their generalizability over users and scenarios.

Along with high advancements in the neural field, capabilities such as Media Pipe have been developed to advance the pre-processing phaseand extraction phase. MediaPipe facilitates effective, real-time detection of hand landmarks, which can be utilized as direct input to deep learning models, alleviating computational requirements and enhancing inference rates. In general, the literature reports a distinct shift towards lightweight, vision-oriented, and deep learning-based methods moving away from hardware limitations and towards scalable real-time sign language recognition solutions.

5.IMPLEMENTATION

5.1 System Architecture

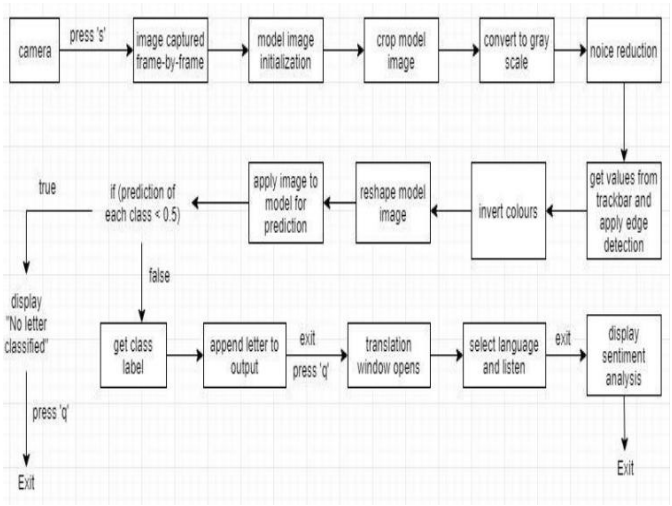

Fig 5.1 System Architecture

The architecture of such system for sign language vision-to-text translation with deep learning consists of different parts that communicate with different technology for accurate recognition and real-time processing. The system architecture has a structured pipeline that begins with data acquisition and ends with producing text output.

The first phase is Data Acquisition in which input sign language in the form of video or images is recorded using a camera. This is achieved either through pre-recorded datasets or live video streams. The raw input is then processed within the Pre-processing Module, where several image enhancement operations like grayscale, background subtraction, noise removal, and resizing of images are undertaken. These pre-processes standardize the input and improve the model to recognize gestures irrespective of

different problems in settings, as well as hand shape variations.

After pre-processing step is done, the system goes to the Feature Extraction part that derives various features from images or video frames by using Convolutional Neural Networks (CNNs). CNNs is used to help identify the hand shape, finger location, and motion paths upon which sign language gestures are based through camera assistance. For better tracking of motion, pose estimation models can be incorporated for identifying important joint movements.

The features are then passed to the Gesture Recognition and Sequence Learning Module, where Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, or Transformer-based models are used to identify the temporal patterns of sign language. The module is such that the system not only detects isolated gestures but also comprehends sequences of gestures, allowing detection of whole words and sentences.

After gesture recognition, the Text Generation and Language Processing Module converts the identified gestures into sensible text. Lastly, the processed text is presented to the user via the User Interface Module. The interface may be a desktop application, mobile application, or web platform, making it accessible on various devices. The system can also incorporate text-to-speech (TTS) functionality to facilitate spoken output, thus improving communication between sign language users and non-signers.

This system architecture is scalable, real-time, and adaptable for various sign languages. Future enhancement can be oriented towards multi-modal learning, data expansion, and low-power device optimization to facilitate wider accessibility and usability.

In the next stage, we have depicted the step-by-step workflow of the envisioned vision-to-text sign language recognition system. This workflow consists of various interrelated processes that altogether convert raw gesture inputs into useful textual output. The system starts with dataset identification, where appropriate sign language datasets are chosen or established depending upon the intended language and application domain.

After capturing the gestures, the system proceeds to the preprocessing stage, involving frame extraction, background subtraction, resizing, normalization, and potentially hand segmentation. Below is the mentioned Data Flow Diagram.

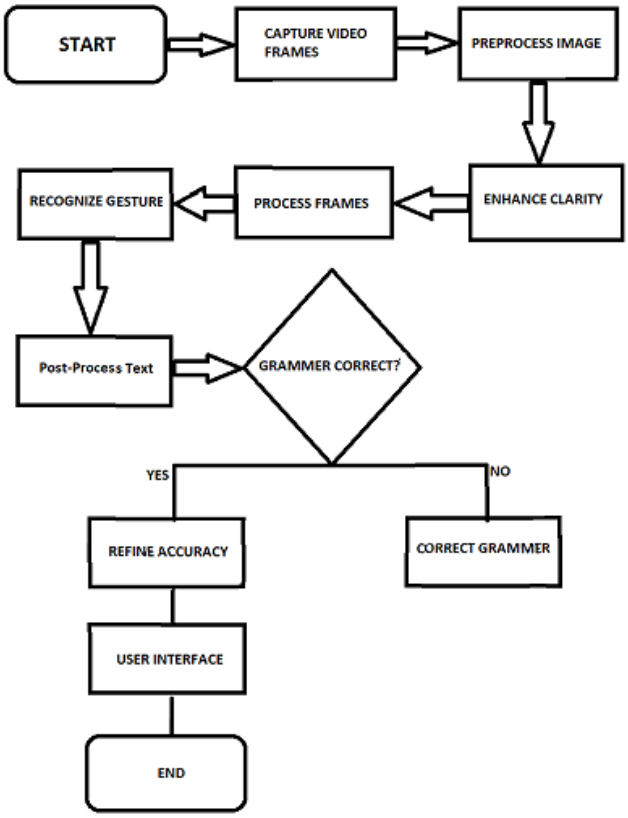

Fig 5.2 Data Flow Diagram

5.2 CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Network (CNN) is a special kind of deep learning model engineered to handle and analyze visual data, thus excelling in performing image recognition activities. CNN's work exactly the same way a human brain receives and processes information visually by reading meaningful patterns within images via multi-layer extraction. CNNs structures include convolutional, pooling, and fully connected layers, and these all contribute significantly to feature recognition and classification.

Convolutional layers employ filters to identify edges, textures, and shapes and progressively more intricate patterns. Pooling layers down sample feature maps, making it more efficient and reducing computation time for the network. The fully connected layers finally make sense of the extracted features and classify the input in terms of learned representations. CNNs have many applications in image classification, object detection, facial recognition, and medical imaging. Their capability to learn automatically and recognize patterns without feature extraction by hand renders them a basic tool in current deep learning for vision-related tasks.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 128, 128, 32)      320

 max_pooling2d (MaxPooling2  (None, 64, 64, 32)        0
 D)

 conv2d_1 (Conv2D)           (None, 64, 64, 32)        9248

 max_pooling2d_1 (MaxPoolin  (None, 32, 32, 32)        0
 g2D)

 flatten (Flatten)           (None, 32768)             0

 dense (Dense)               (None, 128)               4194432

 dropout (Dropout)           (None, 128)               0

 dense_1 (Dense)             (None, 96)                12384

 dropout_1 (Dropout)         (None, 96)                0

 dense_2 (Dense)             (None, 64)                6208

 dense_3 (Dense)             (None, 27)                1755

=================================================================
Total params: 4224347 (16.11 MB)
Trainable params: 4224347 (16.11 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Fig 5.2.1 Convolutional Neural Network Layer

5.3 Detailed Module description

5.3.1 Data Acquisition

There are several methods to obtain data on hand gestures for sign language recognition. One such method is through the use of electromechanical devices, e.g., sensor gloves, that can precisely capture hand movement and position. Even though these devices yield accurate data, they tend to be expensive and may not be user-friendly, thus not being practical for large-scale implementation.

Another method is vision-based gesture recognition, whereby an ordinary camera like a computer webcam records the motion of hands and fingers. It does away with the use of extra hardware and allows a natural interaction between the users and the system with drastically lower costs.

The vision-based hand detection may cause some challenges, such on the extensive variability in hand shape and movement, skin color, and the effect of parameters like camera orientation, and motion velocity. These challenges must be overcome in order to enhance the accuracy and reliability of sign language recognition systems for the people.

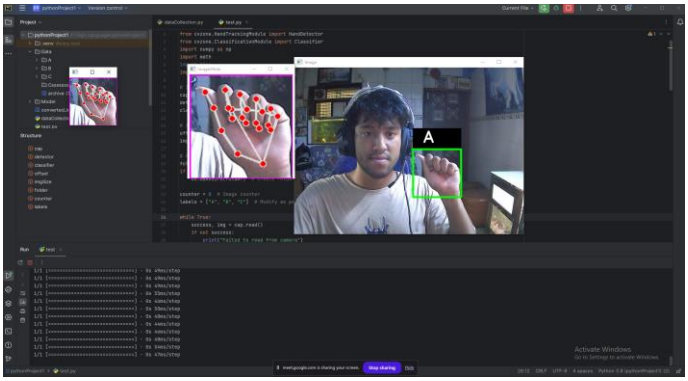

FIG 5.3.1 Data acquisition using webcam

5.3.2 Data Pre-processing and Feature Extraction

In hand detection, the initial step is to locate the hand from the webcam image. For this purpose, the MediaPipe library is used since it is a dedicated library for image processing operations. After the hand detection, Region of Interest (ROI) is found, and the image is cropped accordingly. The cropped image is further transformed into a grayscale image with the help of the OpenCV library, and a Gaussian blur is applied to it to filter out noise and improve the clarity of features.

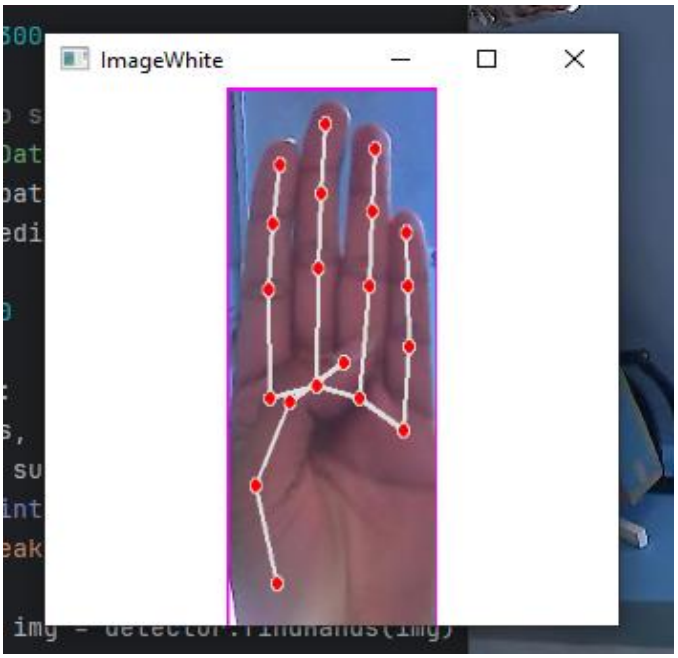

FIG 5.3.2 Hand landmark detection

Once the hand is identified, the system retrieves landmark points describing the fingers' and palm's structure and motion. The points are then projected onto a blank white surface via the OpenCV library to provide an even presentation of gestures and remove background noise and variations of light.

The Media-Pipe library is the important factor to this process since these identifies hand landmarks regardless of the background and lighting in the environment, thereby making the system more versatile and reliable. For training, a collection of 180 skeletal images has been prepared, ranging from alphabet signs A to Z.

These images give a simplified but efficient representation of hand gestures, improving the accuracy and performance of the deep learning model.

6. Implementation and Testing

Here are some Screenshots when user shows some hand gestures in most optimal background as well as in different lightning conditions and system is giving corresponding prediction.

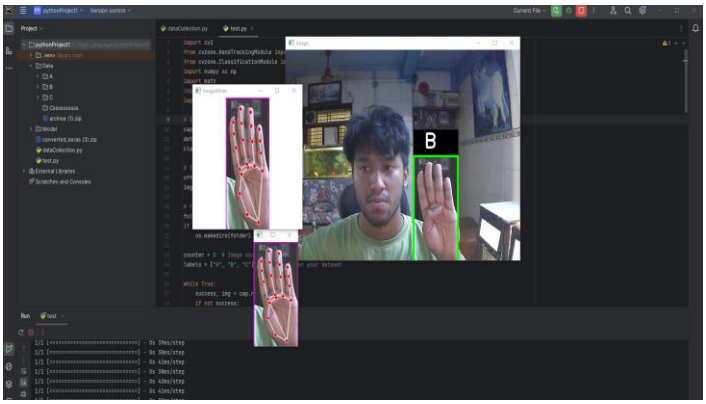

FIG 6.1 Real-time gesture recognition output

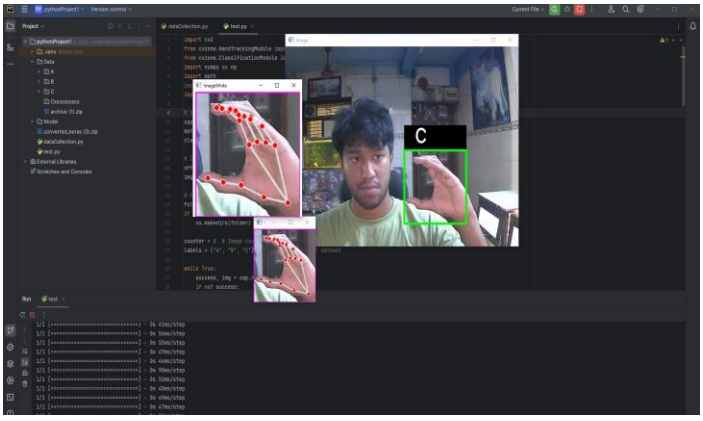

FIG 6.2 System testing under varied conditions

After Implementing the CNN algorithm, we did GUI using python Tkinter and added Suggestions also to make the process smooth for user. We can directly input the suggestions word.

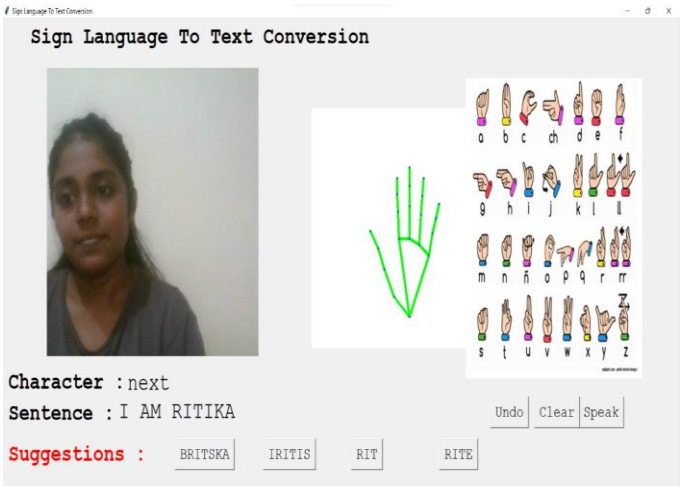

FIG 6.3 GUI with gesture prediction and word suggestions

6.2.2 Future Scope

The future of vision-based sign language translation with deep learning is very promising for enhancing communication between the deaf and hearing populations. With improved deep learning architectures, real-time processing, and multimodal learning, the accuracy and efficiency of sign language recognition systems can be greatly improved. The natural language processing (NLP) can also improve the translated text, and the dialogue can be made more smooth and precise for interpreting with other people. Future developments can be in the form of edge computing and AI-powered wearable devices, enabling sign language recognition on mobile phones without relying on cloud servers.

Further, increasing the dataset to cover different regional sign languages and gestures will make the system more adaptable across different linguistic communities. Advances in gesture tracking through virtual reality (VR) and augmented reality (AR) can further close the divide between sign language users and non-signers and make communication even smoother and more convenient in everyday life. As research and technology advance, sign language recognition systems can become regular tools for use in schools, workplaces, and public services to promote greater inclusivity and access for the deaf community globally.

6. CONCLUSION

Last but not least, we can forecast any alphabet[a-z] with 97% Accuracy (clean background and suitable lightning conditions and without clean background and suitable lightning conditions) by our approach. And if background is clear and lightning condition is good then we achieved even 99% accurate results.

Though deep learning has evolved, issues such as limited training sets, heterogeneity of sign language, and difficulties in deploying the system in real-world applications still exist. But due to the advancements in research and development, the system's accuracy as well as usability will continue to evolve. Future enhancements through multimodal learning, edge computing, and real-time gesture recognition will continue to increase the system's efficiency and accessibility. As technology advances, AI-based solutions such as this will be instrumental in promoting accessibility, education, and social integration for the deaf and hard-of-hearing populations.

In Future we are considering an android app or IOT gadget in which we apply this algorithm for gesture prediction which can assist lots of individuals worldwide.

7. REFERENCES

[1] Pu, J., Zhou, W., Li, H., & Li, W. (2019). Iterative Alignment Network for Continuous Sign Language Recognition. IEEE CVPR, 4165–4174.

[2] Starner, T., & Pentland, A. (1997). Real-Time American Sign Language Recognition Using HMMs. IEEE PAMI.

[3] Pigou, L., Dieleman, S., Kindermans, P. J., & Schrauwen, B. (2014). Sign Language Recognition using CNNs.

[4] Huang, J., Zhou, W., & Li, H. (2015). Continuous Sign Language Recognition with CNN-LSTM. ECCV.

[5] Koller, O., Ney, H., & Bowden, R. (2020). Sign Language Recognition with Transfer Learning Across Languages.