SIGN LANGUAGE RECOGNITION USING ACTION DETECTION

SK. Shabbir Basha
Department of Computer Science
Associate Professor
PBR Visvodaya Institute of Technology and Sciences
Kavali, India
shabbirbasha.sk@visvodayata.ac.in

VVL.Aakanksha
Computer Science and Engineering (Artificial Intelligence)
PBR Visvodaya Institute of Technology and Sciences
Kavali, India
aakankshavemula2003@gmail.com

P.Nagalakshmi
Computer Science and Engineering (Artificial Intelligence)
PBR Visvodaya Institute of Technology and Sciences
Kavali, India
panidapunagalakshmi19@gmail.com

CH.Iswarya
Computer Science and Engineering (Artificial Intelligence)
PBR Visvodaya Institute of Technology and Sciences
Kavali, India
iswaryacheepinapi@gmail.com

Md. Abdul Haneef
Computer Science and Engineering (Artificial Intelligence)
PBR Visvodaya Institute of Technology and Sciences
Kavali, India
haneefmohammad9988@gmail.com

***Abstract*— Sign language detection is one of the uses of computer vision that has grown in importance and effectiveness for people. Research in this field is ongoing. Previous studies used a basic deep learning-based convolutional neural network for static sign detection. In order to identify the action taken by the user, this proposal is based on the continuous real-time detection of image frames using action detection. After identifying key points using mediapipe holistic, which includes face, pose, and hand features, the model employs an LSTM neural network model. The suggested work involves pre-processing the data, generating labels and features, and gathering important value points for testing and training. It uses confusion matrix accuracy to evaluate the model and saves the weights.**

This study aims to bridge the communication gap by introducing a cutting-edge method for translating static and dynamic Indian Sign Language signs into text. Following that, this data is sent wirelessly and is sorted into the relevant text outputs. due to LSTM networks they have been studied and applied to the categorization of gesture data, as they possess the ability to discover enduring relationships. This model showed that LSTM-based neural networks are a good choice for sign language translation, with a 98% classification accuracy.

Keywords—LSTM, Dynamic action recognition, Video to text outputs, Accuracy, MediaPipe, CNN, Open CV

I. INTRODUCTION

The major means of communication for millions of deaf and hard-of-hearing people worldwide is sign language. The deaf community in India primarily uses Indian Sign Language (ISL) for communication. The development of sign language recognition systems to close the communication gap between the hearing and deaf communities has gained momentum with the growth of technology and the rising integration of Deep Learning (DL) into numerous areas.

The development of ISLR systems involves the utilization of various deep learning emerging as a prominent approach due to its ability to automatically learn features from data. These systems typically involve capturing sign language gestures through cameras and processing the data to recognize and interpret the gestures accurately.

Given that sign language includes a vast variety of hand forms, gestures, and facial expressions, one of the main obstacles is the intricacy and unpredictability of sign language gestures. Environmental elements like backdrop clutter and lighting can also have an impact on how well ISLR systems function.

Moreover, the absence of Indian Sign Language-specific benchmarks and standardized datasets makes it difficult to train and assess ISLR models efficiently.

In this work, we present a thorough analysis of current approaches and strategies for the recognition of Indian Sign Language. We analyze and emphasize the advantages, disadvantages, and opportunities for improvement of the approaches like deep learning-based techniques. For the benefit of the deaf community in India, we also address the challenges and potential paths for SLR research, with the goal of advancing this vital technology.

II. LITERATURE SURVEY

Action Recognition has been the subject of numerous approaches and techniques. The following section lists the previous efforts:

1. The authors train neural networks for image classification in a research paper [1]. MNIST models are also used for training complex images. For the images used, this resulted in the creation of a recurrent neural network. The level of classification improvement was so great that even the human eye was unable to distinguish between the images.
2. According to research paper [2], a healthy mix of face and non-facial images are used in training. Using a binary convolutional neural network-developed bi-scale CNN 120 that has undergone auto-stage training for imagenet classification. A state-of-the-art 80% detection rate with only roughly 50 false positives was the outcome of this.
3. In the research paper [3] Under each Hierarchical classifier, there are several datasets. Rejection of the class based on the intermediary stage was performed by the authors. The survey done by the authors considered classification techniques such as Decision Tree (DT), Support Vector Machine (SVM), and Fuzzy Classification under the artificial neural networks.

4. In a research paper [4] Spectral information is combined with spatial information from a sequential trial method. This led to SVM Active Learning Approach for Image Classification Using Spatial Information. The results obtained by the authors in this research paper demonstrate the efficacy of regularization in the spatial domain for active learning purposes.

III. RELATED WORK

A. *Architecture*

The combination of Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) in our suggested architecture for Indian Sign Language Recognition (ISLR) provides a potent framework for collecting both spatial and temporal characteristics inherent in sign language motions. The foundation of feature extraction are CNNs, which are able to automatically identify spatial patterns from input photos of sign language gestures. This is important because it allows for the discernment of nuanced hand shapes, movements, and facial expressions that are typical of Indian Sign Language.

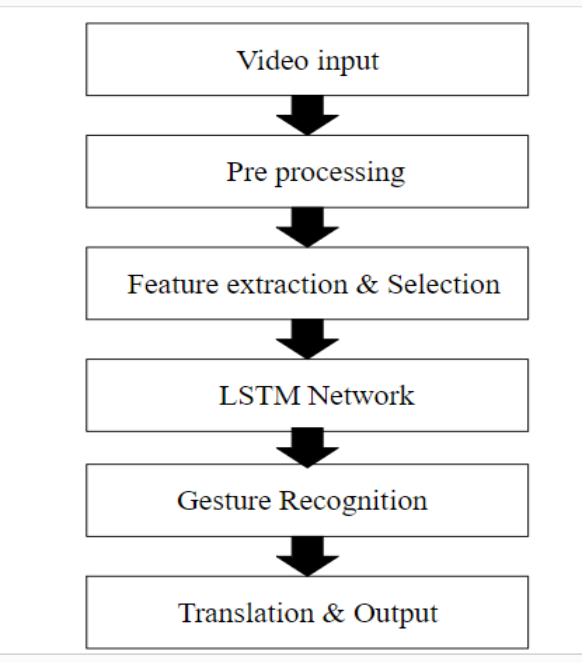

Fig 1 : Architecture for Sign Language Recognition

B. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are essential for accurately evaluating and interpreting visual data in sign language recognition, which allows for the understanding of hand motions and movements. Due to its capacity to automatically learn hierarchical representations of picture data, CNNs are especially well-suited for this task. CNNs are able to extract discriminative features from raw input photos by utilizing many layers of convolutional and pooling processes. This allows them to capture the spatial patterns and textures that are typical of sign language gestures. CNNs can recognize minute differences in hand shapes, hand movements, and face expressions because of this feature extraction method, which is essential for accurately identifying various signs and deciphering their meanings.

C. Long Short Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are an effective tool in sign language recognition because they can simulate the temporal dynamics included in sequential data, such as the movement trajectories of sign language motions. Because LSTM networks can capture long-range dependencies and preserve information over longer data sequences, they are a perfect fit for this task. LSTM networks efficiently address the vanishing gradient issue by using recurrent connections with gating mechanisms, which enables them to gradually recognize and retain patterns in sequential data. LSTM networks are capable of analyzing the temporal evolution of hand gestures and motions across numerous video frames in the context of sign language recognition. This allows the system to comprehend the meaning and context behind various sign sequences.

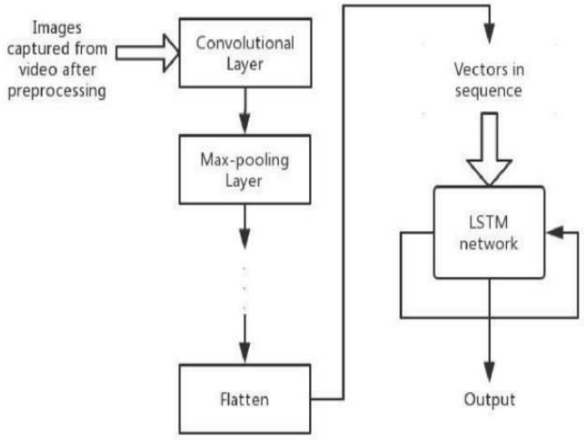

Fig 2 : LSTM Model for Sign Language Recogniiton

D. Mediapipe Holistic

When it comes to real-time video streaming, MediaPipe Holistic offers a comprehensive solution for identifying and evaluating important landmarks and positions on the face, torso, and hands. In order to properly interpret sign language motions, this framework tracks the spatial positions and movements of important body parts using sophisticated machine learning algorithms and computer vision techniques. Due to its holistic approach, MediaPipe Holistic is able to capture the subtle body movements, face expressions, and hand configurations that are typical of sign language communication. Furthermore, MediaPipe's effective and streamlined implementation allows for real-time performance across several devices, which makes it appropriate for deployment in interactive apps and assistive solutions for people who communicate through sign language.

IV. PROPOSED SYSTEM

To improve the accuracy and robustness of gesture identification, we have suggested an Indian Sign Language identification system that combines the Mediapipe and

OpenCV frameworks. By utilizing Media pipe's real-time hand tracking features, our system effectively recognizes and locates hand gestures and motions within video feeds. We extract accurate spatial information on hand configurations and movements, enabling the recognition of unique sign language gestures, by using Mediapipe's keypoint detection and tracking algorithms.

OpenCV is a complete collection of computer vision capabilities that enhance Mediapipe and are useful for processing and analyzing images. We preprocess and improve raw video frames using OpenCV's extensive function library to maximize the input data for gesture detection. We leverage the complementing capabilities of Mediapipe and OpenCV within our ISLR framework to accomplish reliable and fast detection of Indian Sign Language gestures, improving communication and accessibility for the deaf community.

A. Detecting Landmarks

Mediapipe landmark detection technology offers a sophisticated framework for detecting and localizing significant landmarks on hand motions in real-time, revolutionizing the recognition of sign language. By utilizing deep neural network architectures, Mediapipe can precisely encode the spatial configurations and movements present in sign language gestures by precisely identifying and tracking distinguishing landmarks, such as fingertips and palm keypoints. Mediapipe plays a crucial role in promoting accessibility and diversity by enabling people with hearing impairments to express themselves and interact with others through its strong and effective landmark detection capabilities.

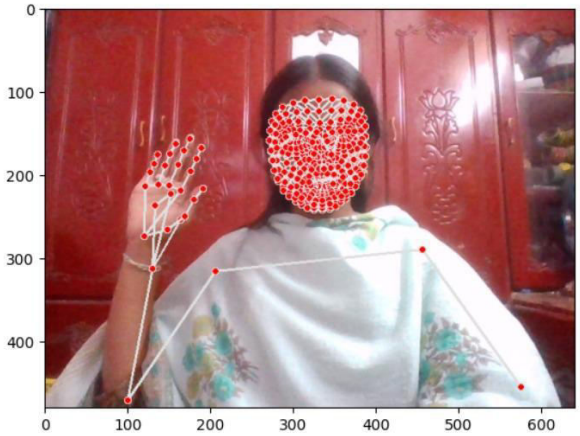

Fig 3 : Landmark Detection through Mediapipe

B. Collecting Key points through frames

Key points or landmarks are extracted and flattened for feature representation and dimensionality reduction in gesture recognition tasks involving human interaction. For instance, while creating folders for data collection, every action—like "hello," "thank you," and "i love you"—has a corresponding directory with several video sequences for testing and training.

There are several films in each sequence, which gives enough data for testing and training the model. The next stage of key point collection involves taking real-time key points from the webcam feed and using MediaPipe's Holistic model to describe the spatial placements of landmarks on the hands, body, and face.

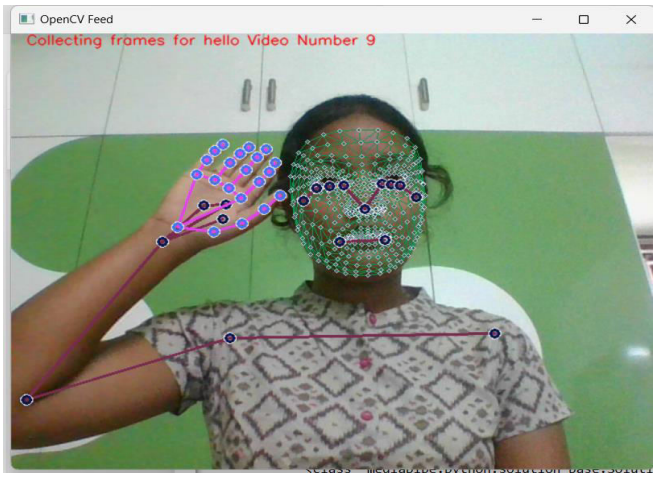

Fig 4 : Collecting frames from the video for ‘hello’

Setting up directories is essential for organizing the collected data systematically. Each action (gesture) will have its own directory, containing multiple sequences (videos) for training and testing the gesture recognition model.

- **Sequences:** Thirty sequences are allocated for each action, totaling 90 sequences.
- **Videos:** Each sequence will contain multiple videos, for training and testing.

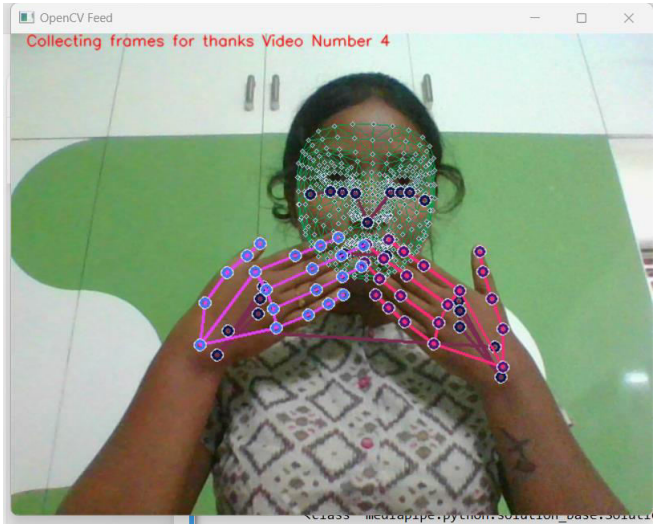

Fig 5 : Collecting frames from the video for ‘thanks’

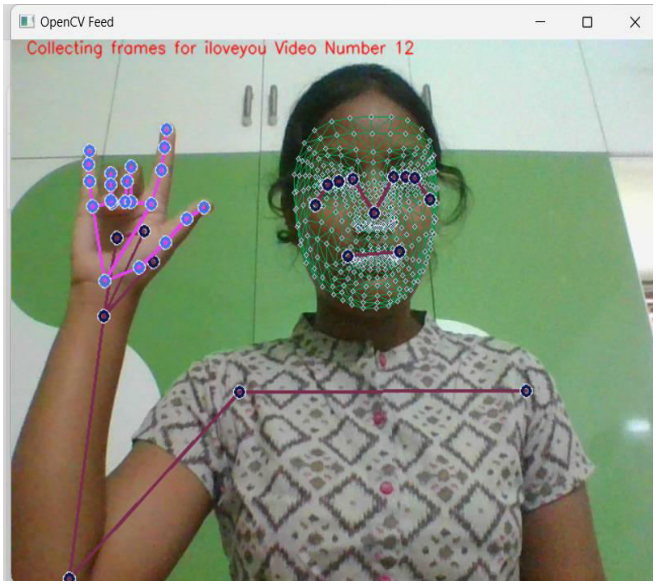

Fig 6 : Collecting frames from the video for ‘iloveyou’

C. *Data Preprocessing*

Using a label mapping, which gives each class or category a distinct numerical identification, labels are transformed into numerical form during the data preprocessing stage. This makes it possible for the machine learning model to efficiently analyze and interpret the labels throughout the training and testing stages. After label conversion, the dataset is arranged into sequences, with each sequence denoting a particular gesture or action. The shape of the sequences array is (180, 30, 1662), which means that there are 180 samples altogether, 30 frames in each sample. There are 1662 characteristics per frame, which are probably the critical spots that were concatenated and smoothed out of the respective video frames.

D. *Building neural networks and Training*

Following preprocessing and organization, the data is split into training and testing sets in order to assess the gesture recognition model's effectiveness. A portion of the data is set aside for the model's training and the remaining amount is used to assess the model's performance, according to the train-test split. The data is split in this instance with a test size of 5%, which means that 95% of the dataset is used for training and 5% is kept aside for testing.

Furthermore, the testing labels' form is given as (9, 3), denoting that the testing set consists of 9 samples, each of which has one-hot encoded size 3 categorical labels. A method for representing categorical data is called "one-hot encoding," in which each label is represented as a binary vector with a 1 in the position corresponding to the label's index and 0s elsewhere.

V. TEST RESULTS AND EVALUATION METRICS

In order to properly evaluate the performance of the trained LSTM model, assessment metrics must be defined prior to using the MediaPipe Holistic model for gesture recognition. For gesture recognition, accuracy, precision, recall, and F1-score are frequently used evaluation metrics. While precision represents the percentage of properly detected positive cases out of all cases projected as positive, accuracy assesses the overall soundness of the model's predictions. The harmonic mean of precision and recall, or F1-score, provides a fair evaluation of the model's performance, especially when there are unequal class distributions.

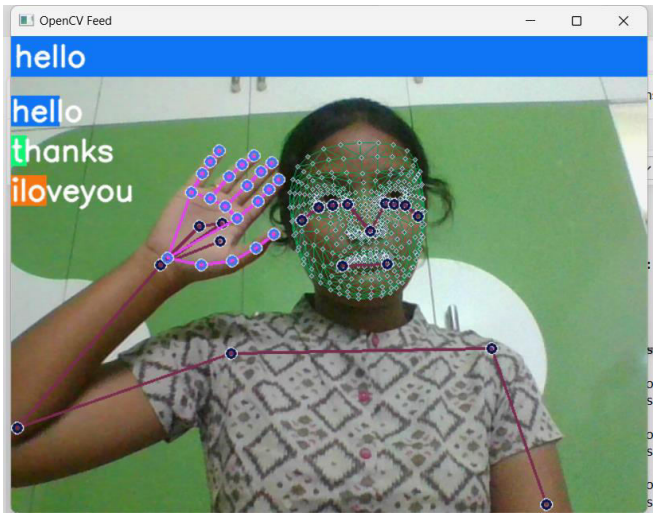

Fig 7 : Real time prediction for ‘hello’

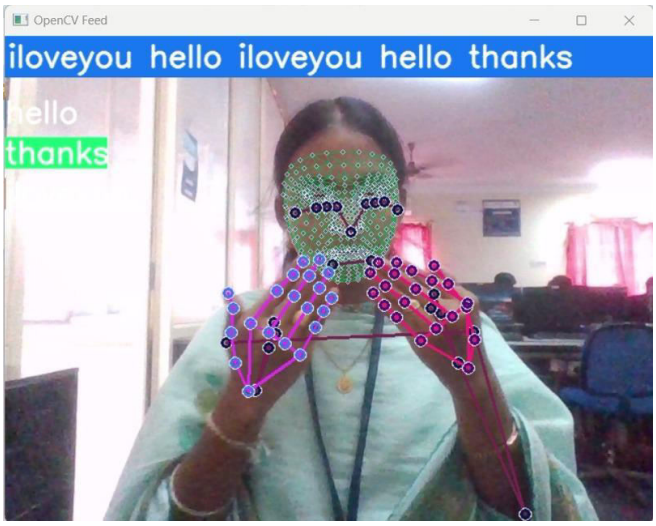

Fig 8 : Real time prediction for ‘thanks’

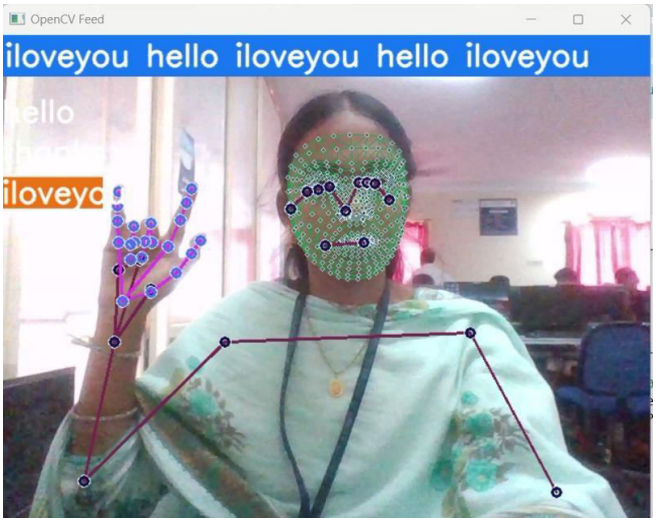

Fig 9 : Real time prediction for ‘iloveyou’

A. *Accuracy*

To guarantee the accuracy and dependability of translating gestures into text or meanings, the recognition

accuracy of sign language gestures is a crucial statistic. It is imperative that sign language recognition systems meet or surpass predetermined benchmarks in order to ensure their efficacy and usability in practical applications. Communication between the hearing and the deaf is made possible by the system's reliable precision, which guarantees that a large variety of sign language motions can be accurately interpreted and translated. Sign language recognition systems can promote better diversity, accessibility, and understanding in society by building user and stakeholder confidence by adhering to specified benchmarks for recognition accuracy.

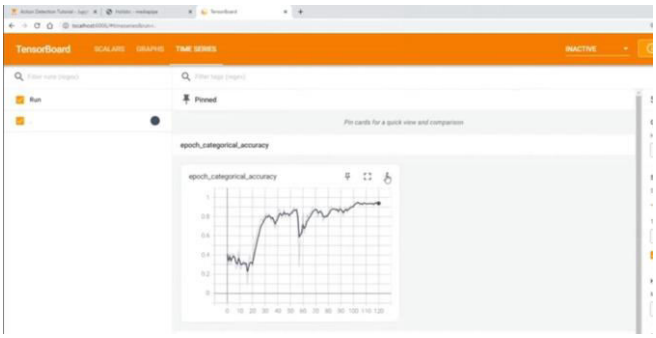

Fig 10 : Epoch Categorical Accuracy

B. Reliability

Ensuring constant recognition and translation accuracy in sign language recognition technology requires the system to operate reliably under a variety of settings, such as changing lighting environments, hand orientations, and motions. The system's versatility guarantees continuous operation in a range of situations due to its ability to adjust and function well in a variety of lighting circumstances, including low light or high glare scenarios. The system's ability to effectively manage these difficulties will enable it to deliver dependable and constant recognition accuracy, improving accessibility.

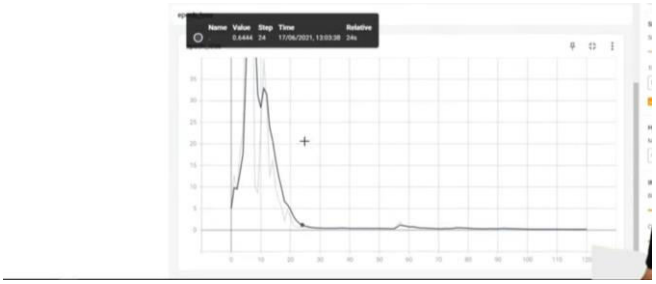

Fig 11 : Epoch loss

C. Interoperability

Ensuring seamless integration with existing communication tools and assistive technologies is crucial for the sign language recognition system to effectively support individuals with speech impairments. By facilitating interoperability and ease of use, the system can seamlessly complement and enhance the functionalities of existing communication aids, such as text-to-speech converters, speech recognition software. Additionally, by leveraging existing assistive technologies, the sign language recognition system can capitalize on established user interfaces and interaction paradigms, further enhancing user acceptance and adoption.

D. *Usability*

Ensuring the accessibility and usability of the sign language recognition system requires designing an intuitive and user-friendly user interface that can accommodate persons with varying degrees of technical expertise and accessibility requirements. Users with different degrees of technical experience may traverse and engage with the system's features with ease thanks to the interface's clear and intelligible visual signals, simple controls, and intuitive navigation menus. Further enhancing inclusivity and allowing people with disabilities or accessibility needs is the provision of accessibility features, such as customizable font sizes, color contrasts, and support for alternate input methods like voice commands or gesture-based controls.

VI. CONCLUSION

To sum up, sign language recognition is a crucial field of study with the potential to completely transform the way deaf and hard of hearing people communicate. Significant progress has been made in computer vision and machine learning, leading to advancements in this domain. Systems for recognizing sign language have been developed using a variety of methods, including motion analysis, deep learning, and hand shape recognition. But even with these improvements, there are still a number of issues that require attention.

The absence of standardization in sign languages is one of the main obstacles to sign language recognition. There are many different sign languages in use today, and each has its own vocabulary, syntax, and grammar. It is therefore a difficult task to develop a universal sign language recognition system that can recognize various sign languages.

The objective is to develop a robust deep learning model for Indian sign language recognition, comprehension, and textual transcription. Background information that showed a notable deficiency in communication between the deaf and silent community and the broader population. Sign language is used by those who are speech-impaired, but other people speak all the time, so these two groups of people speak different languages.

In order to provide accessibility for users of sign language, computing is a crucial tool. In this instance, action recognition is being used to identify sign language movements made in front of a camera or other video recording device. Numerous variations of to increase the obtained model's accuracy, training can be applied. Depending on what is considered necessary, the output is either displayed as a communication tool or utilized as input in other applications.

VII. FUTURE ENHANCEMENTS

The variety of sign language gestures presents another difficulty. Signers can convey meaning through a variety of body language and facial expressions, even with the same sign produced differently by different singers. This unpredictability makes it challenging to create precise and dependable methods for recognizing sign language.

Subsequent research in this area may concentrate on creating sign language recognition systems that are more reliable and accurate, capable of handling variations in sign gestures and recognizing multiple sign languages. Using cutting-edge deep learning methods like reinforcement learning and generative adversarial networks (GANs) may be necessary to increase the system's accuracy and resilience.

Furthermore, studies could be done on creating more organic and user-friendly interfaces for people to use with sign language recognition software. Wearable technology, like gloves or sensors to record facial expressions and sign gestures, may be used in this. or the creation of virtual people who can converse in real time using sign language.

VIII. REFERENCES

1. .. Abraham, Ebey, Akshatha Nayak, and Ashna Iqbal. "Real- time translation of Indian sign languageusingLSTM."2019 global conference for advancement in technology (GCAT). IEEE, 2019.
2. Xiao,Qinkun,etal."Multi-information spatial–temporal LSTM fusion continuous sign language neural machine translation." IEEE Access 8 (2020): 216718-216728.
3. Jayadeep,Gautham,etal."Mudra:convolutional neural network based Indian sign language translator for banks." 2020 4th InternationalConferenceonIntelligent Computing and Control Systems (ICICCS). IEEE, 2020.
4. Basnin, Nanziba, Lutfun Nahar, and Mohammad Shahadat Hossain. "An integrated CNN-LSTM model for Bangla lexical sign language recognition." Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020. Singapore: Springer Singapore, 2020.
5. Sincan, Ozge Mercanoglu, Anil Osman Tur,and Hacer Yalim Keles. "Isolated sign language recognition with multi-scale features using LSTM." 2019 27th signal processing and communications applications conference (SIU). IEEE, 2019.
6. Xu, Biao, Shiliang Huang, and Zhongfu Ye. "Application of tensor train decomposition in S2VT model for sign language recognition."IEEEAccess9(2021):35646- 35653
7. Goel, Pragati, et al. "Real-Time Sign Language to Text and Speech Translation and Hand Gesture Recognition using the LSTM Model."2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT). IEEE, 2022.
8. Chaikaew, Anusorn, Kritsana Somkuan, and Thidalak Yuyen. "Thai sign language recognition: an application of deep neural network."2021 joint international conference on digital arts, media and technology with ECTI northern section conference on electrical, electronics, computer and telecommunication engineering. IEEE, 2021.
9. Mittal, Anshul, et al. "A modified LSTM model for continuous sign language recognition using leap motion."IEEE Sensors Journal 19.16 (2019): 7056-7063.
10. Sonare, Babita, et al. "Video-based sign language translation system using machine learning." 2021 2nd International Conference for Emerging Technology (INCET).IEEE, 2021.