Sign Language Recognition using LSTM and Media Pipe

G. Mallikarjuna Rao

Department of computer science and engineering

Gokaraju Rangaraju Institute Of Engineering and Technology

Hyderabad, India

gmallikarjuna628@grietcollege.com

Cheguri Sowmya

Department of computer science and engineering

Gokaraju Rangaraju Institute Of Engineering and Technology

Hyderabad, India

soumyach2001@gmail.com

Dharavath Mamatha

Department of computer science and engineering

Gokaraju Rangaraju Institute Of Engineering and Technology

Hyderabad, India

mamathadharavath178@gmail.com

P. A.Sujasri

Department of computer science and engineering

Gokaraju Rangaraju Institute Of Engineering and Technology

Hyderabad, India

psujasri181@gmail.com

Soma Anitha

Department of computer science and engineering

Gokaraju Rangaraju Institute Of Engineering and Technology

Hyderabad, India

somaanitha193@gmail.com

Ramavath Alivela

Department of computer science and engineering

Gokaraju Rangaraju Institute Of Engineering and Technology

Hyderabad, India

ramavathalivela99@gmail.com

2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS) | 979-8-3503-9725-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICICCS56967.2023.10142638

***Abstract*—**There are learning aids available for those who are deaf or have trouble speaking or hearing, but they are rarely used. Live sign motions would be handled via image processing in the suggested system, which would operate in real-time. Classifiers would then be employed to distinguish between distinct signs, and the translated output would show text. On the set of data, machine learning algorithms will be trained. With the use of effective algorithms, top-notch data sets, and improved sensors, the system aims to enhance the performance of the current system in this area in terms of response time and accuracy. Due to the fact that they solely employ image processing, the current systems can identify movements with considerable latency. In this project, our research aims to create a cognitive system that is sensitive and reliable so that persons with hearing and speech impairments may utilize it in day-to-day applications

***Keywords*—** *Image Processing, sign motions, sensors, speaking or hearing*

I. Introduction

It can be extremely difficult to talk to persons who have hearing loss. The use of hand gestures in sign language by the deaf and the mute makes it difficult for non-disabled persons to understand what they are saying. As a result, there is a need for systems that can identify various symptoms and notify everyday people of what they mean. It is crucial to create specific sign language applications for the deaf and dumb so they may easily communicate with others who do not understand sign language. The major goal of our initiative is to start closing the communication gap between hearing individuals and sign language users who are deaf or dumb. Creating a vision-based system that can recognize sign language motions in action or video sequences is the main goal of our research. The technique for the sign language gestures was as follows: the video sequences' temporal and spatial properties. Both temporal and geographical aspects have been learned through the use of models. The LSTM model of the recurrent neural network was used to train the model using spatial information from the video series.

The proposed system provides an efficient way to translate sign language into text language with good performance. The system can be used in many applications like engaging the young children with computers by sign-language understanding.

II Literature Study

1.Sign Language Recognition (SLR), which tries to translate Sign Language (SL) into speech or text, aims to enhance communication between hearing-impaired people and able-bodied people. Because sign language is intricate and varies for different individuals, this problem is challenging and has a big social impact. [1] .

2.Many different sign language recognition (SLR) algorithms have been developed by researchers, but they can only distinguish between distinct sign motions. In this research, we propose a modified long short-term memory (LSTM) model for continuous sequences of gestures, also known as continuous SLR, that may be able to recognize a collection of related gestures.[2] .

4.Systems that comprehend sign language instantly translate signs in video feeds to text. Utilizing convolution neural networks (CNNs), feature pooling modules, and long short-term memory networks, a novel isolated sign language recognition model is developed in this study. (LSTMs). [3,4].

5.Deep learning methods can be applied to overcome communication barriers. To identify and detect words in a person's gestures, the model discussed in this paper makes use of deep learning. [5,6] .

6.Hand and body gestures are used to symbolize the vocabulary of dynamic sign language. This approach uses a combination of Media Pipe and RNN models to address problems with dynamic sign language detection. We were able to determine the position, shape, and orientation of the objects by removing important hands, body, and facial parts using Media Pipe.[7,8].

7.Using the sign language datasets and the human key points deduced from the face, hands, and other bodily parts, we develop a sign language recognition system.[9,10].

8. It is still challenging for non-sign language speakers to communicate with sign language users or signers, despite the fact that sign language has lately gained more popularity. Recent developments in deep learning and computer vision have led to promising success in the areas of motion and gesture recognition using deep learning and computer vision-based techniques. [11,12].

III FRAMEWORK DEVELOPMENT

The RNN neural network has been combined with the long short-term memory networks (LSTM) model, which the system uses to anticipate sequences. Numerous sign language recognition (SLR) systems have been developed by researchers, but they can only identify specific sign motions. In the present study, we propose a modified long short-term memory model (continuous SLR) for continuous sequences of gestures that may recognize a collection of related gestures. LSTM networks were researched and employed for the classification of gesture data because they can learn long short-term associations. The created model demonstrated the potential of using LSTM-based neural networks for sign language translation, with a classification accuracy of 98% for 26 motions.

A. Methodology

The method we propose can identify a variety of motions by recording video and converting it into distinct sign language labels. After being categorized and matched to a picture, manually created pixels are then compared to a trained model. Because of this, our system is very good at finding certain character labels. Our proposed system recognizes various actions in video recordings and separates them into discrete frames using sign language. Our method is quite tight in determining specific text labels for characters since the hand pixels are divided and matched to the generated picture before being transferred for comparison with a trained model. Collaborative Communication, which enables users to communicate successfully, is a feature of the suggested system. The proposed system contains an Embedded Voice Module with a User-Friendly Interface to overcome language or speech barriers. The fundamental benefit of this proposed system is that it may be used for communication by both sign language users and verbal speakers. The suggested system is written in Python and uses the YOLOv5 Algorithm, which has modules like a graphical user interface for simplicity, a training module to train CNN models, a gesture module that enables users to create their own gestures, a word-formation module that enables users to create words by combining gestures, and a speech module that turns the converted text into speech. Our suggested approach is intended to alleviate the issues that deaf people in India confront. This system is intended to translate each word that is received.

Here the collection of Gestures are Recognized by using Open CV in the Real Time scenario.The Open CV takes the sequence of frames frequently and gives the desire Output .

B. Data Collection

The Media Pipe Holistic pipeline is used to combine the separate posture, face, and hand models, each of which is optimized for a specific platform. The pose estimation model treats their input as a video frame with a lower, fixed resolution (256x256).

1. Hand Landmarks: Each hand has 21 landmarks.
2. There are 33 Pose Landmarks in total.
3. Face Landmarks: There are a total of 468 landmarks

i. Hand Landmarks

The user experience can be enhanced across a range of technical platforms and disciplines by comprehending the shape and motion of hands.

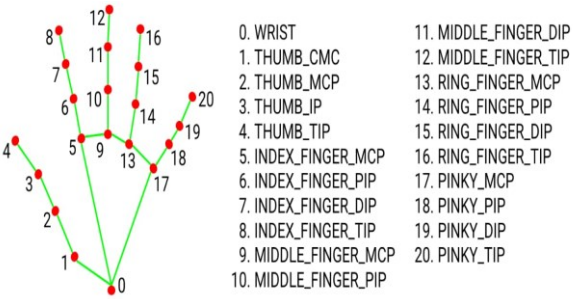

Fig 1 Hand Landmarks

Media Pipe Hands is a high-precision hand and finger tracking solution. In contrast to current state-of-the-art methodologies, which usually require strong desktop workstations for inference, our solution offers real-time performance of many hands scale.

ii Pose Landmarks:

A backdrop segmentation mask and 33 3D landmarks are extracted from RGB video frames using the machine learning (ML) technique known as Blaze Pose. We only consider landmarks at 17 significant locations in the COCO topology. The Media Pipe Pose Landmark Model predicts 33 pose landmarks. The Facial Transform module fills the gap between facial landmark estimation and precise real-time augmented reality (AR) applications. The 3D landmark network receives a clipped video frame as its input. The model outputs the 3D point coordinates and the probability that a face is present and correctly aligned in the input data. By adjusting predictions and iteratively bootstrapping, we can increase the stability and accuracy of our model.

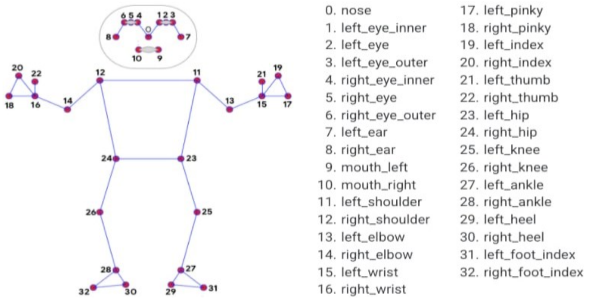

Fig 2: Pose Landmarks

iii.Face Landmarks

Using Media Pipe Facial Mesh technology, 468 3D face landmarks are evaluated in the actual world. Machine learning was used to build the 3D facial surface (ML). It doesn't require a separate depth sensor and simply requires one camera input. The face transform data consists of typical 3D primitives such as a triangular face mesh and a face position transformation matrix.

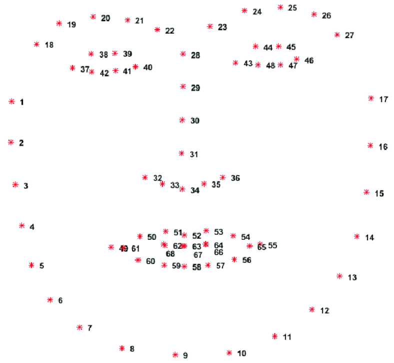

Fig 3:Face Land marks

This face Land Marks are used to Identify and Represent the key points of Human Face marks and it is useful during the capturing of frames in Open CV.

Experimental Design:

The figure 4 describes the system architecture employed in experimental in detail.

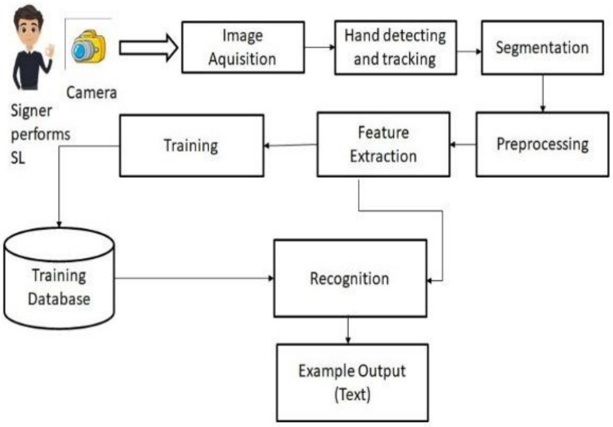

Fig 4. System Architecture

A. Image Acquisition

It is the process of removing a picture from a source, usually one that is hardware-based, for image processing. The hardware-based source for our project is Web Camera. Due to the fact that no processing can be done without a picture, it is the initial stage in the workflow sequence. The image that is obtained has not undergone any kind of processing.

B. Segmentation

It is a method of removing objects or other background details from a recorded image. The segmentation procedure makes use of edge detection, skin color detection, and context subtraction. Order to recognize gestures, the motion and position of the hand must be classified as well as identified

Edge Based Segmentation is used in this project to achieve Segmentation.

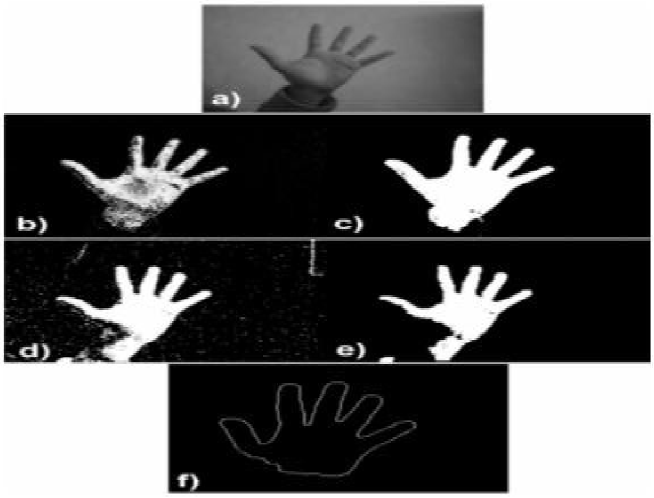

Fig5: Preprocessing process:

C. Preprocessing:

Images need to be processed before they can be used by models for inference and training. This includes, but is not limited to, changes in colour, size, and orientation.

Additionally, preprocessing a model can shorten the training process and speed up inference. Shrinkage of extremely big

input photos will greatly shorten the training period without significantly affecting model performance.
The following are the stages of preprocessing:

1.Morpholical transform:
Morphological processes create an output image of the same size by using the structural characteristics of an input image. By comparing the matching pixel from the input image to its neighbour, each pixel in the output image is given a value. Dilation and erosion are two separate types of morphological changes.

2.Blurring:

One example is the blurring of a picture using a low-pass filter. The term "low-pass filter" in computer vision refers to a method for reducing noise from a photograph while keeping the majority of the image. The blur must be finished before tackling harder tasks, including edge detection.

3.Recognition:
Children who have hearing loss are at a disadvantage since they find it difficult to understand the lectures that are shown on the screen. The American Sign Language was developed to assist these kids in managing their schooling as well as to make daily life easier for them. To assist these kids in learning, we came up with a model that would let them make ASL motions to the camera, which would then interpret them and give feedback on what language was understood. To do this, we combined Media pipe Holistic with OpenCV to determine the essential indicators of the poser with all the values that needed to be collected and trained on the Long Short Term Memory.

4.*Text Output:*
Recognizing diverse postures and bodily gestures, as well as converting them into text, to better understand human behaviour.

D. Feature Extraction

In order to extract preset properties from the already-possessed images, such as shape, contour, geometrical feature (position, angle, distance, etc.), color feature, histogram, and others that are later utilized for categorizing or identifying signs. Feature extraction is a stage in the dimensionality reduction procedure that isolates and arranges a sizable collection of raw data. Class sizes were reduced to more manageable numbers. As a result, processing would be simpler. These massive data sets profusion of variables is their most notable feature. These variables demand a large amount of CPU processing power. The best feature can be extracted from huge data sets via function extraction, which selects and combines variables into functions to reduce the amount of data.

The Feature extraction is achieved by storing key point(face,pose, hand) values in an array using Media pipe.

State Chart Diagram:

An image of a state chart diagram represents a state machine that exemplifies class behavior. By simulating the life cycle of objects from each class, it depicts the actual state changes rather than the procedures or commands that brought about those changes. There are mainly two states in the State Chart Diagram: 1. The Initial condition and the second condition the Final State. Following are a few of the elements of a state chart diagram:

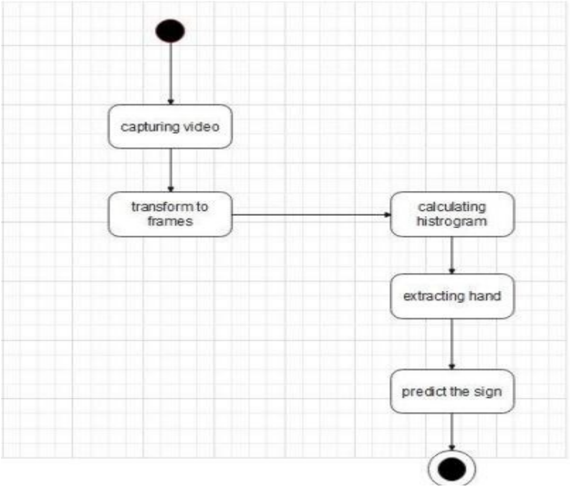

Fig 3 State Chart Diagram

Fig 5. State Chart diagram

State: It is a situation or stage in an object's life cycle in which it encounters a recurring issue, performs an action, or waits for an outcome.

Transition: A "transition" between two states illustrates how an object in the first state acts before transitioning to the second state or event.

An event is a description of a significant occurrence that occurs at a certain time and location.

The state chart diagram given in figure 3 illustrate how the video frame is captured from web cam is processed. It associate extracting image frame and identify the sign represented by the image.

Histogram Calculation are done by using the CV2.calcHist()(in-built function in Open CV).

The signs which are given to the project was predicted by using LSTM model.

Results:

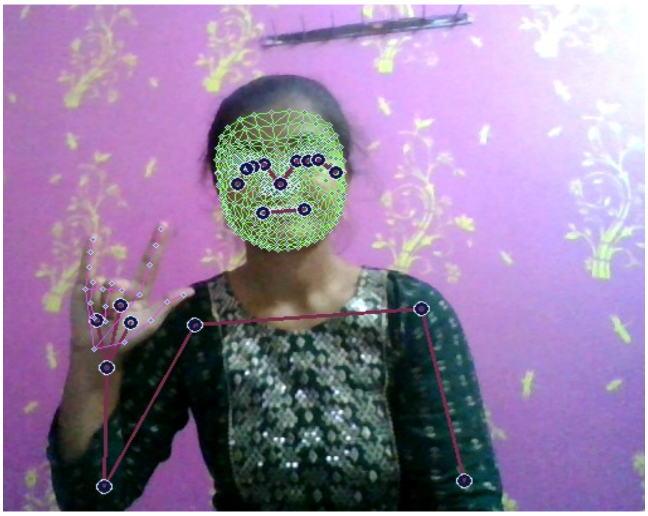

Fig 6. Input for ILoveYou Sign

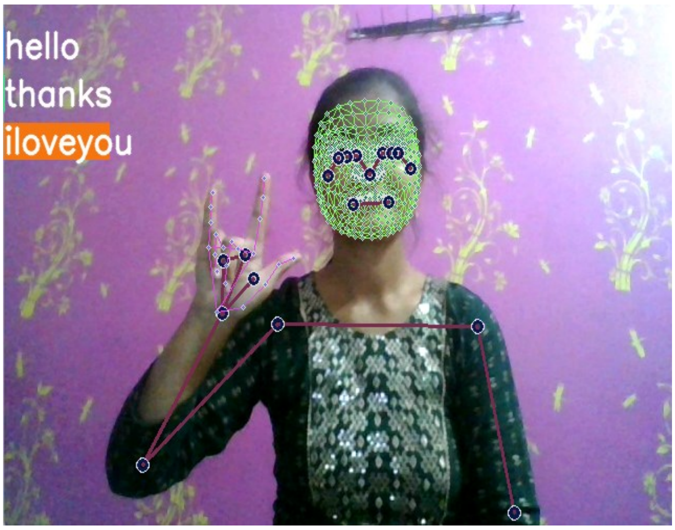

Fig.7 Output for ILoveYou SIgn

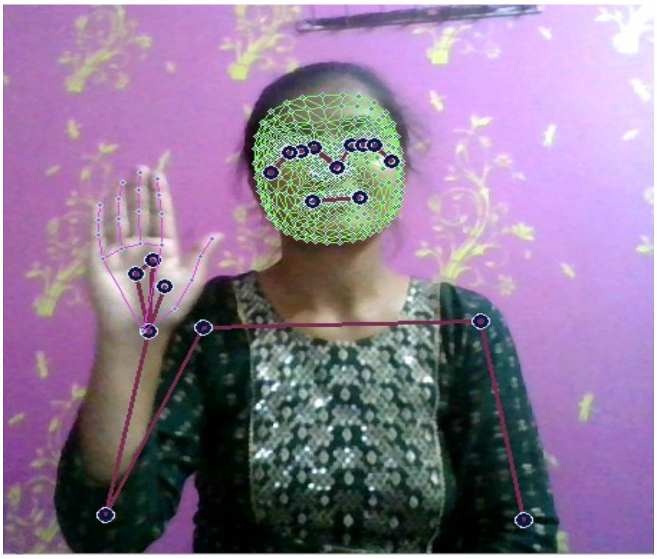

Fig 8 Input for Hello Sign

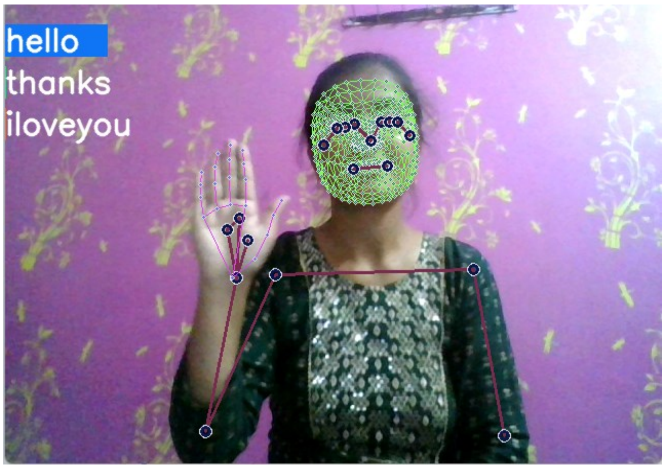

Fig 9 Output for Hello Sign

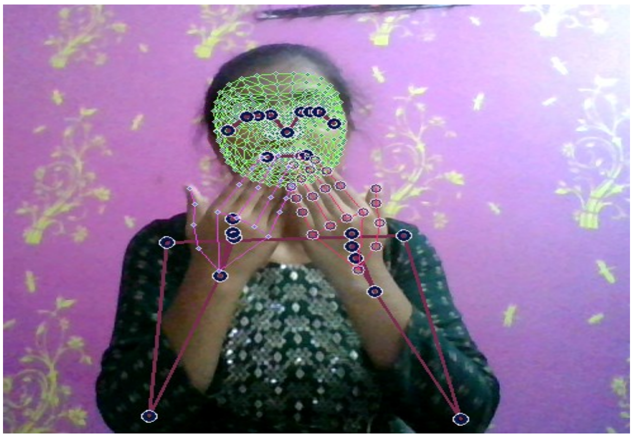

Fig 10 Input for Thanks Sign

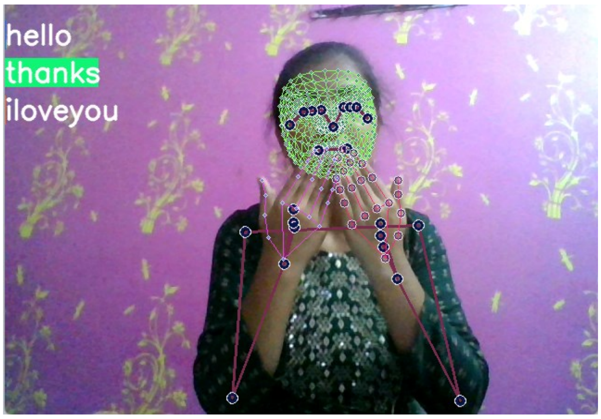

Fig 11 Output for Thanks Sign

Test size of datasets= 10%. because training more set of data will give accurate results/predictions while testing in real time. Train size is 90 %. And while testing it in real time it will give 85 - 90% of accuracy.

IV CONCLUSIONS

There are many potential applications for hand gestures, a potent form of communication, in the area of human-computer interaction. The technique of hand motion recognition using vision has a number of established benefits. Because they have both temporal and spatial features, videos are challenging to analyse. In order to categorise based on the spatial and temporal data, we have employed models. Both attributes were used to classify the data using LSTM. Sign language recognition is a difficult challenge if we consider all the imaginable combinations of gestures that a system of this kind needs to understand and translate.

Having said that, it may be desirable to divide this challenge into more manageable problems, with the method shown here serving as a potential answer to one of them. The system demonstrated that a first-person sign language translation system could be built using simply cameras and convolutional neural networks, despite the fact that it wasn't very efficient. It was found that the model commonly combined various signs, such as U and W. However, after some thought, perhaps it doesn't have to work perfectly as using an orthography corrector or word predictor will increase translation accuracy. The next stage is to analyse the response and look for methods to make the system better.

V Future Scope

For the recognition of single language words and sentences, we can create a model. A system that can recognize changes in the temporal space will be needed for this. By creating a comprehensive offering, we can bridge the communication gap for those who are deaf or hard of hearing. In order to able to translate spoken language into sign language and vice-versa, the systems image processing component needs to be improved. we'll look for any motion related clues. We will also focus on translating the sequence of movements into text, or words and sentences, and then translating that text into audible speech.

VI. ACKNOWLEDGEMENT

We are extremely grateful to Mr. G. Mallikarjuna Rao, our project guide, for his unwavering patience and leadership throughout our project work.

We sincerely thank everyone who has assisted us in meeting our needs for the growth and success of our project work, whether directly or indirectly.

VII References

[1] Liu, Tao, Wengang Zhou, and Houqiang Li. "Sign language recognition with long short-term memory." *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016.

[2] Mittal, Anshul, et al. "A modified LSTM model for continuous sign language recognition using leap motion." *IEEE Sensors Journal* 19.16 (2019): 7056-7063.

[3] Jimmy Jiménez-Salas, Mario Chacón-Rivas, "A Systematic Mapping of Computer Vision-Based Sign Language Recognition", *2022 International Conference on Inclusive Technologies and Education (CONTIE)*, pp.1-11, 2022.

[4] Ozge Mercanoglu Sincan, Hacer Yalim Keles, "Using Motion History Images With 3D Convolutional Networks in Isolated Sign Language Recognition", *IEEE Access*, vol.10, pp.18608-18618, 2022.

[5] Wadhawan, A.; Kumar, P. Sign language recognition systems: A decade systematic literature review. *Arch. Comput. Methods Eng.* **2021**, *28*, 785–81.

[6] Kothadiya, D., Bhatt, C., Sapariya, K., Patel, K., Gil-González, A.B. and Corchado, J.M., 2022. Deepsign: Sign language detection and recognition using deep learning. *Electronics*, *11*(11), p.1780.

[7]Samaan, Gerges H., et al. "MediaPipe's Landmarks with RNN for Dynamic Sign Language Recognition." *Electronics* 11.19 (2022): 3228.

[8] Wadie AR, Attia AK, Asaad AM, Kamel AE, Slim SO, Abdallah MS, Cho YI. MediaPipe's Landmarks with RNN for Dynamic Sign Language Recognition. Electronics. 2022 Oct 8;11(19):3228.

[9] C. Dong, M. C. Leu, and Z. Yin. American sign language alphabet recognition using Microsoft Kinect. In 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 44--52, June 2015

[10] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724--1734, 2014

[11] Bantupalli, Kshitij, and Ying Xie. "American sign language recognition using deep learning and computer vision." *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018.

[12] Wei, Chengcheng, et al. "Deep grammatical multi-classifier for continuous sign language recognition." *2019 IEEE fifth international conference on multimedia big data (BigMM)*. IEEE, 2019.