IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HAND GESTURE RECOGNITION SYSTEM FOR TRANSLATING SIGN LANGUAGE INTO TEXT AND SPEECH

Yojana Gajare[1], Prof. Vina M. Lomte[2], Harsha Bhujbal[3],
Akash Gurav[4], Ishwar Gulanagoudar[5]

[1]Student, [2] Assi. Professor, [3]Student, [4]Student, [5]Student
[1](Dept. of Computer Engineering, RMD Sinhgad School of Engineering, Pune, India)
[2](Dept. of Computer Engineering, RMD Sinhgad School of Engineering, Pune, India)
[3](Dept. of Computer Engineering, RMD Sinhgad School of Engineering, Pune, India)
[4](Dept. of Computer Engineering, RMD Sinhgad School of Engineering, Pune, India)
[5](Dept. of Computer Engineering, RMD Sinhgad School of Engineering, Pune, India)

ABSTRACT

Communication is very essential way of transferring data from one person, place or community to another. Verbal or Oral communication is the way where in individuals engage with others. In today's oral world, a gesture-based language is not mainly popular within the general masses. However, sign language itself is a very evolved language with multiple regional dialects throughout the world. Therefore, to aid with a smoother communication between the signers and non-signer community, technical developments can be introduced. A considerable amount of work has been done in this direction. The basic need of the hour is for a software that could function in real-time and might facilitate real-time conversations among the talking and non-talking world. Conversion of pictures to textual content in addition to speech may be of extraordinary advantage to the speaking and non-speaking masses (the deaf/mute).

In this paper, we introduce a Sign Language Recognition System with the usage of American Sign Language (ASL). The person needs to be capable of seize photographs of hand gestures by using webcam in this analysis, and the gadget need to expect and display the call of the captured photo. The captured photo undergoes collection of processing steps which consist of diverse computer vision strategies together with the conversion to gray-scale, dilation and masks operation. After that it will detect and recognized the sign gesture and translate it into text and speech. In this way, we implemented a virtual dictation signal language translator. Convolutional Neural Network (CNN) is used so one can educate our version and discover the photographs. The basic CNN model has achieved 93.48% of training accuracy and 97.45% of testing accuracy. In similar manner, the proposed model has achieved 97.82% of training accuracy and 99.09% of testing accuracy.

KEY WORDS: **Convolution Neural Networks (CNN), Text-to-speech Algorithms, Machine Learning, Sign Recognition, American Sign Language (ASL).**

I. INTRODUCTION

One of the maximum vital necessities for social survival is communication. Deaf and dumb peoples communicate with each other using sign language, however it's far difficult for non-deaf and dumb humans to understand them. While much study has been done on the recognition of American sign language, Indian sign language varies greatly from American sign language. ISL communicates with two hands (20 out of 26), while ASL communicates with a single hand. Sign language is a way of communicating by using the hand gestures and other parts of the body. According to the World Health Organization, there are around 466 million human beings worldwide who have disabling hearing loss, who frequently use signal language for communication. Sign languages vary from country to country, and we particularly take a look at American Sign Language (ASL) in this paper.

Sign language is a means of conversation used by masses with impaired hearing and speech around the globe. People all over the world use sign language gestures as a way of non-verbal communication to express their thoughts and emotions and to convey information. On the other hand, non-signers find it extremely difficult to process and understand it, which is why skilled and expert sign language interpreters are needed during medical and legal appointments, as well as educational purposes.The need for translating services has risen day by day during the last five years. As a result, they'll offer an easy signal language

interpreting service that may be used, however has enormous restrictions. Recognition System using ASL, uses the Convolution Neural Network (CNN) to translate the pictures into textual contents. ASL is one of the signal languages utilized by deaf and dumb human beings to deliver messages. In our proposed system i.e. *"Hand Gesture Recognition System"* we have use the ASL facts set to recognize alphabets (a-z). First, we convert the films into frames after which pre-processes the frames convert them into grey-scale pictures. After that we construct the Convolution Neural Network (CNN) that classify the frames into 26 different classes which constitute 26 English alphabets. Next, the characteristics calculated during the earlier phase are used to translate the signed gestures into textual content as well as into speech. In this way, we implement a virtual dictation sign language recognition translator system.

II. ALGORITHMIC STUDY

CNN is primarily used for image processing since it includes one or more convolution layers to improve the quality of results. Yann LeCun developed CNN in the 1980s. It is faster than other feed-forward neural networks. It is used in many real-time applications, including Face detection, X-ray image recognition, self-driving or autonomous cars, Analyzing Documents, Cancer detection, Natural language processing (NLP) and many others.

"Convolution" is the combination of two mathematical functions that results in a third function that is used for feature extraction from a particular image. In CNN, the architecture is classified into two parts: first is feature extraction and another is classification. Feature extraction process is something which separates and identifies specific features or characteristics of an image. The classification layer or fully connected layer takes the output of the Convolution layer and predicts which class the image belongs to. CNN consists of three layers, namely the Convolution, Pooling and Fully Connected layers.

2.1 Convolution Layer:

Convolution layer's goal is to take input from the user and apply a set of filters and parameters which can be learned through training and generate the feature map or activation map. Filter sizes are usually smaller than the input image, so we multiply element by element the input image and filter image to generate one matrix.

2.2 Pooling layer:

Pooling layer is used to reduce the size of the activation map. There are three types of pooling: Min pooling, Max pooling, and Average pooling. In Min pooling the pixel value with the lowest value is chosen, for Max pooling the pixel value with the highest value is selected and in Average pooling the block average is taken.

2.3 Fully-Connected Layer:

The fully-connected layer is a feed-forward neural network. In this layer, the input is taken from the activation function or convolution layer and reformed to produce a single vector. This vector is used as input for the next hidden layer. Fully connected output layers give actual output to the user.

2.4 Dense layer:

Dense layer takes input from previous layer and provide to the next layer.

2.5 Dropout:

In a fully connected layer, all features are connected to each other, so you might experience overfitting. To resolve overfitting problem 20%-50% random data is dropped.

2.6 Rectifier Linear Unit (relu) layer:

If it receives any negative input, it returns 0, but if it receives any positive input, it returns that value.

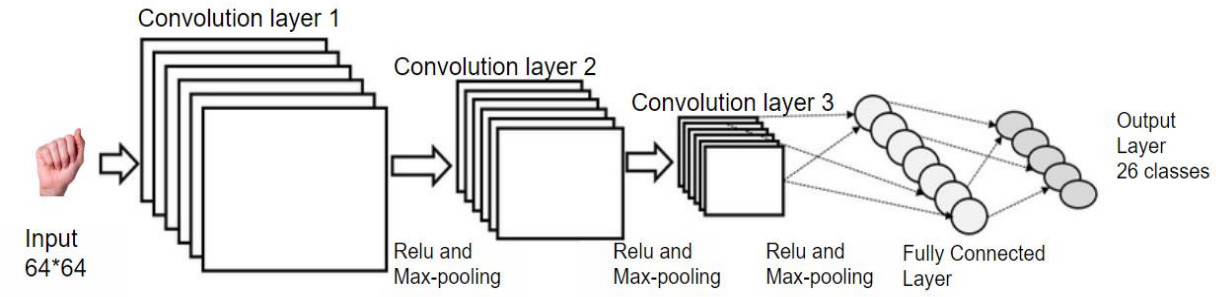

Fig. 1: The CNN Architecture

III. PROPOSED METHODOLOGY

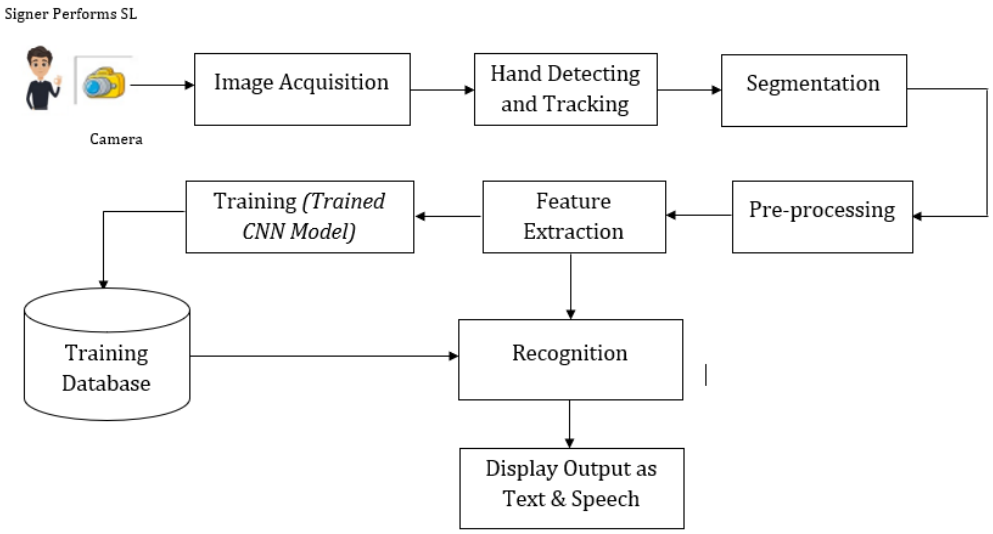

Fig. 2: Proposed Methodology

3.1 Image Acquisition

Image Acquisition is the technique of extracting an image from a source, usually a hardware-based source is used for the process of image processing. In our project, web camera is the hardware-based source for capturing the hand gesture images which is signed by the user or signer. This OpenCV video circulation is used to seize the complete signing duration. The frames are extracted from the circulation and are processed as grayscale images with the measurement of 64*64. This measurement is steady all through the project because the complete dataset is sized precisely the same.

3.2 Hand Region Segmentation & Hand Detection and Tracking

The images which were captured through webcam are further scanned for hand gestures. This is done before the image is fed to the model to obtain the predictions. The segments which contain hand gestures are made more pronounced. This increases the chances of prediction by many folds. The technique of separating objects or signs from the context of a captured image is known as segmentation. Background or context subtracting, skin-color detection, and edge detection are all included in the segmentation process. The motion and location of the hand must be detected and segmented in order to recognize hand gestures.

3.3 Image Pre-processing

Training the raw images as it is might lead to poor performance. That's why for that reasons, simple image processing algorithms can be applied to achieve higher accuracy. Image processing algorithms such as RGB to gray conversion reduce the time required for training purpose and also reduce power consumption. The noise from the images can also be eliminated.

3.4 Feature Extraction

Selection and extraction of important features from images is the most important part in image processing. Images when captured and saved as a dataset usually acquired whole lot of space as they are comprised of a large amount of data. Feature extraction enables us which will resolve this problem by decreasing the data after having extracted the essential features automatically. It additionally contributes in maintaining the accuracy of the classifier and simplifies its complexity. In our case, the features found to be critical are the binary pixels of the images. Scaling the images to 64 pixels has led us to get enough characteristics to effectively classify the American Sign Language gestures.

3.5 CNN Training & Training Options

For this project deep learning is used. Training options are set accordingly before training the database using any CNN architecture. The training options are large batch size, number of the epoch, and learning rate.

3.6 Image Database

The database contains pictures of various hand gestures. These images are taken from various users with multiple repetitions. The resolution of the images may be varying. Different datasets are available for American Sign Language (ASL).

3.7 Recognition

We'll use classifiers in this case. Classifiers are nothing but the methods or algorithms that are used to interpret the signals. Popular classifiers that identify or understand sign language include the K-Nearest Neighbor classifiers, Support Vector Machine (SVM), Artificial Neural Network (ANN), and Principal Component Analysis (PCA), among others. However, in this project, the classifier is CNN. For image classification and recognition CNNs are used due to its high precision. It uses a hierarchical model which builds a network, same as a funnel, and further outputs a fully-connected layer in which all neurons are connected to each other and the output is processed and generated

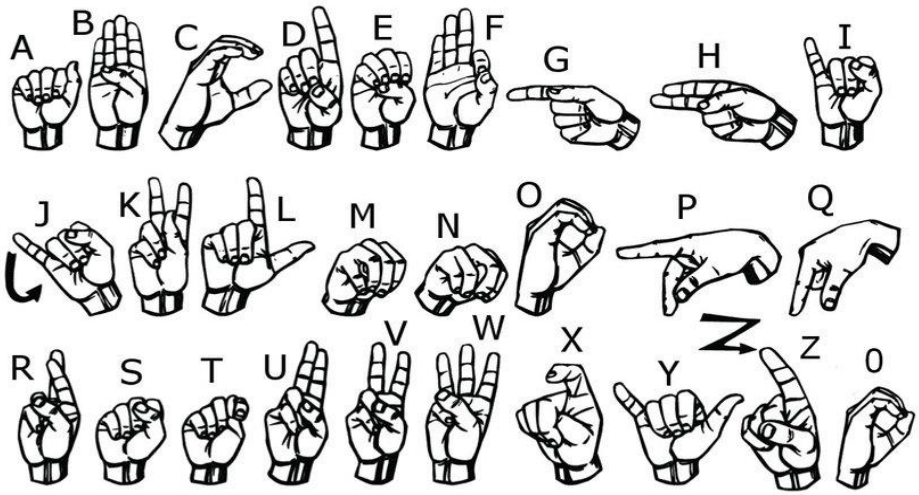

Fig. 3: ASL Alphabet Gesture Symbols that will be in the training data

3.8 Display as Text & Speech

The model accumulates the recognized gesture to words. The recognized words are translated into the corresponding speech using the pyttsx3 library. The text to speech result is a simple and easy work around but is an invaluable feature as it gives a feel of an actual verbal or oral conversation.

IV. SYSTEM ARCHITECTURE

The user must first log in using the appropriate login and password before using the system. The video camera input is processed using the OpenCV library. Hand gestures are captured on video. The captured image is resized to 64*64 pixels. The input image is a colored image that has been transformed to grayscale image. The image obtained is pass to the CNN model. Image pre-processing, feature extraction, and classification are all done in the CNN model. The procedure of feature extraction is used to reduce the size of data, Appropriate filters are applied to image and generate the activation map. The activation map's output is sent into a neural network, which determines which class the inputted image belongs to. The image name will then be displayed on the video streaming. The text is then converted to voice using the Python pyttsx module.

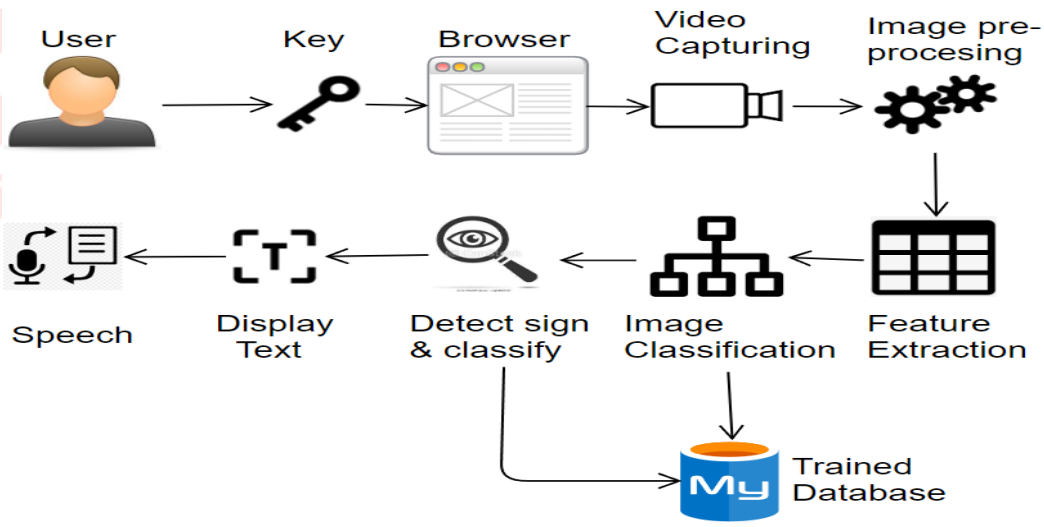

Fig. 4: System Architecture

V. RESULTS

When we train the version, the accuracy and loss in version for validation data will vary based on a variety of factors. Loss should typically reduce as each period progresses, but exactness should increase. However, with validation loss (keras validation loss) and validation accuracy, several scenarios may arise, such as those listed below.

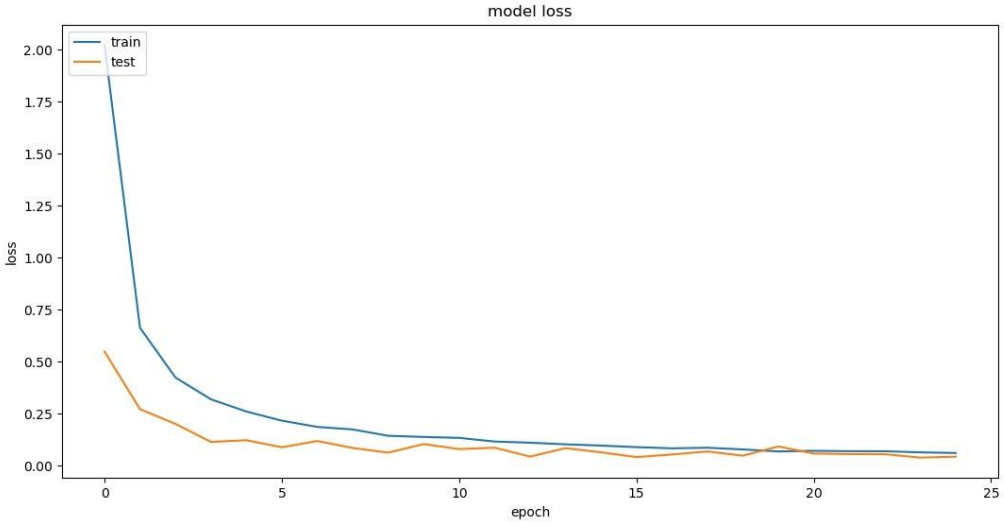

Fig. 5: Training and Testing Loss

1. Validation loss increases, whereas validation accuracy decreases. This implies that the model is cramming values rather than learning them.

2. Validation loss begins to decrease, while validation accuracy begins to rise. This is also fine because it means the model is learning and dealing properly. We obtained the following findings after testing our model: we drew the graph of accuracy and loss with regard to epochs.

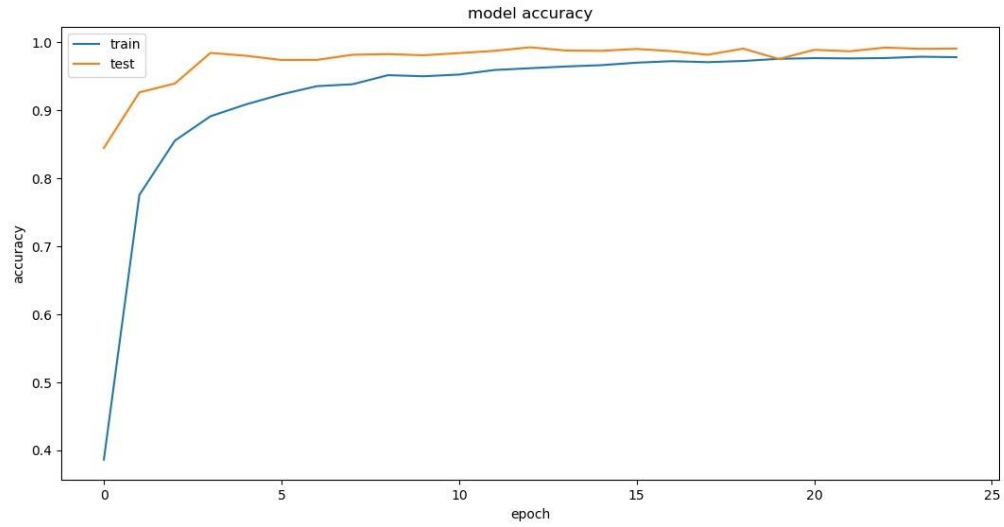

Fig. 6: Training and Testing Accuracy

5.1 Comparison of basic CNN and proposed CNN accuracy models:

CNN	Training Accuracy	Testing Accuracy
Basic CNN	93.48 %	97.45 %
Proposed CNN	97.82 %	99.09 %

5.2 Accuracy Increases per Epoch:

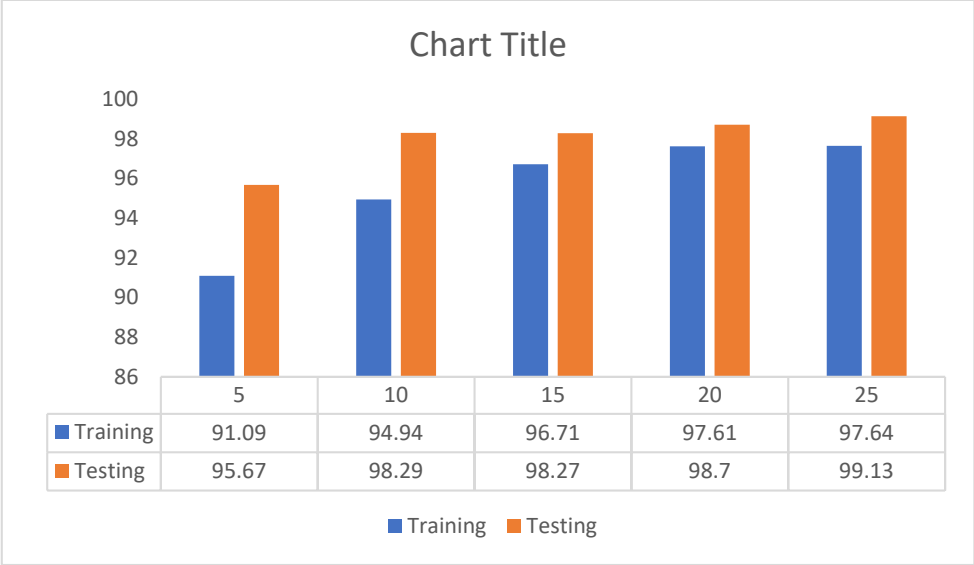

Fig 7: Accuracy Increases per Epoch

5.3 Comparison Chart:

Ref. No.	Paper Title	Algorithm Used	Dataset Used	Accuracy
1	Sign Language to Text and Speech Translation in Real Time Using Convolutional Neural Network.	Convolutional Neural Network (CNN).	American Sign Language (ASL) from Kaggle Dataset	90%
2	Real-time Conversion of Sign Language to Text and Speech.	Support Vector Machine (SVM).	ASL (American Sign Language) Kaggle Dataset.	97%
3	Sign Language to Speech Translation.	Convolutional Neural Network (CNN).	Indian Sign Language (ISL) from Kaggle	85-95%
4	Sign Language Recognition for Speech and Hearing Impaired by Image Processing in MATLAB.	Quadratic SVM.	American Sign Language (ASL) Dataset.	85%
5	Convolutional Neural Network based Bidirectional Sign Language Translation System.	Convolutional Neural Network (CNN), Recurrent Neural Network	American Sign Language (ASL) from Kaggle Dataset	92.68%
6	Proposed Method	CNN & Modified CNN	American Sign Language (ASL) from Kaggle Dataset	Basic CNN- 84.61 Modified CNN- 92.30

5.4 Tested Alphabets:

Letter	Correct Output	Letter	Correct Output
A	YES	Very Good	YES
B	YES	Good Job	YES
C	YES	Yes	YES
D	YES	No	YES
E	YES	Stand Up	YES
F	YES	Gate Open	YES
G	YES	Be in time	NO
H	YES	Do task	YES
I	YES	Any Help	YES
J	YES	Place order	YES
K	NO	Meetings	YES
L	YES	Use Mobile	YES
M	YES	Do your Homework	YES

VI. CONCLUSION and FUTURE WORK

6.1 Conclusions

The American sign language recognition and translator application is a Convolution Neural Network for recognition of hand gestures. This application will help to bridge the gap between general people (normal people) and non-speaking people. This system shows higher accuracy in identifying the sign language characters including words and sentences. It will also capable of converting textual content into speech as well.

6.2 Future Scope

In future work, proposed system can be developed and implemented using Raspberry Pi. Image Processing part can be improved so that the system would be able to communicate in both directions i.e it can be capable of translating general language to sign language. Reduced Background Interference: After conducting relevant research and carrying out various experiments, better approaches can be coined for reducing the background interference in the detection phase. Reduce gesture detection time: The user wait time in between consecutive gesture detection can be reduced, thus making the user interface more smoothly.

REFERENCES

1. "Sign Language to Text and Speech Translation in Real Time Using Convolutional Neural Network", Ankit Ojha; Ayush Pandey; Shubham Maurya; Abhishek Thakur; Dr. Dayananda P; 2020 International Journal of Engineering Research & Technology (IJERT)

2. "Real-time Conversion of Sign Language to Text and Speech", Kohsheen Tiku; Jayshree Maloo; Aishwarya Ramesh; Indra R. 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)

3. "Sign Language to Speech Translation", Aishwarya Sharma; Siba Panda; Saurav Verma, 2020, 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)

4. "Sign Language Recognition for Speech and Hearing Impaired by Image Processing in MATLAB" Parama Sridevi; Tahmida Islam; Urmi Debnath; Noor A Nazia; Rajat Chakraborty; Celia Shahnaz, 2018, IEEE Region 10 Humanitarian Technology Conference (R10-HTC)

5. "Convolutional Neural Network based Bidirectional Sign Language Translation System", Lance Fernandes; Prathamesh Dalvi; Akash Junnarkar; Manisha Bansode; 2020, Third International Conference on Smart Systems and Inventive Technology (ICSSIT)

6. "Hand Gesture Recognition System for Translating Sign Language into Text and Speech [English/Marathi]", Harsha Bhujbal, Vina M. Lomte, Yojana Gajare, Akash Gurav, Ishwar Gulanagoudar