DEEP LEARNING MODEL FOR SIGN LANGUAGE INTERPRETATION USING WEB CAMERA

Gouri Nandan and Dr. Neeba E A

[1]Dept. of Information Technology, Rajagiri School of Engineering & Technology, Ernakulam, Kerala, India. Email:gouri.gourinandan17@gmail.com

[2]Dept. of Information Technology, Rajagiri School of Engineering & Technology, Ernakulam, Kerala, India. Email: neebarset@gmail.com

Abstract— Sign languages are languages that solely utilize gestures to convey meaning. Communication, based on the sign language is a mix of manual explanations and non-manual elements. Sign language recognition framework positively reflects communication between the person who is hard of hearing and world around. It also helps in communicating with machines. One of the most utilized types of gesture based communication is the American sign language (ASL). In the proposed work, the letters are detected from a video frame using convolutional neural network (CNN) and then converted into speech using Google Text-to-Speech (gTTS). The systems are trained with 75% of images and tested with 25% of images from the database.

Keywords— Deep learning, Sign Language Interpretation, Convolutional Neural Network

1. OVERVIEW

The Indian census of 2011 reveals that over one million individuals are deaf. Estimates show that 18 million individuals from the Indian National Association of the deaf have hearing impairment. Among the Indian populace, approximately 1 percent is hard of hearing. Indeed, this is much less, when compared to the fact that 3.5% of the American population and 5% of the global population are hearing impaired. As there is no universal gestural language, different countries like Britain and America have adopted their own versions of the sign language.

Gesture based communication acknowledgment frameworks can facilitate correspondence between two communities. The job of the mediator is to encourage communication between those who are hard of hearing and those who can hear. These services are required in varied environs like schools, offices, hospitals, court rooms and government establishments.

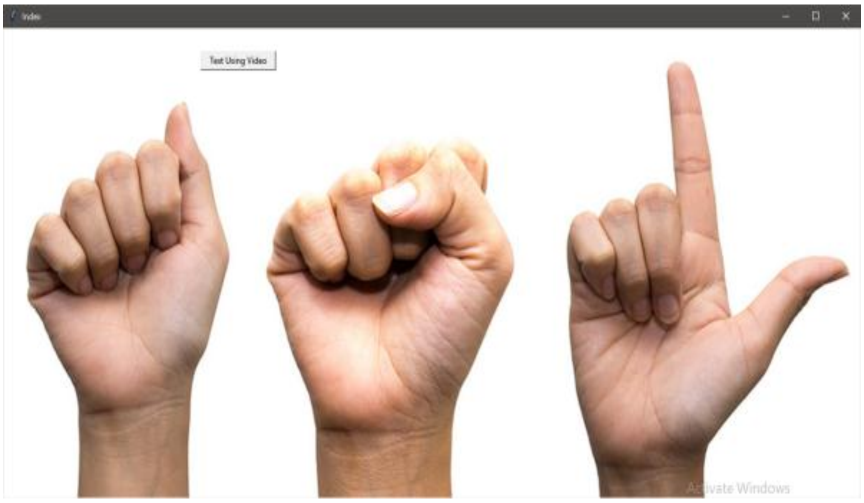


Fig 1 : Example of sign language

Sign language interpretation method is a superior yet economic mode to assist the deaf and mute people by translating their gestures into text and speech in real time. It comprises a simple innovative digital interpreter that works by setting a smart-phone before the user while the application translates gestures or gesture based communication into speech and text. Being affordable and accessible, these interpretation services are in great demand within the deaf community. Besides, varied organizations around the world face many problems in offering their assistance to the hearing impaired.

The sequence is as follows: Section II reviews the literature in the Sign language interpretation. Section III outlines the proposed method. Section IV incorporates the discussion with the results and Section V elucidates the conclusion.

2. RELATED RESEARCHES

Sign language interpretation is a rapidly expanding field due to the need for gesture recognition through affordable technology. Manually monitoring live feeds is not as efficient as automatic detection software. The various methods employed for automatic Sign language interpretation are presented below.

Manar Maraqa et al.[1] discussed in his paper that this assessment strives to present the utilization of varied neural frameworks in human hand motion acknowledgment for static imagery similar to dynamic motions. This centers close to neural systems for enabling Arabic Sign Language hand sign identification. This displays the utilization of feed-forward and intermittent neural systems with its varied designs; both in part and completely repetitive systems. This framework demonstrates an accurateness of 95% for static motion acknowledgment.

Omar Al-Jarrah et al.[2] proposed a framework with Adaptive Neuro-Fuzzy Inference System (ANFIS) using Hand gestures that play a significant role in communication among individuals during their day-to-day lives. Gesture based communication is the essential specialized scheme among those who are hard of hearing. An interpreter is required whenever an individual needs to speak with the deaf community. Herein, the work targets building up a framework for programmed interpretation of motions of the manual letters in the Arabic communication via gestures. The framework doesn't depend on utilizing gloves to achieve the acknowledgment work but manages pictures of exposed hands, w that permit clients to associate with the framework in a characteristic manner. The least-squares estimator and the subtractive clustering algorithm help acknowledge this ambiguous method and training is done via hybrid learning algorithm.Experimentations uncovered this framework's option to identify 30 Arabic letter sets at an exactness of 93.55%.

Kanchan Dabre et al.[3] proposed in the paper that gesture based communication is the essential technique for hearing impaired individuals. Individuals with hearing incapacities face obstacles in speaking with others in the absence of an interpreter. Therefore, the usage of a framework that perceives the communication via gestures would have a huge advantage on the social lives of individuals with hearing difficulties. This paper proposes a marker-free, graphic Indian Sign Language recognition framework. . Besides, the attributes of video images through web cam can be identified by utilizing image processing, computer vision and neural systems.

This methodology will convert the day-by-day video of frequently utilized full sentence gestures first into a text and then into speech. Image processing operations are carried out in series format to identify the hand shape from continuous frames. The Haar Cascade Classifier helps interpret signs with relevant meaning. Lastly, the speech synthesizer is used to convert the displayed text to speech.

Salma Hayani et al.[6] implemented an automatic recognition framework for Arab sign language (ArSL). In this work, a framework dependent on the convolutional neural network and fed with a real dataset perceives numbers and letters of the Arab gesture language. To approve this framework, a relative report that shows the efficacy of the suggested method is contrasted with conventional methodologies based on k-nearest neighbors and support vector machines.

3. PROPOSED METHOD

This proposed system is an approach for sign language interpretation. Firstly, the videos are captured, and then the gestures are identified with a text output. The text is then converted into speech. The stages are described in detail below.

Dataset: For classification, a series of imagery collection is necessary. Several imagery datasets such as Flickr30k, MS COCO, SBU, Pascal, Kaggle etc. can be accessed. Network training is done via the Kaggle dataset with its 87,000 images to train and 87,050 images to test. Here the images are clicked in different light situations and from various angles. These are used for testing as well as training purposes.

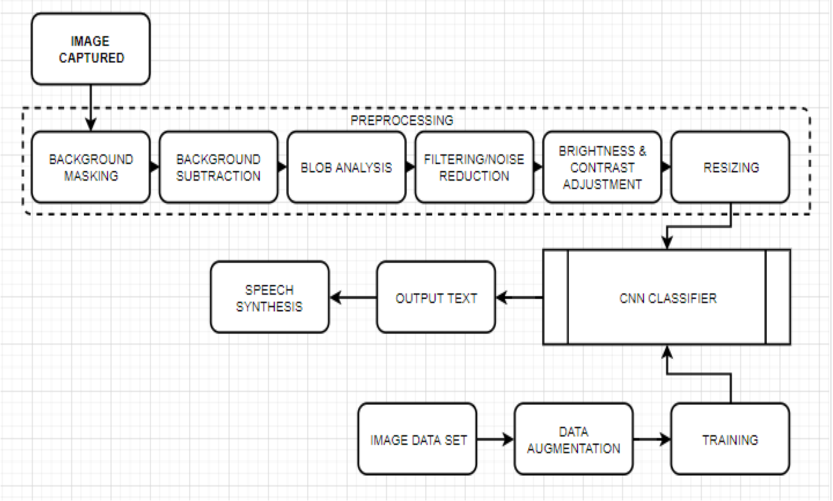


Fig 2 : System architecture

Image Acquisition: A video is filmed by employing a camera and is regenerated into series of images so as to accumulate the principle of image processing.

Pre-processing: This stage comprises separating outline from video stream and performing imagery preparatory steps to obtain image aspects by performing background subtraction, Blob analysis, noise reduction, grayscale transformation, brightness normalisation and scaling operation one at a time.

i. Background Subtraction : This stage includes expelling undesirable foundation subtleties from image frames of video stream and obtaining just hand signs to perform imagery preparatory steps.

ii. Blob Analysis : A blob is a locale having the same properties and pixel values which are consistent or different inside a recommended range. This progression finds locale of enthusiasm for additional processing by discovering every connective piece of the frame and picking the greatest (biggest territory) among them (since the hand is largest by far the biggest). Blob investigation is pertinent in the field of item tracking or item acknowledgment .

iii. Noise Reduction : Decrease in noise is intended to filter the irregularity and commotion by utilizing smooth Gaussian filter. This expels the commotion by the smoothening activity. The Gaussian kernel size used for this is 3.

iv. GrayScale Conversion : This progression changes colour imagery into grayscale which aids more calculations on pixel activities and interrelated signs. The memory space for storing grayscale images is less than that required for coloured ones.

v. Brightness and Contrast Normalization : Imagery obtained in less brighteness have close contrast values. Thus there is a requirement to alter pixel intensity values. Histogram equalization is used to change and standardize contrast and brightness of the processing frame.

vi. Image Scaling : This helps decrease the computational exertion required for picture handling. Each picture will be scaled to 45*45 sizes for additional processing.

Data Augmentation: The primary pre-processed imagery were solely utilized for network training . Subsequently, real-time data augmentation utilized through the training helped improve the network's localization ability. The images were randomly augmented during every epoch.

Convolutional Neural Network: This class of simulated neural networks are used successfully in analysing visual images. It is basically used to resolve tough image-based pattern recognition tasks. CNN has similarity with traditional ANNs in that they comprise neurons that optimize themselves through learning. For each neuron acquires an input and executes an action. The whole of the network will still represent a single score function, which is the weight, from input image vectors to the final result that contains the class score. CNN's are mainly utilized in the field of pattern identification within images.

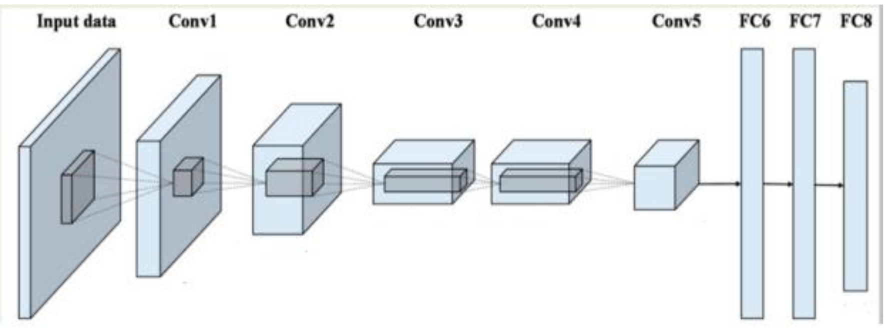


Fig 3 : Basic alexnet architecture

Alexnet: It has 5 convolutional layers with 3 wholly connective layers. It employs the Rectified Linear Unit for the non-linear portion. The advantage ReLu has over sigmoid is that it trains so much faster. It reduces over- fixing with the Dropout layer following every FC layer. Its velocity is 5 times quicker with same precision. This model has 7 hidden layers, 650K units and 60M parameters and hence has good accuracy.

Speech conversion: Instead of speaking out letters, the words are pronounced. The output text is converted to speech using gTTS. This alters text into human-like dialects of over 100 voices in 20+ languages. The speech can be delivered in any of the two accessible audio speeds,i.e., fast or slow.

Layers of Alexnet architecture:

i. Convolution Layer: This main level extricates highlights of the input imagery. The pixels are associated through the study of these features by utilizing little squares of input data via convolution.

ii. Non Linearity: This represents Rectified Linear Unit for non-linear activity wherein the objective is to introduce non-linearity. .

iii. Dropout: It is to keep the system from over fitting.

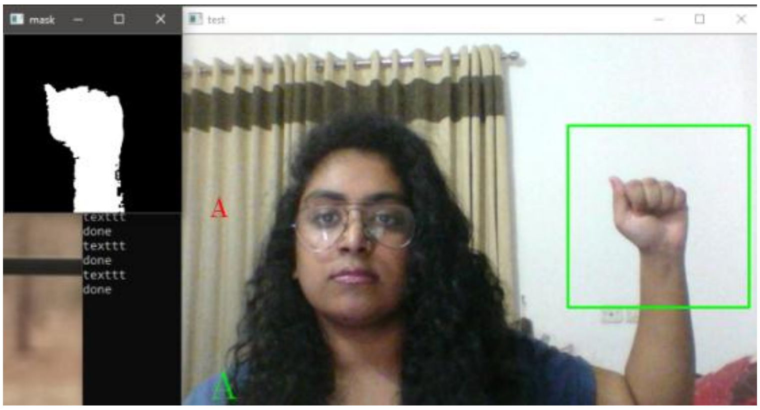
iv. Pooling Layer : This decreases the amount of constraints for imagery that is too enormous.

v. Fully associated : The objective here is to take the consequences of the pooling procedure and use them for classification of images into labels.

vi. Softmax Layer : It permits the yield to be deciphered in a direct manner as a likelihood.

4. RESULTS

The basic working is described in figure 4, wherein the system captures an image of the gesture through webcam.



```
texttt
done
texttt
done
texttt
done
```

Fig 4 : Window capturing an image

The sign language recognition scheme was trained with varied imagery and tested with sample cases. At first, the estimation background frame is subtracted from the video frame and the background image is replaced by grayscale imagery.

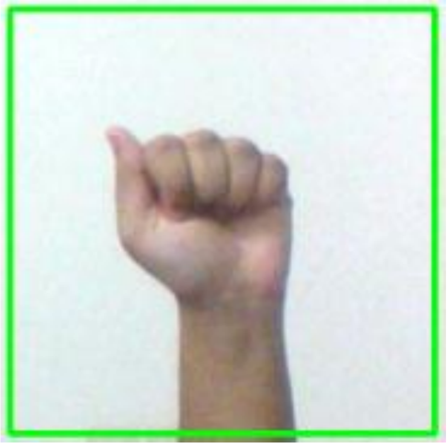


Fig 5 : Input gesture showing letter A

From a grayscale imagery, thresholding can be utilized for making binary images. This is achieved by altering all pixels beneath some threshold measures to zero and all higher up the threshold measure to one.

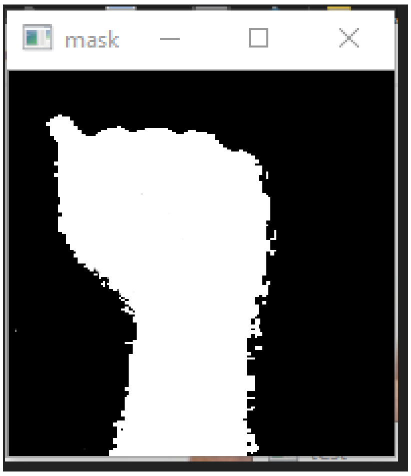


Fig 6 : Masked image of gesture (letter A)

Figure 7 represents the loss/accuracy graph of the model.

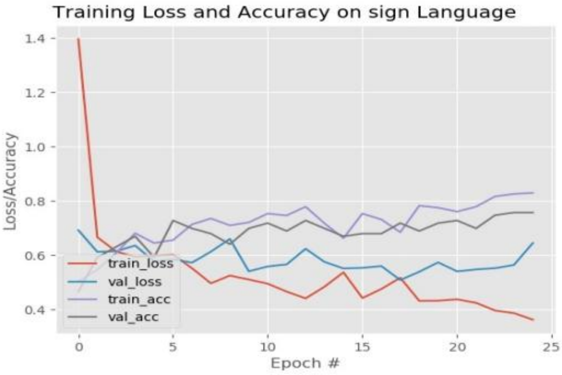


Fig 7 : Loss/accuracy graph

Figure 8 represents the accuracy graph of the model.

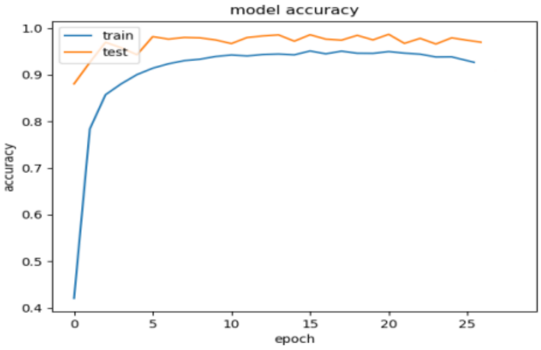
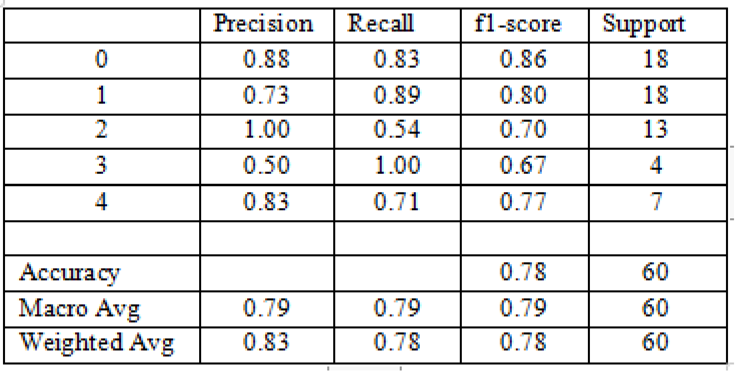


Fig 8 : Model accuracy graph

A categorization score document presented in Table 1 displays the main grouping matrices such as precision, recall, f1-score and support.

Table 1: Classification Score



	Precision	Recall	f1-score	Support
0	0.88	0.83	0.86	18
1	0.73	0.89	0.80	18
2	1.00	0.54	0.70	13
3	0.50	1.00	0.67	4
4	0.83	0.71	0.77	7
Accuracy			0.78	60
Macro Avg	0.79	0.79	0.79	60
Weighted Avg	0.83	0.78	0.78	60

5. CONCLUSION

The deaf community requires the presence of an SL translator when connecting to the world around. However, the absence of SL Interpreters might prevent communication from taking place. Therefore, this project will help the deaf & dumb community to have unhindered communication without depending on them. Gesture identification system can be utilized as communicating media between man and machine like human-computer interaction in VR, gaming and a host of human welfare applications. Besides, this can replace the use of mouse and keyboard by employing gesture applications. It also elucidates different algorithms and outlines the accuracy and efficiency of the proposed method. The gestures made everyday are converted into text and then into the audio format. With their increased accuracy and capability of reducing false interpretations, the methods discussed here can be applied on a large-scale format for sign language interpretation.

References

[1] Manar Maraqa, Farid Al-Zboun, Mufleh Dhyabat, Raed Abu Zitar, “Recognition of sign language using recurrent neural networks”, Journal of intelligent learning systems and applications.

[2] Omar Al-Jarrah halawani, “Recognition of gestures in sign language using neuro-fuzzy systems”, international journal of science engineering and technology research (IJSETR), vol 3, issue 5, May.

[3] Kanchan Dabre, Surekha Dholay, “Machine learning model for sign language interpretation”, International conference on circuits, systems communication and information technology applications (CSCITA).

[4] Bhumika Gupta, Pushkar Shukla and Ankush Mittal, “K-Nearest correlated neighbor classification for indian sign language gesture recognition using feature fusion”, 2016 international

conference on computer communication and informatics (ICCCI-2016), Jan 2016, coimbatore, india.

[5] Soeb Hussain and Rupal Saxena, Xie Han, Jameel Ahmed Khan, Prof. Hyunchul Shin, “Hand gesture recognition using deep learning”, IEEE.

[6] Salma Hayani, Mohamed Benaddy, Othmane El Meslouhi, “Sign language recognition with CNN”, IEEE 2019.

[7] Suharjito, Meita Chandra Ariesta, Fanny Wiryana and Gede Putra Kusuma, “A Survey of hand gesture recognition methods in sign language recognition”, Pertanika J. Sci. & Technol 26(4), 2018.

[8] Rashmi B Hiremath, Ramesh M Kagalkar, “Review paper on sign language recognition techniques”, international journal of computer applications national conference on advances in computing, 2015.