ISSN No. 0976-5697

VOLUME 12 SPECIAL ISSUE 2, JUNE 2021

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

RECOGNITION OF SIGN LANGUAGE USING DEEP NEURAL NETWORK

Pallavi P
School of Computing and IT
REVA University
Bengaluru,India
ppallavi099@gmail.com

Sarvamangala D R
School of Computing and IT
REVA University
Bengaluru,India
sarvamangala.dr@reva.edu.in

Abstract: Speech impairment affects an individual's ability to communicate. People affected by this problem, suffer by the communication barrier with the normal people. Hence, to express their emotions and thoughts people will use sign language for their communication and. Although sign language is common in world, there may be a difficulty for people who do not have knowledge about sign language to communicate with speech impaired people who knows the sign language. Now there is a prominent growth in the field of computer vision and deep learning, which has been tremendous development in the fields of sign and motion recognition. The proposed model focusses on overcoming the challenges in verbal communication between non-sign language speakers and sign language speakers by building a deep learning model to recognize alphabets of American sign language. The model was trained on the dataset collected from Kaggle website which consisted of 26 American Sign Language alphabets. The model achieved 99.3% mean average precision in recognition of sign and average probability value for test image achieved was 0.99. The deep learning method used to build the model achieved more accuracy than the previous solutions.

Keywords: American Sign Language, Deep learning, Neural network, YOLOV3 model, Sign Language recognition.

I. INTRODUCTION

Sign language is a one of the forms of non-verbal communication used by hearing impaired people. Such people make use of sign language as a non-verbal communication to convey their feelings and thoughts. The huge problem is that the normal people cannot understand sign language easily, which builds huge barrier in between non-sign language speakers and sign language speakers. The work focuses 5to solve this problem by using deep learning and neural network models. In sign language there are many forms like American sign language, Indian sign language etc. Primarily our model is built for American sign language alphabets and will be further enhanced to recognize other forms of sign language. Sign languages are not common and they are not equally comprehensible with each other, although there are also prominent resemblances among signed languages. American Sign Language is one of the example for sign languages, which works as the major sign language for the Deaf communities in the US and some parts of Canada. In the US, American Sign Language (ASL) is the third most used language. ASL is an organized and complete visual language that is expressed by movements and motions with the hands as well as facial expression.

Deep convolutional neural networks, the deep learning methods have achieved the best performance in recognition of sign. Detectors based on deep neural networks have been demonstrated for doing a great job in detection and recognition. Deep learning is an arm of machine learning, which is constructed using algorithms that design top level thoughts in data. When the input is given to input layer, then the algorithm will process the input and gives an output and using this output we can modify the succeeding layer. Deep network consists of numerous layers in between the input and output layer. Training the deep neural networks are much tougher. In [6], [8] the author had used a Convolutional Neural Network model, which comprise of many fully connected layers and many convolutional layers. In [9], the author had used 3D ResNet and connectionist temporal classification for feature learning and sequence modelling, respectively. In [11], the author was embedded Convolutional neural networks and Hidden Markov Model to recognize the static signs. In [12], the author had used fully convolutional network for sign recognition. In our work Convolutional neural networks (CNN) is used, which is persuaded by variations of multilayer perceptron and are biological processes which tend to use the slightest quantity of pre-processing.

A CNN comprises of numerous subsampling and convolutional layers and deliberately fully connected layers. The system input to a convolutional layer is a p x q x s image where p x q is the height-width of the image and s is the total no. of channels. The convolutional layer contains n filters of size m x m x s where m is smaller than the image and s is the total no. of channels or lesser and it is different for every kernel. A locally connected structures gives the size of the filters which helps the image creating r feature-maps of size pm+1. CNN follow three ideas i.e., pooling, local receptive fields, and shared weights. Every neuron in the 1st hidden layer will connect itself to a little portion of the system-input neurons. This portion is referred as local receptive field, i.e., a small window on the system-input pixels. Now, across the complete input image slide the local receptive field. In every local receptive field, there is a various hidden neuron in the 1st hidden layer.

The feature map is the map from the system-input layer to the system-hidden layer. Weights and bias that define the feature map are known as the shared weights and shared bias, respectively. The kernel is defined by these shared weights and bias. A pooling layer receives output from each feature map generated from the convolutional layer and produces a final feature map. It makes the output information more planer. Max-pooling is the common approach used for pooling. We have proposed a YOLOV3 model for achieving this objective. YOLO is a 106 layered neural network, and it considers features like color which makes the model to learn, detect, and classify the objects. The system takes image as an input and the system output will contain an image of a hand gesture (sign) with the bounding boxes.

II. LITERATURE SURVEY

In the literature survey, we are going to review the previous system and their approaches to detect and recognize the sign language.

[1] The author used videos and they extracted spatial features and temporal from them. Then they used Inception, a Convolution neural network model for recognition of spatial features. Then they used Recurrent Neural Network model i.e., long short-term memory to train the model on temporal features. In CNN they used softmax layer and pool layer. The data set comprise of hundred American Sign Language signs. In varying lighting conditions, every sign is done 5 times by one signer with different signing speed. The results of this approach, Softmax layer average accuracy is 91.5% and Pool layer average accuracy is 56.5%. Their model fails to obtain great accuracy for various skin tone.

[2] The author proposed the model to recognize the Indian sign language (ISL) gestures by applying artificial intelligence model like CNN. In CNN they used three different architecture namely, max pooling, stochastic pooling, and pooling. They created their own dataset. The dataset contains two hundred ISL sign, which is presented by five native ISL users in five different angles at a rate of thirty frames per second. In that three sets are used for training and two sets are used for testing propose. The result of this approach, the average recognition rate for stochastic pooling is 92.88%. For max pooling, the recognition rate is 91.33% and mean pooling produces a recognition rate of 89.84% respectively. The drawback of this approach is the model only concentrate on Indian Sign Language but not on any other Sign Language.

[3] The author proposed a model to automate the recognition of hand sign language using deep learning with pipeline architecture, 2-D CNN, LSTM, and 3-D CNN from the system input videos. From the 2D input frame they the 3D hand key points using CNN model. Using midpoint algorithm, the estimated key points are connected to build the hand skeleton. They used three different datasets i.e., RKS-PERSIANSIGN with 5 lakh frames and 100 labels, NYU hand pose with 81 thousand frames and 36 labels and First-Person hand action with 1 lakh and 26 labels. The result of this approach, For RKS-PERSIANSIGN dataset they got 96.02% accuracy and for NYU they got 82.10% accuracy. Using this approach, the model accuracy varies for different datasets.

[4] The author proposed a system which is constructed on a skin color modeling technique, i.e., explicit skin color space thresholding. The skin color range is extracted from the predetermined pixels from non–pixels i.e., hand from the background. The images are taken as an input to the CNN for image classification. Keras was used for training of images. The datasets for static SLR were done through, continuous capturing of images using Python. Each class contains 1,200 images, with uniform background and constant light condition, the system developed an average accuracy of 93.67% for testing, of which 93.44% for number recognition, 97.52% for static word recognition, and 90.04% for alphabet recognition.

[5] The author proposed the robust modeling of static signs in the perspective of sign language recognition using deep learning model like CNN. In CNN they used Max-pool layer and softmax layer. The dataset consists of colored images for various static signs. The dataset includes thirty-five thousand images which comprise of three hundred and fifty images in every sign classes. There are hundred different static signs, which contains twenty-three alphabets for English language, 67 commonly used words and 0–10 digits. Here the results are obtained according to the various optimizers, and it has been stated that the proposed model has attained the training accuracy of 99.90% and 99.72% on grayscale and colored images, respectively.

[7] The author proposed the model to recognize the sign language using Hidden Markov Model and Bayesian Classification Combination and they obtained the recognition rate for single hand gesture and double hand gesture i.e., 96.05% and 94.27% respectively.

[10] The author proposed the model to recognize the sign using machine learning model i.e., K-Nearest Neighbors and prototype selection. The training accuracy is 85% and on an average the testing accuracy is 66%.

III. PROPOSED SYSTEM

A. System Overview:

To develop a system that is designed to efficient and accurately perform the task of recognition of sign language using the deep learning technique, YOLOV3.

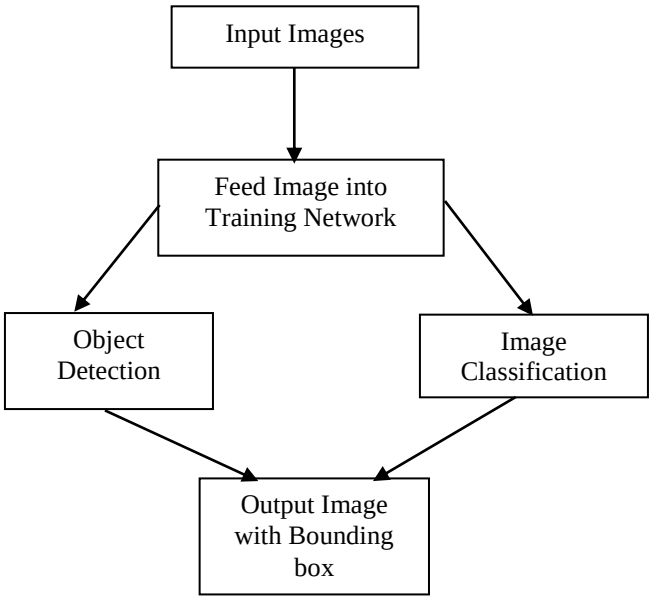

Figure 1 Block diagram for proposed system

The key factor during implementation is the ability to handle the issue like occlusion. In this project we are using Deep Learning methods. Block diagram for the proposed system as shown in Figure 1. The model takes the input as an image and passes it for training using CNN. After processing the image, it detects the object in the image. The output images detect the sign in the image with bounding boxes and gives the label for each bounding box with confidence value. There are two phases in this proposed system i.e., training phase and testing phase as shown in Figure 2 and Figure 3, respectively.

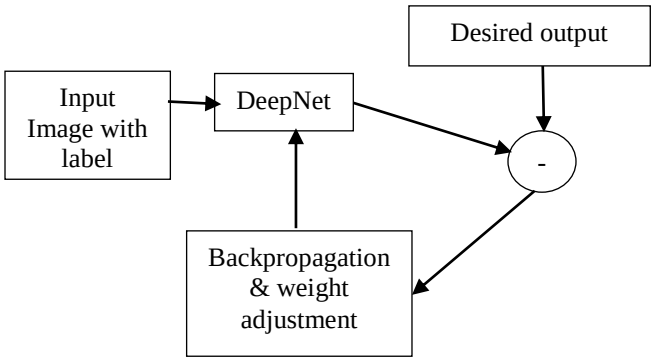

Figure 2 Training Phase

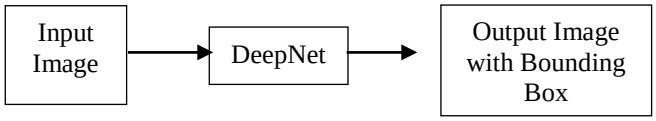

Figure 3 Testing Phase

B. *Dataset:*

The dataset for the detection and recognition of sign is taken from the Kaggle website. Where the dataset contains 26 English alphabet sign which is commonly used in America. For each alphabet sign 200 images are collected. In total 5200 images were used. For training propose 3900 images were used and for testing purpose 1300 images were used. Sample dataset shown in Figure 4.

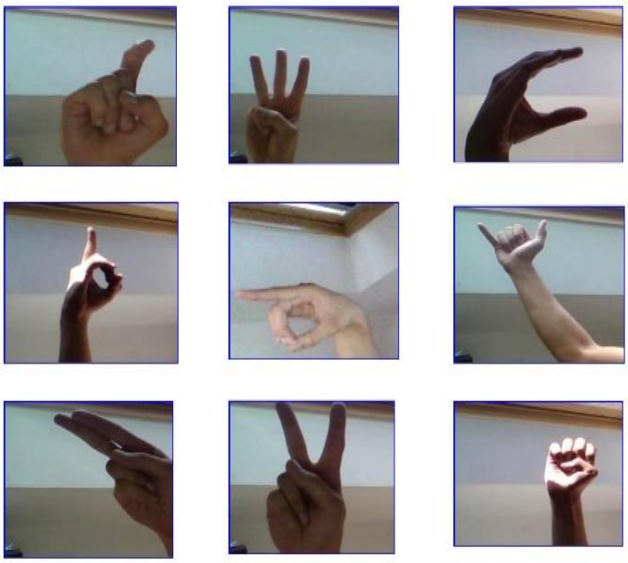

Figure 4 Dataset

C. *Hardware and Software Specification:*

GPU (Graphic Processor Unit), NVIDIA-SMI 440.82, CUDA Version: 10, are the hardware used and python as a programming language, Operating System as Windows 10 i.e., 64bit, Deep Learning Framework Darknet and LabelImg are the software used. Cloud service for the model is Google Collaborator.

D. *Methodology:*

YOLO "You Only Look Once", uses CNN for object detection. This Algorithm uses a single neural network to the entire image that is, the input image will be initially divided into various regions and then the network predicts the bounding boxes and gives the confidence for each region. Only convolutional layers are used by YoloV3 which makes it a fully convolutional network. An algorithm contains three anchors, using these anchors it predicts three bounding boxes for each cell. From these bounding boxes feature is extracted and used in training the model.

Figure 5. represents the YoloV3 architecture. In this architecture there are three stages i.e., Residual Blocks, Detection Layers, and Upsampling Layer. Residual Blocks, it contains repeated conv layers and shortcut paths. with Res Net structure. In YOLOv3, a Res Net structure is the Residual Blocks in the YOLOv3 Architecture, and it is used for feature learning. Detection Layers: yoloV3 will predict across three different scales. It helps in the detection at feature maps of 3 different sizes, having 3 different strides 32, 16, 8 respectively. That is, with an input of 416x416 pixel, to make detections on scales 13x13, 26x26 and 52x52.

Until the first detection layer, the network downsamples the input images where the feature maps of a layer with stride 32 are used for detection. Additional to that, layers will be upsampled by a factor of 2 and feature maps of a prior layers which have same feature map sizes are concatenated with it. The layer with stride 16 will have one more detection done. The same technique is repeated, and the layer of 8 stride will have the final detection.

The Data Flow Diagram for this model as shown in Figure 6. The flow of the system is as follows. Collect the dataset in the form of video and convert those videos into frames, then label those frames which will be fed to yolov3. Darknet-53 is used for training the network. After the training process, it gives a weight file. Testing process is done using those weight files. The processed Image will be displayed.

E. ***Algorithm:***

Algorithm for YoloV3:
Step 1: Collecting the dataset. **Step 2:** Converting videos into image frames using labelImg. **Step 3:** Divide total dataset into training and testing data i.e., 80% and 20% of data respectively. **Step 4:** Configure the iteration and the learning rate for the model. **Step 5:** Train the data using Darknet-53 which is pretrained weight. **Step 6:** For every 1000 iterations save the weight file and using this file update the current training process. **Step 7:** Training the model util the average loss will be less than 0.2. **Step 8:** Using final weight file test the remaining 20% of data and obtain the mAP (mean average precision). **Step 9:** Test the individual image or video using Final weight file. **Step 10:** Output will be an image or a video with labelled bounding boxes with confidence value.

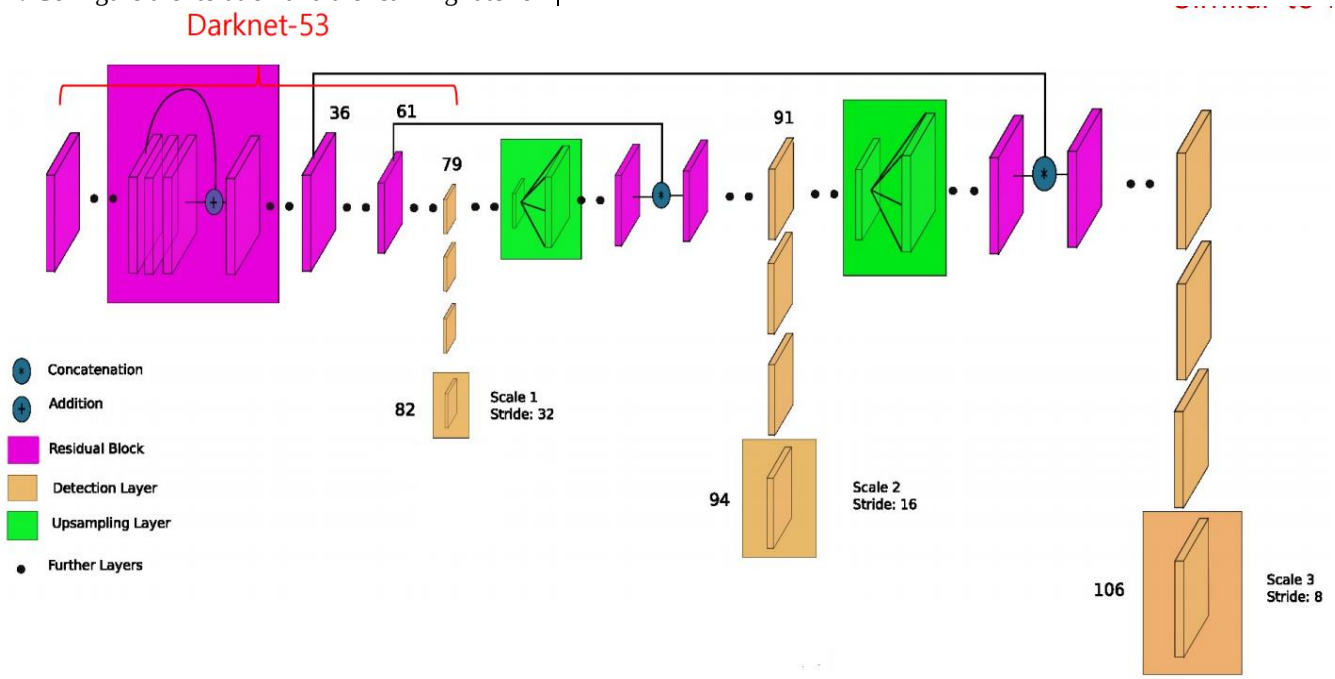

Figure 5 YoloV3 architecture

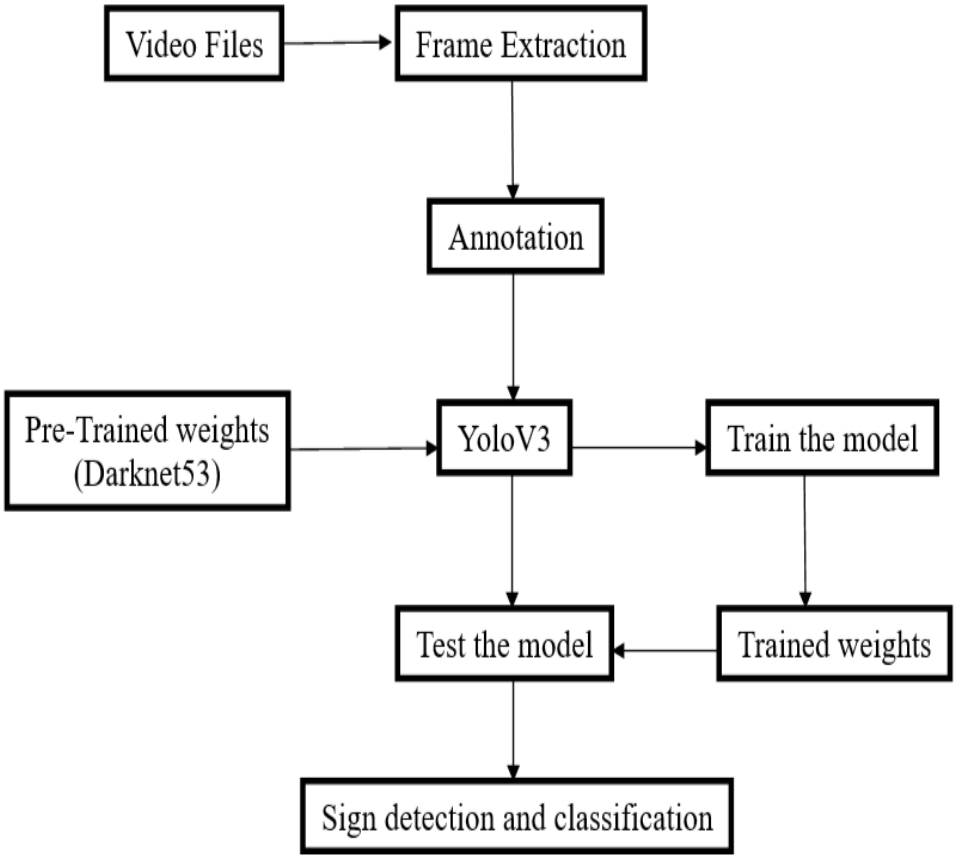

Figure 6 Data Flow Diagram

IV. RESULTS

The model is built for recognition of sign language. The model is trained, and average loss is calculated. Here IOU and mAP (mean average precision) are considered as performance measure.

The prediction percentage and the bounding boxes accuracy depends on Batch size, Learning rate, and Number of training iterations. For Batch size 64, for Learning rate 0.0001 and for number of iterations 1400, the system got 99.3% mAP and average loss is 0.1981. The mAP v/s iteration graph is plotted as shown in the Figure 7.

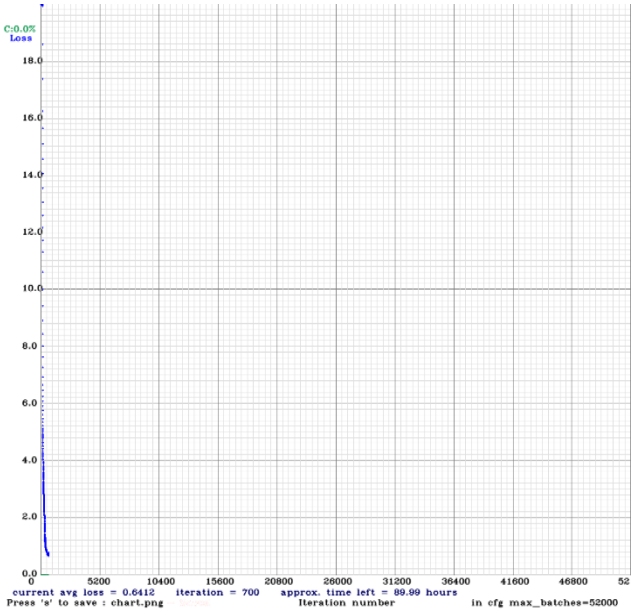

Figure 7 mAP v/s iteration graph

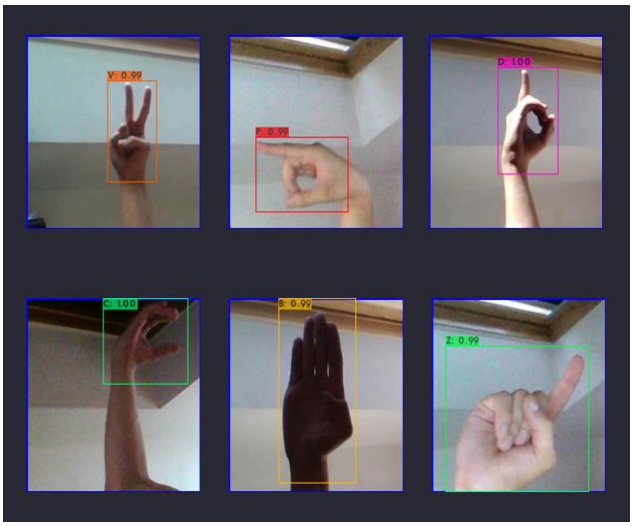

Figure 8 Tested Image with Bounding Box

When the individual images are tested and the resultant image will be generated with labeled bounding box and confidence value, as shown in the Figure 8.

V. CONCLUSION

Sign language is also the way of communication used by people with impaired speech and hearing. The huge problem is that sign language cannot be easily understood by normal people, which builds huge barrier in between non-sign language speakers and sign language speakers. To solve this problem, the model is developed using YoloV3 network to detect and recognize the American sign language. With less training time around 2.5hrs and with 0.1981 average loss the model is trained, and a weight file is generated. Using this weight file, the model is tested on test dataset and mAP is calculated and the value of mAP is 99.3%, which is a good score. When the individual class images are tested, the average probability of those images is 0.99. Improving the system further, results in optimal and hassle-free processes.

VI. REFERENCES

[1] American Sign Language Recognition using Deep Learning and Computer Vision, Kshitij Bantupalli, Ying Xie, 2018 IEEE International Conference on Big Data (Big Data).

[2] Deep Convolutional Neural Networks for Sign Language Recognition, G.Anantha Rao , K.Syamala, P.V.V.Kishore, A.S.C.S.Sastry, Dept. of ECE, K L Deemed to be UNIVERSITY, 2018 IEEE International Conference.

[3] Hand sign language recognition using multi-view hand skeleton, Razieh Rastgooa, Kourosh Kiani, Sergio Escalera, https://doi.org/10.1016/j.eswa.2020.113336 , 2020 Elsevier.

[4] Static Sign Language Recognition Using Deep Learning, Lean Karlo S. Tolentino, Ronnie O. Serfa Juan, August C. Thio-ac, Maria Abigail B. Pamahoy, Joni Rose R. Forteza, and Xavier Jet O. Garcia, December 2019 IEEE International Journal of Machine Learning and Computing.

[5] Deep learning-based sign language recognition system for static signs, Ankita Wadhawan, Parteek Kumar, 2020 Springer Nature.

[6] SignFi: Sign Language Recognition Using WiFi, Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, Woosub Jung, March 2018, ACM Journals.

[7] Independent Bayesian classifier combination based sign language recognition using facial expression,Pradeep Kumara, Partha Pratim Roy, Debi Prosad Dogra, 2018 Springer.

[8] Video-Based Sign Language Recognition Without Temporal Segmentation, Huang J., Zhou W., Zhang, 2018 the AAAI Conference on Artificial Intelligence.

[9] Iterative Alignment Network for Continuous Sign Language Recognition, Junfu Pu, Wengang Zhou, Houqiang Li, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[10] Sign Language Recognition Based on Intelligent Glove Using Machine Learning Techniques, Paul D. Rosero-Montalvo, Pamela Godoy-Trujillo, Edison Flores-Bosmediano, Jorge Carrascal-Garcia, 2019 IEEE conference.

[11] Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs, Oscar Koller, Sepehr Zargaran, Hermann Ney & Richard Bowden, 2019 Springer International Journal of Computer Vision.

[12] Ashwinkumar.U.M and Dr. Anandakumar K.R, "Predicting Early Detection of cardiac and Diabetes symptoms using Data mining techniques", International conference on computer Design and Engineering, vol.49, 2012.

