Real-time Sign Language Recognition using Computer Vision

Jinalee Jayeshkumar Raval
Electronics and Communication Department
Institute of Technology, Nirma University
Ahmedabad, India

Ruchi Gajjar
Electronics and Communication Department
Institute of Technology, Nirma University
Ahmedabad, India

2021 3rd International Conference on Signal Processing and Communication (ICPSC) | 978-1-6654-2864-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICSPC51351.2021.9451709

***Abstract*— Speech impairment is a disability that affects an individual's ability to verbal communication. To overcome this issue sign language is used which is one of the most organised languages. There is definitely a need for a method or an application that can recognize sign language gestures so that communication is possible even if someone does not understand sign language. My paper is an effort towards filling the gap between differently-abled people like deaf and dumb and the other people. Image processing combined with machine learning helped in forming a real-time system. Image processing is used for pre-processing the images and extracting different hand from the background. These images obtained after extracting background were used for forming data that contained 24 alphabets of the English language. The Convolutional Neural Network proposed here is tested on both a custom-made dataset and also with real-time hand gestures performed by people of different skin tones. The accuracy obtained by the proposed algorithm is 83%.**

Keywords—Machine learning, Image processing, Convolutional Neural Network

I. INTRODUCTION

Communication is the key to unlock the doors of the external world and only those who are deprived of it can better understand its importance. Orfield laboratories conducted an experiment and found that a normal person could stay comfortably for only 45 minutes in a silent room. So, imagine being part of such a still world whole life. Deaf and mute people face such issues in their day to day life. Figures of WHO states that nearly 466 million people that comprises approximately 5% of the World's population are with such disabilities and out of which 35 million are children[14].

Sign language is developed for those people so that they can communicate with ease. It is the most structured language where each and every gesture has a specific meaning attached to it. It also has its own grammatical signs in order to connect the words. In this way, they can only communicate with those who have prior knowledge of sign language. But with others, they might find difficulty in communicating. So there's a need for a method to detect sign language by which such a person can communicate with normal people easily.

The sign language detection method can be further classified into two types: i) Isolated sign language recognition ii) Continuous sign language recognition. Isolated sign language recognition deals with individual steady signs. Whereas continuous sign language recognition tracks gestures and decides the signs based upon the movements tracked in back to back frames. Isolated sign language recognition is implemented for this paper. Real-time images are captured and respective detection is shown on the frame itself. Here 24 alphabetical signs except J and Z are used for forming dataset and signs are performed according to the American sign language rule.

Computer vision is providing vision to machines so that they can extract important features from the images captured. Image processing and machine learning can be considered as the subdomains of this vast field. In this paper, machine learning is used along with image processing. Images of the hand are captured and preprocessed for extracting the hand from the background. Images are then scaled down and the dataset is formed after following preprocessing steps. So, the dataset contains images without background and are also scaled down.

A deep learning method called convolutional neural network is used in this paper. The convolutional neural network is very helpful in identifying various features of the images in the spatial domain. Pixels of the images are treated as neurons and then processing is done neuron by neuron. The varying number of kernels are applied at different layers of the convolutional neural network for extracting shapes of fingers. Towards the end, it classifies the images into various groups based upon the features[6]. In this manner, every neuron is linked to a neuron of the next layer and previous layer as well and forms fully connected layers in the network. So here preprocessed images present in the dataset are fed to the convolutional neural network formed and then the model gets to train and tested. Once tested with the dataset image it can identify the signs performed in real-time.

Hence this real-time identification of the images could be very helpful to such differently-abled people. With real-time interaction, they can easily convey their message and the converted text could be displayed on-screen or even it can be utilized for making this system voice-enabled. Fig 1 shows different signs which represent the alphabets of the English language performed in American Sign Language.

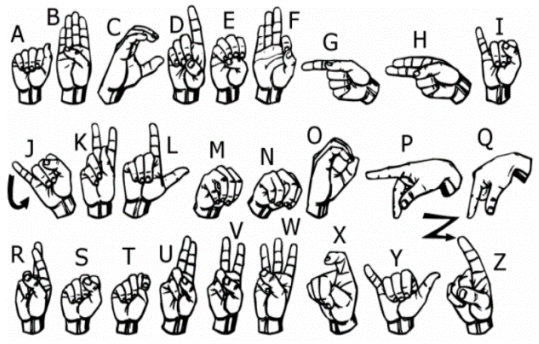


Figure 1. *American Sign Language[11]*

II. RELATED WORKS

Previously a lot of work has been done for recognizing different sign languages. They are implemented by using various methods and components like using camera or some other sensors like Kinect, ultrasonic etc. Out of them some are explained in this section.

A. *Structured Feature Network for Continuous Sign Language Recognition[1]*

This paper provides an insight about the method for recognizing sign language. Gloss is defined as unit of sign language which is composed of gesture, motions and facial expressions. In this paper Sign Language Recognition (SLR) was divided into two major parts namely isolated SLR and continuous SLR. In isolated SLR frame or sign is segmented individually and prediction is done by taking only one running gloss into consideration. By using this method sentences can be formed after recognizing words which are derived from continuous glosses. But forming such final sentence can become a tedious task in some case and that's why it needs some extra models. There for, to overcome such issues Structured Feature Network (SF-Net) was proposed which learns features in a structured manner i.e. from frame to gloss and then to sentence level. Features can be extracted from frames at frame level. The signing gesture and facial emotion are important information for distinguishing different glosses at gloss level. Lastly to align and translate the signing sequence to required sentence, gloss level features need to be re-organized in sentence level so that context information in the sequence can be encoded.

B. *Sign Language Recognition Using Image Based Hand Gesture Recognition Techniques[2]*

In Hand gesture is one of the method used in sign language for communicating. Each gesture in sign language has specific meaning and complex meanings can be derived by combining basic elements. There are two main types of approach in sign language one is vision based and other is sensor based. In this paper vision based approach is followed and is divided into two phases sign detection and sign recognition. Tracking is mainly used for tracking a hand gesture from capture image using Convexity hull algorithm. Finally, recognition is done with the help of features like convex hull and convex defects taken from tracking. Convex hull algorithm which detects convex structures computes maximum and minimum coordinate points by joining those points form bounding rectangle in which contains hull. Then average of all the defects was taken and by which centre of palm was identified and in this manner by extracting features gestures are recognized.

C. *Sign Language Recognition Using Deep Learning on Custom Processed Static Gesture Images[5]*

In this paper sign language detection is done using image classification and machine learning algorithm is used. In machine learning convolutional neural network is employed. Image dataset used in this paper consisted of static sign language gestures taken by an RGB camera. Static sign gestures used to signify alphabets and numbers were taken as dataset and they were detected successfully. Here in this paper the proposed neural network contains neurons belonging to various layers will be individually connected to each other. If an image is having 256 x 256 x 3 pixels with value 0 to 255 then lot of data would be available for processing. So instead of doing this convolutional neural network is used because all pixels do not contain useful information. Convolutional neural networks instead convolve around the input pixels i.e. the image, to reduce its dimensions when it is input to the next layer. Inception model used in this paper stack convoluting, pooling, softmax and related layers parallel to each other instead of stacking on top of each other. Dataset had static images out of which 20% were kept for testing. Data augmentations like cropping, scaling, rotating were done in order to adjust image.

III. PROPOSED ALGORITHM

This section provides an overview about the methodology used in this project and the algorithm used for implementing the project. The approach followed can be simply divided into two segments one is image processing and machine learning. From image processing images are pre-processed and dataset is formed and then these dataset is used to train and test the CNN model formed in machine learning segment.

A. *Image processing*

In the initial portion images of the hand are captured and processed. A box representing Region of Interest is drawn on screen and image between that ROI is only taken into consideration. This is done because that every different image would have equal size and aspect ratio hence uniformity in the dataset would be maintained.

Coloured image having the RGB format is converted to the HSV format.[10] The reason behind this is Red, Green, Blue in RGB are all co-related to the colour luminance i.e., we cannot separate colour information from luminance. HSV or Hue Saturation Value is used to separate image luminance from colour information. So, the HSV colour scheme is used for detecting skin colour of hand in various backgrounds. Then a mask for the skin colour is created. This mask created further passes through various image processing steps like it is blurred for smoothening out the high frequency noise. Then a series of dilation and erosion is applied to it for filling out the gaps and enhancing required features[13]. Afterwards this mask is applied to the image for detecting hand.

After applying mask hand is extracted from the background and then image is converted into a 28x28 image in grayscale. All the pixel values are then normalized by 255 to make calculations in next processes simpler. And then the image is converted to a row matrix of 1x784 for saving them to the database. The image obtained after all these processes is then saved in the CSV file using xlsxwriter library of python which helps in writing data in Excel sheet in CSV format. During real-time detection as well the ROI image is passed through all these steps and converted in a row matrix of 1x784. Then only it is given to CNN algorithm.

These all the steps are performed using OpenCV library and the Python programming language in the Visual Studio Code 2019 programming platform. Fig 3 shows the procedure followed. Fig 2,4,5 shows the images of the hand that are obtained after following the image processing steps.

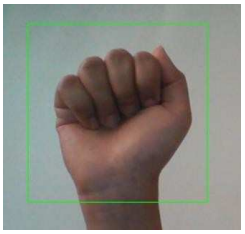


Figure 2. *Region of Interest in green box*

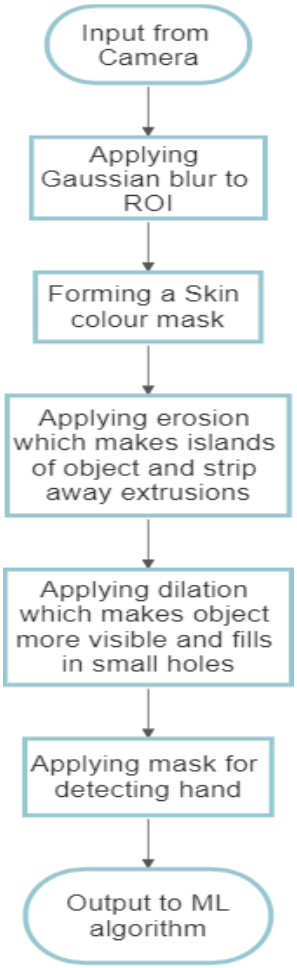


Figure 3. *Image processing steps*

Figure 4. *Normalized Gray scale image which is converted from Coloured image.*

Figure 5. *Grey scale image which is resized to 28x28 and these images are used for forming data set.*

B. Machine learning

Machine learning is used in the experiment for identifying different features of hand and classifying the hand gestures to the corresponding letters. Convolutional Neural Network plays the key role in distinguishing the vivid characteristics of the input image. Neural Network has been evolved by taking inspiration from the way neurons of human brain works. So Neural Network makes use of similar structure and names neuron as perceptron. Neural Network are good with arrays of data but they are unable to identify images when changes occur in spatial domain. Hence, CNN has evolved which performs exceptionally well in identifying visionary information in an image.

In this experiment the CNN layers are used to design a model. Keras library's inbuilt functions are used for stacking layers of CNN[9]. Dataset is read using Pandas library of Python and a 15% data is kept for testing and rest data is utilized for training the model. Data set consists of 240 images of 10 images for each alphabetical sign. Two signs J and Z which requires motion are not part of this dataset.

The algorithm of the CNN model used in this project is described now. Image of each data is in the form of 1x784 matrix. So, this data is again reshaped to form an image matrix of size 28x28x1. This matrix is passed on to the convolution layer containing 32 different kernels for detecting various features. These kernels could be of size 3x3 or 5x5 but I have used 5x5 size here. After convolution 32 different outputs obtained from the convolution layer are pooled. A 2x2 max pooling layer is used which helps in reducing the size of images by removing redundant or not so important data. Along with convolution in the input and hidden layers ReLU activation function is used for removing negative value. A general layout of all these layers are shown in Fig 5.

After moving further more number of kernels are used by convolutional layer like second convolutional layer uses 64 kernels. Towards the end data is flattened to form a fully connected layer. For output layer instead of ReLU, Softmax activation function is used because in simpler terms Softmax provides the probability about the position of the output like probability is 99% that output is 3rd position or class 3. So, human can easily understand such probabilistic data and decide upon output. These steps are repeated according to the epochs. After that accuracy of the method is calculated and displayed. Once tested real-time hand detection starts and according to the sign class of the image is shown.

So, the proposed methodology and the algorithm are described in this section. Here the model which is used for implementing CNN is also explained with a motive to provide some insight about the project.

Another form of outcome is showing the class from which input image belongs to. So, information regarding the class is also displayed during real-time testing and this keeps on changing according to variations in signs. Apart from this end results various other results are also available like the accuracy of the system is shown in Fig 8.

IV. Results

The project has been implemented in Python scripting language. Several libraries used in this project are openCV2, Pandas, Keras, Numpy etc. Visual Studio Code 2019 is used for programming. This program was tested on a computer system having Intel® Core™ i7 8300H processor having clock 2.3 GHz. Model accuracy can be seen in Fig 6.

```
Epoch 98/100
204/204 [==============================] - 2s 8ms/step - loss: 0.0179 - accuracy: 1.0000
Epoch 99/100
204/204 [==============================] - 2s 8ms/step - loss: 0.0131 - accuracy: 1.0000
Epoch 100/100
204/204 [==============================] - 2s 8ms/step - loss: 0.0128 - accuracy: 1.0000
37/37 [==============================] - 0s 2ms/step
Accuracy:  0.837837815284729
```

Figure 6. *Accuracy obtained by the CNN model*

The outcome of the algorithm is available in various forms. Once the testing from the dataset is done then accuracy of the output is displayed. Output images are distributed into various numeric classes. So analogy between the output images and output classes is shown in the Table 1. So during real-time detection classes dispayed can be verified by looking at thus table.

TABLE I LETTERS MENTIONED BY THE CLASS OF THE IMAGE

Class Name	Alphabet	Class Name	Alphabet
1	A	13	N
2	B	14	O
3	C	15	P
4	D	16	Q
5	E	17	R
6	F	18	S
7	G	19	T
8	H	20	U
9	I	21	V
10	K	22	W
11	L	23	X
12	M	24	Y

The accuracy and loss are calculated for both train and test dataset after each epoch and plotted in Fig 7. So, with more epochs the accuracy tends to increase here.

The results displayed below have the class of the detected signs. The results are purposefully taken in different background and lighting conditions to test robustness of the algorithm and how it reacts when the environment changes. Class number is displayed in red on the images using image processing to get the real-time detection results.

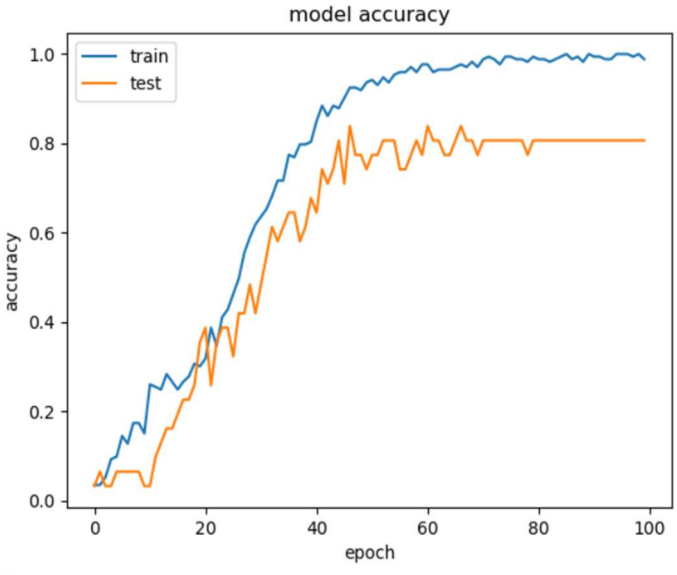
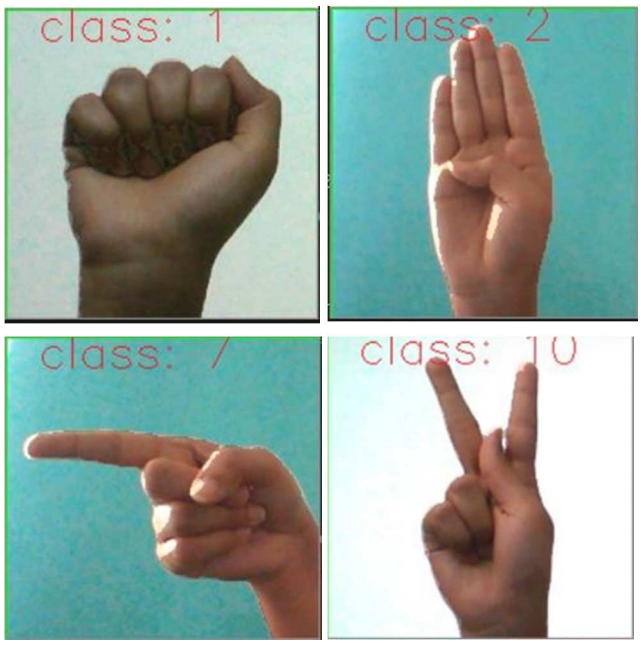
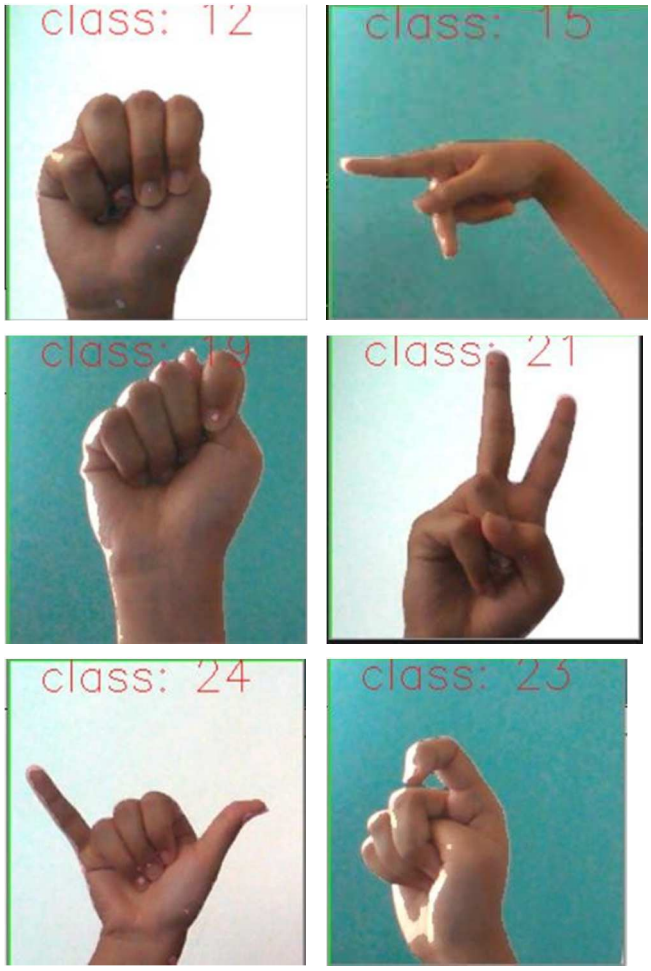


Figure 7. *Model accuracy for train and test data*

TABLE 2 TABLE OF OUTPUT IMAGES





So, here results of some signs can be seen. They are taken in different illuminations and backgrounds. Still, it is able to detect all the signs and gives output accuracy as 83%.

V. Conclusion

The Sign Language Recognition project is done to help the deaf and dumb community in connecting with the outer world, especially to those who do not understand sign language. The project has majorly two part one is image processing for detecting skin colour and eventually hand of the signer. An algorithm was deployed on a laptop with having integrated camera. So, the images captured by the camera were converted into a dataset containing 240 images consisting of 10 images of each sign. So, the dataset contained 24 different signs of the English alphabet following one-handed Indian Sign Language rules. This dataset is divided into two parts test and train giving 15% of the dataset for testing. A Convolutional Neural Network model is designed for extracting features from the corresponding signs. So, 83% accuracy is obtained by testing the model. The previously available datasets were not suitable for my environmental conditions and my skin tone. Therefore, though they were able to give high accuracy while testing from the very same dataset but failed to detect signs performed by my hand in real-time. So. with this dataset model can continuously predict real-time signs and displays the corresponding letter on the screen.

VI. Future Scope

Sign Language Recognition has another segment which is known as Continuous Sign Language Recognition as it deals with taking successive frames in real-time and predict the word by detecting a continuous gesture. Hence, this project can be further extended in this direction and words and after that sentences can be formed according to the continuous gestures performed. Further dataset having images from people with different skin tones and in different lighting conditions is required in order to develop a robust algorithm that can serve the purpose for any kind of people.

References

[1] Zhaoyang Yang, Zhenmei Shi, Xiaoyong Shen, Yu-Wing Tai, "SF-Net: Structured Feature Network for Continuous Sign Language Recognition". *arXiv preprint arXiv:1908.01341v1, 2019*.

[2] Ashish S. Nikam, Aarti G. Ambekar, "Sign Language Recognition Using Image Based Hand Gesture Recognition Techniques", *Online International Conference on Green Engineering and Technologies, 2016.* Available doi: 10.1109/GET.2016.7916786.

[3] G. Rajesh, X. Mercilin Raajini, K. Martin Sagayam, Hien Dang,"A statistical approach for high order epistasis interaction detection for prediction of diabetic macular edema",Informatics in Medicine Unlocked, Volume 20,2020,100362,ISSN 2352-9148.

[4] P. M. Ashok Kumar, Jeevan Babu Maddala, K. Martin Sagayam (2021) "Enhanced Facial Emotion Recognition by Optimal Descriptor Selection with Neural Network", *IETE Journal of Research*, doi: 10.1080/03772063.2021.1902868

[5] Aditya Das, Shantanu Gawde, Khyati Suratwala and Dr. Dhananjay Kalbande. "Sign Language Recognition Using Deep Learning on Custom Processed Static Gesture Images".In: *International Conference on Smart City and Emerging Technology (ICSCET), 2018* . Available doi: 10.1109/ICSCET.2018.8537248.

[6] Vivek Bheda and Dianna Radpour. "Using Deep Convolutional Networks for Gesture Recognition in American Sign Language". In: *CoRR* abs/1710.06836 (2017). arXiv: 1710.06836. url: http://arxiv.org/abs/1710. 06836.

[7] PADMAVATHI . S, SAIPREETHY.M.S, V. "Indian sign language character recognition using neural networks". *IJCA Special Issue on Recent Trends in Pattern Recognition and Image Analysis, RTPRIA*.

[8] https://Www.Ucbmsh.Org/Colleges-In-Dehradun/Courses/Agriculture-Courses-In-Dehradun/Indias-First Sign-Language-Isl-Dictionary [image]

[9] SAKSHI GOYAL, ISHITA SHARMA, S. S. "Sign language recognition system for deaf and dumb people".In: *International Journal of Engineering Research Technology*

[10] doc.opencv.org, "Image Procesing : Getting Started" Available: *https://docs.opencv.org/3.4/d4/d73/tutorial_py_image_processing_begin.html*

[11] *https://Www.Ucbmsh.Org/Colleges-In-Dehradun/Courses/Agriculture-Courses-In- Dehradun/Indias-First-Sign-Language-Isl-Dictionary#Ilightbox[Gallery1939]/0.* [image].

[12] *https://Sds-Platform-Private.S3-Us-East-2.Amazonaws.Com/Uploads/75_Blog_Image_1.Png.* [image].

[13] doc.opencv.org. "Contours : Getting Started" Available: https://docs.opencv.org/3.4/d4/d73/tutorial_py_contours_begin.html

[14] About.almentor.net. 2020. *The Deaf And Mute – Almentor.Net*. [online] Available at: https://about.almentor.net/about/the-deaf-and-mute/#:~:text=The%20Deaf%20and%20Mute%20%E2%80%93%20Facts,whom%2034%20million%20are%20children.