Sign Language Recognition Using Deep Learning and Computer Vision

Article *in* Journal of Advanced Research in Dynamical and Control Systems · May 2020

DOI: 10.5373/JARDCS/V12SP5/20201842

3 authors, including:

Dr.Sabeenian R.S
Sona College of Technology
137 PUBLICATIONS **969** CITATIONS

SEE PROFILE

Sign Language Recognition Using Deep Learning and Computer Vision

R.S. Sabeenian, Department of Electronics and Communication Engineering, Sona College of Technology, Salem, Tamil Nadu, India. E-mail: sabeenian@sonatech.ac.in

S. Sai Bharathwaj, Department of Electronics and Communication Engineering, Sona College of Technology, Salem, Tamil Nadu, India. E-mail: ssb.nsk99@gmail.com

M. Mohamed Aadhil, Department of Electronics and Communication Engineering, Sona College of Technology, Salem, Tamil Nadu, India. E-mail: adhilshereef200@gmail.com

Abstract--- Inability to speak is true disability. Speech impairment is a disability that affects an individual's ability to communicate using speech and hearing. Mode of communication such as sign language is used by people affected by this impairment. There exists a challenge for non-signers to communicate with signers although the sign language is ubiquitous in recent times. There has been a strong progress in the fields of motion and recognition of gestures with the recent advancements in computer vision and deep learning techniques. The major focus of this work is to create a deep learning-based application that offers sign language translation to text thereby aiding communication between signers and non-signers. We use a custom CNN (Convolutional Neural Network) for recognizing the sign from a video frame. MNIST dataset is used.

Keywords--- Machine Learning, Deep Learning, Convolutional Neural Networks, Computer Vision, Sign Language.

I. Introduction

People with impaired speech and hearing uses Sign language as a form of communication. Disabled People use this sign language gestures as a tool of non-verbal communication to express their own emotions and thoughts to other common people. But these common people find it difficult to understand their expression, thus trained sign language expertise are needed during medical and legal appointment, educational and training session. Over the past few years, there has been an increase in demand for these services. Other form of services such as video remote human interpret using the high-speed Internet connection, has been introduced, thus these services provides an easy to use sign language interpret service, which can be used and benefited, yet have major limitations.

To address this, we use a custom CNN model to recognize gestures in sign language. Convolutional neural network of 11 layers is constructed, four Convolution layers, three Max-Pooling Layers, two dense layers, one flattening layer and one dropout layer. We use the American Sign Language Dataset from MNIST to train the model to identify the gesture. The dataset contains the features of different augmented gestures. Introduced a custom CNN (Convolutional Neural Network) model to identify of the sign from a video frame using Open-CV.

Initially, feature extracted dataset is used to train the custom model that has 11 layers with a default image size. Rest of this analysis is organized as follows: Section 2 gives a summary of the performed literature survey; Section 3 talk about the datasets and its specialities. Section 4 overviews the structure of the model introduced. Section 5 highlights point the experiment and observations of this project. At last, Section 6 express the issue faced by model and projected the possible developments in Section 7.

II. Related Work

In [1] the authors used a pretrained model from the google Inception V3. The Inception V3 network is trained with the images from the ImageNet database. The Inception V3 convolutional neural network is re-trained with the ImageNet weights dumped on it to recognize gestures from the image Frame obtained from the video.

Zafar Ahmed Ansari and Gaurav Harit [2] did a tremendous research work to categorize the Indian Sign gestures and have classified one hundred and forty classes that include finger spelling numbers, English alphabets and other common phrases. A Kinect sensor is employed to capture the dataset. 3 channel (RGB) images of dimension 640 x 480 are captured with their depth data. The values of the depth of every pixel is captured. Every pixel of the image maps to a value in the depth data. The hand in the image has least depth since the subjects were standing with their hands stretched out. This is confirmed by pixel masking of the depth values that are more than a certain threshold.

DOI: 10.5373/JARDCS/V12SP5/20201842
ISSN 1943-023X

Received: 15 Mar 2020/Accepted: 18 Apr 2020

The left-out part during this process had the palm with the gesture. However, since this had several consistency issues, this dataset was not used for training the CNN. Unsupervised learning was employed using the K-means clustering algorithm. SIFT mapping and gaussian masks were used to extract features and train the dataset. End accuracy was over 90%.

The sign recognition in [2] is accomplished with PCA (Principal Component Analysis). Recognition with neural networks is also proposed in the paper. The acquired data was with a 3MP camera due to which there was a poor quality. It consists of 15 images per sign. Their results were not satisfactory due to their considerably small dataset. Simple boundary pixel analysis was performed by doing segmentation and separating RGB components. The authors mentioned that a better output can be achieved using Neural networks than the results obtained by combining the fingertip algorithm with PCA.

The paper by Nandy et al. [3] classifies the gestures by splitting the data into segmented features and employs Euclidean distance and K-Nearest Neighbours. Similar work by Kumud et al. [4] shows how to do continuous recognition. The paper proposes extraction of frames from videos, data pre-processing, extracting frames and other features, recognition and optimization. Pre-processing is accomplished by converting the video into RGB frames with same dimension. Skin colour segmentation with the HSV was used to extract skin region and were converted to binary form. Extraction of key frames are done by gradient calculation between the frames, and extraction of features is done by oriental histogram. Classification is achieved by several distance calculation like Euclidean, Manhattan, Chess Board Distance.

III. Sign Language Data Set

The **MNIST** (Modified National Institute of Standards and Technology database) database is a large collection of handwritten digits that is used for training various image processing systems. The data is also extensively used for both process of training and testing in the area of machine learning. This original MNIST image dataset of handwritten digits is a popular benchmark for image related machine learning techniques yet researchers have renewed efforts to update it and develop drop-in replacements that are more challenging for computer vision and original for real-world applications.

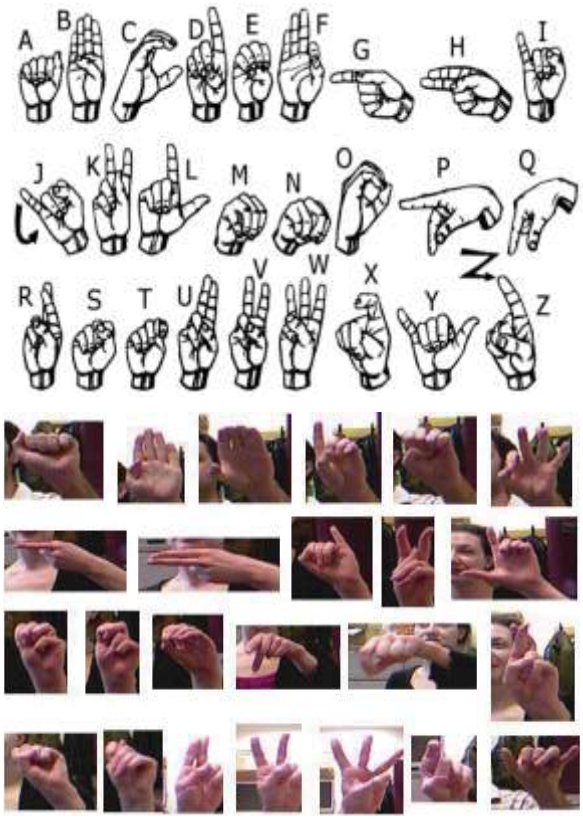

Fig. 1: Sample Gesture from Dataset

DOI: 10.5373/JARDCS/V12SP5/20201842
ISSN 1943-023X

Received: 15 Mar 2020/Accepted: 18 Apr 2020

The MNIST data consists of 60,000 training images, whereas 10,000 testing images. 50 percent of the training set, similarly other set of 50 percent of the test set were taken from NIST's training images, meanwhile the further 50 percent of the training set, similarly other 50 percent of the test set were pull from NIST's testing images. The American Sign Language alphabet and number collections of images of hand gestures produce a multi class level issues having 24 classes of letters (excluding J and Z which require motion). Since J and Z require dynamic gesture, they are not included.

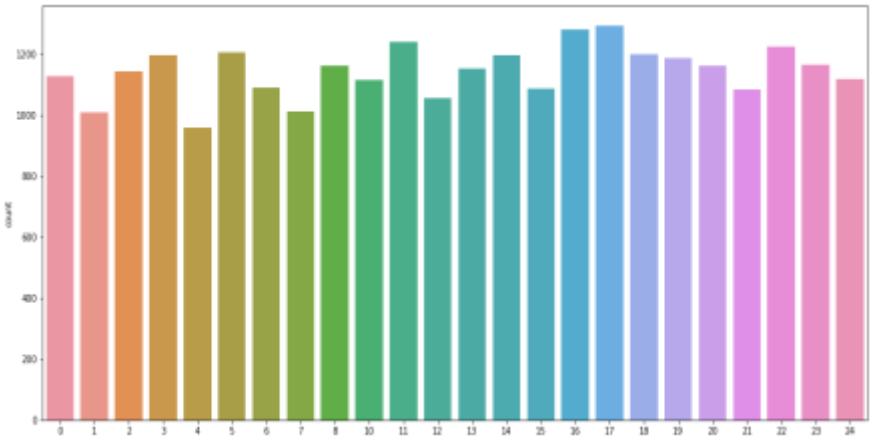

Fig. 2: Count plot of the Dataset

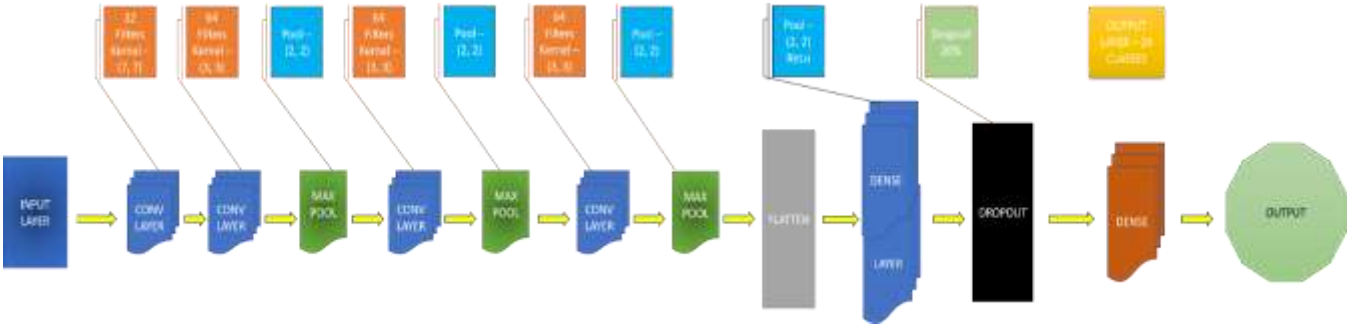

Fig. 3: Custom CNN Model Architecture.

- The training dataset consists 27455 images with 784 (28 X 28) features.
- The validation dataset consists of 7172 images with 784 (28 X 28) features.
- As J and Z are excluded, there are 24 classes in total.
- On an average, each class contains 1000 images.

IV. Network Architecture

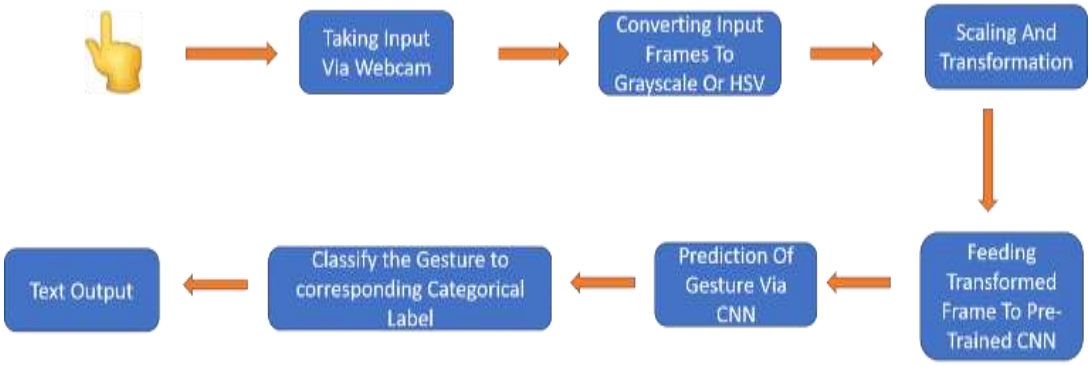

Fig. 4: High Level System Architecture

DOI: 10.5373/JARDCS/V12SP5/20201842
ISSN 1943-023X

Received: 15 Mar 2020/Accepted: 18 Apr 2020

As shown above in Figure 4, Initially the image is segmented from a video input from the Webcam. The frames are dropped from the video with a region of Interest (Threshold Square Box) so as to avoid background conflicts. A custom CNN model with 11 layers is used. The gesture image segmented from the video frame is then converted to a grayscale image. The input image from the webcam is converted to grayscale since the model is trained with the features of the grayscale images i.e. the MNIST dataset is a pre-processed dataset of RGB images that are converted to grayscale. The converted image is then scaled in respect to the size of the images with which the model was trained. The image is fed into the pre-trained custom CNN model post scaling and transformation. The gesture prediction from the CNN model is obtained and post that, it is classified based on the categorical label. The classified gesture is displayed as text.

V. Custom CNN Model Architecture

The CNN model developed consists of 11 different layer,

- Four Convolutional layer
- Three Pooling (Max) layer
- Two Dense Connected, One Flatten and One Dropout layer.

Using the custom CNN model makes it easy in choosing the variety of convolution to utilize (3x3, 5x5) in the model itself. These Initial convolutional layers have a convolution kernel of (7,7) followed by a (3,3) convolutional layer. Each Max Pooling layer consists of a (2,2) pool which is efficient. The Fully Connected Layer after the flattening layer also boards a (2,2) pool with ReLu (Rectified Linear Unit) activation so as to avoid negativities. A dropout layer with a given probability of 20% is included so as to avoid model overfitting as it does drops out 20% of the hidden and visible units from these densely connected layers. The final training of model from scratch produces a considerably high level of 99% accurateness on the training set. The output is taken from the SoftMax Layer with 24 class classification.

VI. Experimental Evaluation

The CNN model is trained with the MNIST ASL dataset. The data set of 27455 training samples of 784 features is used to train the model. The model was trained to minimize loss by usage of cross entropy ADAM [5]. The various model is trained for 10 epochs on a batch size of 128. The model was trained with a learning rate of 0.001 with 0 decay. The validation dataset consists of 7172 samples, reportedly the validation accuracy of the model is greater than 93 %. The following are the training accuracy & loss plots.

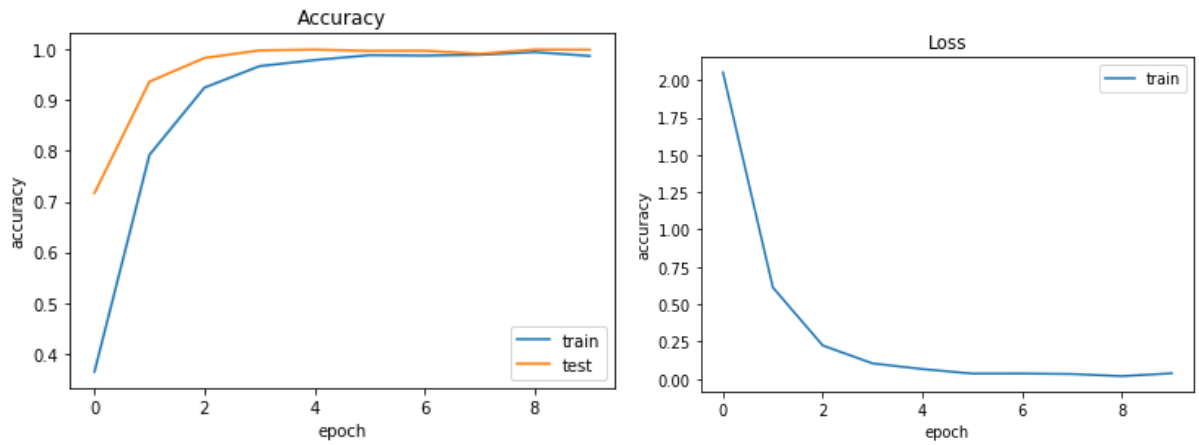

Fig.5: Metrics Plot of Custom CNN Model Architecture

VII. Issues in the Model

The major issue faced is due to the background of the image. As the model is trained with a segmented grayscale gesture image, it doesn't support background subtraction from the image when the frames are dropped from a video. The current model yields better accuracy with a segmented hand gesture which is done by the Open-CV with a Region of Interest (ROI) box implemented in the driver program.

DOI: 10.5373/JARDCS/V12SP5/20201842
ISSN 1943-023X
Received: 15 Mar 2020/Accepted: 18 Apr 2020

The model lacks accuracy with noisy images when it is dropped from the video frame. The model performance was not as expected if a person wears ornaments like ring as the dataset used to train the model was clean without inclusion of any ornaments.

VIII. Scope in Future

Future scope contains yet it is not limited to:

- This developed model can be introduced to other sign languages such as Indian Sign Language, at present it is finite to American Sign Language
- The model can be further trained with a dataset such that it automatically segments the gesture from the captured frame by automatically subtracting the background.
- Tuning and Augmented of the model to identify usual words and expressions
- Additionally, training of the neural network model to well organized identify symbols require two hands
- Incorporate active hand gestures in augmented to the contemporaneous static finger spelling
- Integration of the optimized model to existing AI systems like Amazon Alexa for advancements in visual recognition.

IX. Conclusion

This paper introduces a CNN based approach for the recognition and classification of the sign language using computer vision. Unlike the other approaches, this approach yields better accuracy and considerably low false positives. Other possible extensions to this work include dynamic gesture recognition, [VIII] and are being carried out.

References

[1] Aditya Das, Shantanu Gawde, Khyati Suratwala, Dr. Dhananjay Kalbande (2018, February). Facial expression recognition from video sequences: temporal and static modelling. *Computer Vision and Image Undertaking* 91.

[2] Zafar Ahmed Ansari and Gaurav Harit, "Nearest Neighbour Classification of Indian Sign Language Gestures using Kinect Camera", *in Sadhana*, Vol. 41, No. 2, February 2016, pp. 161-182.

[3] Recognition of Isolated Indian Sign Language Gesture in Real Time, Anup Nandy, Jay Shankar Prasad, Soumik Mondal, Pavan Chakraborty, G. C. Nandi, Communications in Computer and Information Science book series (CCIS, volume 70).

[4] Continuous dynamic Indian Sign Language gesture recognition with invariant backgrounds by Kumud Tripathi, Neha Baranwal, G. C. Nandi at 2015 Conference on Advances in Computing, Communications and Informatics (ICACCI).

[5] Adam: A Method for Stochastic Optimization, Diederik P. Kingma, Jimmy Ba, *Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego,* 2015.

[6] S. Tamura and S. Kawasaki, "Recognition of Sign Language Motion Images", *In Pattern Recognition,* volume 21, pages 343-353, 1988.

DOI: 10.5373/JARDCS/V12SP5/20201842
ISSN 1943-023X

Received: 15 Mar 2020/Accepted: 18 Apr 2020