Convolutional Neural Network for Detection of Sign Language

Dr.Mahesh kaluti[1], Manoj Athreya C.S[2], Manish M.G[3],M.P Mahadeva Aradhya[4],Raghavendra S[5]

[1](Associate Professor, Department of computer science and engineering,
[2,3,4,5](Department of computer science and engineering,
P.E.S College of Engineering, Karnataka, India)

Abstract

Sign languages are languages that use manual communication to convey meaning. This can include simultaneously employing hand gestures, movement, orientation of the fingers, arms or body, and facial expressions to convey speaker's idea. Sign language is an incredible advancement that has grown over the years. Sign language helps the deaf and dumb communities to go on about their daily lives. Unfortunately, there has some drawbacks that has come along with this language. Not everyone knows how to interpret or understand sign language while conversing with a deaf -mute person. To solve this, we need a product that is versatile and robust, which needs to convert sign language into text or written format so that it is understood by everyone.

Keywords *- Convolutional Neural Networks, Machine learning, Computer Vision, Sign language.*

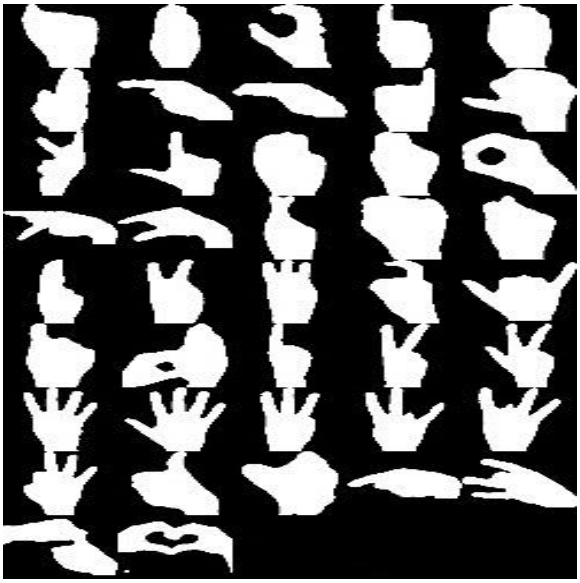


Figure 1: Sign language alphabets

I. INTRODUCTION

Conversation between deaf and dumb community and rest of the world has been in shadow of misconception since ages. The paper is concerned with a product that can eliminate the barrier between the deaf and dumb community and rest. This product should interpret 26 sign language alphabets and 10 sign language digits. The alternative of written communication is cumbersome, impersonal and even impractical when an emergency occurs. In order to overcome this obstacle and to enable mutual communication, we present a sign language recognition system that uses Convolutional Neural Networks (CNN) in real time to translate a video of a user's sign language gestures into text. This is done by taking user input i.e., gestures through web cam, classifying each frame in the video to letter and displaying the most likely word from classification scores wiz, output. However, there are many challenges to be considered including the lighting sensitivity, when a sign ends and the next begins. Our system takes the input from the video stream and we extract individual frames of the video and generate letter probabilities for each using Convolutional neural network.

II. RELATED WORKS

In order to recognize and categorize sign gestures accurately, based on our knowledge in class, we first thought of using basic machine learning techniques such as Support Vector machine, and regular Neural network. Microsoft Kinetic was also a recent advancement that helped taking 3-d depth images of sign MATLAB with PCA was also used in one of our reference papers. We evaluated both pros and cons presented by the papers discussed below by taking complexity, runtime, results, feasibility, flexibility and understandability into consideration. Selectively, we chose some models to implement on our dataset. Eventually, we settled on Convolutional Neural Network for sign language recognition.

In our work, we build on the results of Roel Verschaeren [1]. He proposes a CNN model that recognizes a set of 50 different signs in the Flemish Sign Language with an error of 2.5%, using the Microsoft Kinect. Unfortunately, this work is limited in the sense that it considers only a single person in a fixed environment.

In [2] using Support Vector Machine, the system is able to perform in dynamic and minimally cluttered background with satisfactory result as it relies on skin

color segmentation. For speech recognition Sphinx module is used which maps the spoken alphabet to text with high accuracy. This text is then mapped to a picture if it is a static gesture or a video if it is a dynamic gesture. This system classifies the gesture as static or dynamic by measuring the distance moved by the hand in subsequent frames. For static gestures, the system uses Zernike moments, a well-known shape descriptor in image processing. HSV Segmentation and Finger-tip detection showed satisfactory results in constrained environment, i.e. proper lighting and background with limited skin-colored objects. Static Gesture recognition was carried out on a lexicon of 24 alphabets (a-y, excluding j) and it succeeded with approximately 93% accuracy.

Artificial neural networks approach used in [3] In this paper, converting sign language to text by an automated sign language recognition system based on machine learning was proposed to satisfy this need. Artificial neural Backpropagation algorithm is used in the system. Input Layer was designed to contain 3072 neurons for Raw Features Classifier and 512 neurons for Histogram Features Classifier. Hidden Layer was designed to contain 10 neurons for each classifier. Output Layer had 3 neurons for each classifier. The system gives 70% and 85% accuracy rate from respectively Raw Features Classifier and Histogram Features Classifier. When considered other studies, the obtained results are average results. The recognition rate can be increased by improving processing image step as a future work.

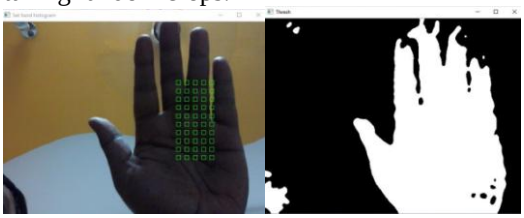
In [4] The propose system was able to recognize single handed gestures accurately bare human hands using a webcam which is MATLAB interface. The aim of this project is to recognize the gestures with highest accuracy and in least possible time and translate the alphabets of Indian Sign Language into corresponding text and voice in a vision-based setup. MATLAB and PCA algorithm. 260 images were included in training and testing database. The images are captured at a resolution of 380X420 pixels. The runtime images for test phase are captured using web camera. Otsu algorithm is used for segmentation purpose. Feature extraction is a method of reducing data dimensionality by encoding related information in a compressed representation Sign reorganization using PCA is a dimensionality reduction technique based on extracting the desired number. A MATLAB based application performing hand gesture recognition for human-computer interaction using PCA technique was successfully implemented. The proposed method gave output in voice and text form.

III. METHOD

The method used belongs to supervised machine learning where stochastic gradient descent was used along with SoftMax activation function at the output layer. Our goal was to train the neural network for proper classification of sign language gestures. The inputs were fixed size high pixel images, 200 by 200 or 400 by 400, being padded and resized to 200 by 200.

A. Data collection and preprocessing

For the data collection, we manually collected some data from Indian Institute of Sign Language. Since the dataset was insufficient for relatively better performance of the model, we were supposed to add some data manually. We captured training images for gestures using the histogram of Gradients approach using the computer vision library. A 10X10 histogram is implemented and only the image in that part of ROI is extracted. The image is converted from BGR to HSV. The processing of images starts from here. The histogram is then calculated for its hue and saturation value from the extracted HSV image. The histogram is then normalized. We use the back-projection operation on the histogram to recognize only the skin color and to avoid background noise [2]. The smoothness of the image is increased by applying gaussian and median blur on the histogram. The gesture is placed inside the histogram as in figure 2 and every gesture is captured.1200 images has been captured and save in a directory. These 1200 images are the rotated along the vertical axis which adds up to 2400 images per gesture. we attempted padding the images with black pixels such that they preserved their aspect ratio upon re sizing. This padding also allows us to remove fewer relevant pixels upon taking random crops.



(a) Input image (b) Processed image

Figure 2: Image processing

B. Convolutional Neural Network

CNNs (based on [7]) are feature extraction models in deep learning that recently have proven to be to be very successful at image recognition [6], [3], [10], [8]. As of now, the models are in use by various industry leaders like Google, Facebook and Amazon. And recently, researchers at Google applied CNNs on video data [11]. CNNs are inspired by the visual cortex of the human brain. The artificial neurons in a CNN will connect to a local region of the visual field, called a receptive field. This is accomplished by performing discrete convolutions on the image with filter values as trainable weights. Multiple filters are applied for each channel, and together with the activation functions of the neurons, they form feature maps. This is followed by a pooling scheme, where only the interesting information of the feature maps are pooled together. These techniques are performed in multiple layers as shown in Figure 3.

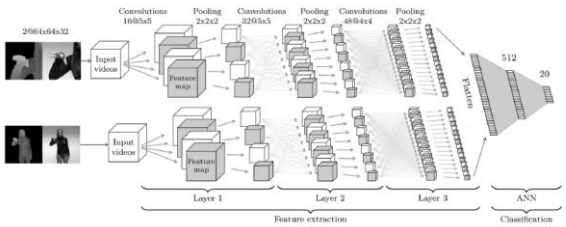


Figure 3: CNN Model

C. Architecture

Most implementations surrounding this task have attempted it via transfer learning, but our network was trained from scratch. Our general architecture was a fairly common CNN architecture, consisting of multiple convolutional and dense layers. The architecture included 2 groups of 2 convolutional layers followed by a maxpool layer and a dropout layer, and two groups of fully connected layer followed by a dropout layer and one final output layer. The activation functions used in both the convolution layer is Rectified linear unit (ReLU) and the activation function for the output layer was the SoftMax function where the loss function is given by,

$$\text{Loss}=\frac{1}{N}\sum_{i=1}^{N}-\log\left(\frac{e^{fi,yi}}{\sum_{j=1}^{C}e^{fi,j}}\right) \quad (1)$$

$$f_j(z)=\frac{e^{zj}}{\sum_{k=1}^{C}e^{z}k} \quad (2)$$

Equation (2) is the SoftMax function. It takes a feature vector z for a given training example and squashes its values to a vector of [0,1]- valued real numbers summing to 1. Equation (1) takes the mean loss for each training example, xi, to produce the full SoftMax loss. Using a SoftMax-based classification head allows us to output values akin to probabilities for each gesture. The input for the model is given through a webcam using the histogram of ordered gradients as discussed above and the corresponding output is displayed on the screen.

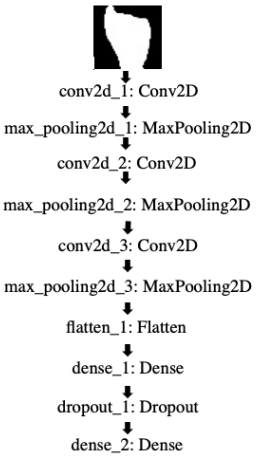


Figure 4: Architectural Model

IV. EXPERIMENTAL RESULTS

The experiment results were highly accurate and up to 99% of accuracy is achieved under standard lightning conditions. Our model took significantly long hours to train the model under standard processor. We can save the trained model, so training again is not necessary until the dataset is altered. The confusion matrix provides the necessary details regarding the labels and the accuracy of the trained model.

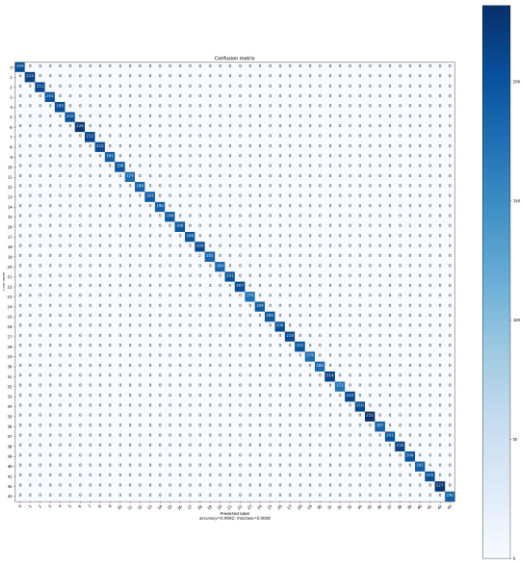


Figure 5: Confusion Matrix

Model accuracy and loss is increased and decreased respectively as the number of epochs increase. We had to stop training near to 20th epoch in order to avoid over fitting.

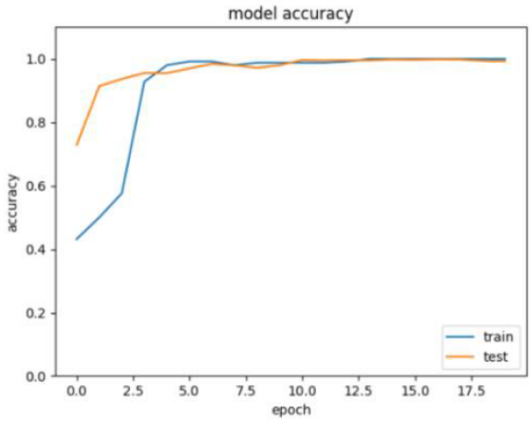
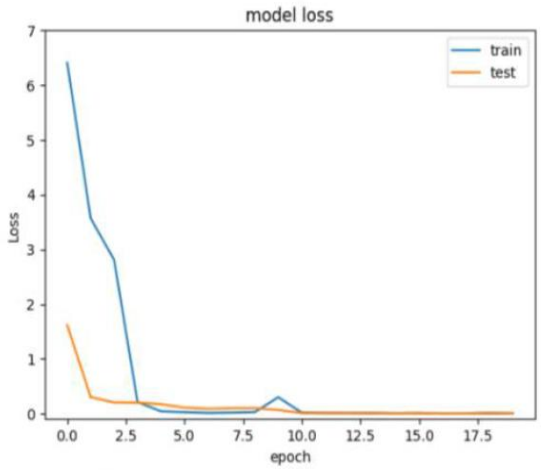


Figure 6: model accuracy
Figure 7: model loss



V. CONCLUSIONS

The earlier sign language translators were not either efficient or economical. Sign language translator gloves and Microsoft kinetic devices are not economical and other machine learning algorithms was not able to achieve the accuracy and efficiency. We implemented and trained the model up to the mark. Even it can be made better with additional dataset and under standard conditions it can achieve up to 100% of accuracy. The training time can also be significantly reduce using the inception V3 model and using higher configurations processor. Based on this we are proposing a novel approach to ease the difficulty in communicating with those having speech and vocal disabilities. Since it follows an image-based approach it can be launched as an application in any minimal system and hence has near zero-cost.

REFERENCES

[1] Verschaeren, R.: Automatische herkenningvan gebaren met de microsoft Kinect(2012)

[2] Anup Kumar, Karun Thankachan, MevinM Dominic, “sign language recognition” 2016 3rd International Conference on Recent Advances in Information Technology(RAIT) IEEE, 11 July 2016

[3] R. Tu¨lay Karayılan and O¨zkan Kılıc¸,“Sign language translation”, 2017 International Conference on Computer Science and Engineering (UBMK) ,IEEE, 02 November 2017.

[4] Shreyashi Narayan Sawant and M. S. Kumbhar , ”Sign Language Translation”, 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies , 26 January 2015

[5] Ciresan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 36423649. IEEE (2012).

[6] J. Jarrett, K., Kavukcuoglu, K.: What is the best multi-stage architecture for object recognition? Computer Vision, 2009 IEEE 12th International Conference on pp. 21462153 (2009) http://ieeexplore.ieee.org/xpls/absall.jsp?arnumber=5459469

[7] Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11) (1998)

[8] *Spec* Goodfellow, I.J., Bulatov, Y., Ibarz, J., Arnoud, S., Shet, V.: Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv preprint arXiv:1312.6082 (2013)

[9] V. G Adithya, P. R. Vinod, Usha Gopalakrishnan, ”Artificial Neural Network Based Method for Indian Sign Language Recognition”, IEEE Conference on Information and Communication Technologies (ICT),pp. 1080-1085, 2017

[10] Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional neural networks. arXiv preprint arXiv:1311.2901 (2013)

[11] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large scale video classiffication with convolutional neural networks. In: CVPR (2014)

[12] Suruchi Bhatnagar, Suyash Agrawal :.”Hand Gesture Recognition for Indian Sign Language: A Review” IJCTT vol. 21, number. 3, pp. 121,mar-2015

[13] Akshata Dabade, Anish Apte ,Aishwarya Kanetkar, Sayali Pisal “Two Way Communication between Deaf & Dumb” IJCTT vol.40 ,number.3 ,pp.114 ,oct-2016