Hand Sign Gesture Recognition Using Convolution Neural Networks

Rishabh Bansal, Saurav Singh Negi, Varnit Jain and Aakanksha Mahajan

Assistant Professor,

Dept. of Computer Science & Engineering,

Panipat Institute of Engineering & Technology, Haryana, India

Abstract

A real-time sign language translator is an important milestone in facilitating communication between the deaf and mute community and the general public. We hereby present the development and implementation of an American Sign Language (ASL) fingerspelling translator based on a convolutional neural network. In this paper, we are trying to propose a method that uses the power of a Convolutional Neural Net to identify and recognize hand signs which are captured in real-time through a laptop's webcam.

1. INTRODUCTION

Sign language is a way of communication between deaf and mute people. While communicating with mute and deaf peoples, those who know sign language, can talk and understand properly. Sign language to text and speech system will be more useful and easy for deaf and mute people to communicate with others more fluently. They can express their feeling with different hand shapes and movements. The task is to convert that shape or their sign language into text or speech. Due to advancements in the field of image processing, an automatic sign language converter system is developed. Few researchers have developed tools to help to convert sign language into text or speech.

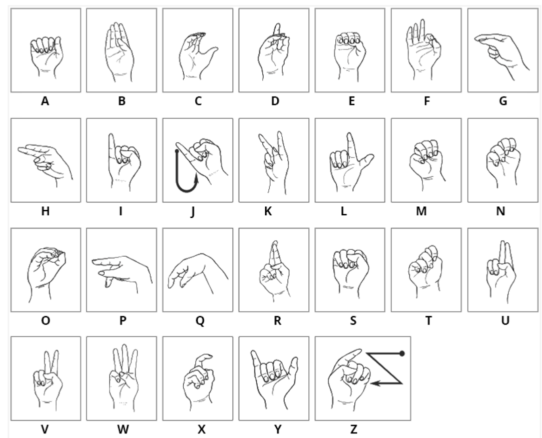


Figure 1. American Sign Language Alphabet

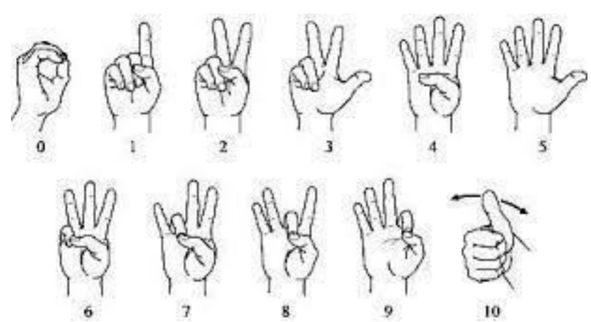


Figure 2. American Sign Language Numbers

Researchers in the field of sign language are broadly categorized in two ways, Data glove & Image processing. In data glove system, user needs to wear glove. Glove consists of flex sensor, accelerometer and motion tracker. Sensor output signals are sending to the computer for processing and analyze the gesture and convert into text or speech. In image processing, image is captured through web camera and then trained via CNN model.

2. Related Work

Convolutional Neural Networks have been extremely successful in image recognition and classification problems, and have been successfully implemented for human gesture recognition in recent years. In particular, there has been work done in the realm of sign language recognition using deep CNNs, with input-recognition that is sensitive to more than just pixels of the images. With the use of cameras that sense depth and contour, the process is made much easier via developing characteristic depth and motion profiles for each sign language gesture . The use of depth-sensing technology is quickly growing in popularity, and other tools have been incorporated into the process that have proven successful. Developments such as custom-designed color gloves have been used to facilitate the recognition process and make the feature extraction step more efficient by making certain gestural units easier to identify and classify . Until recently, however, methods of automatic sign language recognition weren't able to make use of the depth-sensing technology that is as widely available today. Previous works made use of very basic camera technology to generate datasets of simply images, with no depth or contour information available, just the pixels present. Attempts at using CNNs to handle the task of classifying images of ASL letter gestures have had some success [7], but using a pre-trained Google Net architecture.

3. Method

Our overarching approach was one of basic supervised learning using mini-batch stochastic gradient descent. Our task was that of classification using deep convolutional neural networks to classify every letter and the digits, 0-9, in ASL. The inputs were fixed size high-pixel images, 200 by 200 or 400 by 400, being padded and resized to 200 by 200.

3.1 Architecture

there can be many different configurations of a Convolutional Neural Net model. After some trial and error, we settled for a model which consists of a total of 9 layers of which 4 layers are of convolutional layer, 2 maxpool layers, 2 fully connected layers and 1 flatten layer. We have also added some dropout layers which randomly drop some firing neurons so that we could avoid overfitting of the model.

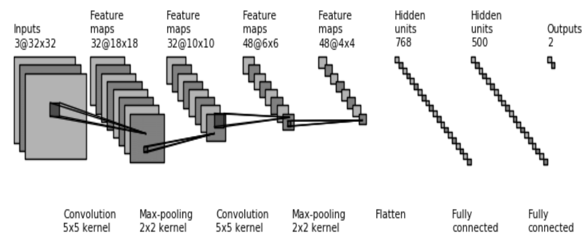


Fig - 3: Model Architecture

4. Data

Data or photos of hand signs were collected using a web camera. We took a total of 1600 photos, 320 photos for each hand sign. Photos of hand signs were taken in various environments with different lighting conditions having slightly different orientation in each photo. So that we could simulate all types of real-world conditions. Given below are the hand signs on which our model will be trained on. Then we flipped each image to create a larger dataset. These images were also used in training dataset.

4.1 **Preprocessing**

Every image in the dataset is converted to an image of 128x128 pixel. After that we have applied random transformation like shear, rotation, zoom, width shift and height shift using ImageDataGenerator class from Keras library. Due to small size of dataset we made a validation split of 80/20. Meaning 80% of images (which is equal to 1280 images) will be used for training the model. And the remaining 20% (320 images) will be used for model validation.

4.2 **Training**

Adam Optimizer was used with a learning rate of 0.0001. Model was trained to 60 epochs. Post training Validation Accuracy was 0.7480 and Validation Loss was 0.6883. We also flipped the images horizontally as we can sign using both hands. While it wasn't extremely effective, we saw that with better and more representative initial training data, augmenting improved the performance more drastically.

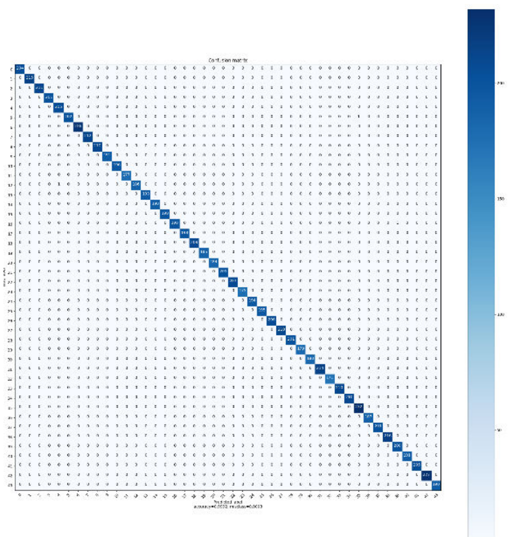
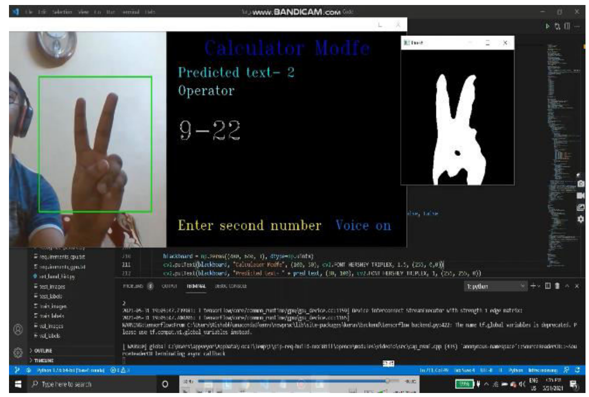
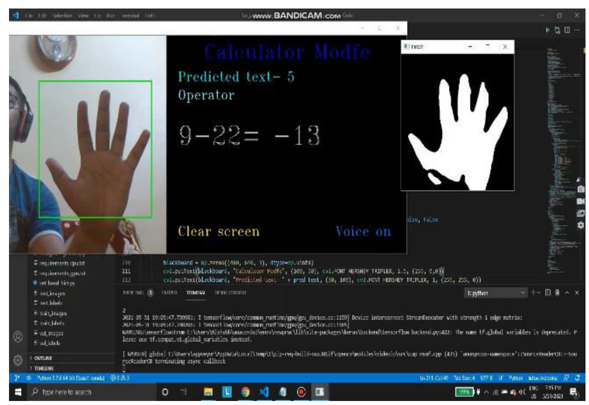


Fig - 4: Confusion Matrix

5. Results

We observed on our self-generated dataset, we have much lower accuracy measures, as was expected since our data was less uniform than that which was collected under studio settings with better equipment. We saw our trained CNN model can be further developed to process the continuous input stream of sign language and convert them into their corresponding sentences. Our model's validation accuracy is 74.80% whereas training accuracy is 91.90%. High training accuracy and low validation accuracy means our model was not able to generalize well on the given data and there might be some degree of overfitting going on in our model.





6. Conclusions

In this paper, we described a deep learning approach for a classification algorithm of American Sign Language. Our results and process were severely affected and hindered by skin color and lighting variations in our self-generated data which led us to resort to a pre-made professionally constructed dataset. With a camera like Microsoft's Kinect that has a depth sensor, this problem is easy to solve [5]. However, such cameras and technology are not widely accessible, and can be costly. Our method shows to have potential in solving this problem using a simple camera, if enough substantial training data is provided, which can be continuously done and added via the aforementioned processing pipeline. Since more people have access to simple camera technologies, this could contribute to a scalable solution.

7. FUTURE WORK

1. Develop the model to process the continuous input stream of sign language and convert them into their corresponding sentences.

2. Develop a mobile app based on this model which will take it to whole another level as it will be much more usable and practical.

REFERENCES

[1] Hong gang Wang, Ming C. Leu and Cemil Oz, "American Sign Language Recognition Using Multidimensional Hidden Markov Models", Journal of Information Science and Engineering, January 2006.

[2] Lionel Pignou. Sander Dieleman, Pieter-Jan Kindermans, Benjamin Schrauwen, "Sign Language Recognition Using Convolutional Neural Networks", March 2015.

[3] Rafiqul Zaman Khan and Noor Adnan Ibraheem, "Hand Gesture Recognition: A Literature Review", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.4, July 2012.

[4] Karthick Arya, Jayesh Kudase, "Convolutional Neural Networks based Sign Language Recognition", International Journal of Innovative Research in Computer and Communication Engineering, Volume 5, Issue 10, October 2017. [5] Vivek and N. Dianna Radpour, "Using Deep Convolutional Networks for Gesture Recognition in American Sign Language", Department of Computer Science, Department of Linguistics, State University of New York at Buffalo, 2017.

[6] Pratibha Pandey and Vinay Jain, "International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 3, March 2015.

[7] Eirini Mathe, Alexandros Mitsou, Evaggeles Spyrou and Phives Mylonas, "Hand Gesture Recognition Using a Convolutional Neural Network", 13th International workshop on Semantic and Social Media Adaptation and Personalization (SMAP), 2018.