# Gesture Recognition of RGB and RGB-D Static Images Using Convolutional Neural Networks

For the write-up on academic paper addressing a problem with neural net architecture, I opted a paper that proposed a solution to recognize sign language by using computer vision. Many advancements are happening in the computer-human interaction field. Researchers are trying to find different technologies that improve this interaction. These advancements include devices and technologies that benefit people with disabilities. Deaf people cannot use voice interaction while interacting with the computer. Deaf people use sign language to express themselves. The author of the paper tried to address this issue by creating a machine learning model using a convolutional neural network. There are many sign languages used; however, the commonly used sign language is American Sign Language in the United States. There are many technologies available to recognize gestures, some focus on full-body, and some focuses on eyes. Many of these technologies need add-on devices. Some of the other machine learning models focused on performance rather than accuracy. There are two categories of the computer-vision based technique Static Gesture Recognition and Dynamic Gesture Recognition.

Convolution neural network is very efficient in tasks that involve image recognition. This model uses the VGG-19 model. VGG-19 is a convolution neural network that is 19 layers deep. This model is pre-trained on one million images from the ImageNet database. Convolution neural network is a machine learning model that resembles the human eye image processing. The neurons of the eye are in the form of layers. The first layer takes image pixels as input. Each pixel brightness is a result of a sensor in the retina. The nodes of the first layer find out quite basic patterns like edges in the bright and dark from these pixels. Each additional layer processes and combines patterns found by the previous layer to find a more complex structure than the prior layer. The final layer gives the output as binary, or probabilities depending on the input labels. The comparison of predicted labels and the observed labels gives the loss. As per the calculated loss, the weights are adjusted on the prior layer. This process is known as back-propagation.

With the popularity of the convolution neural networks, another term became popular known as Transfer Learning. Transfer Learning is the transfer of the already trained model parameters to train another model. The advantages of transfer learning are, it avoids overfitting and reduces the time required to train the new model. The paper mostly focused on the technique to improve Static Gesture Recognition. The proposed algorithm in the paper helps to eliminate the need for feature extraction and reducing computational power requirements. The VGG network model only uses 3X3 convolution layers stacked on each other in increasing depth. All the hidden layers in the VGG19 model are using RELU activation. In the first stage, the VGG19 model is fine-tuned by the author by training top layers with new data from the ASL dataset and freezing all the other layers. In the second stage, to avoid overfitting, the complete model was not trained. Rather, the first sixteen layers frozen, and the remainder of the layers are trained. By using this approach, two separate models trained. One model trained with RGB images and another model with Depth images. Both models merged to increase accuracy. The first model trained with 46000 RGB images and another model trained with 46000 Depth images. For testing, the models 9000 images of each type used.

The average accuracy achieved by the writer is around 95.29 percent on the training data set and about 94.80 percent on the test data set. The model performed quite well on test data when compared with different approaches like SIFT+PLS, H3DF+SVM, and Gabor+RDF. The accuracy of other models was in the range of ’70s.

After going through the paper, I have one question that how much effort and time is saved in training by using the VGG19 pre-trained model compared to the training complete model from scratch?

**Reference:**

**Gesture Recognition of RGB and RGB-D Static Images Using Convolutional Neural Networks** [**https://www.researchgate.net/profile/Ruben\_Gonzalez\_Crespo/publication/335798387\_Gesture\_Recognition\_of\_RGB\_and\_RGB-D\_static\_Images\_using\_Convolutional\_Neural\_Networks/links/5dba0da392851c818019294d/Gesture-Recognition-of-RGB-and-RGB-D-static-Images-using-Convolutional-Neural-Networks.pdf**](https://www.researchgate.net/profile/Ruben_Gonzalez_Crespo/publication/335798387_Gesture_Recognition_of_RGB_and_RGB-D_static_Images_using_Convolutional_Neural_Networks/links/5dba0da392851c818019294d/Gesture-Recognition-of-RGB-and-RGB-D-static-Images-using-Convolutional-Neural-Networks.pdf)