Rabea Ahmed

1711894042

Week 5- Paper Review

CSE499B

**Real-Time Bangla Sign Language Detection with Sentence and Speech Generation**

April 2021

**Abstract**

Sign language is the only medium of communication for the hearing impaired and the deaf and dumb community. Communication with the general mass is thus always a challenge for this minority group. Especially in Bangla sign language (BdSL), there are 38 alphabets with some having nearly identical symbol. This paper proposes a system for BdSL recognition that can interpret BdSL from a sequence of images or a video stream and generate both textual sentences and speech in real-time. We have used YOLOv4 as the object detection model. We have also proposed three new signs for the sentence generation task and built a dataset consisting of 12.5kBdSL images of 49 different classes where 39 are Bangla alphabets, 10 are Bangla digits, and the three new proposed signs

**Introduction**

Approximately 466 million people have hearing disability, which is over 5% of the world’s total population. 0.38% of the total population of Bangladesh have speech and hearing disabilities. Most are deprived of education and fail to compete in the job market due to a lack of facilities. But can have a positive contribution to the nation’s economy if given adequate opportunity to interact with others. Due to their speaking and hearing disability, they can’t communicate through spoken language. So they use the non-verbal means of communication that help them bridge the communication gap with others, also known as sign language where hand gestures and movements are used. Sign languages are completely natural languages with their own grammar and lexicon.

**Literature Review**

For this project, the finger-spelling technique of sign language was used. Finger-spelling is a good way to represent each alphabet of a writing system using different hand formation. These finger alphabet, also known as the manual alphabet, are significantly used for educating students with any sort of hearing loss or deafness. There exist very few words in any language for which a sign is available. In such a case, the finger-spelling of words is the leading alternative for communication for deaf people. Using our proposed technique, finger-spelling can also be used to write documents. For this purpose, we have created a dataset consisting of 12.5k images of 49 different signs taken in various back-grounds and under several lighting conditions. It consists of 10 digits and 36 characters used in standard BdSL along with three proposed signs for generating sentences, which are compound characters, space, and end of sentence.

**Methods**

The system of this project can be divided into four subsidiary sections: Object Detection Process, Word Generation, Sentence Generation, and Speech Generation. Object Detection Model YOLOv4 was used in this system which is primarily divided into three sections- the Backbone Model extracts features from the input image. The Neck Model generates a feature pyramid. The Neck Model helps the object detector to detect the same object at different scales, thus helps with the generalization of the convolution neural network. The Head Model obtains features from the previous section and predicts the bounding-box area and the associated class.

There are multiple models like VGG16, SpineNet, CSPResNeXt50, CSPDarknet53 etc. that can be used as backbone model for the network. For this work as the model backbone, CSPDarknet53 is used which significantly reduces computational cost to extract features from the input image. CSPNet increases.

CNN’s ability to learn even after reducing the number of layers. The computation distribution property of CSPNet reduces the computational bottleneck and increases the utilization rate of each computational unit. Adopting cross-channel pooling, the model compresses the feature maps obtained from the input image during the generation of feature pyramids, which helps to reduce memory consumption on a large scale. As Neck model architecture, FPN or PANet models provides good result, which was used as the object detector. PANet focuses to increase efficiency during the propagation of feature information through the Neural Network.

**Results**

The system has been evaluated with 5500 test images which was not used to train. Four metrics were calculated – Precision, Recall, F1-Score, and Accuracy to assess the performance of the object detection model. Each of the detection classes was considered separately for measurement to clarify performance in each class. Precision is the number of correctly predicted of a class divided by the total number of predictions made by the algorithm of that class. Recall is defined as the total number of correctly predicted classes divided by all positive samples of that class. The F1 score can be seen as a bridge between Precision and Recall which is proved to be a better metric in case of uneven class distribution

**Discussion**

To train the model on 12500 images, Stochastic Gradient Descent(SGD) optimizer was used. It had a learning rate of 0.04, momentum 0.93, optimizer weight decay 0.0005, GIoU loss gain 0.05, class loss gain 0.5, object loss gain1.0, the threshold of IoU 0.20, the threshold of multiple anchors 4.0, focal loss 0.0. With the hyper parameter values, the model took 100 epochs to converge. A learning rate of 0.01 and momentum 1.0 was tried which took around 180 epochs to provide good results. Also, Adam optimizer with a learning rate of 0.001 took 300epochs to converge and provide good results. The image resized to 640x640 before being fed to the model. Eesizing the image to 480x480 was also tried before training the model but it did not provide a good result for detection purposes.

**Conclusion**

People who suffer from physical disability of hearing and speaking faces problems in their daily life regarding communication. This paper aimed to work for the welfare of people with such impairment. The proposed system converts sign language into sentences and later speech in real-time. The system can help to communicate with blind and mute people. Also, it can help the mute and deaf person in day to day communication tasks. The system helps to create a translator for BdSL. As the system captures over 30 frames per second, a person with high efficiency in sign language can also generate sentences in real quick time.

**Reference**

[(PDF) Bangla sign language recognition using concatenated DBs network (researchgate.net)](https://www.researchgate.net/publication/353479114_Bangla_sign_language_recognition_using_concatenated_BdSL_network)