EasyChair Preprint

№ 2312

# Sign Language Recognition with Visual Attention

Shweta Upadhyay, R. K. Sharma and Prashant Singh Rana



January 4, 2020

# Sign Language Recognition with Visual Attention

Shweta Upadhyay, Dr. R.K. Sharma, Dr. Prashant Singh Rana

*Thapar Institute of Engineering and Technology, Patiala*

**Abstract**

Sign Lnaguage Recognition hold significant importance to move towards a globally connected generation, by laying the foundation in the development of support systems for the deaf community. Several CNN based approaches have been explored in the past to tackle the recognition of hand sign gestures. In this work, we implement the techniques that utilize the phenomenon of spatial attention for the classification and recognition of the American Sign Language (ASL) in natural scenario. We experiment on the ASL Alphabet dataset, which is a publicly available dataset, to analyze the performance of the proposed framework.

*Keywords:* Sign Language Recognition, ASL, CNN, Faster RCNN, Visual Attention

## 1. Introduction

With the advancement of technology and diversity in the nature of jobs, more people from various communities are joining in to be educated and to learn new techniques. Sign languages are built ton the notion that vision is the most significant mode of communication among and with the deaf community, thereby, increasing the importance of their successful recognition. At the same time, sign languages are practiced at places and situations where verbal communication is practically impossible, for example, underwater, or at concerts.

In the real-world scenario, signs or hand-gestures can exist in countless forms, places and situations, prompting the researchers to work with their recognition in natural scenario. Complex and inconsistent backgrounds and the highly diverse ways of performing these gestures by different people, may pose numerous challenges while trying to successfully recognize and classify these signs.

Convolutional Neural Networks [1] [2] [3] are one of the most widely used models to solve computer vision problems in the course of present day. To take on the task of recognizing gestures, researchers have advanced the CNNs to various techniques, and achieved substantial results. In this paper, we focus on the concept of visual attention for the task of recognition and classification of the American Sign Language (ASL) in natural scenario. We then analyze the outcomes and present a comparison with the existing approaches of sign language recognition.

The subsequent paper is laid out as follows: Section 2 presents a review of the published literature in the area of this study and the state-of-the-art for the task of sign language recognition. Section 4 demonstrates the methodology to administer and implement the proposed approach. Section 5 displays and analyses the qualitative and quantitative results in details. Finally, section 6, concludes the study and yields recommendations for the future work.

## 2. Related Work

In this section, we discuss the recent works in the field of recognizing sign language gestures and the spatial attention.

### *2.1. Sign Language Recognition*

[4] Proposed an Adaline Network based framework for detecting sign languages. It firstly introduced an advanced feature set over the previous ones and secondly, suggested MAdaline network, an extension for Adaline Network. It reduced the system complexity by eliminating the step of cropping input images to remove the irrelevant background.

[5] Implemented ASL fingerspell translator using CNNs. The authors used the GoogleNet architecture trained on ILSVRC2012 and ASL datasets, and further used transfer learning sign language recognition.

[6] Introduced a hybrid CNN-HMM (Hidden Markov Model) architecture which combined the sequence modelling of HMMs with discriminative abilities of a CNN and thus enabled the model to deal with sequence data for training and evaluation of sign language recognition.

[7] Proposed a 3D CNN to perform classification-detection of real time hand gestures. This work introduced a multimodal dynamic hand gesture dataset and used connectionist temporal classification to predict class labels.

The system achieved accuracy of 88.4% and state-of-the-art performance on SKIG and ChaLearn2014 datasets.

### *2.2. Spatial Attention*

[8] introduced attention based technique to train the model for the task of machine translation. Visual attention allows the model to learn to automatically focus on salient features of the given input image for generating corresponding words in the output sequence. The authors implemented an encoder-decoder based model and provided detailed visualizations to thoroughly explain the working mechanism. Two variants of attention namely: 'hard' and 'soft' are incorporated on 3 datasets.

[9] presented an empirical study on the structure and working of spatial attention mechanisms and their deep neural network based applications. The authors examined how different inuencing factors and computing methods affect the performance. They broadly arranged the attention computing methods into two major categories, i.e., encoder-decoder attention and self-attention to suit the study and analysis purpose.

## 3. Proposed Approach

Prior work has illustrated the use of numerous CNN variants to tackle the problem of recognition of sign languages. In this paper, we propose to implement visual attention based approach that works inside the proposed regions to focus on the target object.

There are many existing algorithms for object recognition including RCNNs, YOLO etc. In this work, we employ Faster-RCNN [10] as the baseline for our detection model, since it provides state-of-the-art performance for various detection tasks.

The Faster-RCNN model reuses CNN results which are used to calculate the image features, for generating region proposals i.e., a single CNN is used for generating region proposals as well as classification. Therefore, the entire process needs only one CNN to be trained and the region proposal process is almost cost free.

The proposed architecture (Figure 2) is similar to that of faster-RCNN, where the input image volume is passed through CNN, returning feature map of that image. RPN (Region Proposal Network) is applied to generate region proposals for the target object and RoI pooling brings all proposals to same size. Finally, fully connected layer is applied which has softmax classifier

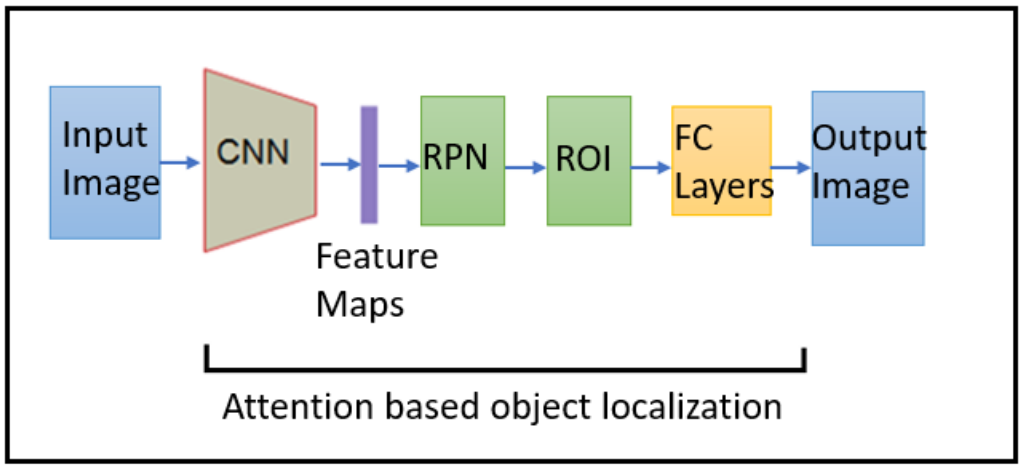


Figure 1: Proposed Framework

and linear regression layer. Visual attention mechanism for localization of target object (hand gestures for ASL) in the proposed regions. The attention regions are re-scaled to the size of the bounding boxes.

## 4. Experiments

### *4.1. Dataset Description and pre-processing*

For this work, we have evaluated our proposed model over the ASL Alphabet dataset [11], publicly available for use.

The ASL Alphabet dataset , consists 87,000 hand-gesture images. It includes 29 sets of sign classes, each containing about 3,000 images. The classes constitute of 26 alphabet classes which represent, A-Z english alphabets in the ASL, and 3 special characters classes of ASL, namely 'SPACE', 'DELETE' and 'NOTHING', which are extremely useful in the real-time applications. The test data-set merely contains 29 images, one for each of the classes, to encourage using real-time test data. But for this work, we separated 300 images from each of training classes to be added to the test set of that class. Hence, making 2,700 images in each train folder and 8729 images in the test dataset.

### *4.2. Implementation Details*

All experiments are carried out on Xeon E5 processor with GTX 1070 Ti GPU. Deep learning library tensorflow is used for the implementation of the

model. We have used Adam optimizer for the training module. The initial learning rate is 0.0001.

## 5. Results

This section presents the qualitative and quantitative outcomes of the implemented algorithms and the experiments performed.

Figure 2 and 3 show the qualitative results recognition of sign languages hand gestures performed in real time.

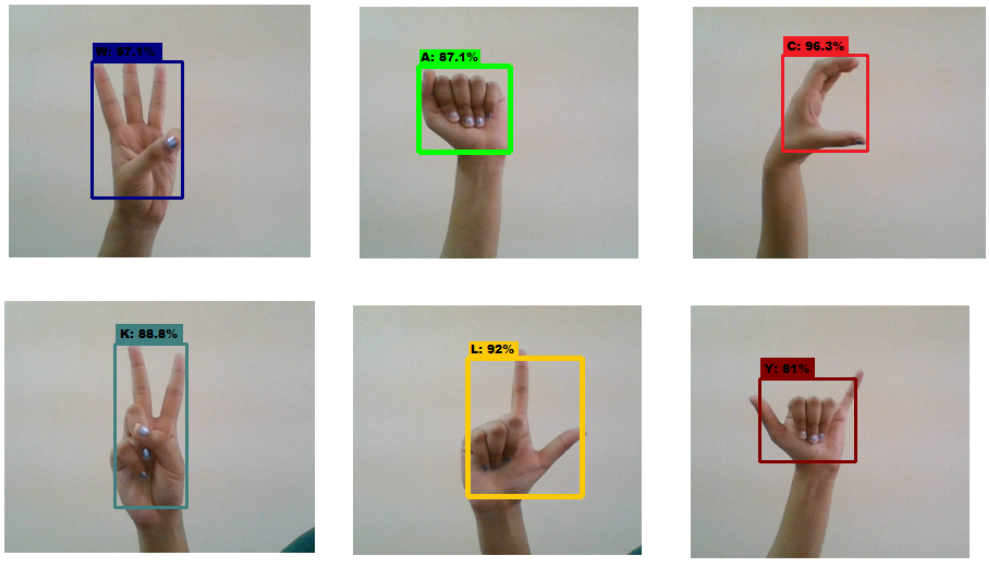


Figure 2: Real time recognition results for alphabets: W, A, C, V, L and Y (Top-left to bottom-right)

During the course of the experiment, each class was trained separately to evaluate class-wise training performance. Table 1 provides a comparision of the training accuracy percentages for Faster-RCNN and our attention based RCNN. It is noted that a significant improvement over the Faster-RCNN is shown by its attention based variant. The attention mechanism helps improve the outcomes, especially in the classes where the accuracies are are lower, for example, classes S and T.

Table 2 displays the overall test performance accuracy in percentage of Faster-RCNN and our proposed method over the ASL Alphabet dataset.

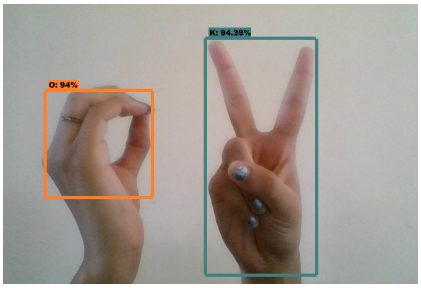


Figure 3: Recognition of alphabets O and V

Table 1: Training Accuracy (in percentage) for signs A to Z

| **Symbol** | **Faster-RCNN** | **Attention Faster-RCNN** |
|---|---|---|
| A | 88.70 | 91.24 |
| B | 96.33 | 93.48 |
| C | 90.73 | 96.76 |
| D | 89.01 | 92.49 |
| E | 78.42 | 87.03 |
| F | 90.21 | 91.20 |
| G | 94.00 | 89.93 |
| H | 92.28 | 94.88 |
| I | 95.84 | 95.82 |
| J | 92.14 | 95.01 |
| K | 86.90 | 94.57 |
| L | 93.78 | 95.65 |
| M | 91.34 | 94.02 |
| N | 92.99 | 96.33 |
| O | 96.31 | 97.68 |
| P | 86.22 | 93.66 |
| Q | 87.90 | 90.15 |
| S | 78.30 | 84.59 |
| T | 82.05 | 89.81 |
| U | 91.00 | 92.45 |
| V | 83.77 | 92.12 |
| W | 92.21 | 94.00 |
| X | 86.25 | 93.28 |
| Y | 94.46 | 95.40 |
| Z | 92.55 | 96.98 |
| SPACE | 98.02 | 98.09 |
| DELETE | 96.91 | 97.43 |
| NOTHING | 90.99 | 95.10 |

Table 2: Sign Language Recognition performance for ASL Alphabet dataset.

| Model | Overall Performance |
|---|---|
| Faster RCNN | 89.72 |
| Faster RCNN with attention (proposed) | 94.87 |

## 6. Conclusion

The task of recognising hand gestures or signs in natural scenario requires strong discriminating features. In this paper, we proposed a novel approach to take on this task. The presented model is based on the concept of visual attention to localize the object of interest. The experiments infer that our proposed method shows a significant improvement over Faster-RCNN for recognizing ASL Alphabet data signs. We look forward to improving this task of sign language recognition in terms of accuracy and complexity and implementing systems for their successful understanding in future.

## References


[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (2015) 436.

[2] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[3] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[4] S. Saha, R. Lahiri, A. Konar, A. K. Nagar, A novel approach to american sign language recognition using madaline neural network, in: 2016 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2016, pp. 1–6.

[5] B. Garcia, S. A. Viesca, Real-time american sign language recognition with convolutional neural networks, Convolutional Neural Networks for Visual Recognition 2 (2016).

[6] O. Koller, O. Zargaran, H. Ney, R. Bowden, Deep sign: Hybrid cnn-hmm for continuous sign language recognition, in: Proceedings of the British Machine Vision Conference 2016, 2016.

[7] P. M. X. Y. S. Gupta, K. K. S. T. J. Kautz, Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks, CVPR, 2016.

[8] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International conference on machine learning, 2015, pp. 2048–2057.

[9] X. Zhu, D. Cheng, Z. Zhang, S. Lin, J. Dai, An empirical study of spatial attention mechanisms in deep networks, arXiv preprint arXiv:1904.05873 (2019).

[10] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[11] Kaggle, Asl alphabet, https://www.kaggle.com/grassknoted/asl-alphabet (2018).