Proceedings of IOE Graduate Conference, 2017
Volume: 5 ISSN: 2350-8914 (Online), 2350-8906 (Print)

A Novel Approach to Traffic Sign Localization and Recognition Based on Single Shot Multibox Detector

Sushma Shrestha [a], Sanjeeb Panday [b], Deepesh Lekhak [c]

[a, b, c] *Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, Tribhuvan University, Nepal*
Corresponding Email: [a] sanushr7@gmail.com, [b] sanjeeb77@hotmail.com, [c] deepeshlekhak@gmail.com

Abstract
The localization and recognition of traffic signs which are in the vicinity of the vehicle accurately on real time plays a key role in Intelligent Transportation Systems particularly in safety. In this work, a novel approach for traffic sign localization and recognition in the cluttered images is designed by using Single Shot Multibox Detector method (SSD). The training, testing and validation was performed on German Traffic Sign Detection Benchmark(GTSDB). When performance of designed system were compared with state –of- art methods Maximally Stable Extremal Regions(MSERs), Sliding Window Algorithm(Windows) and Support Vector Machine with Histogram Oriented Gradient Features using Convolutional Neural Network(SVM+HOG+CNN) for GTSDB, designed system in SSD showed best result in both aspects of accuracy and speed. The mean average precision(mAP) and timing of the system using SSD based on VGG net were 0.83 and 10 fps respectively whereas mAP and timing for system using SSD based on ZF net were 0.81 and 15fps respectively. The mAP for MSERs was 0.52 and for Windows was 0.49. The timing of system using Windows was 1 fps, with MSERs was 6 fps and with SVM + HOG + CNN method was 6 fps. On the basis of timing, system with ZF net was faster compared to system with VGG net.

Keywords
CNN, GTSDB, Traffic Sign Localization, Traffic Sign Recognition, SSD

1. Introduction

Traffic signs are signs erected at side of or above roads to provide important information to road users. They are simple and rigid objects, limited by shape with eye-catching colors designed to be noticeable to humans so that the information they carry is easy to understand. However, large number of accidents occur day by day as drivers fail to pay attention to traffic sign in time. There is a need for designing a system for the localization and classification of traffic signs in the vicinity of the vehicle to maximize the safety of passengers inside the vehicle and all other traffic participants.

It has mainly two distinguishable phases. First phase is localization which can also be considered as image segmentation problem where there are two classes: traffic sign and background class.In Traffic Sign Localization, it's necessary to know not only whether there is a traffic sign in an image, but also its position. Second phase is recognition phase in which discovered traffic sign candidates are classified into specific predetermined classes or super-classes using some classification techniques.

Traffic Sign Recognition is getting considerable interest in recent years and the interest is driven by the market for intelligent applications such as autonomous driving, advanced driver assistance systems (ADAS), monitoring violation of traffic rules, mobile mapping etc. Most of previous traffic recognition methods were based on color-based methods and/or shape-based techniques [1],[2]. Color-based methods are commonly used because traffic signs are usually red, yellow and blue. These methods are sensitive to changes in illumination, color of traffic signs is likely to fade. Shape based methods exploit the shape of traffic signs, such as circle, triangle and square to detect traffic signs, which are more robust to changes in illumination. Varying techniques have been used for shape based segmentation such as canny edge detection, Hough transform, template matching, radial base symmetry and

corner detection.

Ching Hao Lai et al. [3] used a traditional template based shape recognition method to detect red circle and red triangle traffic signs. Feature extraction based methods had used SURF [4], salient region features [5], key point detectors [6] for extracting features from images to classify traffic signs. Wahyono, L. Kurnianggoro et al. [7] had used SVM classifier with HOG features to classify traffic signs. However, these methods could not perform well with difficulties small signs, sign with similar appearance, partial occlusions, lighting variations and different traffic signs perspectives etc.

Ciresan et al. [8] participated in GTSRB benchmark competition. They used a multilayer perceptron (MLP). M. M. Lau et al. [9],Yang et al. [10]and S. Junget et al. [11] had used biologically inspired convolution neural networks for recognition of traffic signs in images. Convolutional neural networks are translational invariant and provides better performance irrespective of size, shape or orientations of objects in images. These methods provided good performance in well cropped images of traffic signs. Experiments were not done for cluttered images.

Moving towards in Object Detection in cluttered image by CNNs, in OverFeat [12], Sermanet et al. observed that convolutional networks are inherently efficient when used in a sliding window fashion, as many computations can be reused in overlapping regions. Another widely used strategy for object detection using CNNs is to first calculate some generic object proposals and perform classification only on these candidates. They can largely narrow the search ranges. R-CNN [13] was the first to use this strategy, but it is very slow for two reasons. Firstly, generating category independent object proposals is costly. Selective search [14] takes about 3s to generate 1000 proposals for the Pascal VOC 2007 images; the more efficient EdgeBoxes approach [15] still takes about 0.3 s. Secondly, it applies a deep convolutional network to every candidate proposal, which is very inefficient. Girshick et al. later proposed Fast R-CNN [16], which uses a softmax layer above the network instead of the SVM classifier used in R-CNN. Ignoring object proposal time, it takes 0.3 s for Fast R-CNN to process each image. To overcome the bottleneck in the object proposal step, in Faster R-CNN [17], Ren et al. proposed region proposal networks (RPNs) which use convolutional feature maps to generate object proposals. Liu, Wei, et al. introduced SSD [18], a fast single-shot object detector for multiple categories. SSD compares favorably to its state-of-the-art object detector counterparts Faster R-CNN [17], YOLO [19] in terms of both accuracy and speed. However, the performance of all of these object detection networks was evaluated on PASCAL VOC and ILSVRC, where target objects occupy a large proportion of the image.

2. Concept

The task of localizing traffic signs in cluttered images and classifying those signs automatically need to be done on real time and accurately. Authors have addressed this by designing an efficient method using Single Shot Multibox Detector. A key concept is the use of multi-scale convolutional bounding box outputs attached to multiple feature maps at the top of the network. This representation completely eliminates proposal generation and subsequent pixel or feature resampling stages and encapsulates all computation in a single network.The system was designed using VGG-16 Net and ZF Net architecture. These architecture for CNN were modified to implement SSD technique.Then extra convolutional feature layers (decreasing in size progressively) are added to the network. Therefore, these allow predictions of detections at multiple scales. The output of the system is predicted number of bounding box for traffic sign detected in the images and traffic sign class with probabilities of each class.

3. Research Methodology

3.1 Data Collection

German Traffic Sign Detection Benchmark (GTSDB) was used for traffic sign detection and localization task. It contains 900 images (1360 x 800 pixels) in PPM format.The images contain zero to six traffic signs.The sizes of the traffic signs in the images vary from 16x16 to 128x128. Traffic signs may appear in every perspective and under every lighting condition.

Annotations are provided in CSV files where fields are separated by a semicolon (;). They contain the following information: 1. Filename 2. Traffic sign's in the image (leftmost image column , upmost image row, rightmost

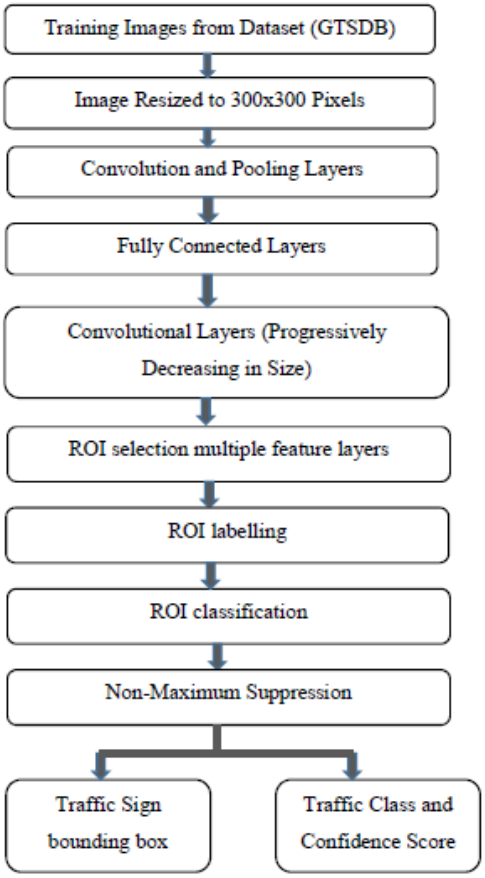

Figure 1: Block Diagram of Traffic Sign Localization and Recognition System

image column , down most image row of the ROI) 3. traffic sign's class ID

Distribution of traffic signs is uneven. All together there are 852 signs in train dataset and 361 traffic signs in test dataset. Table 1 shows the number of traffic signs in train and test dataset category-wise.

3.2 Single Shot Multibox Detector

Single Shot MultiBox Detector (SSD) is based on a feed-forward convolutional network that produces a fixed-size collection of bounding boxes and scores for the presence of object class instances in those boxes, followed by a non-maximum suppression step to produce the final detections. A standard architecture VGG-16 structure is used which is referred as base network and add some extra convolutional feature layers as auxiliary structure to the network to produce

Table 1: Number of Traffic Signs in different 4 categories

Traffic Sign Category	Train Dataset	Test Dataset
Prohibitory	413	163
mandatory	114	49
Danger	156	63
Other	169	86
Total Signs	852	361

detections. These layers decrease in size progressively and allow predictions of detections at multiple scales. The convolutional model for predicting detections is different for each feature layer.

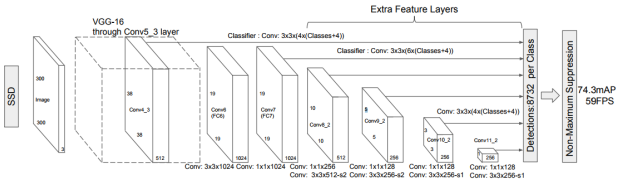

Figure 2: Single Shot Multibox Detector

SSD is simple relative to methods that require object proposals because it completely eliminates proposal generation and subsequent pixel or feature resampling stages and encapsulates all computation in a single network. It uses a mechanism for choosing Region of Interests (ROIs), training end-to-end for predicting class and boundary shift for that ROI.

Intersection over Union

Intersection over Union (IOU), which is also known as Jaccard Overlap, is an evaluation metric used to measure the accuracy of an object detector using two sets of bounding boxes i.e., ground-truth bounding boxes and predicted bounding boxes.

Computing IOU can be determined via:

$$IOU = \frac{Areaof Overlap}{Areaof Union} \quad (1)$$

Multiple feature map layers with different resolutions were used for generating ROIs. Some ROIs are labelled as positive and some negative depending on jaccard overlap after ground box has scaled appropriately taking

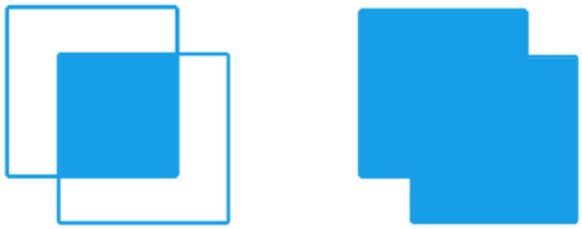

Figure 3: Area of Overlap (Left) and Area of Union (Right) between bounding boxes

resolution differences in input image and feature map into consideration.Any ROI that matched to Ground Truth for a class after applying appropriate transforms and having Jaccard overlap greater than threshold (0.5) is positive.

Single convolution kernel of 3*3 receptive fields were used to predict for each ROI, the 4 offsets (centre-x offset, centre-y offset, height offset, width offset) from the Ground Truth box for each RoI, along with class confidence scores for each class. As there were 44 classes (including background), so there were (44+4) filters for each convolution kernels that looks at a ROI. So summarily, convolution kernels look at ROIs (which were default boxes around each pixel in feature map layer) to generate 48 scores for each RoI.

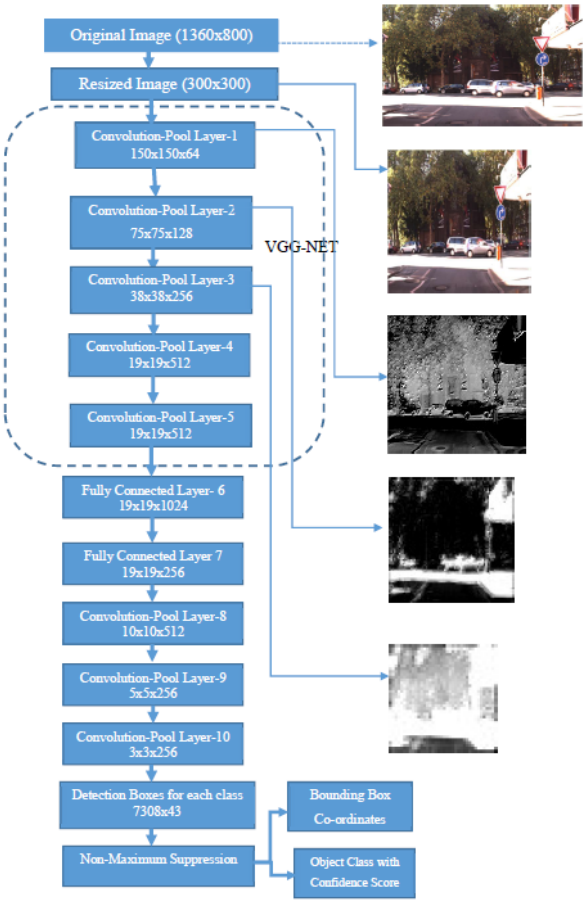

Figure 4: Architecture of SSD Implementation

Training

After pairing ground truths and default boxes, and marking the remaining default boxes as background, the objective function of SSD is formulated as:

$$L(x,c,l,g) = \frac{L_{\text{conf}}(x,c) + \alpha L_{\text{x,l,g}}}{N} \quad (2)$$

where N is the number of matched default boxes. If N = 0, the loss is set to 0. The weight term α is set to 1 by cross validation. The overall objective loss function is a weighted sum of the localization loss (loc) and the confidence loss (conf).

An input image and ground truth boxes for each sign were used during training. The loss function and back propagation were applied end to-end. During training which default boxes correspond to a ground truth detection was determined and the network was trained accordingly. For each ground truth box,selection was done from default boxes that vary over location, aspect ratio, and scale.At beginning, each ground truth box was matched to the default box with the best Jaccard overlap, then default boxes were matched to any ground truth with Jaccard overlap higher than a threshold (0.5). This allows the network to predict high scores for multiple overlapping default boxes rather than requiring it to pick only the one with maximum overlap.

Non-maximum Suppression

After training the network, at the time to put detector to use, one particular problem arises that multiple default boxes can be matched to a single ground truth box if the threshold is passed which is not the desired output of the system. Non-maximum suppression (NMS) is a post-processing algorithm used in the design of system for merging all detections that belong to the same object. The image was scanned along the image gradient direction, and if pixels were not part of the local

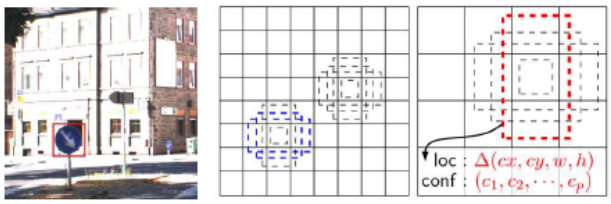

Figure 5: SSD framework

maxima they were set to zero. This has the effect of suppressing all image information that is not part of local maxima. Most confident boxes were added to the final output. If a candidate box highly overlapped (Jaccard overlap >0.5) any box of the same class from the final output, the box was ignored.

Figure 6: Non-Maximum Suppression

4. Result and Analysis

To evaluate the effectiveness of the proposed traffic sign localization and recognition approach comparative experiments were carried out using German Traffic Sign Detection Benchmark. The dataset were divided into train set of 540 images, validation set of 60 images and test set of 300 images. SSD Based on VGG net with different added layers was designed. The original images of 1360x800x3 size were resized to 300x300x3. Batch generator method was used to generate given batch size of images from actual data with some variance in some characteristics of the images.

Training Results

The training set from German Traffic Sign Dataset was used for training and validation with 10-fold validation. For training of the system, default boxes of size 4x4 and 8x8 were used. Default boxes with intersection-over-union greater than 0.5 were considered as positive samples and other were considered as negative samples. Traffic Sign Detection System was trained using following experimental parameter setup.

Table 2: Parameters of Training

S.N.	Parameter	Details
1	Image Size Resized	300x300x3
2	Basic Learning Rate(lr)	0.001
3	Decay	0.9
4	Dynamic Learning Rate	(lr)*decay**epoch
5	IoU Threshold	0.5
6	Batch Size	32

Batch size of 32 is used for training. Adam optimizer is used to optimize the training process. The training and validation loss are shown in figure 7.

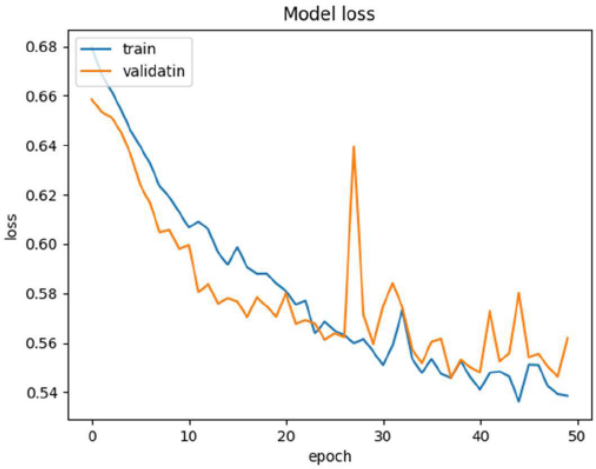

Figure 7: Loss Function of Training and Validation

Evaluation

After Training, model was used to evaluate the testing dataset of GTSDB. Some sample outputs from the system visualized in figure 8:

Detection accuracy of the system was calculated using mean average precision (mAP).Test Result in terms of mAP is shown in the table 3 The mean average precision of the system was 0.81 on ZF Net and 0.83 on VGGNet. Though this is not much different performance but it is clear that VGG performs better compared to ZF net. The comparative analysis of the results of some state-of-the-art detection algorithm with this system on German

Figure 8: Traffic Sign localization and Recognition on some GTSDB images

Table 3: Test Result of Traffic Sign Detection

Category	mAP SSD (VGGNet)	mAP SSD (ZFNet)
Prohibitory	0.879	0.833
mandatory	0.842	0.807
Danger	0.856	0.834
Other	0.7577	0.768
mAP	0.833	0.81

Traffic Sign Dataset was shown in table 4.

The designed system was evaluated for different IOU threshold value. The recall vs IoU analysis is shown in following table 5 The designed system for traffic sign recognition and localization showed best performance when IOU >0.5 is taken for the experiment and drops gradually requiring higher overlap.

The timing of the system for detection of traffic signs in images using the system is compared with other system in table 6 System designed in thesis work outperforms the previous methods. Also the SSD designed with ZF net as basic structure is faster compared to SSD design with VGG-16 net.

The precision – recall curve , i.e. PR curve, is created by plotting the precision (PPV) against the recall.A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall). The figure showed that the designed classifier tends to be at the upper right corner. As both curves obtained are close to the perfect precision-recall curve, therefore they have a better performance level.

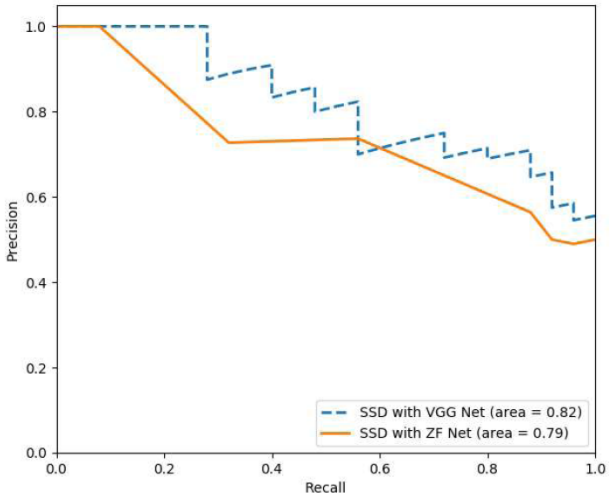

Figure 9: PR curve of detection in Traffic Signs

The area under the curve for VGG Network is 0.82 while for ZF Network is 0.79. The higher the precision recall area under curve is, the better the model is. Therefore, VGG net showed the better performance.

Some Nepali traffic signs were also collected and tested on designed system. Outputs from system on some of

Table 4: Experimental results of the different detection methods

Method	MSERs	Windows	SSD VGGNet	SSD ZFNet
Total number of signs	296	296	361	361
Detections of traffic signs correctly	260	283	355	351
Number of False Alarm	231	288	71	76
Number of misses	36	13	6	10
Precision	52.95%	49.56%	83.33%	81.06%
Recall	87.84%	95.61%	98.34%	97.23%
F-measure	65.07%	65.28%	90.21%	88.41%

Figure 10: Localization and Recognition of some Nepali Traffic Signs

Table 5: IoU vs Recall

S.N.	IoU Threshold	Recall (SSD VGGNet)	Recall (SSD ZFNet)
1	0.5	0.9834	0.9723
2	0.6	0.8637	0.866
3	0.7	0.6802	0.659

Table 6: Timing of the different detection methods

S.N.	Method	Timing
1	Windows	1 fps
2	MSERs	6 fps
3	SVM + HOG +CNN	6 fps
3	SSD (VGG Net)	10 fps
4	SSD (ZF Net)	15 fps

those sample images having traffic sign in Nepal are shown in figure 10.

5. Conclusion

In this work, a novel approach for traffic sign localization and recognition was designed by using SSD. VGG Net architecture and ZF Net architecture for CNN were modified to implement SSD technique. The system was trained and evaluated using GTSDB. Performance of designed system were compared with detection algorithms MSERs, sliding window classifier and SVM with HOG and CNN method for GTSDB. The mAP for MSERs was 0.52 and for Windows was 0.49. The timing of system using Windows was 1 fps, with MSERs was 6 fps and with SVM + HOG + CNN method was 6 fps. Whereas, the designed system yield mAP of 0.83 and 10 fps for VGG Net while 0.81 mAP and timing of 15 fps for ZF net. The results showed that designed system outperformed state-of-art methods. On the basis of accuracy VGG Net performed well compared to ZF net (though they did not yield much different performance) and based on timing, system designed with ZF net is faster compared to system with VGG net. Among 3 different values of IOU threshold

(0.5, 0.6, 0.7) used in experiment, IOU >0.5 yielded the better performance. Images with Nepali traffic signs were also collected and tested on designed system. System also performed well on Nepali test images.

Acknowledgments

The authors are thankful to Department of Eletronics and Computer Engineering, Pulchowk Campus, IOE, TU for all the support and guidance in this research work.

References

[1] Qiong Wang and Xinxin Liu. Traffic sign segmentation in natural scenes based on color and shape features. In *Advanced Research and Technology in Industry Applications (WARTIA), 2014 IEEE Workshop on*, pages 374–377. IEEE, 2014.

[2] Wenju Li, Haifeng Li, Tianzhen Dong, Jianguo Yao, and Lihua Wei. Improved traffic signs detection based on significant color extraction and geometric features. In *Image and Signal Processing (CISP), 2015 8th International Congress on*, pages 616–620. IEEE, 2015.

[3] Ching-Hao Lai and Chia-Chen Yu. An efficient real-time traffic sign recognition system for intelligent vehicles with smart phones. In *Technologies and Applications of Artificial Intelligence (TAAI), 2010 International Conference on*, pages 195–202. IEEE, 2010.

[4] Md Zainal Abedin, Prashengit Dhar, and Kaushik Deb. Traffic sign recognition using surf: speeded up robust feature descriptor and artificial neural network classifier. In *Electrical and Computer Engineering (ICECE), 2016 9th International Conference on*, pages 198–201. IEEE, 2016.

[5] Keren Fu, Irene YH Gu, and Anders Ödblom. Traffic sign recognition using salient region features: A novel learning-based coarse-to-fine scheme. In *Intelligent Vehicles Symposium (IV), 2015 IEEE*, pages 443–448. IEEE, 2015.

[6] Monika Lasota and Marcin Skoczylas. Recognition of multiple traffic signs using keypoints feature detectors. In *Electrical and Power Engineering (EPE), 2016 International Conference and Exposition on*, pages 535–540. IEEE, 2016.

[7] Laksono Kurnianggoro, Joko Hariyono, Kang-Hyun Jo, et al. Traffic sign recognition system for autonomous vehicle using cascade svm classifier. In *Industrial Electronics Society, IECON 2014-40th Annual Conference of the IEEE*, pages 4081–4086. IEEE, 2014.

[8] Dan CireşAn, Ueli Meier, Jonathan Masci, and Jürgen Schmidhuber. Multi-column deep neural network for traffic sign classification. *Neural Networks*, 32:333–338, 2012.

[9] Mian Mian Lau, King Hann Lim, and Alpha Agape Gopalai. Malaysia traffic sign recognition with convolutional neural network. In *Digital Signal Processing (DSP), 2015 IEEE International Conference on*, pages 1006–1010. IEEE, 2015.

[10] Yi Yang, Hengliang Luo, Huarong Xu, and Fuchao Wu. Towards real-time traffic sign detection and classification. *IEEE Transactions on Intelligent Transportation Systems*, 17(7):2022–2031, 2016.

[11] Seokwoo Jung, Unghui Lee, Jiwon Jung, and David Hyunchul Shim. Real-time traffic sign recognition system with deep convolutional neural network. In *Ubiquitous Robots and Ambient Intelligence (URAI), 2016 13th International Conference on*, pages 31–34. IEEE, 2016.

[12] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[14] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[15] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pages 391–405. Springer, 2014.

[16] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[19] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.