On-road object detection using Deep Neural Network

Huieun Kim, Youngwan Lee, Byeounghak Yim, Eunsoo Park, Hakil Kim
Department of Information and communication engineering, INHA University
Computer Vision Laboratory
Incheon, South Korea
{hekim, yw.lee, bhyim516, espark}@inha.edu, hikim@inha.ac.kr

***Abstract*— Industrialization of transportation system has derived serious accidents that resulted in thousands of deaths. To solve the problem, vision based object detection for autonomous vehicle and advanced driver assistance system has been researched. In this study, we provide experimentations of object detection and localization in on-road environment using deep neural network. We compared the detection accuracy among object classes and analyzed the recognition results with fine-tuned Single shot multibox detector on KITTI dataset. This work improves the performance of original detection model by increasing precision of overall detection about 6%, especially about 10% in pedestrian and cyclist.**

Keywords—deep neural network; object detection; localization; advanced driver assistance system;

I. Introduction

Since Industrialization of transportation, demands for vehicle have been increased. As the number of vehicle grows annually, proportional accidents involving fatal crashes occurs. According to [1], the number of road traffic accidents that plateaued since 2007 have killed 1.25 million people on the road. Vulnerable road users such as Pedestrians, cyclists and motorcyclists make up half of these fatalities. To solve the traffic problems, global organizations research driver assistance systems with practical sensors that can provide convenience to drivers. Global companies like Volvo and Honda proposed systems for advanced driver assistance systems(ADAS) while Google invented autonomous vehicle.

ADAS and autonomous vehicle are the most representative technology that can benefit drivers by providing driving environment information or autonomous drive by extension. ADAS such as forward collision warning system(FCWS), lane departure warning system and night vision system recognize the objects on the road in various conditions and report the surrounding situation to drivers. On the other hand, Autonomous vehicle integrates the information from each system and control the driving by itself.

Both of ADAS and autonomous vehicle depend on vision based detection. Cameras, light detection and ranging(LIDAR) are used for visualization of surroundings [2]. Among the vision sensors, cameras feature lower cost in both price and computation compared to other active sensors. Recognizing on-road objects with camera in real-time has been important issue between object detectors.

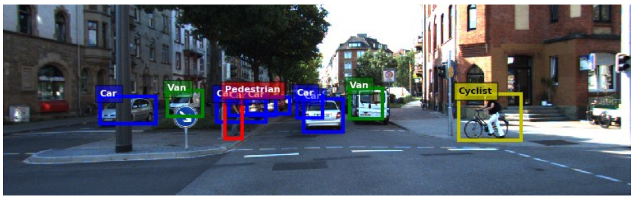


Fig. 1. End-to-end object detection result of SSD

Therefore, this paper suggests an on-road object detection using SSD [3] that overcomes the problems mentioned above and makes possible robust object detection in real-time. SSD is one of the detection mechanisms based on deep neural network. From input images, it generates appearance features using convolutional layers. In addition, it trains object location in 2D image coordinate by computing loss of object box position in training step. As shown in Fig.1, SSD provides detection results for various classes.

However, SSD has drawback of missing small objects with its grid approach. Therefore, this paper suggests to adjust SSD on on-road object detection and fine-tune the model to be robust on localizing small objects such as pedestrian. Then, we compare the accuracy of fine-tuned SSD and the original model as well as the state of the art, YOLO.

The remainder of this paper is organized as follows. Section 2 mentions the related works that covers the trend of object detection. Then, section 3 introduces the fine-tuned SSD which is the object detection method used in this work. After that, section 4 describes the object experiment results on KITTI dataset [4]. Finally, in section 5, we analyze the experiment results and draw conclusion about limitations and future works of the work.

II. Related Work

On road object detection methods have been studied last for several years. Especially the researches about detecting active objects such as pedestrian and vehicles has been issued actively in computer vision. For last several years, object detection mechanism has been evolved by applying Convolutional Neural Network in more practical way while improving the performance.

978-1-5090-2743-9/16/$31.00 ©2016 IEEE

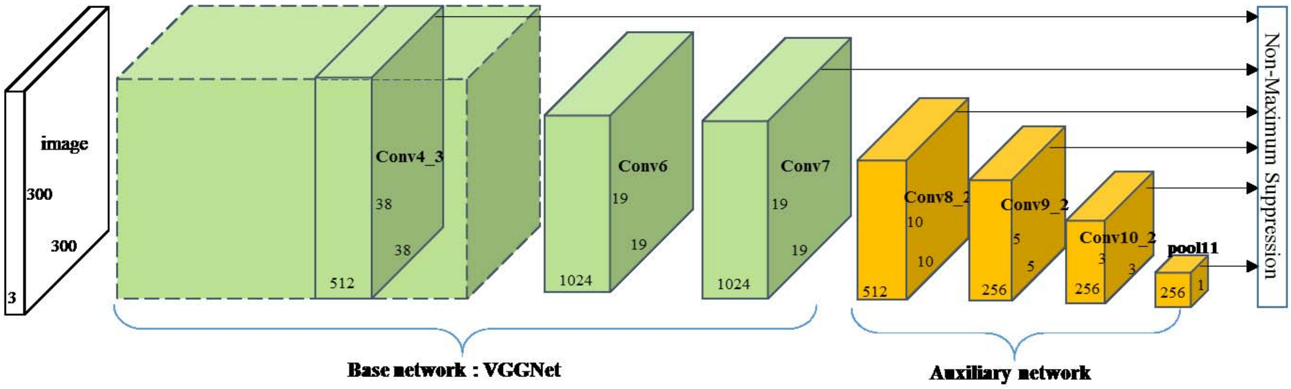


Fig. 2. SSD network architecture: Base network(VGGNet) + Auxiliary network.

Previous object recognition methods applied region proposal methods based on sliding window to localize object [5]. Such method can be useful for searching object in every location in the image. However, searching the unnecessary regions where objects never appear is unavoidable and it causes additional computational load with false positive detections.

In Overfeat [6], sliding window fashion is extremely time consuming so there was an approach generating all possible object regions first and then classifying on each region [7]. In that manner, reducing the computation on proposal network has been issued for real-time object detector.

Unlike other end-to-end detectors like Faster R-CNN [7] with additional region proposal mechanism, SSD makes the regions using grid structure. With this approach, SSD is faster than Fast R-CNN by 41 FPS (Frames Per Second) [3]. Furthermore, it generates object regions in feature maps with various scales, which compensates the disadvantage of YOLO [8] that only derives object regions in only one feature map.

In this paper, we use the model based on SSD and fine-tune the model with KITTI dataset which is composed of on road environment object classes. We experiment various fine-tuned SSD models and analyze the result to find the most effective way to improve the detecting performance in road environment.

III. ON ROAD OBJECT DETECTOR - SSD

A. Network structure

SSD is based on end-to-end convolutional network which predicts fixed-sized bounding boxes and scores for the presence of objects in those default boxes. Fig.2 shows the SSD network architecture that consists of the base network and an auxiliary network. The base network is for good quality classification using VGGNet [9] which is well known for image classification and the auxiliary network is used for detection steps. By adding convolutional feature layers to the end of the base network and decreasing in size for multiple scales, the auxiliary network can predict detections at multiple feature maps. A set of convolutional filters can be adopted for producing either a class score for a category and a bounding box offset relative to the default box coordinate in those added feature layers as illustrated in Fig.2, followed by a non-maximum suppression which decide the final detections.

B. Training

The overall loss term is defined by a weighted sum of the localization loss and the confidence loss:

$$\mathrm{L(x,c,l,g)} = \frac{1}{N}\Big(L_{conf}(\mathrm{x,c}) + \alpha L_{loc}(\mathrm{x,l,g})\Big) \qquad (1)$$

where N is the number of matched default boxes, and the localization loss is based on the smooth L1 loss [7] between the predicted box (l) which is matched to any ground truth(g) with jaccard overlap higher than a 0.5 and the ground truth box g. Then the center of the bounding box, its width, and height was regressed. The confidence loss is based on the softmax loss for multiple classes confidences(c).

C. Fine-tuning

The SSD benchmark model is fine-tuned from pre-trained model on Pascal VOC images. Since the detection rates of small objects such as pedestrian and cyclist are relatively low compared to vehicle, this paper solved the problem with data compensation and augmentation. We added more pedestrian dataset and fine-tuned the original model to make the program robust on pedestrian detection. As a result, accuracy in both pedestrian and cyclist has increased for almost 10%.

Then, we applied additional aspect ratio of the default box that is used for object localization. SSD samples random patches from the input image and then resize them into particular size. In this process, height ratio of resized ground truth of objects becomes much bigger than the ratio of width in the object like

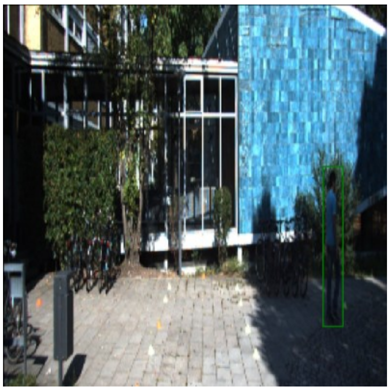


Fig. 3. Patch and ground truth from input frame.

Fig.3. Then, default boxes in training process can't cover the resized ground truth area and derive into low detection score in objects like pedestrian or cyclist. By adding additional aspect ratio that is longer in height, we could compensate the detection accuracy of pedestrian.

Additional data augmentation strategies applied in this work are like below. randomly sampling a patch which is overlapped with the object by 0.1, 0.3, 0.5, 0.7, or 0.9 is used and the patch size is from 0.1 to 1 of the original image, followed by resized to fixed size and horizontally flipped. This study uses VGGNet16 for base network pre-trained on the ILSVRC CLS-LOC dataset [10] and fine-tunes the model using SGD with 0.9 momentum and initial learning rate 10^{-3}, 0.0005 weight decay, and batch size 32.

IV. Experiment results

A. Datset

This study evaluate the proposed method on the KITTI dataset [4] that consists of 9 classes such as car, van, truck, pedestrian, person sitting, cyclist, tram and misc. The model is trained on the 4987 training images and evaluated on the 2494 validation images. In this paper, we compared the detection accuracy on three classes - car, pedestrian and cyclist – that are most relevant on the road.

B. Analysis

1) Accuracy

In this work, we compared the evaluation result of object detection based on proposed fine-tuned SSDs with original SSD model and YOLO, the state of the art algorithm. There are three fine-tuned model, which are SSD-ASP4, SSD-PED3, SSD-ASP4-PED3. SSD-ASP4 and SSD-PED3 are SSD model with VGG base network. SSD-ASP4 is trained with additional aspect ratio for default box and SSD-PED3 is a fine-tuned SSD model with compensated pedestrian dataset. Our proposed model is VGG-ASP4-PED3, an integrated model of SSD-ASP4 and SSD-PED3.

Fig.4 shows the recall of the proposed models, benchmark model and YOLO. In Fig.4, recall rate is evaluated based on IoU (Intersection of Union) overlap threshold. YOLO has relatively lower accuracy than other SSD models. On the other hand, SSD related methods show higher recall rates in Fig.4. In addition, table 1 reports the average precision of three models.

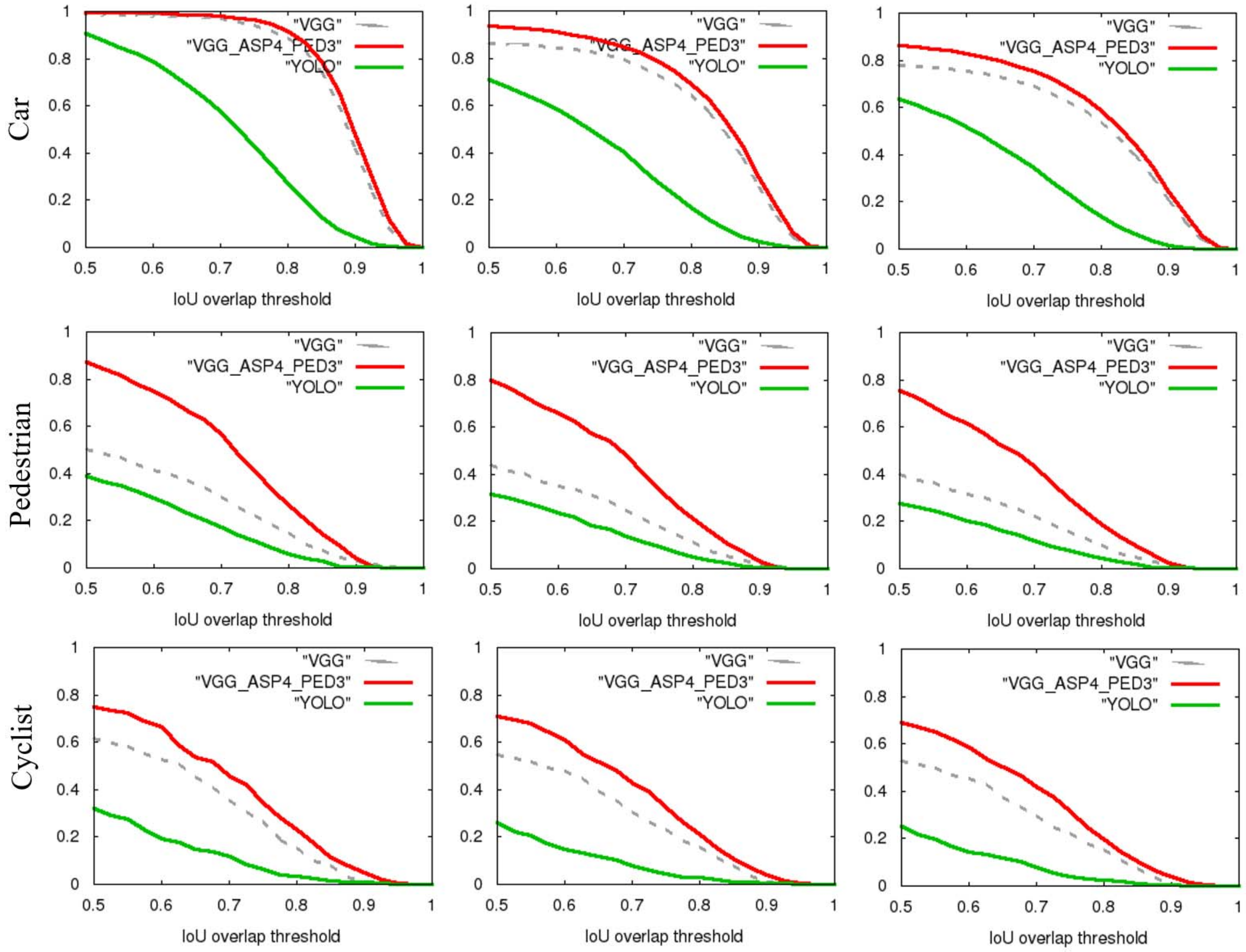
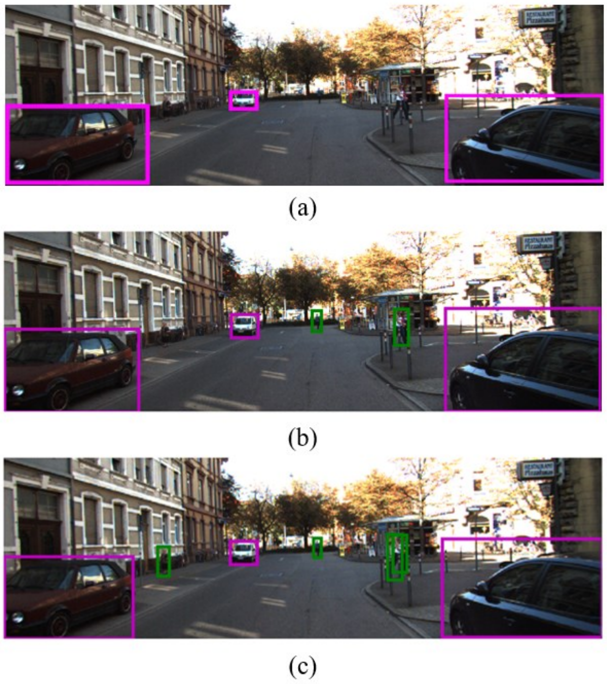


Fig. 4. Recall vs. IoU overlap threshold



(a)

(b)

(c)

Fig. 5. Object detection and classification result of (a) YOLO, (b) benchmark SSD and (c) proposed fine-tuned SSD. Box colored magenta denotes car while the green box indicates pedestrian.

Among the three methods, SSD with data compensation and additional aspect ratio of default boxes has almost 10% higher average precision in pedestrian and cyclist detection.

Fig.5 shows the detection and classification results. The proposed method achieved nice performance in object detection. especially it localized pedestrian and cyclist better than the original SSD model and the state of the art algorithm. However, missing pedestrian and cyclist is still remained in some frames.

2) *Processing time*

In terms of practical issues and commercialization, the processing time is very critical factor for ADAS. Thus, this study tested detection on a PC equipped with NVIDIA Geforce GTX TITAN X GPU which is a state-of-the-art GPU for deep learning. SSD model achieved 29.4 FPS, which is very impressive performance and verify the feasibility for real-time system.

V. Conclusion

In this study, we have provided experimentation of object detection and localization in driving environment using deep neural network. We fine-tuned SSD model on KITTI dataset and analyzed the recognition results. From the experiment, we confirmed that fine-tuning SSD on road dataset using data augmentation can improve the detection result. To adjust end-to-end detector more practically in vehicle industry, research on deep compression for reducing memory of the model has to be accompanied. While vision-based object detection has matured over decades, more practical understanding of the on-road environment will remain an active area of research.

VI. Acknowledgement

This work was supported by the Industrial Technology Innovation Program "10052982, Development of multi-angle front camera system for intersection AEB" funded by the Ministry of Trade, industry & Energy (MI, Korea).

References

[1] T. Toroyan, M. M. Peden, and K. Iaych, "Global status on road report 2015," *World Heal. Organ.*, vol. 19, no. 2, p. 150, 2013.

[2] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S. Kammel, J. Z. Kolter, D. Langer, O. Pink, V. Pratt, M. Sokolsky, G. Stanek, D. Stavens, A. Teichman, M. Werling, and S. Thrun, "Towards fully autonomous driving: Systems and algorithms," *IEEE Intell. Veh. Symp. Proc.*, pp. 163–168, 2011.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "SSD: Single Shot MultiBox Detector," *Arxiv*, pp. 1–15, 2015.

[4] a Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Rob. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[5] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[6] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks," *arXiv Prepr. arXiv*, p. 1312.6229, 2013.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Nips*, pp. 1–10, 2015.

[8] D. Impiombato, S. Giarrusso, T. Mineo, O. Catalano, C. Gargano, G. La Rosa, F. Russo, G. Sottile, S. Billotta, G. Bonanno, S. Garozzo, A. Grillo, D. Marano, and G. Romeo, "You Only Look Once: Unified, Real-Time Object Detection," *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 794, pp. 185–192, 2015.

[9] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *ImageNet Chall.*, pp. 1–10, 2014.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

TABLE I. Average Preicsion of object detectors

KITTI	**Car**			**Pedestrian**			**Cyclist**		
	Easy	***Moderate***	***Hard***	***Easy***	***Moderate***	***Hard***	***Easy***	***Moderate***	***Hard***
YOLO	26.5	23.1	19.8	35.5	32.3	28.3	25.8	25.2	24.3
SSD	85.6	74.9	67.2	53.5	50.1	48.1	46.6	52.5	51.6
SSD-ASP4 [Ours]	85.4	75.3	67.6	53.6	49.5	47.8	50.8	55.8	54.6
SSD-PED3 [Ours]	86.4	75.4	67.5	52.5	51.4	50.0	51.4	55.7	54.3
SSD-ASP4+PED3 [Ours]	**86.7**	**76.0**	**70.7**	**61.3**	**60.3**	**58.2**	**58.8**	**60.5**	**59.2**