Journal of KIIT. Vol. 16, No. 10, pp. 93-100, Oct. 31, 2018. pISSN 1598-8619, eISSN 2093-7571 93
http://dx.doi.org/10.14801/jkiit.2018.16.10.93

Multi Scale Object Detection Based on Single Shot Multibox Detector with Feature Fusion and Inception Network

Md Foysal Haque*, Dae-Seong Kang**

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (NO.2017R1D1A1B04030870).

Abstract

Inception layers are most effective and constructive architecture in current object detection field. Moreover, beside Inception layers, feature fusion layers also shows significant performance for object detection. Single Shot Multibox Detector(SSD) is one of the most convenient and fastest algorithms in the current object detection field. SSD uses only a single convolutional neural network to detect the object from the input. Moreover, SSD algorithm shows high accuracy to detect multi-scale objects but detecting small objects SSD network performance is not satisfactory. In this paper, we added feature fusion layers and Inception layers with SSD to increase its performance. In this paper, our main focus to improve the features of different layers by feature fusion and also minimize the calculation cost by Inception. Moreover, here in SSD, we add extra layers to improve its performance without any affecting on its speed. By adding Inception layers and feature fusion layers, SSD network can catch more information without increasing the complexity of the network and it also able to detect more small objects. The new proposed algorithm can detect small and big objects with satisfactory accuracy.

요 약

현재 객체 탐지 분야에서 가장 효과적이고 실용적인 구조는 Inception layer이다. 또한 Inception layer 외에도 Feature fusion layer는 객체 탐지에서 중요한 성과를 보여준다. SSD(Single Shot Multibox Detector)는 현재 객체 탐지 분야에서 가장 유용하고 빠른 알고리즘 중 하나이다. SSD는 하나의 Convolutional neural network 만을 사용하여 입력 영상으로부터 객체를 탐지한다. 또한 SSD는 다중 스케일 객체들의 탐지에 높은 정확도를 나타낸다. 그러나 작은 객체들의 탐지에서 SSD 네트워크의 성능은 만족스럽지 않다. 본 논문에서는 SSD의 성능 향상을 위해 SSD와 함께 Feature fusion layer와 Inception layer를 추가한다. 본 논문에서의 주요 초점은 Feature fusion에 의해 다른 계층들의 성능을 향상시키고 Inception layer에 의한 연산 비용을 최소화하는 것이다. 또한 이 SSD에서는 속도에 영향을 주지 않고 성능을 향상시키기 위해 여분의 layer를 추가한다. Inception layer와 Feature fusion layer를 SSD에 추가하면 네트워크의 복잡성을 증가시키지 않고도 더 많은 정보를 얻을 수 있고 더 작은 객체도 감지할 수 있다. 제안하는 알고리즘은 작거나 큰 객체들을 더 만족스러운 정확도로 탐지할 수 있다.

Keywords
convolutional neural network, object detection, SSD, inception, feature fusion

* Dept. of Electronic Engineering, Dong-A University
- ORCID: https://orcid.org/0000-0003-0634-9783
** Professor, Dept. of Electronic Engineering, Dong-A University
- ORCID: https://orcid.org/0000-0003-0186-2430

· Received: Aug. 13, 2018, Revised: Sep. 04, 2018, Accepted: Sep. 07, 2018
· Corresponding Author: Dae-Seong Kang
Dept. of Dong-A University, 37 NaKdong-Daero 550, beon-gil saha-gu, Busan, Korea,
Tel.: +82-51-200-7710, Email: dskang@dau.ac.kr

Ⅰ. Introduction

Object detection is making its position strong in current computer vision field and day by day it getting more popular. To classify, recognize and detect the object lot of algorithms are developed and day by day their speed and accuracy is improving. Object detection systems use several approaches to detect objects such as: hypothesize bounding boxes, resample pixels or features for each of box, and also apply a high-quality classifier. This pipeline has persuaded on detection benchmarks since the most selective search work[1] through the current major results on PASCAL VOC, and ILSVRC detection mostly all are based on R-CNN[2] and Faster R-CNN[3]. Moreover, a lot of improved methods based on R-CNN[2], such as fast R-CNN[4], Faster R-CNN[3] and R-FCN[5]. These networks emerged high accuracy in the object detection area, but their network structures are relatively complex. Object detection and classification methods used mostly now on the facial recognition system[6], automatic car system and smart shopping system[7]. R-CNN[2] algorithm detection speed is not satisfied, to detect objects faster, the improved algorithm build and this algorithm is Faster R-CNN[3]. It detects objects by using region proposal method. This network structure relatively very complex and another complexity of this network is training time. It takes a long time for training. To make an easy way to detect object WeiLiu et al proposed the Single Shot Multi-Box Detector (SSD)[8] based on the idea of the Faster R-CNN anchor method[3]. It is mainly based on the traditional classification networks VGG[9] and introduced extra layers as feature extraction layers, these layers are able to detect multi-scale objects. However, to detect small objects, its performance is not very satisfying. The most challenging task on object detection is to maintain feature maps size and computational cost. To solve this problem many methods are developed but their accuracy is not satisfactory. However, to scale feature maps is a challenging task in object detection. Feature pyramid detection methode[10] introduced fuse the feature make scale that comes from different layers with different sizes. It also concatenated to scale feature maps comes from different layers by down-sampling and gather all as new feature pyramid. To solve this problem feature fusion and Inception is added in SSD. Inception inspired from the GoogLeNets[11]. Inspection block and the dipper residual network, which makes it more accurate at positioning. SSD have training problems, such as gradient disappearance and over-fitting. However, considering the tradeoff between performance and speed, here introduced the feature fusion and Inception structure in extra layers after VGG16[9] by increasing the type of convolution kernels. As a consequence, the scope of the receptive field is expanded, which increase the sensitivity of the model to small objects without losing large objects. Fig. 1 shows the structure of Inception network.

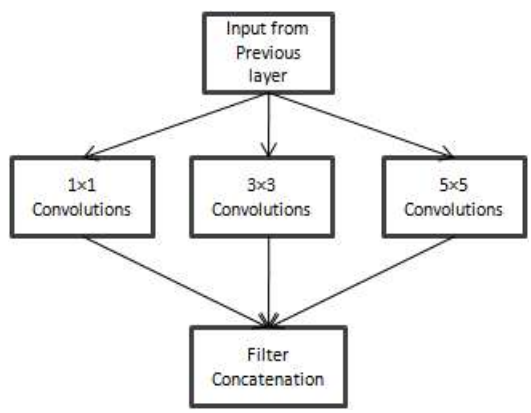


Fig. 1. Architecture of inception network

Ⅱ. Related Algorithm

In the field of deep neural network for object detections have a lot of methods. The deep learning methods for object detection can be divided into two kinds, including the region-proposal based methods and the regression-based methods. The first kind includes R-CNN and faster R-CNN which generate object boxes in the initial stage, and next use deep neural

networks for classification and also point regression on the next step.

2.1 SSD Network

SSD networks only use a single network to generate bounding boxes and simultaneously classification by regression-based method. SSD can achieve real-time processing on GPU. Detecting objects with SSD network, it divided the input image into grids and each mesh predicts the confidence and the point of two object boxes. However, at the top of network SSD work along with a lot of default boxes with feature maps, using this feature maps it can detect and identify different objects from various scale and aspect ratios. Fig. 2 shows the architecture of SSD network.

By adding several extra convolutional layers, SSD achieves multi-scale effects. The extra additional convolution layers feature maps plays a major role to produce bounding boxes by influencing making its confidence strong to aim objects.

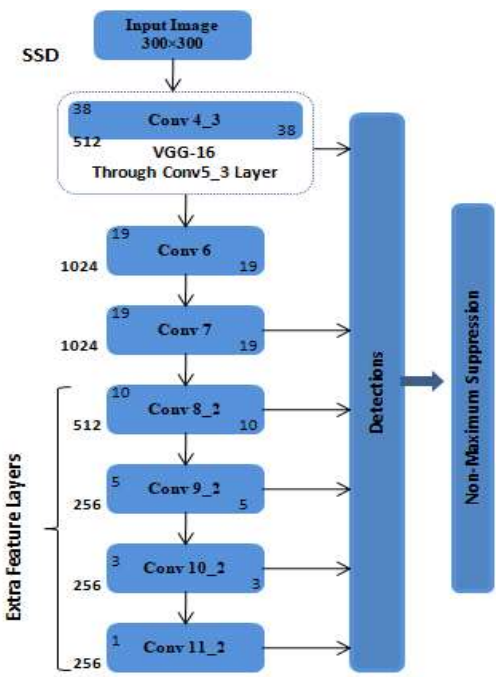


Fig. 2. Architecture of SSD network

The size of added extra layers are comparatively very small and they can carry small scale of information on small objects. Moreover, when SSD network moves to classify any objects deeply the information of small objects gets less. That is why its performance not satisfactory when input carries a lot of objects. Its performance not satisfactory when input have lot of objects. Moreover, it cannot detect small objects. However adding Inception layers in SSD, it can detect small objects.

2.2 Inception Network

Inception layers[11] are generally convolutional layers, that we chose as on our system requirement Inception layer are have different size like 1×1, 3×3 and 5×5. These convolution layers use like as convolution filter Inception layer are used on the network to improve the system accuracy without affecting its speed and other characteristics. Inception layers implement in a specific portion of a system. It take input from system and classifies the data as required and after finishing all process it passes output to the next step for further process. It uses bottleneck method to minimize the computation cost that is most important characteristics of Inception network.

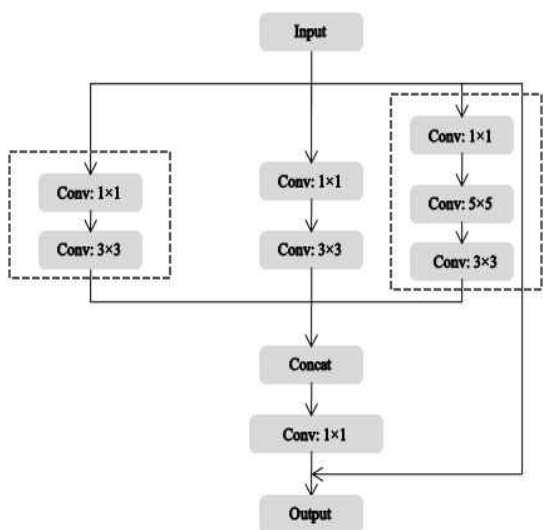


Fig. 3. Architecture of improved inception network

Fig. 3 shows the improved Inception architecture. From Figure we can see at left side we add 3×3 convolutional network and the right side one 3×3 convolutional layer replaced by 5×5 convolutional layer.

2.3 Cost Calculation of Inception Network

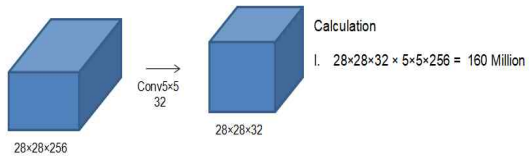
Fig. 4 shows the computation cost of a normal convolutional layer and Inception layer. Here for calculation, we take one 5×5 convolution layer and another one is 1×1 convolution layer (Inception layer).

From Fig. 4 we can see 1×1 convolution layer is most cost-effective than 5×5 convolution layer because 1×1 convolutional layer use bottleneck theory to minimize the computation cost and help to save the spaces of our system. The output of 1×1 convolution layer pass through another 3×3 convolutional filter, this filter again classifies data and collect more specific reliable information for further use. Moreover, using Inception layer, it helps to classify input very deeply and as a result it able to help current system to detect more small object from the input. Inception layers are mainly focused on searching the position of optimization for increasing the performance of the system.

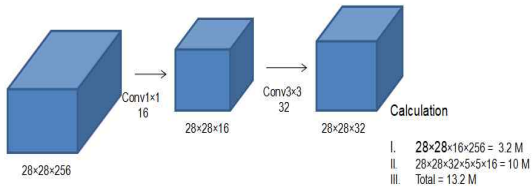
2.4 Feature Fusion

Fig. 5 shows the architecture of feature fusion network. Feature fusion set of convolutional layers and batch normalizer. In object detection system every layer have feature maps, these feature maps are different resolution and size. Moreover, with this feature maps network not able to confidence about detecting the actual object location. Feature fusion layers process selected layers feature maps to scale them as same the size that we required.

To resize all feature same size, feature fusion network merges the feature maps by up-sampling and down-sampling tasks. After completing scaling of all feature, all feature maps sum into a batch normalizer and pass through to the next step for detection. In order to scale all feature maps same size it need to up-sampling and down-sampling[12]. Fig. 6 shows the process of feature maps resizing at the feature fusion network.



(a) Normal convolution layer



(b) Inception layer

Fig. 4. Comparison of computation cost

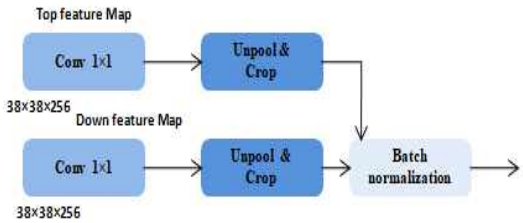


Fig. 5. Architecture of feature fusion

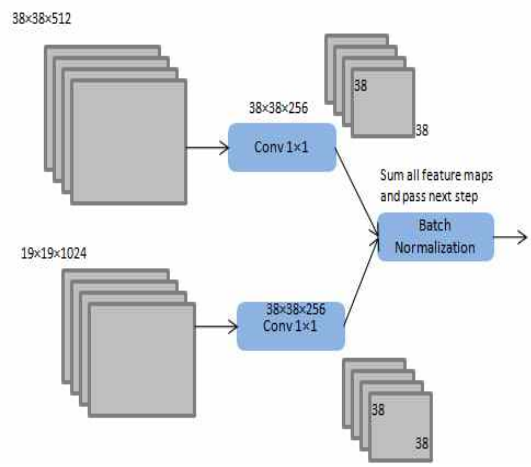


Fig. 6. Process of feature maps resizing at feature fusion network

III. Proposed Algorithm

In this section, we describe SSD network with feature fusion and Inception layers. Inception and feature fusion layers are not familiar for object detection task. In this experiment, we found Inception is most cost-effective method for detecting both small and large objects form input image and feature fusion layers are very effective to scale feature maps. Moreover, these two features make SSD network more strong to detect small and big objects. To compare with traditional methods deep neural network have more complex model structure. This complex model can train a large number of data to get comparatively better performance. The performance of deep neural network can be improved by some structural modification like increasing their depth and width. Moreover, for this reason, the parameter increases, which may conduct to over-fitting and increase the computation of the network. The use of dense components in Inception architecture to conduct approximate the optimal local infrequent structure[13]. The Inception structure exploits the high performance of the dense matrix, although it keeps the sparsity of the network. Inception blocks can acquire a lot of information without increasing its network complexity.

Fig. 7 shown the proposed architecture of SSD network with feature fusion and Inception layers. In this paper, we used SSD network. SSD network mainly based on VGG16. In VGG network add some extra layers that become SSD network. Fig. 2 shows the architecture of SSD network. In our proposed method, we added some extra feature layer in SSD network. To detecting the object here use conv4_3, conv6, conv7, conv8_2, conv9_2, and conv10_2 layer. The top of the network we add feature fusion layers and the last of the network we add Inception layers. Feature fusion layers are generally 1×1 convolution layer with batch normalizer. These layers are shows very efficient result to scale feature maps. However, with conv4_3 we add first Feature fusion layer and the second one with conv6. Both are use for down-sample the feature maps. The output of these two layers passes through to a batch normalizer to sum up all feature maps and finally it passes to conv7. After finishing scaling all feature maps, the next task is to minimize the computation cost. To minimize the computation cost we use here Inception network. Inception basically one type of convolution kernel, like 1×1, 3×3 and 5×5 kernel. The feature maps of the extra layers produced the objects location offset and confidence by small convolution operation. In SSD network, conv9_2 and conv10_2 are replaced with Inception layers.

These layers are really useful for reducing the computation cost and also help to retrain more object to the network. Inception block makes the features layers as same as original layers. Moreover, each of the convolution layers is replaced by three Inception convolutional layer.

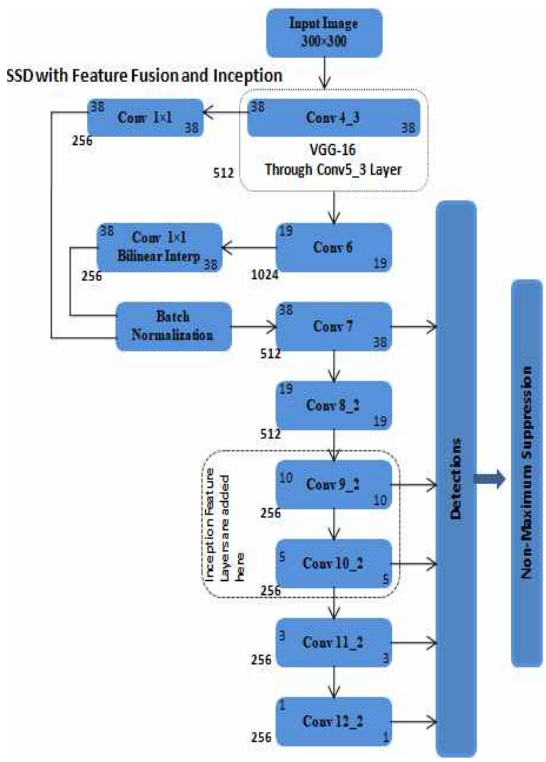


Fig. 7. Architecture of SSD network with feature fusion and inception layers

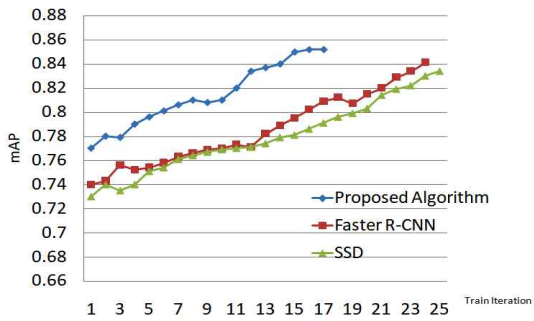
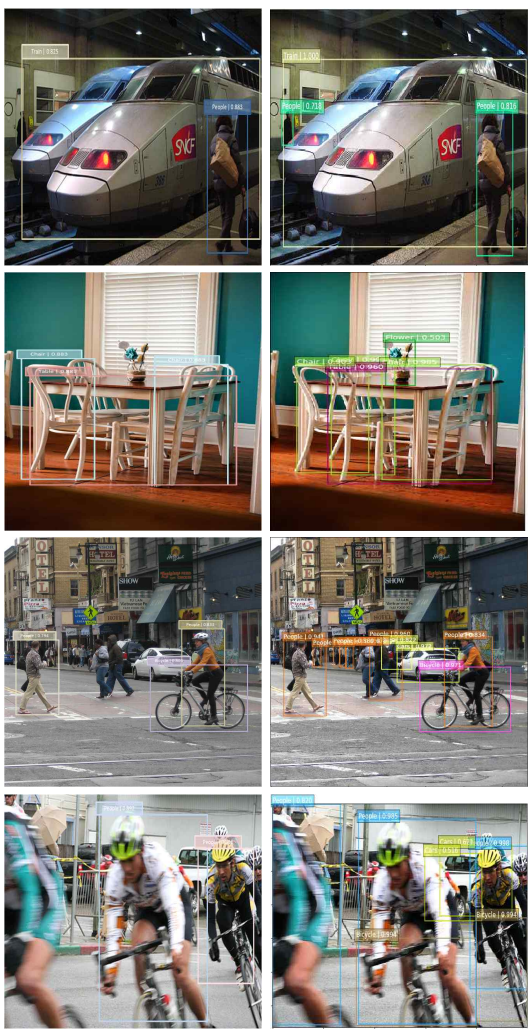


Fig. 8. mAP comparison of proposed algorithm, faster R-CNN and conventional SSD network



(a) (b)

Fig. 9. Detection results of proposed SSD and conventional SSD, (a) SSD detection results, (b) Proposed SSD detection results

Ⅳ. Experiments

In our experiment, we used VGG16 network for recognition and regression. SSD use a single network with the input size 300×300. We use PASCAL VOC2007 and VOC2012 dataset for object detection. We use SGD with the initial learning rate set to 0.001, momentum set to 0.9, weight decay set to 0.0005, batch size 32 and input resolution as 300×300. In this experiment, we make two network SSD & proposed new SSD to compare their output results. This experiment, we use PASCAL VOC, it contains 16,551 trainable images and 4,952 test images over 20 different categories. To train improved SSD model we use the same strategy as used to train SSD model.

Moreover, this experiment at frist we train the model by learning rate 10^{-4} for 90k iterations, then continue the training by learning rate by 10^{-5} for 40k iterations.

Ⅴ. Results

The performance of SSD with feature fusion and Inception architecture is shown in Fig. 9. From the figure, we can see the proposed method can detect several types of object with high quality. Proposed algorithm achieved high accuracy to detects small objects and also minimize the computational cost that is most significant requirements for object detection task. Inception modules minimize the computational cost and feature fusion helps to confidence about the target localization of objects by scaling feature maps. As a result, the proposed algorithm able to detect small objects than SSD network.

To compare with SDD network, SSD has few localization errors, specify the proposed method can detect objects better whereas it learns directly to regress the object's shape and initially classify the objects categories it uses two different steps. However, in SSD network it is more confused to detect similar object categories. Moreover, it has another major lack

it cannot detect small objects. When we add feature fusion and Inception layer it can detect small objects beside big objects.

Table 1. Detection results of PASCAL VOC test

Method / Class	SSD with Feature Fusion and Inception Network	Fast R-CNN [3]	SSD [6]
	mAP		
	85.2	84.1	83.4
Car	91.2	88.1	83.4
People	94.3	92.1	87.6
Cat	94.2	93.9	93.4
Bike	90.6	89.2	87.5
Bicycle	91.4	90.6	89.2
Train	92.9	92.4	88.3
Table	80.7	80.1	78.9
Horse	92.1	91.8	88.9
Cow	90.4	89.7	83.6

VI. Conclusions

In this paper, we proposed an improved Single Shot Multibox Detector with feature fusion and Inception layers, this model shows conspicuous accuracy to detect multi-scale objects. To compare between conventional SSD and new proposed improved SSD, conventional SSD performance is not satisfactory to detect small objects and multi-scale. However, in proposed method feature fusion layers helps SSD network to make more confidence about the target by merging all feature maps and Inception layers help SSD network to go deeper to retrain the network and is able to detect small objects. In the future, we will continue our research to improve the model to increase accuracy for detecting small objects.

References

[1] J. R. R. Uijlings., K. E .A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition", International Journal of Computer Vision, Vol. 104, No. 2, pp. 154-171, Sep. 2013.

[2] R. Girshick, J. Donahue, T. Darrel, and J. Malik, "Region-Based Convolutional Network for Accurate Object Detection and Segmentation", IEEE Trans. Pattern Anal. Mach. Intell., Vol. 38, No. 1, pp. 142-158, May 2016.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, No. 6, pp. 1137-1149, Jun. 2017.

[4] R. Girshick, "Fast R-CNN", Internation Conference on Computer Vision 2015, pp. 1440-1448. Apr. 2015.

[5] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks", Advances in Neural Information Processing Systems 29 (NIPS 2016), pp. 379-387, May 2016.

[6] J. W. Kim and P. K. Rhee, "High Efficiency Adaptive Facial Expression Recognition base on Incremental Active Semi-Supervised Learning", JIIBC, Vol. 17, No. 2, pp. 165-171. Apr. 2017.

[7] J. K. Sung, S. M. Park, S. Y. Sin, Y. B. Kim, and Y. G. Kim, "Deep Learning-based Product Image Classification System and Its Usability Evaluation for The O2O Shopping Mall platform", JIIBC, Vol. 17, No. 3, pp. 227-234. Jun. 2017.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector", European Conference on Computer Vision, pp. 21-37, Dec. 2015.

[9] K. Simonayan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", International Conference on Learning Representations 2015, pp. 1-14, Apr. 2015.

[10] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid

Networks for Object Detection", Computer Vision and Pattern Recognition, pp. 936-944, Apr. 2017.

[11] C. Szegedy, Wei Liu, yangqing Jia, and P. Sermanet, "Going deeper with convolutions", in IEEE Conference on Computer Vision and Pattern Recognitions on Learning, pp. 1-9, Sep. 2015.

[12] B. Hao and D. S. Kang, "Research on Image Semantic Segmentation Based on FCN-VGG and Pyramid Pooling Module", Journal of KIIT, Vol. 16, No. 7, pp. 1-8, Jul. 2018.

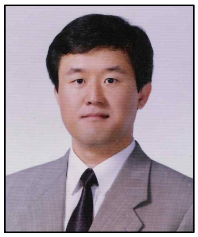
[13] K. H. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park, "PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection", CoRR abs/1608.08021, pp. 1-7, Sep. 2016.

Authors

Md Foysal Haque

2015 : B.Sc. in Electrical & Electronic Engineering, University of Information Technology & Sciences(UITS).
2018 ~ 2020 : M.Sc. in Electronic Engineering, Dong-A University.
Research Interests: Digital Image Processing and Pattern Recognition

Dae-Seong Kang



1994 : Ph.D. in Electrical Engineering, Texas A&M University.
1995 ~ Present : Professor at Dong-A University.
Research Interests: Image Processing and Compression