Journal of KIIT. Vol. 19, No. 2, pp. 29-37, Feb. 28, 2021. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2021.19.2.29

PPCNN: Object Detection using Fine-grained Feature Extraction and Localization

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

Recently, deep convolutional neural network achieved excellent performance on computer vision. Deep convolutional neural network is renowned for classifying object information precisely, and it also employs simple convolutional architecture to reduce the complexity of the layers. VGGNet(VGG network) designed with the concept of deep convolutional neural network. The VGGNet achieved enormous performance in classifying large-scale images. The network is designed with a stack of small convolution filters that makes the network structure very straightforward, but the network has some localization errors. We proposed unique network architecture, PPCNN (Pyramid Pooling Convolutional Neural Network), to reduce the localization error and extract high-level feature maps. The proposed network is composed of an improved VGGNet and a U-shape feature pyramid network. The network introduced to extract and collect small feature information of the object and detect small objects from the source image. The proposed approach achieved higher accuracy in localization and detection tasks.

요 약

최근, 심층 컨볼루션 신경망은 컴퓨터 비전에서 뛰어난 성능을 달성했다. 심층 컨볼루션 신경망은 객체 정보를 정확하게 분류하는 것으로 유명하고 계층의 복잡성을 줄이기 위해 단순한 컨볼루션 구조를 사용하고 있다. VGGNet은 심층 컨볼루션 신경망의 개념으로 설계되었다. VGGNet은 대규모 이미지 분류에서 엄청난 성능을 달성했고 일련의 작은 컨볼루션 필터로 네트워크 구조를 매우 단순하게 설계되었지만, 약간의 위치 인식 오류가 있다. 본 논문에서는 위치 인식 오류를 줄이고 높은 수준의 특징맵을 추출하기 위해 독특한 네트워크 구조인 피라미드 컨볼루션 신경망(PPCNN)을 제안한다. 제안된 네트워크는 개선된 VGG 네트워크와 U자형 특징 피라미드 네트워크로 구성되어 있으며, 소스 이미지에서 객체의 작은 특징 정보를 추출 및 수집하고 작은 객체를 검출하기 위한 네트워크이다. 제안하는 방식은 위치 인식 및 검출 작업에서 더 높은 정확도를 달성했다.

Keywords
deep learning, object detection, image pyramid pooling convolutional neural netwrok

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

· Received: Jan. 20, 2021, Revised: Feb. 18, 2021, Recepted: Feb. 21, 2021
· 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

Detecting small objects is one of the most complex tasks in computer vision. Two essential tasks carried to eliminate the problem at first improve the detection modules recognition process by collecting exact object location information with distinguishing objects foreground information. After that, predict accurate bounding boxes for multi-scale objects to solve the localization error. These processes are quite tricky for a detector because it has to justify a lot of close and positive information. The detector classifies all nearest close information of all objects to confirm the exact object location with the object class.

Deep neural networks show enormous performance in recognizing and detect objects. SSD [1], VGG [2], ResNet [3], R-CNN [4], Fast R-CNN [5], Faster R-CNN [6], R-FCN [7], and CoS [8], these networks are constructed based on deep convolutional neural network. These networks use lots of different methods to classify and detect objects. These networks are constructed with complex convolutional architecture that increases the layer computation complexity. An improved PPCNN (Pyramid Pooling Convolutional Neural Network) architecture is introduced in this paper to eliminate the layers' complexity and extract high-level feature maps.

The network constructed with an improved VGGNet [9] and u-shape feature pyramid network. The feature pyramid network work as a helper network for the VGGNet. The FPN (Feature Pyramid Network) [10] uses the end-to-end method to extract high-level feature maps. The proposed FPN uses the upsampling process to collect exact object information. The main goal of this paper, improve detection efficiency by using fine-grained feature extraction and localization. An improved architecture proposed that composed of VGG with a feature pyramid network that achieved significant performance to detect small objects.

Ⅱ. Related Work

2.1 Overview of FPN

Feature pyramid method applies multi-scale feature representation that the core basis of various successful computer vision applications. The method examines feature information through different feature layers, and this property boasts deep learning methods to achieve higher accuracy in computer vision tasks. Moreover, the network is composed based on deep convolutional networks. The deep convolutional structures adopt complex convolutional structure for feature extraction and classification task.

However, using Deep Convolutional Networks to feature representations through featuring images imposes a large computation weight. Recent works on human pose estimation, image segmentation, and object detection [11]-[16] introduce feature connections in deep networks that connect internal feature layers in different scales to address this issue. The before-mentioned methods efficiently improve feature representations so that they are semantically strong and contain high-resolution feature information.

Recent works have studied how to improve multi-scale feature presentations; in [17], extend the idea of building more robust feature pyramid representations by applying multiple U-shape modules subsequent to a backbone model. In [18], combines features with different scales and accepts feature at each scale through a global attention operation.

Despite the fact that most architecture designs of pyramid architecture remain lightweight linked to the backbone model. In extension to manually design the feature pyramid, in [19], propose to discover the associations throughout the gating tool for visual counting and dense description predictions. In this paper, we introduced pyramidal feature representations that use a compound of scalable feature examination space, including the neural architecture algorithm to

overcome the small object localization and detection exploration space of pyramidal architecture.

2.2 Structure of FPN

FPN [9] constructed with simple hierarchal convolutional pyramid layers. The convolutional layers are designed with a stack of simple convolution filters that extracts features from low to high levels. To construct the FPN network follows two methods, one is a bottom-up pathway, and the second one is a top-down pathway. These two are connected laterally. The output of the backbone network forwardly feeds with the bottom-up pathway. It computes a single feature hierarchy extract feature maps at several scales with step-up. The last layers of the bottom-up pathway generally enrich the feature maps. Then the output passes to a top-down pathway to compute the feature at high resolution by using the upsampling process. The upsampling process carries from top to down to the network and laterally from the bottom-up pathway. It enriches the feature maps with high resolution.

Feature pyramid network improves the feature quality for the backbone network. It extracts feature maps, including a different scale that makes the network confident about object classifications for classification and detection tasks. Classification accuracy is dependent on the feature extraction process.

III. Proposed Algorithm

3.1 PPCNN

The pyramid pooling network allows multi-scale features from different convolutional layers as inputs and extracts output feature maps in identical scales, as shown in Fig. 1.

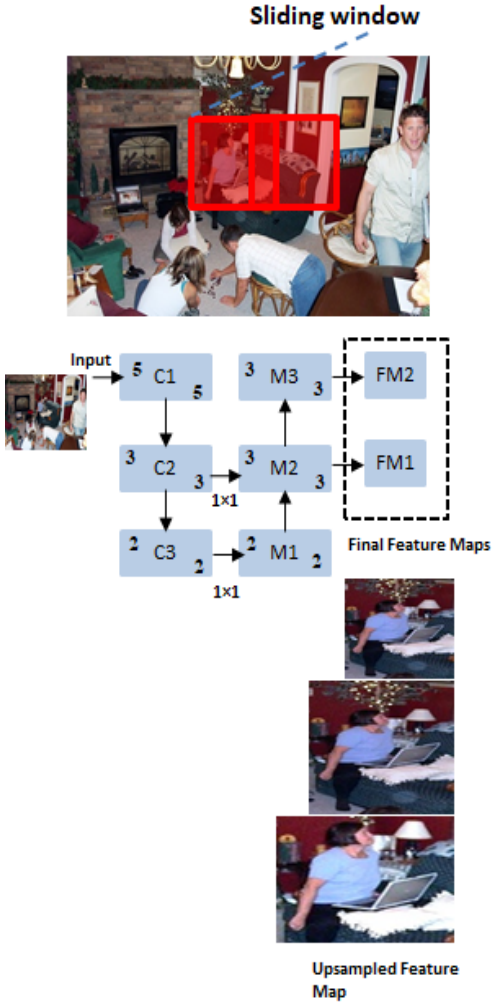

Fig. 1. Proposed architecture of improved feature pyramid network

The network architecture followed the design by VGGNet, which uses the mid-layer in each feature layer as the inputs to the first pyramid pooling network. The output of each pyramid pooling network directly preserves in a fully-connected layer. We use as inputs features in 3 different scales {C1, C2, C3} with corresponding feature stride of {32, 64, 128} pixels. The feature scaling conducts by simply applying stride 2 and stride 4 max-pooling.

The input features are then collected from the mid-layers and passed to the pyramid pooling network consisting of a series of merging cells that include upsampling connections. The pyramid network then outputs augmented multi-scale feature representations

{M1, M2, M3}. Considering pyramid pooling networks comprise feature layers in identical scales, the architecture of the FPN can be accumulated repeatedly for better efficiency.

We proposed the improved network architecture constructed with VGGNet amidst the U-shape feature pyramid network to extract high-level feature maps. The feature extraction process of the proposed feature pyramid network is shown in Fig. 2. The VGGNet uses straight forward architecture to classify objects. The VGGNet one of the successful image classifier; it performs classification by three fully connected layers. The first two fully-connected layers preserve feature data; each layer consist of 4096 channels. The third fully-connected layer contains 1000 channels that classify the objects by comparing them from the previous two fully-connected layers.

VGGNet has localization and feature extracting issues. To overcome the problem, the improved VGGNet architecture introduced among with improved feature pyramid network. The improved VGGNet architecture layers depth and channels modified to improve the detection performance. To rearrange the layer depth and channel dimension, we consider the computational complexity. Moreover, improved architecture designed with ideal frame size and pooling size to generate proper frame caption for feature extracting network. The proposed feature pyramid network uses simple feed-forward convolutional layers that can extract multi-scale high-level feature maps for the backbone network. The output of conv2 passes through the input of the first feature pyramid network (PPCNN_1), and the output of conv4 passes to the PPCNN_2. The primary feature data passes hierarchy through the convolutional filter to upsample the feature data to generate high-level feature maps. The convolutional layers are bilaterally interconnected with each other. The architecture follows the same procedure of bottom-up and top-down pathways as conventional FPN. It extracts high-quality feature maps, and all feature information preserves on FC2 (fully-connected layer 2).

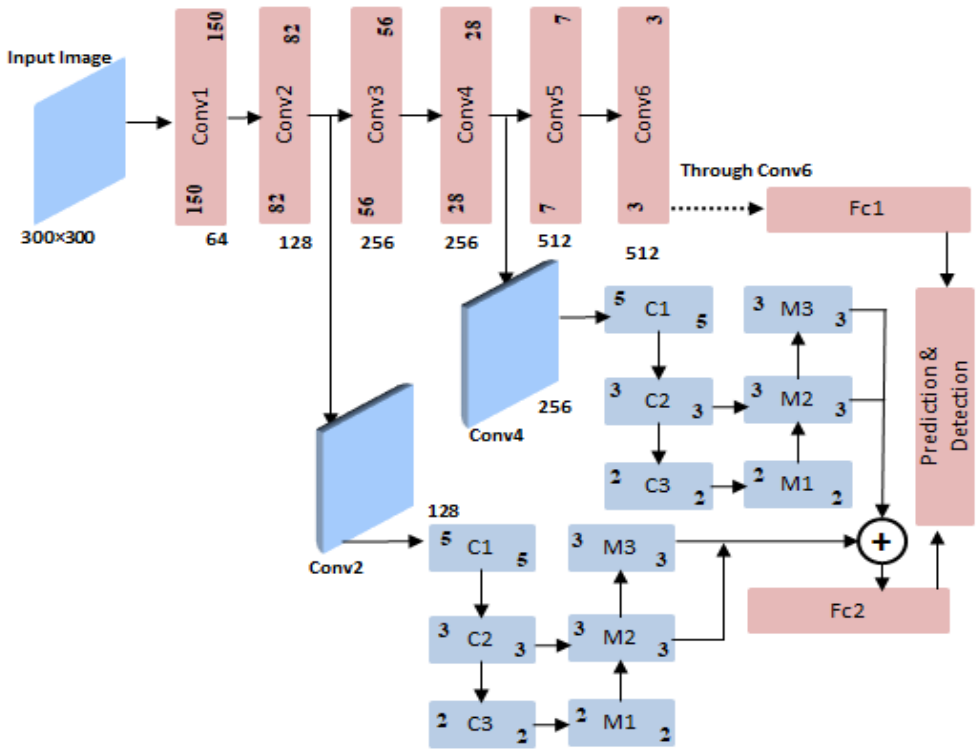

Fig. 2. Feature extracting process using proposed PPCNN

However, the VGGNet continues its convolutional process. The improved pyramid pooling convolutional neural network look over the source image and selectively search for specific object location.

$$P_n^{ij} = \frac{1}{(H_n \times W_n)} \cdot \sum_i^{H_n} \sum_j^{W_n} f_n(i,j) \quad (1)$$

In equation (1), the feature maps of the n^{th} pyramid level represent as f_n, the height and weight of the feature maps is H_n, and W_n, also, P_n^{ij} denoting the average pooling feature data of the PPCNN network.

The proposed VGGNet architecture introduced to reduce the training loss and improve the object localization performance. The image downsamples through 3×3 convolutional filters among stride 2 to generating feature maps. The network continues the convolutional process and extracts multi-scale feature maps through different convolutional layers. These features pass to the fully-connected layer for the further detection task. After that, the second feature pyramid network added with conv4 of the VGGNet to collect more high-grade feature data. The second pyramid network upsamples this feature information and collects the localization task's robust object location information. It carried a bottom-up convolutional process to upsample the feature data, and then the top-down process executes top-down process generates final feature maps for the next task.

These two pyramid network connected with the fully-connected layer. The fully-connected layer preserves the high-level feature maps to classification and detection. Predicting objects and generate the bounding-box NMS (Non-Maximum Suppression) [20] used in the proposed architecture. It examines feature data from fully-connected layers by selective search and predicts bounding-boxes to locate the exact object. To detect the exact object from feature information, it continuously searches and generates windows and score the windows. Then it checks all windows scores and picks the maximum window values to detect the object.

$$IOU = p_r(C/O) * p_r(O) * C_{pred}^{truth} \quad (2)$$
$$C = Object\,Class$$
$$O = Object$$

In eqn. (2), the bounding boxes are predicts by examining each grid cell and also considers the class probability C.

IV. Experiments

To carry the experiment VGGNet used as the backbone network and feature pyramid network for the feature extractor. The proposed feature pyramid pooling network combined with an improved VGGNet. Furthermore, it reduces VGGNet localization error and improves the feature quality.

4.1 Experimental Environment

The essential experimental parameters are shown in Table 1.

Table 1. Experimental parameters

Experimental parameters	Set value
Input size	300×300
Batch size	32
Weight decay	0.0009
Learning rate	0.001

The proposed model is constructed in tensorflow backend. The batch size is set to 32, and the learning rate sets to 0.001. The network composed of lightweight 3×3 convolutional layers carries the main roles in this experiment; these convolutional filters are used to construct pyramid pooling layers. PASCAL VOC 2007 [21] used to train the proposed network and evaluate the proposed method performance used MS COCO [22] dataset.

V. Results

The proposed model assumes a unique feature learning and extraction method that uses an enhanced feature scaling process. The feature extraction model examines feature information by using the pyramid pooling network to provide the important feature information to the detection module to detect the objects.

Detecting small objects is the most difficult and challenging task. The experimental results of the proposed method shown in Fig. 3. Compare to the evaluation result with the conventional network the proposed method achieved better detection accuracy in the multi-scale detection task. It successfully detected multi-scale objects that are remarkable.

The proposed architecture successful in identifying small objects from the MS COCO dataset. It conferred significant performance to detect multi-scale objects. The performance of the proposed method is shown in Tables 2, 3, and Table 4.

In Table 2, the evaluation matrics confirm the detection accuracy of the individual class. However, to perform a more deeply inspection, the proposed network was tested with a robust detection module, and the proposed model achieved a higher mAP(mean Average Precision) in the detection task shown in Table 3. Finally, the multi-scale detection test performed to check the detection accuracy of the specific size of objects (small, medium, and large objects), also in this task, the proposed network achieved better detection accuracy, shown in Table 4.

Table 2. Detection results on MS COCO dataset

Method	Object class	mAP	
		Specific object mAP	Average mAP
Proposed method (PPCNN)	People	89.4	**86.2**
	Train	83.7	
	Bicycle	85.2	
	Car	90.4	
	Table	83.2	
	Horse	84.9	
	Bike	83.7	
	Airplane	86.2	

Table 3. mAP comparison between the proposed network and different object detectors

Method	Base network	Input resolution	FPS	Boxes	mAP
PPCNN (Proposed model)	**VGG**	**300×300**	**13**	**4163**	**86.2**
R-FCN [7]	ResNet -101	1000×600	9	300	80.5
ION [23]	VGG	1000×600	1.3	4000	79.2

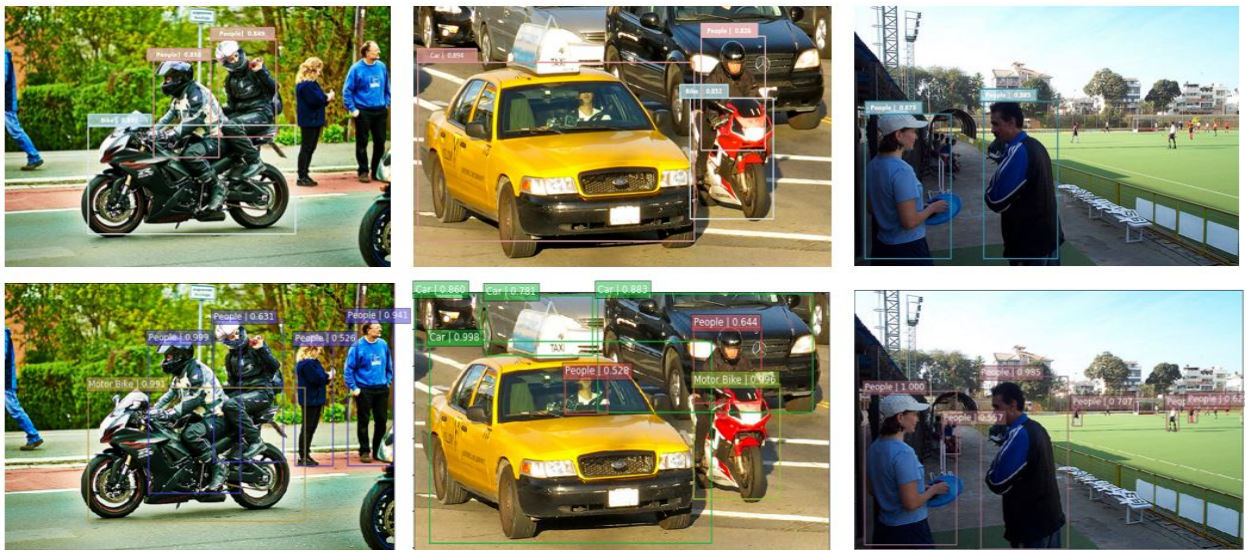

Fig. 3. Experimental results of conventional VGG network and proposed PPCNN (VGG network with u-shape feature pyramid network) on MS COCO dataset. The top row contains results of the conventional VGG network, and the bottom row contains the detection results of the proposed network

Table 4. Multi-scale detection results on MS COCO evaluation

Method	Average precision, IOU			Average precision, area		
	0.5:0.95	0.5	0.75	S	M	L
Proposed method	**37.2**	**51.3**	**39.2**	**17.2**	**40.3**	**51.8**
SSD 300 [1]	23.2	41.2	13.4	5.3	23.2	39.6
Faster R-CNN [6]	21.9	42.7	18.1	4.1	19.6	34.5
ION [23]	33.1	55.7	34.6	14.5	35.5	47.2

Evaluate the multi-scale detection performance; the network considers the minimum and maximum IoU for exact object localization and detection. However, it is crucial to localize small visual contents from the low-resolution image, but the proposed network showed the enhanced performance to localize small objects.

Ⅵ. Conclusions

In this paper, we build an end-to-end PPCNN model for object detection from pyramid pooling network sensing images. First, we extract image features by VGGNet, increasing the size of receptive fields while keeping the feature map's spatial size. Then we use an effective pyramid pooling network to fuse multi-level information, sighting to get a feature map that includes valuable information and inadequate target localization information. Subsequent simple pyramid pooling modules increase the combination of scattered target information and context information at distinctive scales. Finally, the detection task is carried out by optimizing the feature weighted balance loss, which accelerates the model's convergence under the premise of ensuring IoU. The network obtained mAP of 86.2 on the MS COCO dataset that is better than the conventional object detection frameworks. The proposed network also achieved enormous performance to detect multi-scale objects. The improved feature pyramid network architecture achieved high-level feature maps and reduced the backbone networks localization error.

References

[1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, Scott Reed, C. Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector", European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, pp. 21-37, Oct. 2016.

[2] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", ICLR, San Diego, CA, USA, arXiv:1409.1556, Apr. 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA, pp. 770-778, Jun. 2016.

[4] R. B. Girshick, J. Donahue, T Darrel, and J. Malik, "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 38, No. 1, pp. 142-158, Jan. 2016.

[5] R. Girshick, "Fast R-CNN", IEEE International Conference on Computer Vision, Santiago, Chile, pp. 1440-1448, Dec. 2015.

[6] 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.

[7] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks", NIPS, arXiv:1605.06409, pp. 379-387, Jun. 2016.

[8] J. Han and D.S. Kang, "CoS: An Emphasized Smooth Non-Monotonic Activation Function Consisting of Sigmoid For Deep learning", The

Journal of Korean Institute of Information Technology, Vol. 19, No. 1, pp. 1-9, Jan. 2021.

[9] M. F. Haque, H. Y. Lim, and D. S. Kang, "Object Detection Based on VGG with ResNet Network", In International Conference on Electronics, Information, and Communication (ICEIC), Auckland, New Zealand, IEEE, pp. 1-3, Jan. 2019.

[10] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection", Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, pp. 2117-2125, Jul. 2017.

[11] G. Ghiasi and C. C. Fowlkes, "Laplacian Pyramid Reconstruction and Refinement for Semantic Segmentation", In European Conference on Computer Vision, Amsterdam, The Netherlands, pp. 519-534, Oct. 2016.

[12] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN", In Proceedings of the IEEE International Conference on Computer Vision , Venice, Italy, pp. 2961-2969, Oct. 2017.

[13] A. R. Siyal, Z. Bhutto, S. M. S. Shah, A. Iqbal, F. Mehmood, A. Hussain, and S. Ahmed, "Still Image-Based Human Activity Recognition with Deep Representations and Residual Learning", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 11, No. 5, pp. 471-477, 2020.

[14] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation", In European Conference on Computer Vision, The Netherlands, pp. 483-499, Oct. 2016.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", In Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, pp. 234-241, Oct. 2015.

[16] J. M. Lee and D. S. Kang, "Start Point Detection Method for Tracing the Injection Path of Steel Rebars", The Journal of Korean Institute of Information Technology, Vol. 17, No. 6, pp. 9-16, Jun. 2019.

[17] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2det: A Single-shot Object Detector Based On Multi-level Feature Pyramid Network", In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, USA, Vol. 33, No. 01, pp. 9259-9266, Jul. 2019. https://doi.org/10.1609/aaai.v33i01.33019259

[18] Z. Bhutto, M. Z. Tunio, A. Hussain, J. Shah, I. Ali, and M. H. Saikh, "Scaling of Color Fusion in Stitching Image", International Journal of Computer Science and Network Security, Vol. 19, No. 4, pp. 61-64, Apr. 2019.

[19] R. J. L. S. D. Ooro-Rubio, and M. Niepert, "Learning Short-cutconnections for Object Counting", arXiv preprint arXiv:1805.02919, Nov. 2018.

[20] A. Neubeck and L. V. Gool, "Efficient Non-Maximum Suppression", 18th International Conference on Pattern Recognition, Hong Kong, China, Vol. 3, pp. 850-855, Aug. 2006.

[21] M. Everingham, L. J. V. Gool, C. K.I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge", International Journal on Computer Vision, Vol. 88, No. 2, pp. 303-338, Jun. 2010.

[22] T. Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollar, "Microsoft COCO: Common Objects In Context", European Conference on Computer Vision, Zurich, Switzerland, pp. 740-755, Sep. 2014.

[23] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Network", Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, pp. 2874-2883, Jun. 2016.

Authors

Md Foysal Haque

2018 ~ 2020 : M.Sc. in Electronic Engineering, Dong-A University.
2020 ~ Present : Ph.D. in Electronic Engineering, Dong-A University.
Research Interests : Digital Image Processing, Computer Vision 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