IEIE Transactions on Smart Processing and Computing, vol. 9, no. 1, February 2020
https://doi.org/10.5573/IEIESPC.2020.9.1.022

Real-time Robust Object Detection Using an Adjacent Feature Fusion-based Single Shot Multibox Detector

Donggeun Kim, Sangwoo Park, Donggoo Kang, and Joonki Paik

Graduate School of Advanced Imaging Science, Multimedia, and Film, Chung-Ang University / Seoul 06974, Korea

***** Corresponding Author: Joonki Paik, paikj@cau.ac.kr

Received June 27, 2019; Revised September 16, 2019; Accepted November 5, 2019; Published February 28, 2020

** Regular Paper*

***Abstract*:** A single shot multibox detector (SSD) is used as a baseline for many object detection networks, since it can provide sufficiently high accuracy in real time. However, it cannot deal with objects of various sizes, because features used in an SSD are not robust to multi-scale objects. To solve this problem, we present an improved feature pyramid for using multi-scale context information. The proposed feature pyramid fuses only adjacent features of the conventional SSD to achieve high accuracy without decreasing the processing speed. Our detector, with a 320×320 input, achieved 79.1% mean average precision (mAP) at 63 frames per second on a Pascal Visual Object Classes Challenge 2007 test set using a single Nvidia 1080 Ti graphics processing unit. This result shows better performance than existing SSDs.

***Keywords*:** Object detection, Feature pyramid, Pascal VOC, SSD

1. Introduction

Object detection is one of the most important research areas in computer vision. In recent years, many object detectors have adopted convolutional neural networks (CNNs) to achieve both high accuracy and fast processing speeds [1, 5, 7, 8, 10].

Deep learning–based object detection methods can be classified into one-stage and two-stage approaches. The two-stage method uses sliding windows and various anchor boxes to search proposals for highly accurate object detection at the cost of a slow processing speed. On the other hand, the one-stage method considers frame detection as a regression problem, which provides higher speed and lower accuracy than the two-stage method. The one-stage method is particularly suitable for real-time applications, such as intelligent surveillance systems and advanced driver assistance systems (ADASs).

The region-based convolutional neural network (R-CNN) is the basic model for various two-stage methods that generate object candidates using an external proposal algorithm [10]. Fast R-CNN uses region of interest (ROI) pooling to make each proposal in a single CNN model [4]. Faster R-CNN makes the two-stage detection method an end-to-end model by replacing the proposal extractor with a region proposal network (RPN) [7]. The region-based fully convolutional network (R-FCN) replaces ROI pooling with position-sensitive ROI pooling, which effectively reduces the channel and improves both speed and accuracy [5].

You Only Look Once (YOLO) [1] and the single shot multibox detector (SSD) [8] are the most popular one-stage detection methods. Both YOLO and the SSD are designed for real-time object detection while maintaining high average precision. YOLO divides the input image into multiple grid cells of size $s \times s$, and each grid cell predicts bounding boxes across all classes. YOLO version 2 (YOLOv2) removes fully connected layers from the original YOLO, and applies anchor boxes for higher accuracy [12]. YOLO and YOLOv2 are not robust to small objects. To solve that problem, YOLOv3 performs the generation of bounding boxes on three different scale features [13].

The processing speed with an SSD is as fast as YOLO while providing detection accuracy as high as a two-stage method, such as Faster R-CNN. An SSD extracts multi-scale feature maps from one CNN, and predicts multi-scale objects using the feature maps. To make the SSD more robust to multi-scale objects, the feature pyramid has been reconstructed. The deconvolutional SSD (DSSD) uses a deconvolution layer to build high-level semantic feature pyramids such as the feature pyramid network (FPN) [11].

In this paper, we propose a novel feature pyramid structure generated by simply connecting a conventional

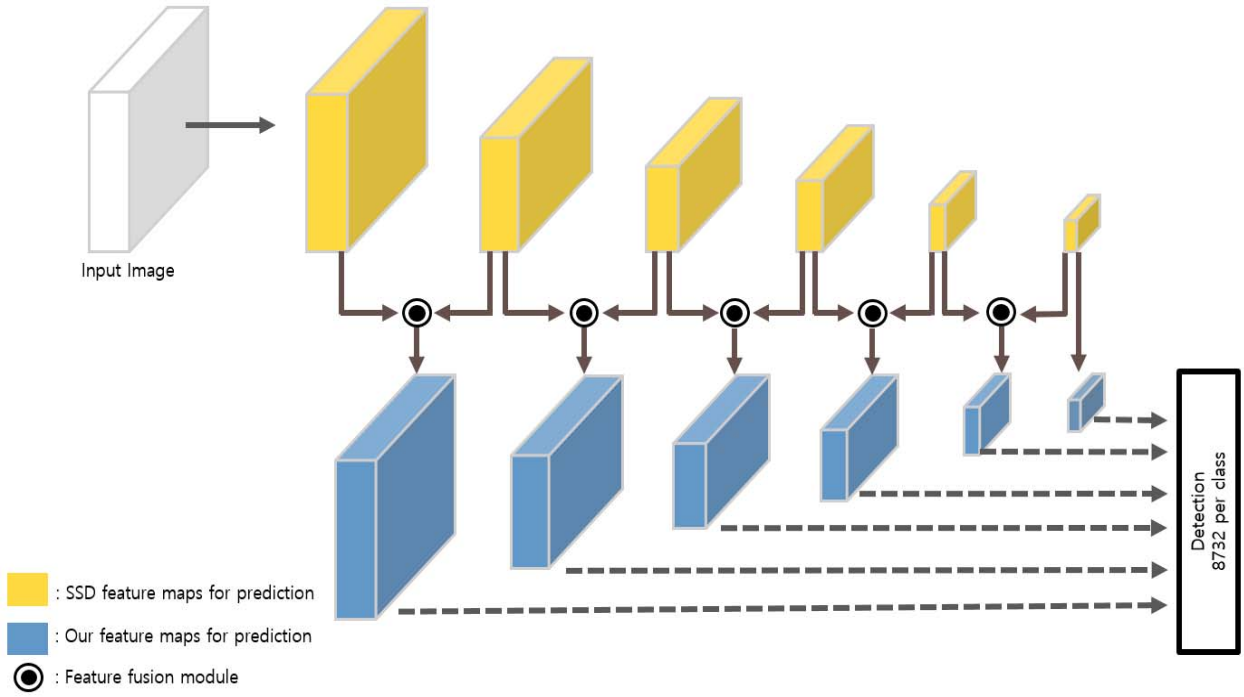

Fig. 1. Our proposed network.

SSD's prediction feature layers. We note that constructing the top-down structure from the SSD to feature layers, like the DSSD, is not suitable for using context information by making the pyramid deeply. We find out that a simple structure that connects to adjacent feature layers can effectively use context information. The proposed method outperforms other improved models of the SSD.

This paper is organized as follows. In Section 2, we explain the related work on other various and important SSD-based detection methods. In Section 3, we propose a novel version of the SSD; we then show experimental results and offer an evaluation of the proposed algorithm in Section 4. Section 5 concludes the paper and suggests future work.

2. Related Works

An SSD [8] is a simple CNN model that performs fast real-time detection. Also, an SSD uses multi-scale features to improve accuracy, so the SSD has been used as the baseline for many detectors. However, an SSD has a limitation in that it is not robust to multi-scale objects, especially small objects.

Various studies based on the SSD have noted that the simple feature pyramid structure from the plane network is not robust to multi-scale objects, and they proposed the following feature pyramid reconstruction to overcome that. A feature fusion SSD (FSSD) reconstructs a feature pyramid with only one feature fusion, which increases accuracy by adding weight to the SSD [6]. The rainbow SSD (RSSD) uses rainbow concatenation consisting of pooling and concatenation to improve accuracy with only a marginal extra cost [3].

The DSSD uses a top-down structure on the SSD, which transfers semantic information to the bottom layer, making it robust to small objects [9]. However, there is another limitation: losing detailed information about the bottom layer by overlapping semantic information from the top layer to the bottom layer many times.

The feature-fused SSD proposes to add context information to the SSD through a multi-level feature fusion method [15]. The feature map including context information is generated by combining the Conv4_3 layer, which mainly contains the location information of the object, and Conv5_3, which contains some background noise and detailed information.

3. The Proposed Method

As mentioned above, many studies have attempted to solve the scale variance problem by improving the feature pyramid of the SSD. Our goal is also to efficiently reconstruct the feature pyramid so it is robust to multi-scale objects. Our strategy is to connect only the nearest of the feature maps generated in the SSD, making it possible to utilize the context information without losing detailed information.

3.1 Feature Map Fusion

Our proposed network uses context information by fusing adjacent feature maps separately. The DSSD uses a deconvolution module (DM) and a prediction module (PM) to create a top-down structure. However, the combination of many modules makes the network complicated, resulting in a trade-off between speed and accuracy. The feature pyramid (by fusing the adjacent features we propose) is simple in structure, making it

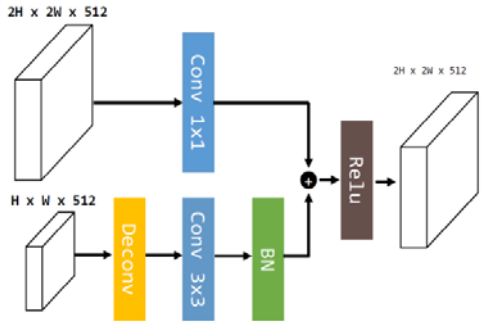

Fig. 2. Feature fusion module.

possible to efficiently utilize context information without making the network heavy.

Fig. 1 represents our overall network. Our network is generated by fusing feature maps, which are extracted from the feature pyramid of the conventional SSD. Feature map fusion is performed by the feature fusion module that we propose.

Unlike the top-down structure of a DSSD, our model consists of fusion of adjacent feature maps. The top-down structure superimposes semantic information from the top layers on the bottom layer, thus losing detailed information necessary to detect small objects. We note that fusion of features only in adjacent layers is robust to context without losing detailed information.

3.2 Feature Fusion Module

Fig. 2 shows the feature fusion module, which integrates information from different scale features. Our feature fusion module was built with reference to the DM of a DSSD. The DM consists of three convolution, three batch normalization, one deconvolution, two rectified linear unit (ReLU), and element-wise product operations. We note that this module is heavy, and makes the algorithm slow. We designed the feature fusion module by using one deconvolution, one batch normalization, one ReLU, and two convolution operations, which is simpler than a DM and sufficient to transfer context and important detailed information to the output feature.

The deconvolution layer is used to make the size of the low-resolution feature maps and the high-resolution feature maps the same. Each convolution layer transforms the previous feature maps so that the fused feature map has both context and detailed information. The batch normalization layer is to normalize features. After the element-wise sum and the ReLU operations, the output feature is generated.

3.3 Training

We follow the same strategy used in the SSD [8]. First, in the matching strategy, we match each ground truth box to the default box that has a higher Jaccard overlap than a set threshold (0.5). The default boxes with high confidence losses are selected so the ratio of negatives to positives is adjusted to 3:1. Then, we minimize the localization loss and confidence loss. For robust training, data augmentation is performed, which inflates data by random cropping, random photometric distortion, and random flipping.

4. Experimental Results

We evaluated our model and other models on the Pascal Visual Object Classes (VOC) Challenge 2007 detection benchmark [2]. This dataset consists of about 5k images in 20 object categories for testing. If the intersection over union (IOU) between the predicted bounding box and the ground truth is higher than 0.5, it is correct. We used mean average precision (mAP) as an actual metric for evaluating detection performance.

We trained our model with VOC 2007 'trainval' and VOC 2012 'trainval'. We used a single Nvidia 1080 Ti GPU, set the batch size to 32, and set the input size to 320x320. The initial learning rate was 0.001, divided by 10 for every 80k, 100k, and 120k iterations, and the total iterations was 140k. We set the weight decay to 0.0005 and applied stochastic gradient descent (SGD) with momentum 0.9. We used a pretrained Visual Geometry Group VGG16 on ImageNet as the backbone [14].

Table 1 presents the experiment results comparing our model with other models on Pascal VOC 2007. Our model320 achieved a 79.1% mAP, which is 1.9% higher than SSD300, and was the highest value among other SSD-based models with input size 300. With high-dimension input (i.e., 512 × 512), our model presents about 1.2% and 0.2% higher performance than SSD512 and RSSD512, respectively. The mAP of our model512 is 0.5% lower than DSSD513. However, our model increased speed from 5.5 frames per second (FPS) to 33 FPS using the 1080 Ti, and our model is superior, considering the tradeoffs. In addition, our model achieved a higher mAP and FPS than other two-stage methods.

Fig. 3 shows some detection examples from Pascal VOC 2007 and a comparison of our model with a conventional SSD. The SSD does not work well to find crowded, occluded, and small objects, such as people and vehicles in a road scene. However, in these cases, our model distinguishes these objects from others, and shows high-quality detection results.

5. Conclusion

In this paper, we proposed a novel feature pyramid that is generated by fusing the feature maps generated in the SSD. The proposed feature pyramid is robust to multi-scale objects using context information. Experiments on the Pascal VOC 2007 detection benchmark proved that our model improves the conventional SSD and is an efficient detector that is fast and accurate.

In the future, we plan to employ our model to other challenge object detection datasets, e.g., pedestrians and vehicles, which have small and crowded objects.

Table 1. Comparison of Speed & Accuracy on PASCAL VOC 2007.

Method		Backbone	Input size	FPS	mAP (%)	GPU
Two-stage	Fast [4]	VGG-16	~ 1000 x 600	0.5	70.0	K40
	Faster [7]	VGG-16	~ 1000 x 600	7	73.2	Titan X
	OHEM [16]	VGG-16	~ 1000 x 600	7	74.6	Titan X
	R-FCN [5]	ResNet-101	~ 1000 x 600	9	80.5	Titan X
One-stage	SSD321 [8]	ResNet-101	321 x 321	11.2	77.1	Titan X
	SSD300 [8]	Vgg-16	300 x 300	46	77.2	Titan X
	SSD300 [8]	Vgg-16	300 x 300	85	77.2	1080 Ti
	RSSD300 [3]	VGG-16	300 x 300	35	78.5	Titan X
	DSSD321 [9]	ResNet-101	321 x 321	9.5	78.6	Titan X
	SSD512 [8]	VGG-16	512 x 512	19	79.8	Titan X
	SSD513 [8]	ResNet-101	513 x 513	6.8	80.6	Titan X
	RSSD512 [3]	VGG-16	512 x 512	16.6	80.8	Titan X
	DSSD513 [9]	ResNet-101	513 x 513	5.5	**81.5**	Titan X
Our model 320		VGG-16	320 x 320	63	**79.1**	1080 Ti
Our model 512		VGG-16	512 x 512	33	81.0	1080 Ti

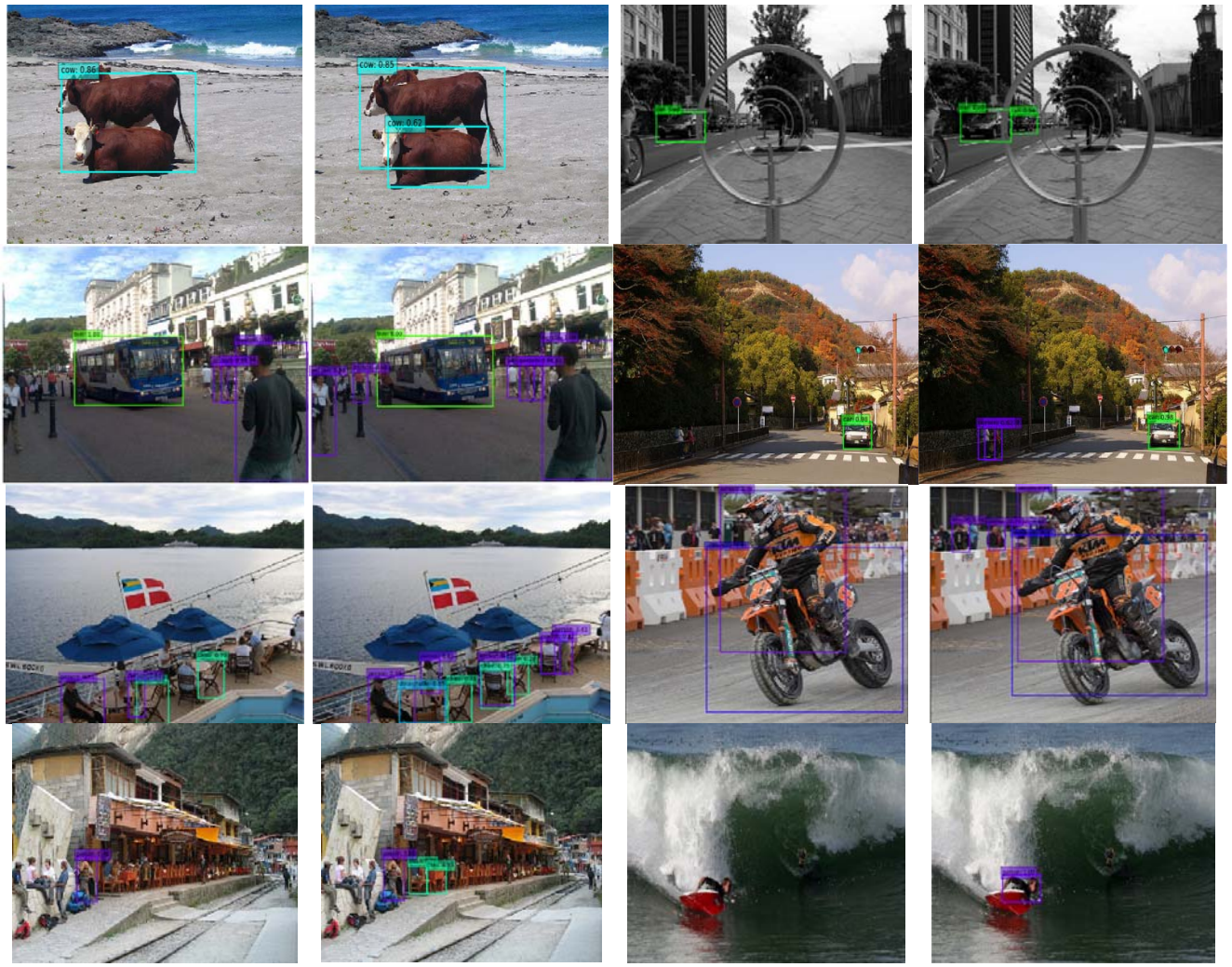

Fig. 3. Detection examples on the Pascal VOC 2007 dataset from the SSD300 and our 320. Both models were trained with a VOC07+12 dataset, and VGG16 was used as the backbone network. Columns are formatted in pairs, with the left side the result of the conventional SSD, and the right side the result from our network.

Acknowledgement

This work was partly supported by an Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (2017-0-00250, Intelligent Defense Boundary Surveillance Technology Using Collaborative Reinforced Learning of Embedded Edge Camera and Image Analysis) and by the ICT R&D program of MSIP/IITP (2014-0-00077, development of global multi-target tracking and event prediction techniques based on real-time large-scale video analysis).

References

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection”, In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 779–788, 2016. Article (CrossRef Link)

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge”, International Journal of Computer Vision, 88(2):303–338, June 2010. Article (CrossRef Link)

[3] Jeong, H. Park, and N. Kwak, “Enhancement of SSD by concatenating feature maps for object detection”, arXiv preprint arXiv:1705.09587 (2017). Article (CrossRef Link)

[4] R. Girshick, “Fast r-cnn”, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 1440–1448, 2015. Article (CrossRef Link)

[5] J. Dai, Y. Li, K. He, and J. Sun, “R-fcn: Object detection via regionbased fully convolutional networks”, In NIPS, 2016. Article (CrossRef Link)

[6] Li, Zuoxin, and Fuqiang Zhou, "FSSD: feature fusion single shot multibox detector", arXiv preprint arXiv:1712.00960 (2017). Article (CrossRef Link)

[7] Shaoging Ren, Kaiming He, Ross Girshick, Jian Sun, “Faster R-CNN: Toward Real-Time Object Detection with Region Proposal Networks”, Advances in Neural Information Processing Systems 28 (NIPS), 2015. Article (CrossRef Link)

[8] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg, “SSD: Single Shot MultiBox Detector”, European Conference on Computer Vision, pages 21-37, September 2016. Article (CrossRef Link)

[9] C. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, “DSSD : Deconvolutional single shot detector”, arXiv preprint arXiv:1701.06659 (2017). Article (CrossRef Link)

[10] 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. Article (CrossRef Link)

[11] T. Lin, P. Doll´ar, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, “Feature pyramid networks for object detection”, Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. Article (CrossRef Link)

[12] Redmon, Joseph, and Ali Farhadi, "YOLO9000: better, faster, stronger", Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. Article (CrossRef Link)

[13] Redmon, Joseph, and Ali Farhadi, "Yolov3: An incremental improvement", arXiv preprint arXiv:1804.02767 (2018). Article (CrossRef Link)

[14] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, arXiv preprint arXiv:1409.1556 (2014). Article (CrossRef Link)

[15] Cao, Guimei, et al, "Feature-fused SSD: fast detection for small objects", Ninth International Conference on Graphic and Image Processing (ICGIP 2017). Vol. 10615. International Society for Optics and Photonics, 2018. Article (CrossRef Link)

[16] A. Shrivastava, A. Gupta, and R. B. Girshick, “Training region-based object detectors with online hard example mining”, Proceedings of the IEEE conference on computer vision and pattern recognition, pages 761–769, 2016. Article (CrossRef Link)

Donggeun Kim was born in Incheon, Korea, in 1992. He received a BSc in Information and Communications Engineering from Sunmoon University, Korea, in 2017. Currently, he is pursuing an MSc in Image Science at Chung-Ang University. His research interests include object detection and deep learning.

Sangwoo Park was born in Incheon, Korea, in 1989. He received a BSc in Electric and Electronic Engineering from Soon Chun Hyang University, Korea, in 2015. Also, he received an MSc in Image Science at Chung-Ang University, Korea, in 2017. Currently, he is pursuing a PhD in Image Science at Chung-Ang University. His research interests include image translation and object detection.

Donggoo Kang was born in Seoul, Korea, in 1992. He received a BSc in Financial Economics from Seokyeong University, South Korea, in 2018. Currently, he is pursuing an MSc in Image Processing at Chung-Ang University. His research interests include salient object detection and pose estimation.

Joonki Paik was born in Seoul, South Korea, in 1960. He received BSc in Control and Instrumentation Engineering from Seoul National University, in 1984, and an MSc and a PhD in Electrical Engineering and Computer Science from Northwestern University, in 1987 and 1990, respectively. From 1990 to 1993, he joined Samsung Electronics, where he designed image stabilization chipsets for consumer camcorders. Since 1993, he has been a member of the faculty with Chung-Ang University, Seoul, Korea, where he is currently a Professor with the Graduate School of Advanced Imaging Science, Multimedia, and Film. From 1999 to 2002, he was a Visiting Professor with the Department of Electrical and Computer Engineering at the University of Tennessee, Knoxville. Since 2005, he has been the Director of the National Research Laboratory in the field of image processing and intelligent systems. From 2005 to 2007, he served as the Dean of the Graduate School of Advanced Imaging Science, Multimedia, and Film. From 2005 to 2007, he was the Director of the Seoul Future Contents Convergence Cluster established by the Seoul Research and Business Development Program. In 2008, he was a full-time Technical Consultant for the System LSI Division of Samsung Electronics, where he developed various computational photographic techniques, including an extended depth of field system. He has served as a member of the Presidential Advisory Board for Scientific/Technical Policy with the Korean Government, and is currently serving as a Technical Consultant for the Korean Supreme Prosecutor's Office for computational forensics. He was a two-time recipient of the Chester Sall Award from the IEEE Consumer Electronics Society, and received the Academic Award from the Institute of Electronic Engineers of Korea, and the Best Research Professor Award from Chung-Ang University. He has served the Consumer Electronics Society of the IEEE as a member of the Editorial Board, as Vice President of International Affairs, and as Director of the Sister and Related Societies Committee.