Journal of Korea Multimedia Society Vol. 24, No. 3, March 2021(pp. 416-422)
https://doi.org/10.9717/kmms.2020.24.3.416

Activity Object Detection Based on Improved Faster R-CNN

Ning Zhang[†], Yiran Feng[††], Eung-Joo Lee[†††]

ABSTRACT

Due to the large differences in human activity within classes, the large similarity between classes, and the problems of visual angle and occlusion, it is difficult to extract features manually, and the detection rate of human behavior is low. In order to better solve these problems, an improved Faster R-CNN-based detection algorithm is proposed in this paper. It achieves multi-object recognition and localization through a second-order detection network, and replaces the original feature extraction module with Dense-Net, which can fuse multi-level feature information, increase network depth and avoid disappearance of network gradients. Meanwhile, the proposal merging strategy is improved with Soft-NMS, where an attenuation function is designed to replace the conventional NMS algorithm, thereby avoiding missed detection of adjacent or overlapping objects, and enhancing the network detection accuracy under multiple objects. During the experiment, the improved Faster R-CNN method in this article has 84.7% target detection result, which is improved compared to other methods, which proves that the target recognition method has significant advantages and potential.

Key words: Human activity object; Faster R-CNN; Dense-Net; Soft-NMS

1. INTRODUCTION

As an important technology of human-computer interaction, human behavior detection has always received extensive attention from computer vision. However, in the real environment, there are some complicated backgrounds, the human body is covered by objects, and the human body moves in various poses, which make the task of human behavior detection more difficult.

The R-CNN (regions with CNN features)[1] model is a classic algorithm applied to object detection tasks. The algorithm idea of this model is to first read in the image and then generate approximately 2,000 category-independent candidate regions for the input image. Use a convolutional neural network to obtain feature vectors of the same length from each candidate area, and then use a support vector machine (SVM)[2] to detect and classify objects in each candidate area[3]. The R-CNN model uses image affine deformation to calculate the input of a convolutional neural network with a fixed size for each candidate window, regardless of the window shape.

Fast R-CNN[4] and Faster R-CNN[5] developed on the basis of R-CNN are more effective in object detection. The Fast R-CNN model is based on the R-CNN model and combines the characteristics of the SPP-Net[6] network to improve the speed of training and testing, and at the same time improve

※ Corresponding Author : Eung-Joo Lee, Address: (48520) 428, Sinseon-ro, Nam-gu, Busan, Korea, TEL : +82-51-629-1143, FAX : +82-51-629-1143, E-mail : ejlee@tu.ac.kr

Receipt date : Jan. 19, 2021, Approval date : Mar. 15, 2021

[†] Dept. of Information and Communication Engineering, Tongmyong University, Busan, Korea
Dept. of Duoyuan Technology Co., Ltd. Shenyang, China (E-mail : jangneyong0829@hotmail.com)

[††] Dept. of Information and Communication Engineering, Tongmyong University, Busan, Korea
(E-mail : 345509139@qq.com)

[†††] Dept. of Information and Communication Engineering, Tongmyong University, Busan, Korea

the accuracy of model detection. The Fast R-CNN model mainly solves the three problems of R-CNN and SPP-Net: slow test speed, slow training speed and large training space.

This paper introduces an improved Faster R-CNN network for identification and localization of human activity objects. Modification is made on the Faster R-CNN framework based on the features of Human activity object, the original feature extraction module is replaced with densely connected network (Dense-Net)[7], and the multi-level features of fusion objects are extracted to add expressive power to the features. Meanwhile, Soft-NMS is used instead of the original proposal merging strategy, and an attenuation function is designed to enhance the object box localization accuracy. Furthermore, used 2017MS COCO Test-dev data sets to train and test the algorithm, and tested in real scenes and got ideal results.

2. FASTER R-CNN ARCHITECTURES

As a current mainstream two-stage detection network, Faster R-CNN is a combination of RPN and Fast R-CNN, which enables output of detection categories and box positioning at each stages[8,9]. Depending on network architecture, the Faster R-CNN can be divided into three parts: the basic feature extraction network, the RPN and the detection network. The specific steps of the algorithm are described below. Fig. 1 presents the algorithmic framework.

3. IMPROVED FASTER R-CNN

3.1 Dense Block Network

Although a deeper network allows extraction of deeper semantic information, there will be an inevitable increase in parameters with the deepening of network[10,11]. As a result, a series of problems are brought to the network optimization and the experimental hardware. The data-sets built specifically for the shellfish classification and detection algorithm herein have small sample sizes, so that the network training easily leads to over- fitting. The use of Dense-Net as the feature extraction network helps solve the above problems[12].

As a novel network architecture, Dense-Net draws on the ideas of Res-Net. The most intuitive difference between the two architectures lies in the varying transfer functions for various network blocks.

$$x_l = H_l(x_{l-1}) + x \tag{1}$$

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]) \tag{2}$$

As is clear from (2) describing the transfer function of the Res-Net, the l-th layer output of the network equals the nonlinear variation of $(l-1)$

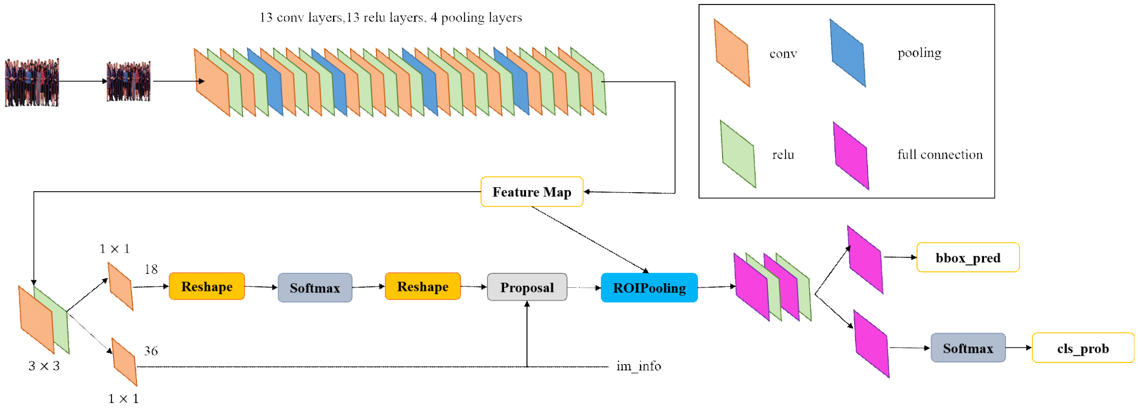

Fig. 1. Faster R-CNN Architectures.

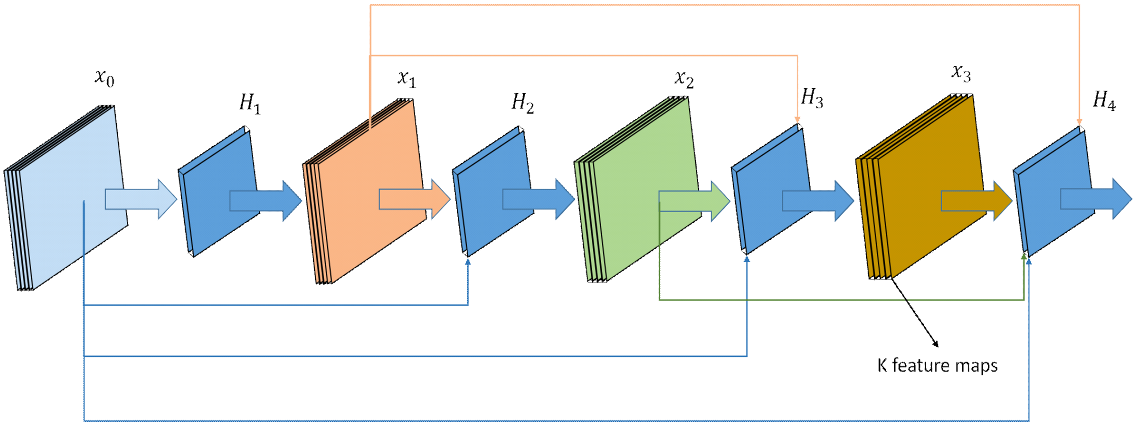

Fig. 2. Dense Block Architectures.

-th layer output plus the $(l-1)$-th layer output. Contrastively, the l-th layer output of a Dense-Net block is the set of nonlinear transformations output by all the previous layers. Fig. 2 depicts the Dense Blocks of the Dense-Net.

3.2 Non-Maximum Suppression

In essence, non-maximum suppression (NMS) aims to search for the local maximums and suppress the non-maximum elements, which is an important step of detection process[13,14]. Faster R-CNN generates a series of detection boxes $B = b_i, \ldots, b_N$ in an image, as well as the corresponding box score set C_i. NMS algorithm selects the detection box M in the object detection process prior to the maximum score, which is then subjected to intersection over union (IoU) computation with the remaining detection boxes. The detection box M will be suppressed if the result is greater than the set threshold N_t. The NMS algorithm formula is as follows:

$$C_i = \begin{cases} C_i & IoU(M, b_i) < N_t \\ 0 & IoU(M, b_i) \geq N_t \end{cases} \tag{3}$$

where the IoU is computed by the following formula:

$$IoU = (A \cap B)/(A \cup B) \tag{4}$$

where A and B represent two overlapping detection boxes.

As is clear from (3), the NMS algorithm zeroes with the detection box that is adjacent to M and greater than the threshold. If an object under detection appears in the overlapping region, the NMS algorithm will fail to detect the object, thereby reducing the accuracy of detection model.

To address this problem, the conventional NMS algorithm is replaced with the Soft-NMS, where an attenuation function is designed based on the IoU between adjacent detection boxes instead of setting their scores to zero, thereby ensuring accurate identification of adjacent objects. The Soft-NMS algorithm is expressed as:

$$C_i = \begin{cases} C_i & IoU(M, b_i) < N_t \\ C_i(1 - IoU(M, b_i)) & IoU(M, b_i) \geq N_t \end{cases} \tag{5}$$

To accomplish the shellfish classification and detection in real contexts, modifications are made on the front-end feature extractor and the tail-end regressor of the Faster R-CNN detection algorithm.

4. EXPERIMENTAL ANALYSIS

4.1 Data Sets Making and Processing

The Microsoft Common Objects in Context (MS COCO) dataset contains 91 common object categories with 82 of them having more than 5,000 labeled instances. In total the dataset has 2,500,000 labeled instances in 328,000 images. In contrast to the popular Image-Net dataset[15], COCO has fewer categories but more instances per category. This can aid in learning detailed object models capable of precise 2D localization. The dataset is also significantly larger in number of instances per category than the PASCAL VOC[16] and SUN[17] data sets. Additionally, a critical distinction between our dataset and others is the number of labeled instances per image which may aid in learning contextual information. MS COCO contains considerably more object instances per image (7.7) as compared to Image-Net (3.0) and PASCAL (2.3). In contrast, the SUN dataset, which contains significant contextual information, has over 17 objects and "stuff" per image but considerably fewer object instances overall.

4.2 Results Comparison and Analysis

In the test data, pictures of human behavior were randomly selected for testing, and the test results are shown in Fig. 3 to 4, respectively. Fig. 3 shows the test results of the unimproved algorithm. The human behaviors from left to right from top to bottom are horse riding (0.895), using a computer (0.824), playing musical instruments (0.781), calling (0.894), reading (0.797), cycling (0.846), jumping (0.897), taking pictures (0.734), among them, the position accuracy of the detection is in the brackets. Fig. 4 shows the test results of the improved algorithm. The human behaviors from left to right from top to bottom are horse riding (0.937), using a computer (0.954), playing musical instruments (0.892), calling (0.957), reading (0.922), cycling (0.936), jumping (0.979), taking pictures (0.894), among them, the position accuracy of the detection is in the brackets. Comparing the randomly selected test data, the last three actions in Fig. 3 are reading,

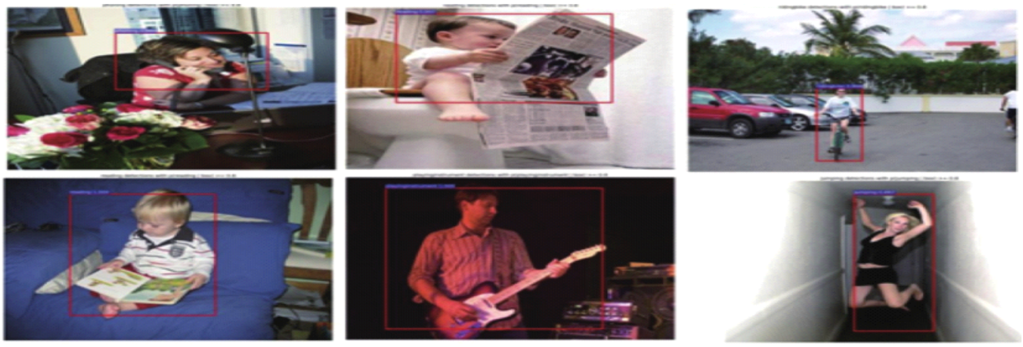

Fig. 3. The Original Algorithm Detection Results.

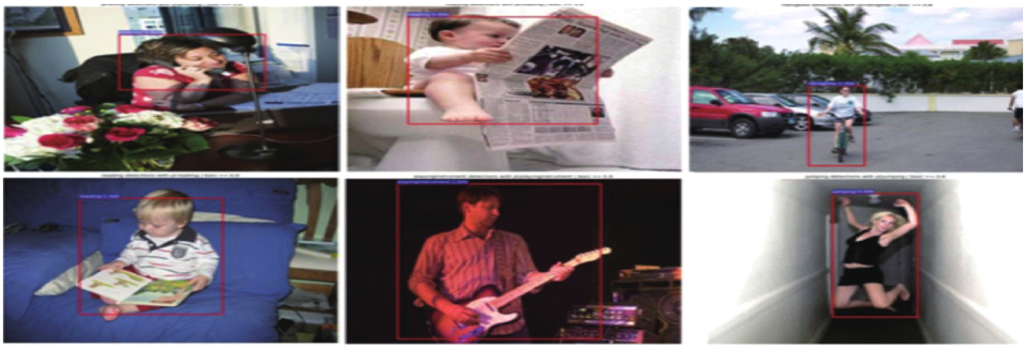

Fig. 4. The Improved Algorithm Detection Results.

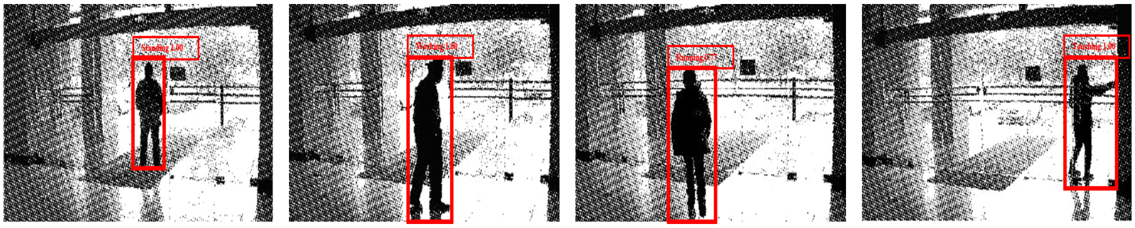

Fig. 5. Training Process of Loss and Testing Process of Validation Accuracy.

taking photos, and riding a bicycle. When using VGGNet, the algorithm did not detect the actions in the figure. In Fig. 4 for the same picture and the same behavior, the improved algorithm accurately recognizes the three actions, and the actions in the picture are more accurate than the original algorithm. Compared with the two algorithms, the improved algorithm can not only accurately identify all human behaviors in the sampled test images, but also has improved accuracy.

The improved Faster R-CNN algorithm has a better recognition effect on the three actions of playing computer, horse riding and cycling, although the recognition effect of reading and playing musical instruments is slightly worse than other actions. However, compared with the original algorithm, the accuracy of identifying categories and locations has been significantly improved. Moreover, the recognition effect of the improved Faster R-CNN algorithm has been significantly improved, and the average classification effect and position accuracy have reached 84.7%, which proves the effectiveness of the improved algorithm for human behavior detection tasks.

In this paper, also have some sampling tests were performed in real scenarios, and the test results are shown in Fig. 5. The human behaviors from left to right in the figure are standing (1.000), walking (0.956), running (0.774), and touching (1.000). The position accuracy of detection is in parentheses.

According to Fig. 6, it can be seen that the improved Faster R-CNN quickly stabilizes after the 1500 iteration process, which further improves the efficiency of target detection. And maintain a steady trend in accuracy.

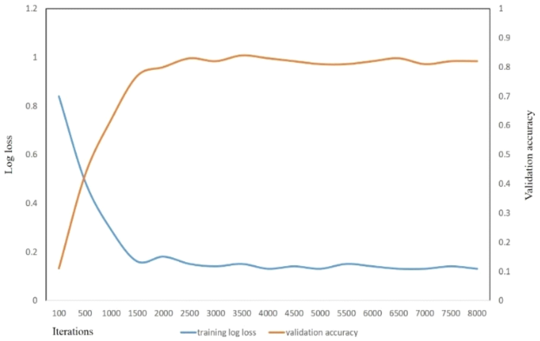

Fig. 6. Validation Accuracy in the Processing of Iterations.

This article adds a comparative experiment of R-CNN, Fast R-CNN and Faster R-CNN. The comparative analysis shows that Improved Faster R-CNN first has a very large improvement in recognition speed, and the average detection time for each picture is 0.14s, to achieve a faster detection effect; and the CNN training network selects a deeper VGG16 network model, and the recognition accuracy reaches 84.7%. Table 1 and Fig. 7 show the test performance comparison of R-CNN, Fast R-CNN, Faster R-CNN and Improved Faster R-CNN methods.

Table 1. The Test Performance Comparison.

Algorithm	Accuracy [%]	Duration [s]
R-CNN[1]	77.10	13.40
Fast R-CNN[4]	77.50	4.60
Faster R-CNN[5]	81.67	0.76
Improved Faster R-CNN	84.70	0.14

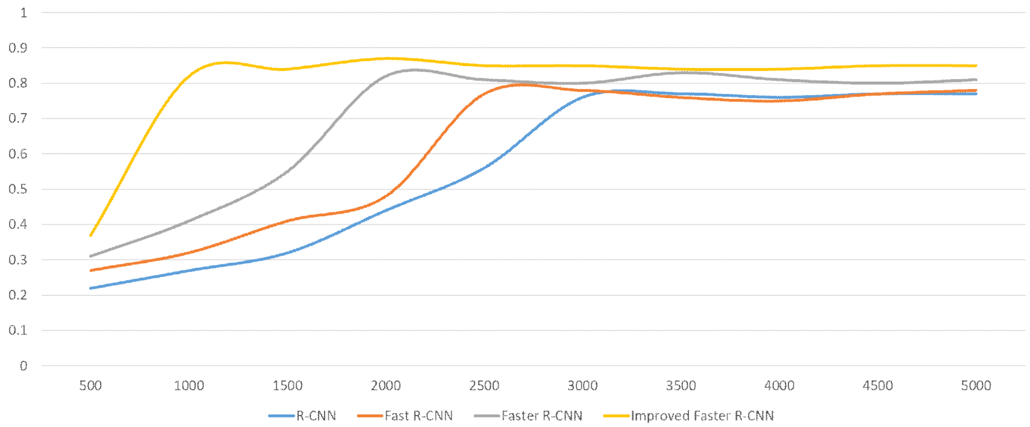

Fig. 7. The Rate of Comparison Tests for Different Methods.

5. CONCLUSION

As a modification based on Faster R-CNN, the algorithm uses Dense-Net as the feature extraction network, where the dense connection between blocks allows effective utilization of the shallow and deep layer features, thereby enhancing the shellfish detection accuracy. Meanwhile, the proposal merging strategy is optimized by using Soft-NMS instead of the original algorithm, thereby adding precision to the proposals. Furthermore, shellfish data sets are built in real contexts, and then augmented to improve the robustness of the training model. The proposed detection algorithm can achieve multi-object shellfish detection in daily life, and has preferable accuracies in complicated scenarios like illumination influence, partial occlusion and complex background, which exhibits a good detection performance.

REFERENCE

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, Issue 1, pp. 142-158, 2016.

[2] L. Hang, *Statistical learning method*, Beijing, Tsinghua University Press, pp. 36-58, 2012.

[3] Z. Wenda, X. Yuelei, and N. Jiacheng, "Image Target Recognition Method Based on Multi-Scale Block Convolutional Neural Network," *Journal of Computer Applications*, Vol. 36, No. 4, pp. 1033-1038, 2016.

[4] R. Girshick, "Fast R-CNN," *Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile*, pp. 1440-1448, 2015.

[5] R. Shaoqing, H. Kaiming, and R. Girshick, "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks,“ *Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada*, pp. 91-99, 2015.

[6] H. Kaiming, Z. Xiangyu, and R. Shaoqing, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland*, pp. 346-361, 2014.

[7] J.H. Park and E.-J. Lee, "Human Activity Recognition Based on 3D Residual Dense Network," *Journal of Korea Multimedia Society*, Vol. 23, No. 12, pp. 1540-1551, 2020.

[8] Y. Liu, X. Wu, and G. Xue, "Multi-target Real-time Detection for Road Traffic Signs Based on Deep Learning," *Journal of Guangxi*

Normal University(Natural Science Edition), Vol. 38, No. 2, pp. 96-106, 2020.

[9] X. Wu and X. Ling, "Facial expression recognition based on improved Faster RCNN," *CAAI Transactions on Intelligent Systems*, DOI:10.11992/tis.201910020.

[10] B. Chen, T. Zhao, and J. Liu, "Multipath Feature Recalibration Densenet for Image Classification," *International Conference of Machine Learning and Cybernetics*, pp. 1-10, 2020.

[11] Q. Liu, X. Xiang, and J. Qin, "Coverless Image Steganography Based on DenseNet Feature Mapping," *EURASIP Journal on Image and Video Processing*, Article No. 39, 2020.

[12] Z. Li, Y. Lin, and A. Elofsson, "Protein Contact Map Prediction Based on ResNet and Dense Net," *BioMed Research International*, pp. 1-12, 2020.

[13] B. Chen, W. Chen, and X. Wei, "Characterization of Elastic Parameters for Functionally Graded Material by a Meshfree Method Combined with the NMS Approach," *Inverse Problems in Science and Engineering*. Vol. 26, No. 4, pp. 601-617, 2018.

[14] Y. Zhou, Y. Zhang, X. Xie, and S.Y. Kung, "Image Super-Resolution Based on Dense Convolutional Auto-Encoder Blocks," *Neurocomputing*. Vol. 6, pp. 98-109, 2020.

[15] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, doi: 10.1109/CVPR.2009.5206848.

[16] 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*, Vol. 88, No. 2, pp. 303-338, 2010.

[17] J. Xiao, J. Hays, K.A. Ehinger, A. Oliva, and A. Torralba, "SUN Database: Large-scale Scene Recognition from Abbey to Zoo," *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, doi: 10.1109/CVPR.2010.55 39970.

Zhang Ning

received his B. S. at Dalian Polytechnic University in China (2008-2012), and received his master degree and doctor degree at Busan Tongmyong University in Korea (2014-2016, 2016-2020). Currently, he is working as a full-time researcher at the ICT Convergence Research Institute of Sookmyung Women's University in Korea, and concurrently as an information technology research engineer at Shenyang Duoyuan Co., Ltd. in China. His main research areas are image processing, computer vision and intelligent pattern recognition.

Feng Yiran

is a doctoral candidate majoring in electronic information and Communication at Tongmyong University, South Korea. He received the B.E. degree and M.E. degree from Dalian Polytechnic University, China in 2013 and 2015 respectively. He works as a teacher at Dalian Polytechnic University and member of Korea Multimedia Association. He is mainly engaged in image recognition and robot technology research.

Lee Eung-Joo

received his B. S., M. S. and Ph. D. in Electronic Engineering from Kyungpook National University, Korea, in 1990, 1992, and Aug. 1996, respectively. Since 1997 he has worked with the Department of Information & Communications Engineering, Tongmyong University, Korea, where he is currently a professor. From 2005 to July 2006, he was a visiting professor in the Department of Computer and Information Engineering, Dalian Polytechnic University, and from Dec 2018 he was appointed honorary professor of Dalian Polytechnic University, China. His main research interests include biometrics, image processing, and computer vision.