Development of Specific Area Intrusion Detection System using YOLO in CCTV Video

GeunTae Kim, Yeonghun Lee, Kyounghak Lee, HyungHwa Ko

Abstract*: Recently, many things are monitored with CCTV. It needs manpower to monitor. Therefore, we propose a system that stores the object's data when the object enters a specific area. Generally, You Only Look Once (YOLO) method is used for object detection. We propose a system that increases the recognition rate by using new dataset because the recognition rate may not be accurate when using existing VOC weight, and can apply to intrusion detection. The Proposed model with the training data created on the particular conditions precisely detects proper objects and reduces effect by a variance of frame and brightness. The object was found more precisely with newly trained weights instead of using VOC weight. The dataset is made to fit to the specific situation, and it can be used for genera situation. When using the VOC weight, object detection was affected by the frame change and the brightness change of the light, but the effect was reduced by using new weight.

***Index Terms*: CCTV, CNN, Intrusion Detection, Object Classification, Object Detection, YOLO**

I. INTRODUCTION

These days, CCTV has sprung up in all the streets and alleys. It is primarily used to prevent crimes. We still require supervising CCTV screen by ourselves. That is, people are watching screen and making decision. Now when an event occurs, we must search around a time when the event is occurred in person. In this paper, we propose a system that records information on the database when some object intrudes specific area.

The Convolutional Neural Network (CNN) used for recent image processing field shows great result. The CNN conducting only classification in the past has been improved a lot and now it is possible to detect object. The object detection needs localization with classification. Consequently, it allows us to find what object is and where object is in the image and video. CNN started getting paid attention by AlexNet[1], which won in ImageNet Challenge in 2012. Initial CNN is LeNet[2], which was developed in 1998 by Yann LeCun to recognize handwritten zip codes. Passing through dark-age for around 10 years after 1998, AlexNet was emerged to induce lots of study. Since then, a various CNN models have been developed. YOLO V1[4] had been published in 2016 and was upgraded to YOLO V3[5] in 2018. The model used in this paper is YOLO V2[3] created in 2017. The reason why we use V2 is the fast processing property for real time image. In case of V3, mAP is 57.9 and processing frames per second(fps) is 20. On the other hands, mAP of V2 is 48.1 and frame rate is 40. Therefore, we adopted the model V2 that is two times faster than V3.

The proposed system implements Object detection and Intrusion detection on CCTV using YOLO. At first, it detects cars, trucks, buses, motorcycles and save information of those objects as a log file form. After then, it reconstructs the video to find the moment that intrusion of object is detected or the time that we choose, using stored log file and the video. Furthermore, it draws the bounding box of intruding object and tracks the object using information of log file. We address this stage as Reconstruction Process. Users can use this result by Reconstruction Process with database.

When using VOC weight offered by YOLO, it has limitation of detecting objects previously defined and not detecting the object we want because of CCTV environment suitable for our study. Thus, we created dataset with CCTV footage on common road, using YOLO Mark[6] tool. As a result, this dataset enables recognition rate to be improved and object to be detected as classified.

II. MATERIALS AND METHODS

We use VOC weight with YOLO V2 so that detect only two objects, cars and human in previous work. In this work, we got new weight by training the dataset we made. With this new weight, we made the system to detect four objects, such as car, truck, bus and motorcycle.

A. YOLO V2

YOLO V2 is one of the object detection system that is faster and more robust system than previous YOLO V1 with basic dataset such as PASCAL VOC dataset[7] and COCO dataset[8]. In PASCAL VOC 2007 dataset, mAP of YOLO V2 is 76.8 and frame rate is 67fps. In COCO dataset, mAP is 78.6 and frame rate is 40fps. This shows better performance than Faster Region Convolutional Neural Network (Faster R-CNN[9]) and Single Shot Detection (SSD)[10] based on ResNet[11]. YOLO system was motivated to improve R-CNN[12], Fast R-CNN[13], and Faster R-CNN and makes faster and more precise. Especially YOLO removed the region proposal method in R-CNN series and made it one Convolutional neural network so that make it faster.

Revised Manuscript Received on May 23, 2019.

GeunTae Kim, Dept. of Electronics and Communications Eng, Kwangwoon Univ., Seoul, Korea.

Yeonghun Lee, Dept. of Electronics and Communications Eng, Kwangwoon Univ., Seoul, Korea.

Kyounghak Lee, IACF, Kwangwoon Univ., Seoul, Korea.

HyungHwa Ko, Dept. of Electronics and Communications Eng, Kwangwoon Univ., Seoul, Korea.

YOLO V2 removed Fully Connected Layer (FC Layer) of YOLO V1. This is expressed as table 1, and it can be confirmed that a 1x1 Convolutional Layer is used instead of the FC Layer.

Table 1. Darknet-19

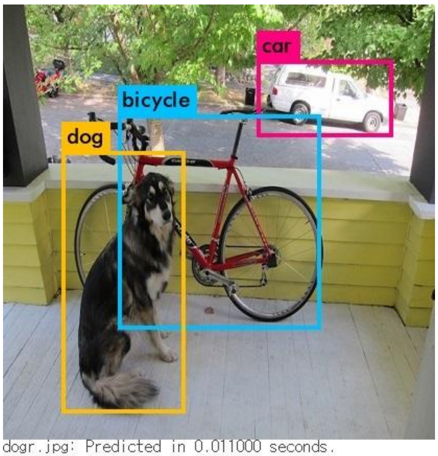
Type	Filters	Size/Stride	Output
Convolutional	32	3x3	224*224
Maxpool	N/A	2x2/2	112*112
Convolutional	64	3x3	112*112
Maxpool	N/A	2x2/2	56*56
Convolutional	128	3x3	56*56
Convolutional	64	1x1	56*56
Convolutional	128	3x3	56*56
Maxpool	N/A	2x2/2	28*28
Convolutional	256	3x3	28x28
Convolutional	128	1x1	28x28
Convolutional	256	3x3	28x28
Maxpool	N/A	2x2/2	14x14
Convolutional	512	3x3	14x14
Convolutional	256	1x1	14x14
Convolutional	512	3x3	14x14
Convolutional	256	1x1	14x14
Convolutional	512	3x3	14x14
Maxpool	N/A	2x2/2	7x7
Convolutional	1024	3x3	7x7
Convolutional	512	1x1	7x7
Convolutional	1024	3x3	7x7
Convolutional	512	1x1	7x7
Convolutional	1024	3x1	7x7
Convolutional	1000	1x1	7x7
Avgpool	N/A	Global	1000
Softmax	N/A	N/A	N/A

Learning process is divided into two sections. The first step is to learn dataset with Darknet-19 as classification network. The network input of 448*448 sizes is passed through preprocessing. In case of ImageNet dataset, output of convolutional layer is 1000, which is the number of class. Learning is progressed by Darknet-19, which consists of 19 convolutional layers shown in table 1. In the second step, it is learned by deformed Darknet-19, which removed Convolutional, Avgpool, Softmax layer and added four Object Detection layers. It is possible to detect object and region box after learning with the newly added networks. The reason why learning steps are divided is because of difficulty of learning both boxing the region and classification at the same time. Through these two learning steps, the front layers are fine-tuned.

There are two methods improving accuracy of classification. The first one is to use high resolution image. Images with 448*448 size exploit much more information than smaller size of image in previous model, whose size is 256*256[1], 224*224[4] or 300*300[9]. The second method is fine tuning through 2 learning steps. As explained in architecture part, it is learned twice using two networks of object classification and detection, except the last Convolutional, Avgpool and Softmax layer substituted with object detection layers. So that, the front part is learned for classification and performed better in object detection. It makes easier learning of detector and improves mAP by 4%[3]. As shown in table 2, YOLO V2 has the best performance in PASCAL VOC 2012 test data set. The result shows only mAP of classification of vehicles we are interested. For classification of object in CCTV video, class information of detected objects is saved for every single frame. Figure 1 shows a test result in common images with VOC weight. It denotes each class name and classification probability of detected objects.

Table 2. PASCAL VOC2012 test detection results[3]

Method	mAP	Bike	Bus	Car
YOLOV2	73.4	82.0	79.8	76.5
YOLO	57.9	67.2	68.3	55.9
SSD300	72.4	80.1	79.4	76.1
Faster R-CNN	70.4	67.2	77.5	75.9



```
dogr.jpg: Predicted in 0.011000 seconds.
car: 76%
bicycle: 79%
dog: 89%
```

Figure 1. Classification Probability for an Test Image

Object Detection of YOLO V2 is accompanied with the Anchor Box which shows the location of the object in the image. Input image is resized to 416*416 and made an odd-sized feature map. YOLO V2 uses Grid cell to generate N boxes for each cell and merges it to have one box per one object, using IOU of box and probability of object in each cell. YOLO V1 generates two boxes but YOLO V2 generates five boxes and merges. Figure 2 denote average IOU graph according to the number of clustering.

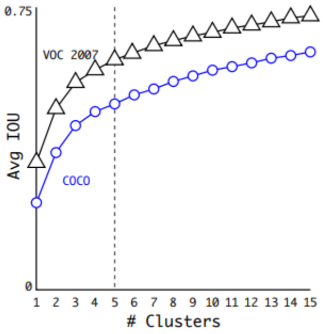


Figure 2. Average IOU over various Numbers of Cluster[3]

B. Dataset

Accuracy of CCTV image from different place and angle tends to be lower because VOC dataset or COCO dataset used in YOLO which was learned by popular images. To increase the accuracy, we made a dataset by collecting images related CCTV footage and derived from video of CCTV. Total number of images is 1,900, where 700 images are collected by crawling and 1,200 images are generated from CCTV video.

We take two CCTV video of different time zone in consideration of the impact by varied time. But night video was excluded because it is impossible to recognize. Figure 3 denote an image sample from each two-different time zone of CCTV videos. The image shows various vehicles passing by a crossroad. To figure out to classify every class correctly, the images including all classes, like cars, trucks, buses, motorcycles, which was not participated for training, are used as shown in the right side in Figure 3.

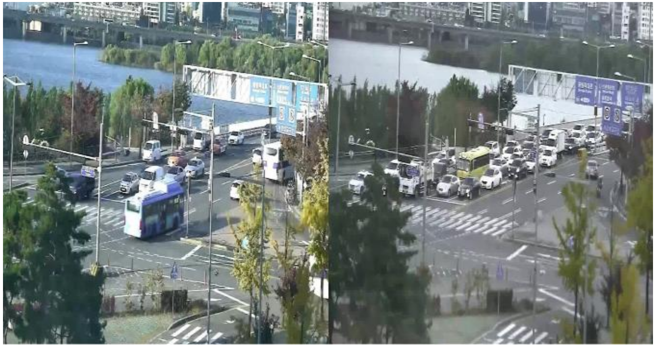


Figure 3. CCTV Image Sample

Dataset must be prepared to train YOLO system to detect 4 classes of object. YOLO training requires data containing configure file, dataset information, class and position of the object. The tool that can make this is YOLO Mark. The tool helps to make file with format we want, but it still takes a lot of time to edit every single image. Log data file has data of each object with class, position, width and height. Class number means that 0 is cars, 1 is motorbikes, 2 is trucks, and 3 is buses.

C. Intrusion Detection

In Intrusion Detection, when bounding box of detected object and bounding box of specific area is overlapped, it is called intrusion. The bounding box of specific area depict on left and right in the image of Figure 4. The Intrusion Detection is conducted by IoU.

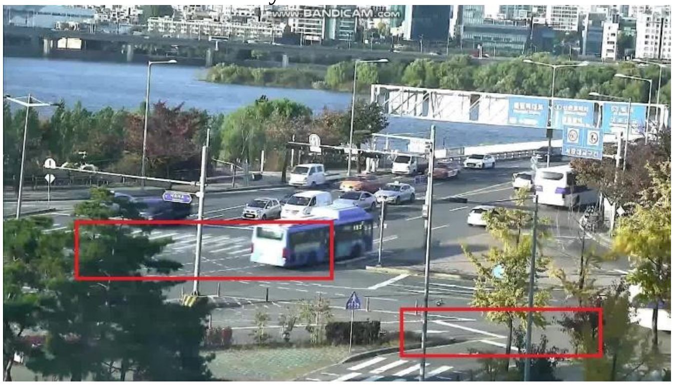


Figure 4. Specific Area Box on Sample Image

Intersection of Union (IoU) is the measure to know how much the boxes are overlapped. It is calculated by dividing the Area of Overlap by the Area of Union as shown in equation (1) and Figure 6 shows the equation by (a plane) figure. This model measures whether the object enters in specific area by IoU and save it into data file.

$$\mathrm{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \qquad (1)$$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

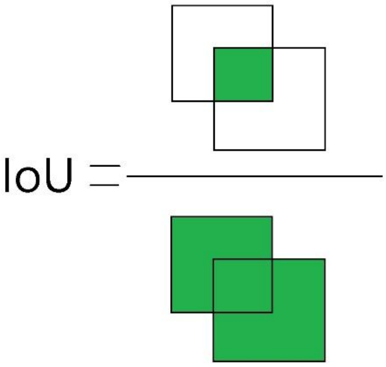


Figure 5. Explanation Of IoU

III. RESULTS AND DISCUSSION

The specific area intrusion system was implemented using the methods described earlier. There is an improvement in recognition rate using new weight. Figure 6 denote results of system using existing weight, and recognition rate of right image in Figure 6 is about 60%. Figure 7 denote results of system using new weight, and recognition rate of right image in Figure 7 is about 80%. There is an improvement in recognition rate of about 20%. In addition, it can detect objects that were not found in case of using previous VOC weight.

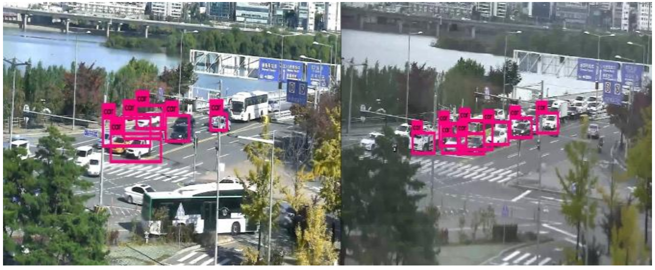


Figure 6. Detection Samples using VOC weight

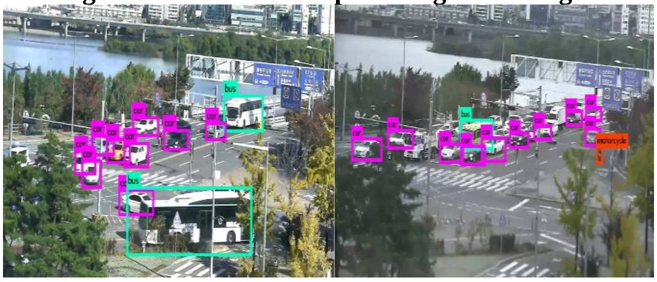
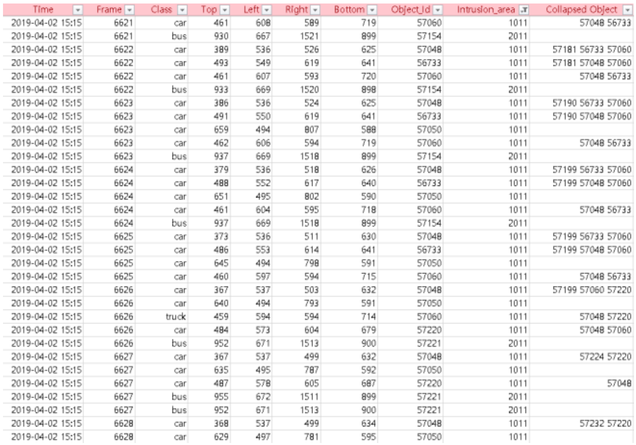


Figure 7. Detection Samples using New weight

In Figure 7, many objects that could not find in Figure 6 are found, and detectable areas are extended. The specific area shown in Figure 4 was not depicted in the Detection Sample, but if an object enters in the area, the information is saved as log file. The log file has information of time when object was detected, frame number, class, class id, location and area data. Figure 8 denote sample of log data file.

Intrusion area, of the data shown in log file, indicates the area that the object was intruding. One object was detected per line, and it has class and location data of the object. Based on this information, when viewing the video again, the reconstruction process is carried out. Figure 9 is a block diagram representing flow of this system. At first, when Image of CCTV is received, YOLO model carries out object detection on the frame and save data as a log file. After that, it implements reconstruction process to make bounding box of detected objects, using saved data and video. In common mode, every detected object is shown on the video. Whereas, Object which was intruded in the specific area is shown in special case. Figure 10 shows the result of reconstruction by expressing only objects in a specific area.



Time	Frame	Class	Top	Left	Right	Bottom	Object_Id	Intrusion_area	Collapsed Object
2019-04-02 15:15	6621	car	461	608	589	719	57060	1011	57048 56733
2019-04-02 15:15	6621	bus	930	667	1521	899	57154	2011	
2019-04-02 15:15	6622	car	389	536	526	625	57048	1011	57181 56733 57060
2019-04-02 15:15	6622	car	493	549	619	641	56733	1011	57181 57048 57060
2019-04-02 15:15	6622	car	461	607	593	720	57060	1011	57048 56733
2019-04-02 15:15	6622	bus	933	669	1520	898	57154	2011	
2019-04-02 15:15	6623	car	386	536	524	625	57048	1011	57190 56733 57060
2019-04-02 15:15	6623	car	491	550	619	641	56733	1011	57190 57048 57060
2019-04-02 15:15	6623	car	659	494	807	588	57050	1011	
2019-04-02 15:15	6623	car	462	606	594	719	57060	1011	57048 56733
2019-04-02 15:15	6623	bus	937	669	1518	899	57154	2011	
2019-04-02 15:15	6624	car	379	536	518	626	57048	1011	57199 56733 57060
2019-04-02 15:15	6624	car	488	552	617	640	56733	1011	57199 57048 57060
2019-04-02 15:15	6624	car	651	495	802	590	57050	1011	
2019-04-02 15:15	6624	car	461	604	595	718	57060	1011	57048 56733
2019-04-02 15:15	6624	bus	937	669	1518	899	57154	2011	
2019-04-02 15:15	6625	car	373	536	511	630	57048	1011	57199 56733 57060
2019-04-02 15:15	6625	car	486	553	614	641	56733	1011	57199 57048 57060
2019-04-02 15:15	6625	car	645	494	798	591	57050	1011	
2019-04-02 15:15	6625	car	460	597	594	715	57060	1011	57048 56733
2019-04-02 15:15	6626	car	367	537	503	632	57048	1011	57199 57060 57220
2019-04-02 15:15	6626	car	640	494	793	591	57050	1011	
2019-04-02 15:15	6626	truck	459	594	594	714	57060	1011	57048 57220
2019-04-02 15:15	6626	car	484	573	604	679	57220	1011	57048 57060
2019-04-02 15:15	6626	bus	952	671	1513	900	57221	2011	
2019-04-02 15:15	6627	car	367	537	499	632	57048	1011	57224 57220
2019-04-02 15:15	6627	car	635	495	787	592	57050	1011	
2019-04-02 15:15	6627	car	487	578	605	687	57220	1011	57048
2019-04-02 15:15	6627	bus	955	672	1511	899	57221	2011	
2019-04-02 15:15	6627	bus	952	671	1513	900	57221	2011	
2019-04-02 15:15	6628	car	368	537	499	634	57048	1011	57232 57220
2019-04-02 15:15	6628	car	629	497	781	595	57050	1011	

Figure 8. Sample of Log Data

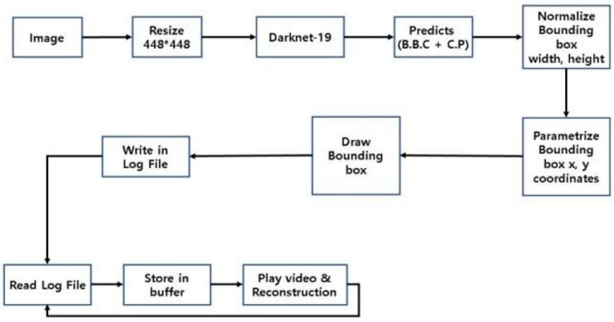


Figure 9. Overview of Proposed Intrusion Detection System Block Diagram

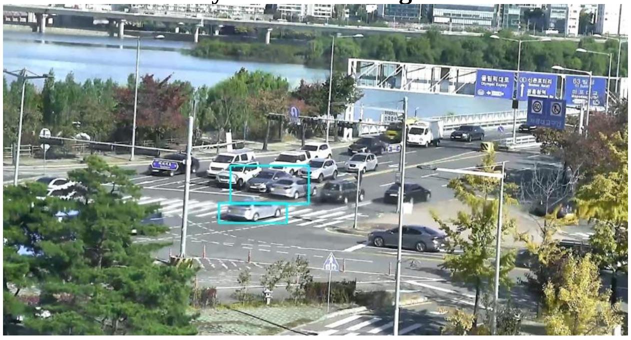


Figure 10. Example of Reconstruction Result

IV. CONCLUSION

This paper proposes a system to carry out object classification and detection using YOLO V2 and to save area information when object enter the specific area.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

As a result, our new dataset for CCTV improves performance of detection and classification, instead of using VOC weight. The proposed system reduces workload and saves time than the existing observation system because it informs the intrusion time among the stored data. By experiment, Object Detection and Intrusion Detection were proved to be well operating.

ACKNOWLEDGMENT

The present Research has been conducted by the Research Grant of Kwangwoon University in 2019.

REFERENCES

1. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems (NIPS)*. 1097-1105
2. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. (1998) Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE*. 2278-2324
3. J. Redmon and A. Farhadi. (2017) Yolo9000: Better, Faster, Stronger. *IEEE Conference on In Computer Vision and Pattern Recognition* (ICVPR). 6517–6525.
4. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. (2015) You Only Look Once: Unified, Real-time Object Detection. *arXiv* preprint arXiv:1506.02640.
5. J. Redmon and A. Farhadi. (2018) Yolov3: An Incremental Improvement. *arXiv*.
6. AlexeyAB. Yolo Mark : Open source neural networks in python. https://github.com/ lexeyAB/Yolo_mark
7. Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A. (2014) The PASCAL Visual Object Classes Challenge - a Retrospective. *International Journal of Computer Vision*.
8. Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár. (2014) Microsoft COCO: Common Objects in Context. *In CVPR*, arXiv:1405.0312
9. S. Ren, K. He, R. Girshick, and J. Sun. (2015) Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. *arXiv* preprint arXiv:1506.01497.
10. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.- Y. Fu, and A. C. Berg. (2016) SSD: Single Shot Multibox Detector. *In European conference on computer vision*, Springer. 21–37.
11. K. He, X. Zhang, S. Ren, and J. Sun. (2016) Deep Residual Learning for Image Recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 770–778.
12. R. Girshick, J. Donahue, T. Darrell, and J. Malik. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *In CVPR.* arXiv:1311.2524
13. R. Girshick. (2015) Fast R-CNN. in IEEE International Conference on Computer Vision (ICCV).

AUTHORS PROFILE

GeunTae Kim received his Bachelor Engineering in Electronics and Communication Eng. from Kwangwoon University in 2018. He is currently in master's degree course from Kwangwoon University. His research interest is in Image Processing, Object Detection, Face Emotion Recognition and Style Transfer using Deep Learning.

Yeonghun Lee received his Bachelor Engineering in Electronics and Communication Eng. From Kwangwoon University in 2019. He is currently in master's degree course from Kwangwoon University. His research interest is in Computer Vision.

Kyounghak Lee is Associate Professor at Kwangwoon University. He received Bachelor's Degree, Master's Degree and Ph.D in Kwangwoon University. He was senior researcher at KEIT from 1994 to 2011 and assistant professor at Namseoul University from 2011 to 2016. His research interest is in vision system, AI and se platform.

HyungHwa Ko is Professor at Kwangwoon University. He received Bachelor's Degree, Master's Degree and Ph.D in Seoul National University. He was researcher at LG from 1979 to 1980, chair of office of information support in Kwangwoon University from 2001 to 2003 and dean of college of Electronics and Information Eng. from 2008 to 2010. His research interest is in Image Compression, Video Compression, Document Analysis and Face Recognition.