

A YOLO-based Separation of Touching-Pigs for Smart Pig Farm Applications

Jihyun Seo*, Jaewon Sa*, Yunchang Choi*, Yongwha Chung*, Daihee Park*, Hakjae Kim**

*Dept. of Computer Convergence Software, Korea University, Sejong, South Korea

**Class Act Co., Ltd., Seoul, South Korea

goyangi100@korea.ac.kr, sjwon92@korea.ac.kr, ycc4477@korea.ac.kr, ychungy@korea.ac.kr, dhpark@korea.ac.kr, krunivs@gmail.com

Abstract—For specific livestock such as pigs in a pigsty, many surveillance applications have been reported to consider their health for efficient livestock management. For pig surveillance applications, separating touching-pigs in real-time is an important issue for a final goal of 24-hour tracking of individual pigs. Although convolutional neural network (CNN)-based instance segmentation techniques can be applied to this separation problem, their collective performance of accuracy-time may not satisfy the required performance. In this study, we improve the collective performance of accuracy-time by combining the fastest CNN-based object detection technique (i.e., You Only Look Once, YOLO) with image processing techniques. We first apply image processing techniques to detect touching-pigs by using both infrared and depth information acquired from an Intel RealSense camera, then apply YOLO to separate the touching-pigs. Especially, in using YOLO as an object detector, we consider the target object as the boundary region of the touching-pigs, rather than the individual pigs of the touching-pigs. Finally, we apply image processing techniques to determine the final boundary line from the YOLO output. Our experimental results show that this method is effective to separate touching-pigs in terms of the collective performance of accuracy-time, compared to the recently reported CNN-based instance segmentation technique.

Keywords— Smart farm, Computer Vision, Touching-Pigs Separation, Convolutional Neural Network, YOLO

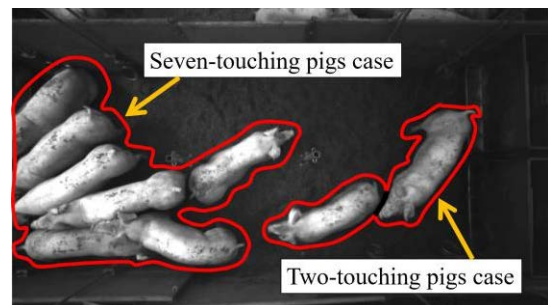
I. INTRODUCTION

Early detection and management of problems related to health and welfare is important when caring for group-housed livestock. In particular, the care of individual animals is necessary to minimize potential damage caused by infectious disease or other health and welfare problems. However, it is almost impossible for a small number of farm workers to care for individual animals in a large livestock farm.

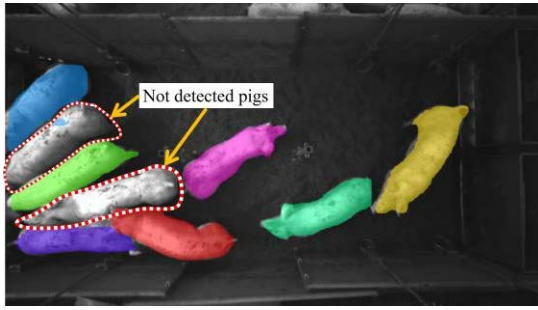
Several studies have recently used surveillance techniques to automatically monitor livestock [1]-[4]. In this study, we focus on video-based pig monitoring applications with an Intel RealSense camera [5]. Although the top-view sensor

can provide both infrared and depth information, the depth information acquired from the low-cost sensor is inaccurate, compared to the infrared information. Therefore, we need to apply image processing techniques carefully in using both infrared and depth information in order to separate touching-pigs in a crowded room accurately.

Recently, many deep learning-based techniques [6] have been reported to solve various computer vision applications, such as image classification [7], object detection [8], semantic segmentation [9], and even instance segmentation [10]. In fact, segmentation is one of the most challenging problems in computer vision applications. Especially, because the appearance of each pig in the crowded pig room is similar, the accuracy of instance segmentation with current segmentation techniques is not satisfied. For example, Mask R-CNN [10] has been recently proposed as a CNN-based instance segmentation technique, and widely used for various computer vision applications [11]-[13]. As shown in Figure 1, however, the accuracy of segmentation with Mask R-CNN may not be satisfied (i.e., some pigs are totally missed). Furthermore, even with GPUs, the execution time of Mask R-CNN cannot satisfy the real-time requirement of 30 frames per second (fps), which is very important in 24-hour continuous pig monitoring. Finally, preparing the learning data for instance segmentation is very time consuming because we need to specify all the boundary pixels of touching-pigs.



(a) The result of semantic segmentation (i.e., pig detection)



(b) The result of instance segmentation (i.e., individual pig detection) with Mask R-CNN [10]

Figure 1. Illustration of segmentation results for pig room monitoring

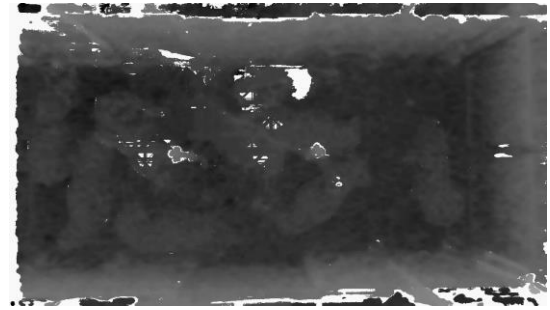
In this study, we propose an *accurate* and *real-time* method to separate touching-pigs. We first apply image processing techniques to detect touching-pigs by using both infrared and depth information acquired from an Intel RealSense camera, then apply YOLO (i.e., one of the fastest CNN-based object detection techniques) [8] to separate the touching-pigs. Especially, in using YOLO as an object detector, we consider the *target object* to be detected as the “boundary region of the touching-pigs,” rather than the “individual pigs of the touching-pigs.” Also, compared to “instance segmentation,” *preparing the learning data* for YOLO-based “object detection” (i.e., specifying only the bounding box of each touching boundary) is less time consuming, and thus we can prepare more learning data for YOLO under the same preparation time. Finally, we apply image processing techniques to determine the final boundary line from the YOLO output.

In the following, we use the term *segmentation* and *separation* interchangeably, and the remainder of this paper is structured as follows. Section 2 explains background of this study, and Section 3 describes the proposed method to separate touching-pigs in the crowded pig room. The experimental results are presented in Section 4, and the conclusions are presented in Section 5.

II. BACKGROUND

The present study contributes to our final goal of 24-h automatic pig behavior analysis by focusing on identifying individual pigs based on pig separation. Previous studies performed segmenting and tracking [14], [15], but the mean times between tracking failures were less than a few minutes. Clearly, typical tracking algorithms can segment and track each isolated moving pig correctly. When multiple moving pigs are very close to each other (i.e., we designate these adjacent “moving” pigs as **touching-pigs**), and the tracking algorithms cannot identify each pig, and thus tracking failures occur. Furthermore, the higher the pig room density, the more difficult it will be to segment the pigs in the room [16]. Therefore, the key problem when segmenting and

tracking individual pigs continuously during automatic analysis of pig behavior is to separate touching-pigs in a crowded environment without ID switch.



(a) Depth image



(b) Infrared image



(c) Result of Otsu [18] from the depth image



(d) Result of Otsu [18] from the infrared image

Figure 2. Difficulty of separating touching-pigs obtained from an Intel RealSense [5]

As explained in Section 1, we consider a low-cost depth camera Intel RealSense. However, the accuracy of the depth data measured by the low cost camera decreases quadratically as the distance increases [17]. Also, it is not easy to detect pigs accurately from the infrared image due to light fluctuations, cluttered backgrounds, varying surface status of each pig caused by manure, etc. Considering these difficulties, it is very challenging to separate individual pigs using the low-cost camera.

For example, Figure 2a and Figure 2b respectively show a depth image and an infrared image acquired from an Intel RealSense camera. We can clearly see the difference between the two gray images. Also, the results of binarization such as Otsu [18] cannot be used directly for pig detection, as shown in Figure 2c and Figure 2d. By carefully utilizing both images as well as balancing image processing techniques and the fastest CNN-based object detection technique (i.e., YOLO), we can separate touching-pigs accurately in real-time. To the best of our knowledge, this is the first YOLO report on considering the target object to be detected as the “boundary region of a group object”, rather than the “individual objects of a group object.”

III. PROPOSED APPROACH

In this study, we assume that the size of each pig is similar. We first detect each pig by using the depth information, and then each moving pig is detected using frame difference method. By analyzing the size of each connected component of the frame difference result, we can recognize each connected component as a single pig or a group of pigs, and we try to separate the adjacent moving pigs (i.e., touching-pigs) for high-level vision tasks. It is well known that each pig sleeps most of the time, and the moving activity of each pig (measured with frame difference) was observed with a probability of less than 1% on the average [3]. Algorithm 1 displays the overall procedure of the proposed method. The details of the preprocessing steps for the depth images (e.g., noise removal and background subtraction) can be found in [4].

Our segmentation method consists of two modules (YOLO Processing Module and Image Processing Module), and Figure 3 presents an overview of our method for touching-pigs detected using the RealSense camera. We first apply YOLO to the touching-pigs to obtain the bounding boxes (BBs) from the YOLO Processing Module. Although YOLO has been widely used for detecting objects, we consider the target object to be detected as the “boundary region of a group object,” rather than the “individual objects of a group object.” That is, we train YOLO to detect the boundary region of each touching-pig. Once YOLO generate a boundary BB for each touching-pig at test time, we conduct

final segmentation within the boundary BB using the Image Processing Module.

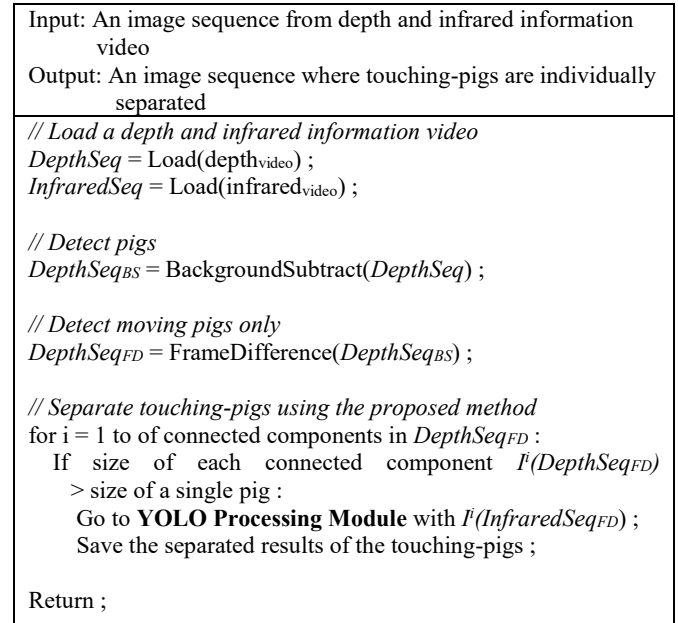


Figure 3. Overall algorithm with the proposed method

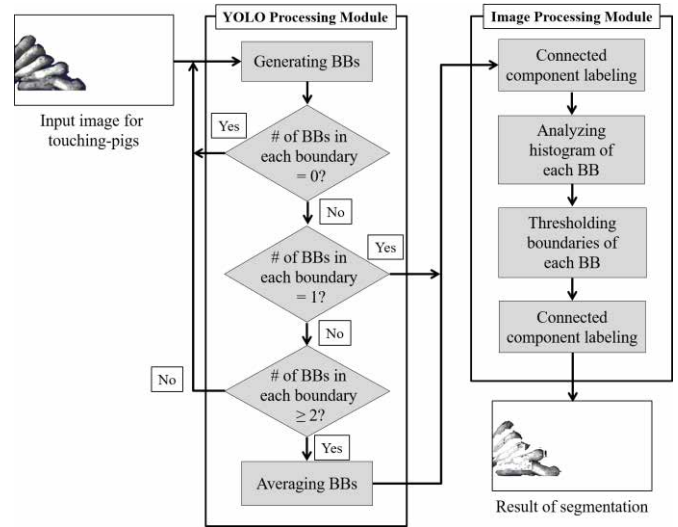


Figure 4. Overview of the proposed modules

A. YOLO Processing Module

YOLO is a CNN-based object detection technique, and it uses the grid method, which allows efficient object detection in real time. In particular, this study uses the YOLO9000 technique. The input image contains only touching-pigs, as we assumed earlier, and it is applied to YOLO. Then, we can obtain the image with bounding boxes (BBs) for the possible boundary region of each touching-pig. Note that, we assume

YOLO was trained with images with ground-truth based BBs for the boundary region of each touching-pig, rather than the individual pig of each touching-pig.

In YOLO, the input image is equally and arbitrarily divided into set $S \times S$ grid cells, and B BBs are generated through each grid cell with x coordinate, y coordinate, width, height and probabilities of existence of each object (i.e., a boundary region of a touching-pig). In other words, through this step, each cell can generate a large number of BBs with object probability information that is higher than the threshold value (denoted as *BB_probability_th*). Also, YOLO is well known is one of the fastest techniques for object detection. In this study, we set S , B , and *BB_probability_th* as 13, 5, and 1, respectively.

After generating BBs, we need to evaluate the generated BBs. For example, if one BB encloses another BB, then the two BBs are averaged. Also, if no BB is generated for a touching-pig, then YOLO is applied again with lower *BB_probability_th*. This step is repeated until at least one BB is generated for the touching-pigs.

B. Image Processing Module

In the Image Processing Module, we separate the inner pixels of BB boxes were generated from YOLO. The pixels of the boundaries between touching-pigs are slightly darker compared to other parts of the pigs such as their body. With these properties of the pigs, thus, the touching-pigs can be separated by using histogram analysis to correctly find the boundaries. In addition, the touching-pigs can also be separated precisely through the number of pixel labels from the connected component analysis.

IV. EXPERIMENTAL RESULTS

A. Experimental Environment and Dataset

In our experiment, the training for the pig detection in the pig room with YOLO was performed in the following environment: Intel Core i7-7700K 4.20 GHz, 32 GB RAM, Ubuntu 16.04.2 LTS, NVIDIA GeForce GTX1080 Ti 11 GB VRAM, and OpenCV 3.2 [19]. Furthermore, the test to separate the touching-pigs through both image processing module and the detection model trained from YOLO was derived in the same environment.

In order to collect both depth and infrared sequences in a pig room, we first installed an Intel RealSense D435 camera on a ceiling above 3.2 m from the floor in a 4.8 m $\times$ 2.0 m pig room located in Chungbuk National University, Korea. Hereafter, we collected both 24-h depth and infrared sequences in the pig room where nine pigs were raised. The video sequences had properties of a resolution of 1280 $\times$ 720 and 30 fps, respectively. Here, we extracted one minute video sequence from both the depth and infrared video sequence, in which the most vigorous activities of the pigs were recorded

in daytime because of the appearance of a person for feeding them. We used the one minute sequence to test separation of touching-pigs.

Using these video sequences, we first conducted background subtraction for detecting foreground (i.e., pigs). Then, moving pigs in the pig room were detected by using frame difference method and only touching moving pigs were extracted through a threshold according to the detected pigs' sizes. Finally, we obtained 1900 touching-pig cases from 6150 video frames, and these frames were used to train the touching cases with YOLO, where the parameter of training was set as follows: learning rate of 0.001, weight decay of 0.0005, momentum of 0.9, and activation function of leaky ReLU with 6150 images.

For comparison of the separation performance of our proposed method, we trained Mask R-CNN [10] for separation of the touching-pigs with 954 pig cases from 106 video frames. Note that, compared to YOLO-based object detection, it was more time consuming to prepare the learning data for Mask R-CNN-based instance segmenter because Mask R-CNN required to specify all the bounding pixels of each pig. Therefore, we could prepare less learning data for Mask R-CNN, compared to YOLO. We set the parameters to train the touching-pigs frames for separation with Mask R-CNN as follows: learning rate of 0.0001, weight decay of 0.0001, and momentum of 0.9 with 106 frames.

B. Evaluation of Separation Performance

From the 1900 touching-pig cases, we increased the training frames by conducting additional augmentation, such as rotating objects in each frame, to effectively learn YOLO model. Then, we compared the separation performance of our proposed approach with Mask R-CNN with 120 frames extracted from the one minute sequence. Through our proposed approach, we confirmed that detecting a boundary region of a group object was better for separating touching-pigs rather than detecting individual objects of a group object. For example, when the pigs adjoined horizontally, detecting individual objects using YOLO poorly separated the touching-pigs because of a long boundary line. On the other hand, detecting a boundary region of a group of object (i.e., our proposed approach) could separate more precisely the touching-pigs because it detected the long boundary on the touching-pigs.

We compared the separation performance of our proposed approach with Mask R-CNN. The separation results of some touching-pigs' cases are shown in Figure 4.

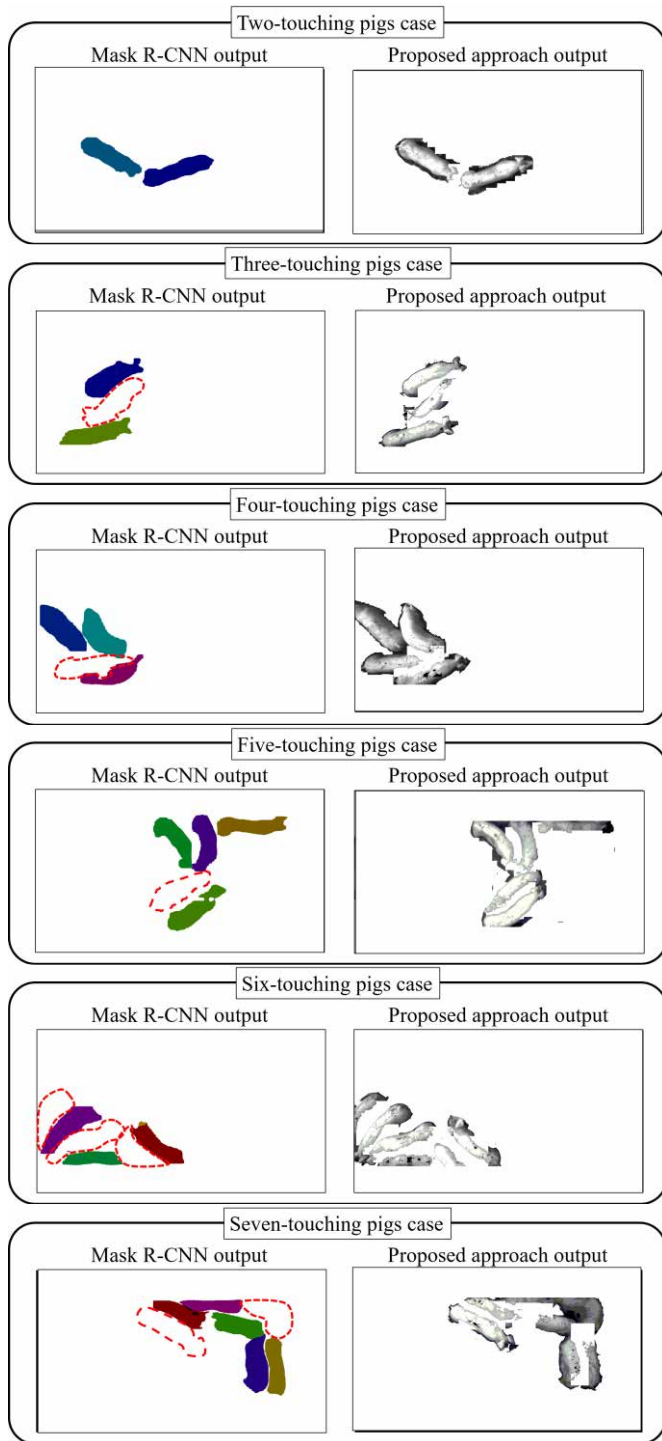


Figure 5. Comparison of separation results with Mask R-CNN [10] (undetected pigs were shown as dashed lines.) and the proposed method

For evaluation of the separation performance for each approach, both separation approaches were compared with ground-truth in “pixel-level”. We defined the separation

accuracy (denoted as SA) as follows: G denoted the ground-truth area (i.e., an area of a pig marked on the ground-truth), F denoted the false area (i.e., an area of a pig that did not belong to the ground-truth), and C denoted the correct area (i.e., an area of a pig which was identified correctly). Then, we also formulated $SA = C / (G + F)$ and respectively measured the accuracy each separation approach for touching-pigs. Because of omission of touching pigs during separation and a small amount of data, it exerted a great influence on accuracy. Compared with the Mask R-CNN-based segmenter, our proposed method separated the touching-pigs well with the average accuracy of 85%. Furthermore, we confirmed that the execution time of our proposed method was faster more than nine times compared to the time of Mask R-CNN-based segmenter.

Especially, for the final goal of 24-hour continuous monitoring, ID switch is the most critical factor. For all the test cases of touching-pigs, the proposed method could separate all touching-pigs without any ID switch. Furthermore, the high-speed execution of the segmentation can have a better chance in producing a complete vision system for higher-level analysis in real-time.

TABLE 1. COMPARISON OF ACCURACY AND EXECUTION TIME (MASK: MASK R-CNN[12])

Case	# Cases		Acc. (Avg.%)	Time (Avg. msec)	Acc./Time
Two-touching pigs	75	Mask	88.33	253.18	0.35
		Proposed	87.39	26.19	3.34
Three-touching pigs	35	Mask	81.59	253.88	0.32
		Proposed	78.52	28.76	2.73
Four-touching pigs	10	Mask	97.42	254.83	0.38
		Proposed	93.05	27.38	3.40
Five-touching pigs	11	Mask	94.75	255.11	0.37
		Proposed	91.57	28.47	3.22
Six-touching pigs	8	Mask	77.57	253.89	0.31
		Proposed	75.14	28.87	2.60
Seven-touching pigs	3	Mask	84.39	255.34	0.33
		Proposed	81.30	29.25	2.78
Total	142	Mask	87.34	254.37	0.34
		Proposed	85.00	28.15	3.02

V. CONCLUSIONS

The touching-pigs separation in a surveillance camera environment is an important issue to efficiently manage pig farms. Even with the recently reported CNN-based instance segmentation techniques, however, touching-pigs cannot be separated accurately in real-time, due to the complicated patterns of touching-pigs having similar appearance.

In this study, we focused on separating touching-pigs in real-time using both the fastest CNN-based object detection technique and image processing techniques, in order to analyze individual pigs with an ultimate goal to achieve 24-hour continuous monitoring. The touching-pigs were

detected by using image processing techniques with both infrared and depth information acquired from an Intel RealSense camera. Then, YOLO generated bounding boxes for each touching pattern. Especially, in using YOLO as an object detector, we considered the target object as the boundary region of the touching-pigs, rather than the individual pigs of the touching-pigs. Finally, we applied image processing techniques to determine the final boundary line from the YOLO output

Based on the experimental results for 120 touching-pigs frames over one minute obtained during the daytime (i.e., when most pigs were moving), we confirmed that the collective performance of accuracy-time of the proposed method was much better (i.e., more than eight times) than the recently reported CNN-based instance segmentation technique (i.e., Mask R-CNN [10]). By carefully balancing the tradeoffs between the computational workload and accuracy, we could propose a light-weight method with an acceptable accuracy for real-time video monitoring.

ACKNOWLEDGMENT

This research was supported by the Basic Science Research Program through the NRF funded by the MEST (2018R1D1A1A09081924) and the Leading Human Resource Training Program of Regional Neo Industry through the NRF funded by the MSIP (2016H1D5A1910730).

REFERENCES

- [1] T. Banhazi, H. Lehr, J. Black, H. Crabtree, P. Schofield, M. Tschärke, and D. Berckmans, "Precision Livestock Farming: An International Review of Scientific and Commercial Aspects," *International Journal of Agriculture and Biological Engineering*, vol. 5, pp. 1–9, 2012.
- [2] M. Tschärke and T. Banhazi, "A Brief Review of the Application of Machine Vision in Livestock Behaviour Analysis," *Journal of Agricultural Informatics*, vol. 7, pp. 23–42, 2016.
- [3] Y. Chung, H. Kim, H. Lee, D. Park, T. Jeon, and H. Chang, "A Cost-Effective Pigsty Monitoring System based on a Video Sensor," *KSII Trans. Internet Inf. Sys.*, vol. 8, pp. 1481–1498, 2014.
- [4] J. Kim, Y. Chung, Y. Choi, J. Sa, H. Kim, Y. Chung, D. Park, and H. Kim, "Depth-based Detection of Standing-Pigs in Moving Noise Environments," *Sensors*, vol. 17, pp. 2757, 2017.
- [5] *Intel RealSense D400 Series (D435)*, Intel, 2018.
- [6] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent Advances in Convolutional Neural Networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.
- [7] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.
- [8] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *arXiv*, arXiv:1612.08242, 2016.
- [9] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.
- [10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.
- [11] Y. Zhang, Y. Zhang, S. Li, and J. Zhang, "Accurate Detection of Berthing Ship Target based on Mask R-CNN," in *Proc. of the 2018 International Conference on Image and Video Processing, and Artificial Intelligence*, vol. 10836, pp. 1083602, 2018.
- [12] T. Zhao, Y. Yang, H. Niu, D. Wang, and Y. Chen, "Comparing U-Net Convolutional Network with Mask R-CNN in the Performances of Pomegranate Tree Canopy Segmentation," in *Multispectral, Hyperspectral, and Ultraspectral Remote Sensing Technology, Techniques and Applications VII*, vol. 10780, pp. 107801J, 2018.
- [13] R. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense Human Pose Estimation in the Wild," *arXiv*, arXiv:1802.00434, 2018.
- [14] P. Ahrendt, T. Gregersen, and H. Karstoft, "Development of a Real-Time Computer Vision System for Tracking Loose-Housed Pigs," *Computers and Electronics in Agriculture*, vol. 76, pp. 169–174, 2011.
- [15] S. Zuo, L. Jin, Y. Chung, and D. Park, "An Index Algorithm for Tracking Pigs in Pigsty," in *Proc. of ICITMS*, pp. 797–803, 2014.
- [16] M. Kashiha, C. Bahr, S. Ott, C. Moons, T. Niewold, F. Tuytens, and D. Berckmans, "Automatic Monitoring of Pig Locomotion using Image Analysis," *Livestock Science*, vol. 159, pp. 141–148, 2014.
- [17] T. Mallick, P. Das, and A. Majumdar, "Characterization of Noise in Kinect Depth Images: A Review," *IEEE Sensors Journal*, vol. 14, pp. 1731–1740, 2014.
- [18] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans. on Sys. Man, and Cybernetics*, vol. 9, pp. 62–66, 1979.
- [19] Open Source Computer Vision, OpenCV. [Online]. Available: <http://opencv.org>. (accessed on 18 Dec. 2016)



Jihyun Seo(Korea) is currently a Bachelor student in the Dept. of Computer Convergence Software at Korea University, Korea.



Jaewon Sa(Korea) is currently a PhD student in the Dept. of Computer Convergence Software at Korea University. He received the BS and MS degrees in computer science from Korea University, Korea, in 2015 and 2017, respectively.



Yunchang Choi(Korea) is currently a Master student in the Dept. of Computer Convergence Software at Korea University. He received the BS degree in computer science from Korea University, Korea in 2015.



Yongwha Chung(Korea) received his BS and MS degree in Electronic Engineering from Hanyang University, Korea, in 1984, and his PhD degree in computer science from the University of Southern California, USA, in 1997. He joined Korea University in 2003, where he is currently a professor in the Dept. of Computer Convergence Software.



Daihee Park(Korea) received his BS degree in mathematics from Korea University, Korea, in 1982, and his PhD degree in computer science from the Florida State University, USA, in 1992. He joined Korea University in 1993, where he is currently a professor in the Dept. of Computer Convergence Software.



Hakjae Kim(Korea) received his BS and MS degree in computer science from Korea University, Korea in 2007 and 2009. He was a researcher of the team of a study of serverplatform in ETRI from 2009 to 2012. He is currently a CEO of Classact Inc.