Implementation of Deep-learning Algorithm for Obstacle Detection and Collision Avoidance for Robotic Harvester

Li Yang[a], Iida Michihisa[a,*], Suyama Tomoya[a], Suguri Masahiko[a], Masuda Ryohei[a]

[a] Graduate School of Agriculture, Kyoto University, Kitashirakawa Oiwake-cho, Sakyo-ku, Kyoto, 606-8502, Japan

Abstract: Convolutional neural networks (CNNs) are the current state of the art systems in image semantic segmentation (SS). However, because it requires a large computational cost, it is not suitable for running on embedded devices, such as on rice combine harvesters. In order to detect and identify the surrounding environment for a rice combine harvester in real time, a neural network using Network Slimming to reduce the network model size, which takes wide neural networks as the input model, yielding a compact model (hereafter referred to as "pruned model") with comparable accuracy, was applied based on an image cascade network (ICNet). Network Slimming performs channel-level sparsity of convolutional layers in the ICNet by imposing L1 regularization on channel scaling factors with the corresponding batch normalization layer, which removes less informative feature channels in the convolutional layers to obtain a more compact model. Then each of the pruned models were evaluated by mean intersection over union (IoU) on the test set. When the compaction ratio is 80 %, it gives a 97.4 % reduction of model volume size, running 1.33 times faster with comparable accuracy as the original model. The results showed that when the compaction ratio is less than 80 %, a more efficient (less computational cost) model with a slightly reduced accuracy in comparison to the original model was achieved. Field tests were conducted with the pruned model (80 % compaction ratio) to verify the performance of obstacle detection. Results showed that the average success rate of collision avoidance was 96.6% at an average processing speed of 32.2

* Corresponding author. Tel.: +81-75-753-6166. E-mail address: iida@elam.kais.kyoto-u.ac.jp.

FPS (31.1 ms per frame) with an image size of 640 ×480 pixels on a Jetson Xavier. It shows that the pruned model can be used for obstacle detection and collision avoidance in robotic harvesters.

Keywords: robotic combine harvester; deep learning; human detection; image cascade network; network slimming.

1 Introduction

To ensure the safety and precision operation of autonomous combine harvesters it is important to identify obstacles quickly and accurately in the surrounding paddy. When a combine is working in a paddy, it should avoid colliding with paddy field ridges and humans, and it should also go along the navigation line between harvested and unharvested areas. Our laboratory has developed algorithms to determine the path between harvested and unharvested areas (Cho et al., 2014a; Cho et al., 2014b), the identification of ridges (Takagaki et al., 2013), and the detection of humans in the field (Hisae et al., 2017). However, the paddy field environment is complex, and with many different objects (the harvested area, unharvested area, ridges, and humans) need to be detected simultaneously. Traditional image recognition methods are based on hand-crafted features, such as HOG, LBP and Haar features (Yao et al, 2015, Singh, et al., 2015, Cabrera et al., 2011). Since it is tedious to design features manually and susceptible to the effects of light, vibration, and dust, it is difficult to mine deep-feature information and obtain accurate results. One approach to addresses these challenges is semantic segmentation (SS), which can realize pixel-by-pixel identification in an image.

Recently, SS has become a popular approach for a variety of computer vision tasks in agriculture. For example, Yang et al. (2017) employed a SS method to recognize lactating sows. Milioto et al. (2018) proposed a SS model for crop and weed. McCool et al. (2017) proposed an approach for training SS that can be used to derive compact models for robotic platforms.

These research results indicate that SS can be used to achieve good results for processing agricultural images (Kamilaris et al., 2018), thereby reducing manual preprocessing and subsequent processing to obtain the final segmentation result directly from the original input image (Tang et al., 2016). However, the large computational cost of SS models still makes it difficult to apply to embedded devices in real-time. Our objective is to make a SS model compact to implement for embedded devices and apply it for obstacle detection of robotic combine harvester.

To achieve this objective, on the one hand, many scholars have proposed different real-time SS models. For example, Yu (Yu et al., 2018) proposed a bilateral segmentation network, which used affluent spatial details and large receptive field to improve the speed and accuracy of SS. Wang (Wang et al., 2019) designed an asymmetric encoder-decoder architecture for SS. Zhao (Zhao et al., 2018) proposed ICNet, which uses an image cascade to speed up the SS method. On the other hand, many methods to compress large CNNs have been developed for fast inference. These include low-rank approximation (Denton et al., 2014), network quantization (Chen et al., 2015; He et al., 2015) and binarization (Rastegari et al., 2016; Courbariaux et al., 2016), weight pruning (Han et al., 2015), dynamic inference (Huang et al., 2017), etc. Network Slimming is a simple yet effective compaction approach (Liu et al., 2017), and more importantly, it is convenient to obtain the pruned model just by modifying the number of corresponding channels in the configuration files.

Considering the speed and accuracy in the CamVid (Atlas, 2018), the network used for rice field images were based on ICNet. This method incorporates effective strategies to accelerate network inference speed without sacrificing much performance (Zhao et al., 2018). In this study, a ICNet that maintains a high accuracy was trained first with paddy field images. Paddy field image are, however, not common in public data sets, such as the CamVid Dataset (Brostow et al., 2009). When the ICNet that performs well with public datasets is applied to paddy field images, high segmentation accuracy was obtained. Since the network was designed manually, the importance of each component in the network cannot be determined before training. During training, it could learn the importance of each component through adjusting the weights in trainable layers automatically. After training, some connections and computations in the model would become redundant or non-critical (Ye et al., 2018). Consequently, the redundant or non-critical connections and computations in the network can be removed without significant degradation in performance (Ye et al., 2018). Based on this assumption, we removed these redundant parameters in the model while ensuring similar accuracy, thereby increasing the speed of the model.

Since Network Slimming method is a simple yet effective compaction approach (Liu et al., 2017), the pruned SS models were obtained based on this method in the convolutional layers of ICNet. To this end, we enforced channel-level sparsity of convolutional layers by imposing L1 regularization on channel scaling factors γ in batch normalization (BN) layer (the latter in formula (3)), then removed the less informative channels in the convolutional layers, which correspond to the small γ to obtain the pruned models. The models and methods were introduced firstly in Section 2; then the pruned models were evaluated on test dataset and in the field. Then the results and discussion were presented in Section 4. Finally, we made a conclusion in Section 5.

2 Materials and Methods

2.1 Semantic segmentation model

In order to achieve SS in real-time, ICNet was used for paddy field images in this study, and its structure is shown in Fig. 1. In this figure, numbers in parentheses are feature map size ratios to the full-resolution input (640 $\times$480 pixels). Operations are highlighted in brackets. The final $\times$4 upsampling in the bottom branch is only used during testing. The ICNet takes cascade image inputs (i.e., medium- and high-resolution images), and it adopts a pyramid pooling module (PPM) and cascade feature fusion (CFF) unit in Fig. 2. It was trained with cascade label guidance. Different-scale (e.g., 1/16, 1/8, and 1/4) ground truth labels were utilized to guide the learning stage of low, medium and high-resolution input.

As shown in Fig. 2a, the PPM fuses four different pyramid scale features, and 'POOL' means pooling layer in the figure. First, it separates the feature map into different sub-regions by using an operation called adaptive average pool. Then upsampling the low-dimension feature maps to get the same size feature as the original feature map via bilinear interpolation. Finally, different levels of the features are summed as a final pyramid pooling global feature. To combine cascade features from different resolution inputs, 2 CFF units were used in the ICNet. Details of the structure is shown in Fig. 2b, the sizes of feature maps F1 and F2 are C1 × H1 × W1 and C2 × H2 × W2, respectively, and the resolution of the label is 1 × H2 × W2, where H2 = 2 × H1. It combines feature maps F1 and F2. In order to enhance the learning of F1, auxiliary label on the upsampled feature of F1 is applied.

2.2 Network Slimming algorithm

The algorithm used in this paper to prune the network model was based on the principle of Network Slimming method (Liu et al., 2017). The method could remove the less important connections with small weights in each convolution layer. As we know that, batch normalization (BN) layer performs the following transformation after each convolution layer in the model:

$$\hat{z} = \frac{z_{in} - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \tag{1}$$

$$z_{out} = \gamma\hat{z} + \beta \tag{2}$$

where z_{in} and z_{out} are the input and output of a BN layer, μ_B and σ_B are the mean and standard deviation values of input activations over B, B denotes the current minibatch, γ and β are trainable affine transformation parameters (scale and shift).

As γ in BN layers corresponds to a specific convolutional channel, γ was used for channel scaling factors. The approach imposes L1 regularization (the latter part in formula (3)) on the channel scaling factors γ in BN layers for each channel. Pushing the values of channel scaling factors towards zero with L1 regularization enables insignificant channels to be identified. The network weights and these channel scaling factors were trained with sparsity regularization (the latter part in formula (3)). The training objective of our approach is given by

$$L = \sum_{(x,y)} l(f(x,W),y) + \lambda \sum_{\gamma \in T} g(\gamma) \tag{3}$$

where (x, y) denotes the training input and target, W denotes the trainable weights, the first sum-term corresponds to the normal training loss of a CNN, T denotes the gradient of each convolution layer, $g(\cdot)$ is a sparsity-induced penalty on the channel scaling factors, and λ balances the two terms. In our experiment, we chose $g(s)=|s|$, which is known as L1-norm and widely used to achieve sparsity. Subgradient descent was adopted as the optimization method for the non-smooth L1 penalty term. The channel scaling factors act as the agents for channel selection. As they were jointly optimized with the network weights, the network can automatically identify insignificant channels, which can be safely removed without greatly affecting the generalization performance. Channels with small factors γ removed (all their incoming and outgoing connections), then we could get the pruned network model.

2.3 Dataset for semantic segmentation models

All the images in the training set, validation set and test set were derived from experimental videos from December 2016 to August 2019. A detailed description of the above data set is shown in Table 1, which includes sample number, rice variety, field type, weather, camera angle, camera depression angle, etc. Since some of the scenes in the video are not related to the field scene, and sometimes some areas in the image are not clear enough, so some clear images of appropriate size were cut out from the original images. Then the cut images were rotated (±15°), and flipped horizontally. Finally, a total of 5000 images (jpeg format) were obtained. The size of all images was 640×480 pixels, and the mean value of the RGB channels of the images were 0.485, 0.456, and 0.406, and the standard deviation were 0.229, 0.224 and 0.225, respectively, when these images were transformed to the range of [0, 1]. According to a previous field trial video, a training set and test set were prepared, including 4,000 and 1,000 images, respectively, which were selected up from the data set of 5,000 images mentioned before; Then the data was normalized to reduce the negative effects of uneven brightness.

2.4 The procedure of getting pruned (segmentation) model

During the training process, a stochastic gradient descent method was used for backward propagation of the learning phase to obtain the best network parameters. The initial learning rate was 0.02, and the decay coefficient of the learning rate was 0.5. The decay frequency was 10 epochs, with the batch size of 4. The regularization parameter λ was 0.0001, with a penalty factor 0.0001 to perform channel-level sparsity regularization. When the current loss function converged and stabilized, training was halted.

After sparsity training, we removed channels with a global threshold γ1 across all layers except for CFF, which was defined as a certain percentile of all the needed scaling factor values. Such as 5-th percentile, corresponding to a 5 % compaction ratio. Then the compaction ratio is defined as

$$\text{compaction ratio} = \frac{C_1}{C_2} \times 100\% \tag{4}$$

where C_1 is the numbers of removed channels, C_2 is the numbers of channels in the original network. Two convolution layers before the 'sum' operation in the CFF unit were required to have the same channel number. To match the feature channels of the 2 layers, we iterate through the layers and perform the same percentile compaction operation to generate a pruning mask for these connected layers, respectively. The percentile is same as the percentile used for the global threshold $\gamma 1$.

The channel pruning procedure is shown in Fig. 3. ICNet was initially trained with channel-level sparsity regularization; sequentially, pruned ICNet was obtained by pruning feature channels to a certain ratio according to their scaling factors in the ICNet; After channel pruning, a fine-tuning operation was performed on pruned models to compensate potentially temporary degradation in segmentation accuracy. The Network Slimming (training with sparsity regularization, pruning, and fine-tuning) was repeated several times. The model was pruned 10 % each time. In our experiments, we directly retrain using the same training hyper-parameters as the initially training of ICNet.

2.5 Experimental condition in field test

In order to evaluate the performance of the pruned model, the pruned model which with the compaction ratio of 80 % was used in the field test for human detection. Tests were conducted in actual paddies, the places were in Kisosaki, Kuwana District, Mie, Japan (35°05'19.9"N, on Aug. 24-25, 2019 and Nantan City, Kyoto, Japan (35°02'35.4"N, 136°46'16.8"E), on Sep. 22, 2019. The weather was sunny on Aug. 24-25, 2019 and cloudy on Sep. 22, 2019. Fig. 4 shows the main devices used in this study. The base machine was a four-row head-feeding combine harvester ER470 (Kubota, Osaka, Japan). Our laboratory has developed an autonomous harvesting system based on ER470 (Iida et al., 2017), which could follow a target path based on its absolute position and orientation, planning a counterclockwise spiral path in a rectangular paddy field. An Intel RealSense D435 (Intel, Santa Clara, USA) camera was mounted on the front of the harvester to capture color images and depth data in real-time. It was mounted at a height of 1.65 m above the ground, and with its lens facing down at an angle of 28 ° to the horizontal. A Jetson Xavier (NVIDIA, Santa Clara, USA) was used for running segmentation models. Light levels were measured under different conditions with a digital light meter KG-75 (Kaise, Nagano, Japan).

Since the pruned models could segment 5 classes (harvested area, unharvested area, ridges area, human and background) and the combine could harvest rice automatically, the test was conducted in this way. When the combine automatically harvested along the target path at a speed of 1.0 m/s, a human would appear at different times on the target path in front of the combine at different distances. In these conditions based on segmentation results and the distance obtained by the depth camera, the combine took three actions timely, either stopping, slowing down or continuing to work. The principal flow of the test algorithm for automatic rice harvesting is shown in Fig. 5. When the combine begins harvesting, it captures color images and depth data from the D435 camera, then inputs the RGB images into the segmentation model. Based on the segmentation results, if there is a human in the image, it calculates the center of the human area and gets the corresponding distance to the center from the depth data. Then according to the distance between the combine and the human, it sends a control signal to stop, slow down or continue to work, to the combine's electronic control unit through an RS-232 serial port. Two tests were conducted, in Test 1, a human appeared on the target path in front of the combine at different distances. In Test 2, no human appeared on the target path in front of the combine.

3 Results and Discussion

3.1 Comparison of segmentation performance

To evaluate the robustness of the models, each of the pruned models were validated on the test set. Fig. 6 presents the mean intersection over union (IoU) at different compaction ratios.

Based on the data in Fig. 6, the following results were found. As the compaction ratio continues to increase, there is a small loss in the accuracy of the model. In our experiments, the fine-tuned pruned model could even achieve higher accuracy than the original unpruned model in some cases (compaction ratio: 61.4 %). However, when the compaction ratio is greater than 80 %, the accuracy of the model seriously degrades. When the compaction ratio is less than 80 %, compelling results are achieved in comparison to the original counterpart.

When a combine harvester is working in the field, the harvested area, unharvested area, ridge area, and human area occupy different ratios at different stages. Fig. 7 shows the segmentation result for each class at different compaction ratios. It shows that models are more inclined to predict pixels in the image as the harvested area and the unharvested area. This may be due to data imbalance, because the harvested area and unharvested area occupy a bigger part than the ridge area and human area for most images. It shows that when the compaction ratio is less than 80 %, the mean IoU for each class at different compaction ratios is close to the original counterpart. When the compaction ratio is greater than 80 %, the mean IoU for each class decreases quickly.

3.2 Inference run-time performance

All these models were tested with an image size of 640 ×480 pixels, Table 2 shows the frames per second (FPS) on the Jetson Xavier and model volume of different pruned models. Because the accuracy of the model drops sharply when the compaction ratio is greater than 80%, only models with a compaction ratio of less than 80% were measured. The run-times were achieved using CUDA 10.0.117 and cuDNN 7.3.07. As can be seen from Table 2 as the compaction ratio increases, the size of the model volume decreases and the speed of the model increases. When 80 % of the channel been pruned, the model has a 97.4% reduction of model volume size, and ran 1.33 times faster with comparable detection accuracy to the original model. It can be known from segmentation performance and the inference run-time performance that when the compaction ratio is less than 80 %, the Network Slimming method could be used for decreasing the computational cost of the ICNet for field image segmentation.

3.4 Results and discussion in field test

Because the combine harvester traveled counter-clockwise during the harvesting, as shown in Fig. 8, the noise levels in the acquired images differed as the light conditions changed depending on the direction of movement. All of the test scenes were categorized into four scenes (A, B, C, D) according to the direction of harvester movement. In all scenarios, based on the segmentation results of the model, the harvester would take the appropriate action (stop, slow down or continue to work). Table 3 shows the results of Test 1 by using the pruned model. Because a human always appeared on the target path in front of the combine at different distances during Test 1, if the harvester slowed down, stopped and then continued to work, it was regarded as a successful result.

The results show that the average success rate of collision avoidance was 96.6% at an average processing speed of 32.2 FPS (31.1 ms per frame). The evaluation results show that the proposed method is effective for human segmentation and collision avoidance regardless of the movement direction of the combine harvest or the light conditions experienced, as shown in Fig. 9. However, as shown in the last column of Fig. 9, the human is not successful segmentation when the camera is backlighted (dataset B). Because the camera in scene B was in backlight mode, the sunlight affected the image quality obtained by the camera, which reduced the accuracy of model segmentation. Finally, it made the success rate in scene B lower than that in other scenes.

Table 4 shows the result of Test 2. Because no humans appeared on the target path in front of the combine in Test 2, the harvester should continue to work normally, so we focused on the number of false results in this test. If the harvester slowed down or stopped, it was regarded as a false result. The result in Table 4 indicate that the number of false detection was small under various light conditions. However, the segmentation is not successful when the camera is backlighted (first column in Fig. 10), or the shadow of rice is similar to that of a human (second column in Fig. 10).

It can be known from two field tests that when the camera is in a backlight mode or some objects are visually similar to a human in the image, the SS model that only relied on a color image as input still has the probability of false detection. Since thermal images and Lidar data are less affected by light than color images, which could provide additional information for making detection. So, our future work is to fuse the thermal image or Lidar data for further improving the accuracy of detection.

4 Conclusion

1) Network Slimming based on ICNet was proposed and evaluated as a means to compact the semantic segmentation model. It directly imposes sparsity-induced regularization on the scaling factors in batch normalization layers, and unimportant channels in convolutional layers can thus be automatically identified during training.

2) The pruned models, which were achieved through channel pruning of the convolutional layers, substantially decreased the computational cost of ICNet, with a slightly reduction in accuracy. When the compaction ratio is 80 %, it gives a 97.4 % reduction of model volume size, running 1.33 times faster with comparable detection accuracy as the original model.

3) A pruned model (with 80 % compaction ratio) was then tested in the field to validate the feasibility of the method. Results showed that the average success rate of collision avoidance was 96.6% at an average processing speed of 32.2 FPS (31.1 ms per frame) with an image size of 640×480 pixels on a Jetson Xavier. Results demonstrate that with channel reduction of the convolutional layer in the ICNet, a pruned (segmentation) model can be

successfully used in a rice combine harvester for obstacle detection and collision avoidance in real time.

References

Atlas ML. 2018. Real-Time Semantic Segmentation on CamVid. Accessed March 15, 2020. https://paperswithcode.com/sota/real-time-semantic-segmentation-on-camvid.

Brostow, G. J., Fauqueur, J., Cipolla, R. 2009. Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters, 30(2), 88-97.

Cabrera, R.R., Tuytelaars, T., Gool, L.V., 2011. Efficient multi-camera detection, tracking, and identification using a shared set of haar-features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 65-71.

Chen, W., Wilson, J., Tyree, S., et al., 2015. Compressing neural networks with the hashing trick. In International Conference on Machine Learning, pp. 2285-2294.

Cho, W., Iida, M., Suguri, M., et al., 2014a. Vision-based uncut crop edge detection for automated guidance of head-feeding combine. Engineering in Agriculture, Environment and Food, 7(2), 97-102.

Cho, W., Iida, M., Suguri, M., et al., 2014b. Using multiple sensors to detect uncut crop edges for autonomous guidance systems of head-feeding combine harvesters. Engineering in Agriculture, Environment and Food, 7(3), 115-121.

Courbariaux, M., Hubara, I., Soudry, D., et al., 2016. Binarized neural networks: Training deep neural networks with weights and activations constrained to + 1 or -1. arXiv preprint arXiv:1602.02830.

Denton, E. L., Zaremba, W., Bruna, J., et al., 2014. Exploiting linear structure within convolutional networks for efficient evaluation. In Advances in neural information processing systems, pp. 1269-1277.

Huang, G., Chen, D., Li, T., et al., 2017. Multi-scale dense convolutional networks for efficient prediction. arXiv preprint arXiv:1703.09844, 2.

Han, S., Pool, J., Tran, J., et al., 2015. Learning both weights and connections for efficient neural network. In Advances in neural information processing systems. pp. 1135-1143.

He, K., Zhang, X., Ren, S., et al., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision, pp. 1026-1034.

Hisae, T., Masuda, R., Suguri, M., et al., 2017. Human detection in paddy field by image processing (in Japanese). Journal of the Japanese Society of Agricultural Machinery and Food Engineers, 79(1), 49-58.

Iida, M., Harada, S., Sasaki, R., et al., 2017. Multi-combine robot system for rice harvesting operation. ASABE paper No.1700321. DOI:10.13031/aim.201700321

Kamilaris, A., Prenafeta-Boldu, F.X., 2018. Deep learning in agriculture: a survey. Computers and Electronics in Agriculture, 147, 70-90.

Liu, Z., Li, J., Shen, Z., et al., 2017. Learning efficient convolutional networks through Network Slimming. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2736-2744.

McCool, C., Perez, T., Upcroft, B., 2017. Mixtures of lightweight deep convolutional neural networks: applied to agricultural robotics. IEEE Robotics and Automation Letters, 2(3), 1344-1351.

Milioto, A., Lottes, P., Stachniss, C., 2018. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs. IEEE International Conference on Robotics and Automation, Brisbane, Australia, pp. 2229-2235.

Rastegari, M., Ordonez, V., Redmon, J., et al., 2016. Xnor-net: Imagenet classification using binary convolutional neural networks. In European Conference on Computer Vision, pp. 525-542.

Singh, S., Kaur, A., Taqdir, A., 2015. A face recognition technique using local binary pattern method. International Journal of Advanced Research in Computer and Communication Engineering, 4(3), 165-168.

Takagaki, A., Masuda, R., Iida, M., et al., 2013. Image processing for ridge/furrow discrimination for autonomous agricultural vehicles navigation. IFAC Proceedings Volumes, 46(18), 47-51.

Tang, H., Wang, W., Gimpel, K., et al., 2016. End-to-end training approaches for discriminative segmental models. Spoken Language Technology Workshop, pp. 496-502.

Wang, Y., Zhou, Q., Liu, J., et al., 2019. LEDNet: A Lightweight Encoder-Decoder Network for Real-Time Semantic Segmentation. arXiv preprint arXiv:1905.02423.

Yang, A., Xue, Y., Huang, H., et al., 2017. Lactating sow image segmentation based on fully convolutional networks. Transactions of the Chinese Society of Agricultural Engineering, 33(23), 219-225. (in Chinese with English abstract)

Yao, S., Pan, S., Wang, T., et al., 2015. A new pedestrian detection method based on combined HOG and LSS features. Neurocomputing, 151, 1006-1014.

Ye J, Lu X, Lin Z, et al. 2018. Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers[J]. arXiv preprint arXiv:1802.00124.

Yu, C., Wang, J., Peng, C., et al., 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 325-341.

Zhao, H., Qi, X., Shen, X., et al., 2018. Icnet for real-time semantic segmentation on high-resolution images. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 405-420.

Figures

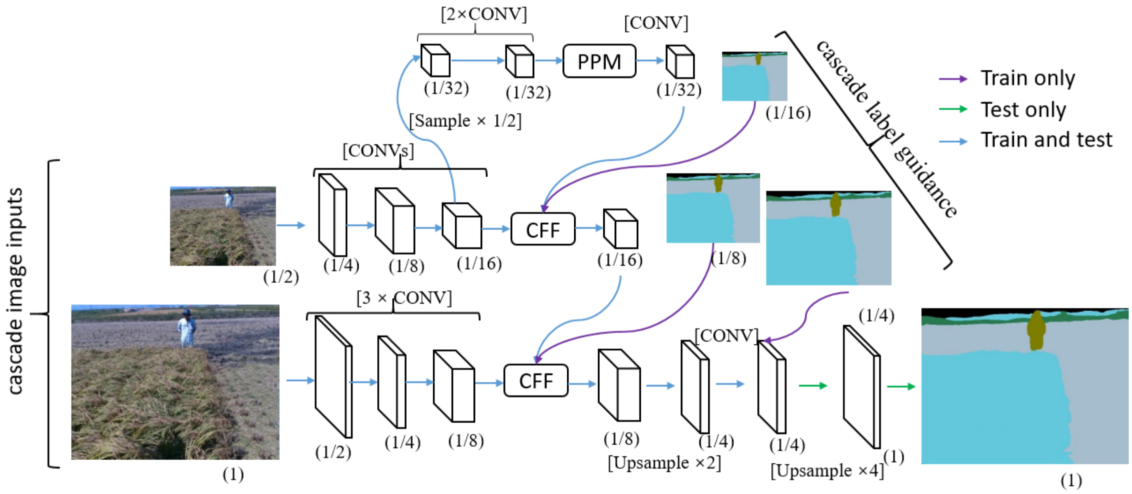
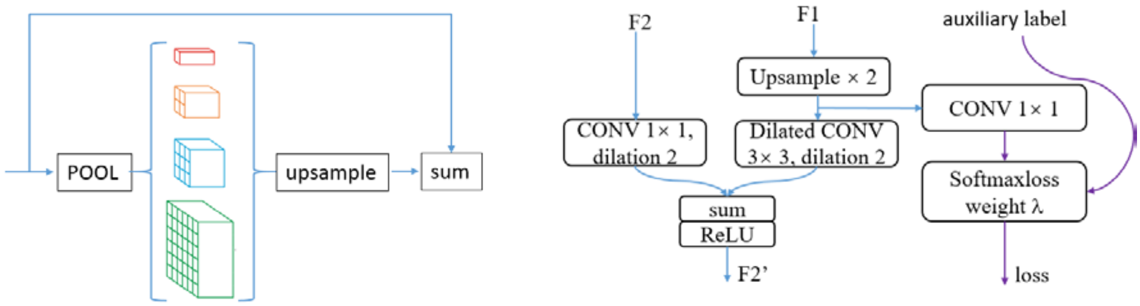


Fig. 1 ICNet structure for image segmentation.



(a) Pyramid pooling module(PPM) (b) Cascade feat fusion unit(CFF)

Fig. 2 Pyramid pooling module (PPM) and cascade feat fusion (CFF) unit in ICNet.

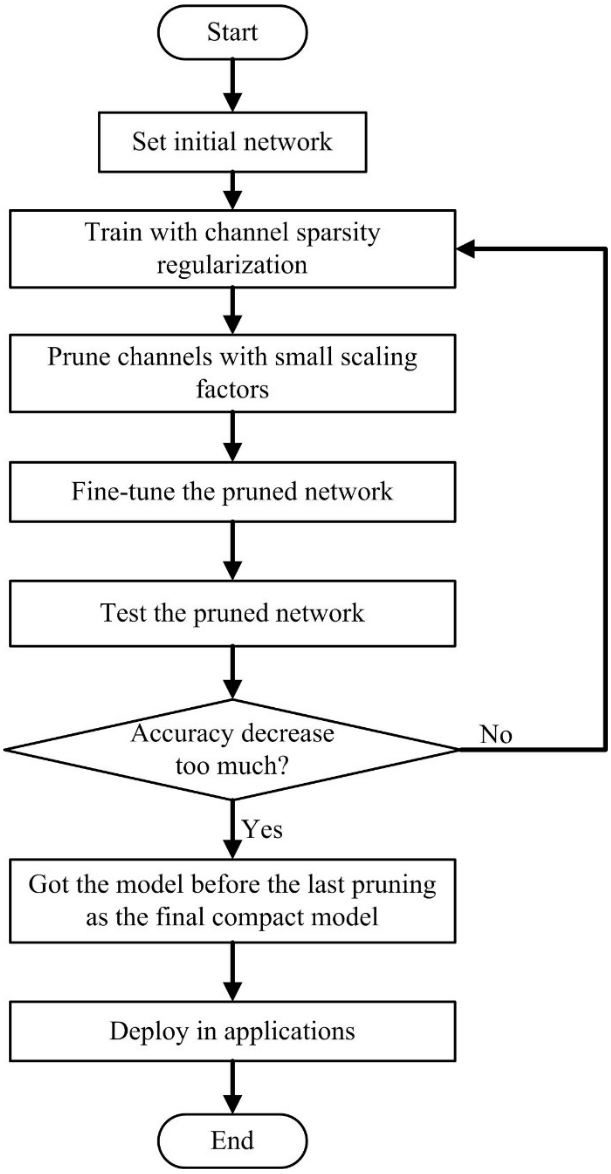


Fig. 3. An iterative procedure of getting efficient segmentation model through sparsity training and channel pruning.

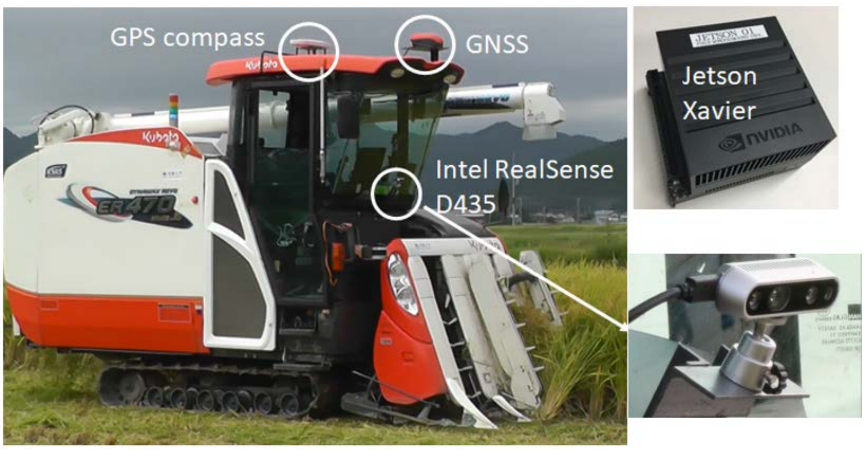


Fig. 4. Robotic combine harvester and devices installed.

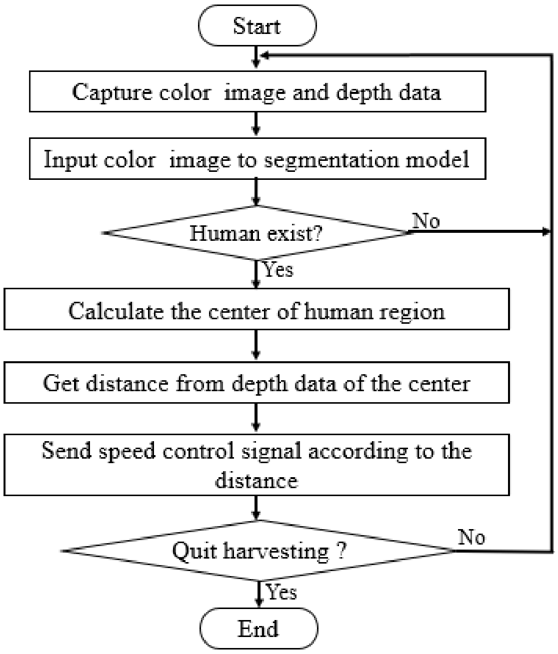


Fig. 5. The principal flow of the test algorithm for automatic rice harvesting.

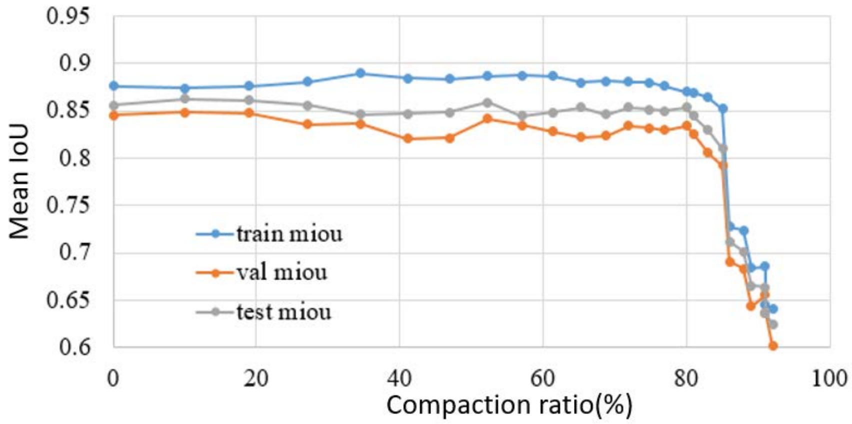


Fig. 6. The mean IoU at different compaction ratio.

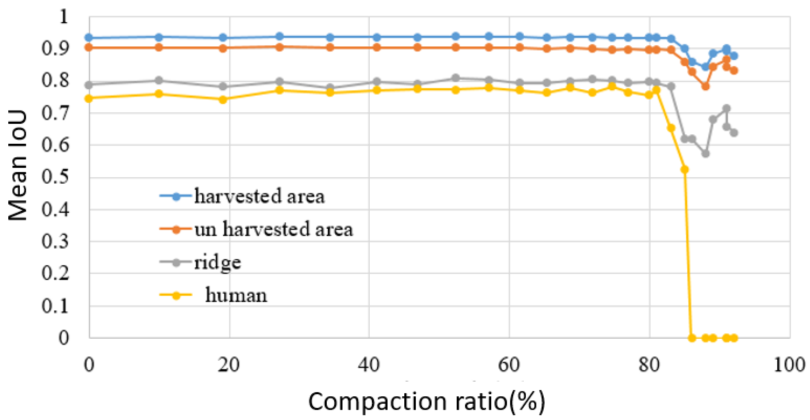


Fig. 7. Segmentation accuracy of mean IoU for each class at different compaction ratio.

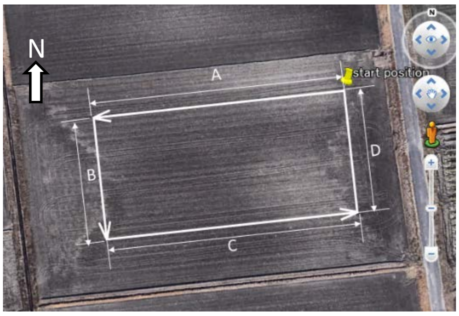


Fig. 8. Movement direction of the robotic combine harvester in paddy field in Kisosaki.

Fig. 9. Examples of images(top) and segmentation results (bottom) of the model in Test 1.

Fig. 10. Examples of images (top) and outputs (bottom) from SS model in Test 2.

Tables

Table 1 Description of the dataset.

Items	Description of dataset source 1	Description of dataset source 2
Camera	GoPro HERO5	Intel RealSense D435
Time	Morning, Noon	Afternoon
Weather	Cloudy, Sunny	Cloudy, Sunny
Place	Nantan, Japan; Narita, Japan;	Narita, Japan; Kizu, Japan;
Source image Size(width × height)	1920 × 1080	640 × 480
Rice variety	KoshiHikari, Husakogane	Husakogane, HinoHikari
Source sample Number	700	550
Field type	paddy field	paddy field
Camera height	1.75 m	1.75 m
Camera depression angle	The lens is facing down and at an angle of 15 degrees to the horizontal	The lens is facing down and at an angle of 15 degrees to the horizontal

Table 2 The frames per second (FPS) on the Jetson Xavier and model volume of different pruned models.

compaction ratio (%)	FPS on Xavier	inference time (ms)	Volume size of model parameter file (MB)
0	24.2	41.3	30.8
10.0	26.2	38.2	26.7
19.0	28.4	35.2	21.8
27.1	29.0	34.4	17.5
34.4	29.4	34.0	13.8
41.0	29.5	33.9	10.8
46.9	30.5	32.8	8.5
52.3	30.6	32.7	6.6
57.1	30.5	32.8	5.1
61.4	30.5	32.8	4.1
65.3	30.7	32.6	3.1
68.8	30.8	32.5	2.4
71.9	32.0	31.3	1.8
74.7	31.3	31.9	1.4
77.0	32.0	31.3	1.0
80.0	32.2	31.1	0.8

Table 3 Results of Test 1 by the pruned models.

Illumination [lx]	Movement direction	Number of times human appeared	Number of successes
32500 ~ 54850	A	10	10
	B	5	4
	C	10	10
	D	5	5
63210 ~ 79610	A	10	10
	B	5	4
	C	10	10
	D	5	5

Table 4 Results of Test 2 by the pruned models.

Illumination [lx]	Movement direction	Travel distance[m]	Number of failures
32500 ~ 54850	A	200	0
	B	100	0
	C	200	0
	D	100	0
63210 ~ 79610	A	200	0
	B	100	1
	C	200	1
	D	100	0