Detecting Ripeness of Strawberry and Coordinates of Strawberry Stalk using Deep Learning

Seo-jeong Kim
Korea Electronics Technology Institute
Jeonju-si, Republic of Korea
scott3554@keti.re.kr

Sunghwan Jeong
Korea Electronics Technology Institute
Jeonju-si, Republic of Korea
shjeong@keti.re.kr

Heegon Kim
Jeonnam Agricultural Research & Extension services
Naju-si, Republic of Korea
khg7136@korea.kr

Sooho Jeong
Jeonnam Agricultural Research & Extension services
Naju-si, Republic of Korea
aosi274@korea.kr

Ga-yun Yun
Jeonnam Agricultural Research & Extension services
Naju-si, Republic of Korea
gayun526@korea.kr

Keunho Park
Korea Electronics Technology Institute
Jeonju-si, Republic of Korea
root@keti.re.kr

Abstract— The ripeness of strawberries is an important criterion for determining whether strawberries can be picked or not, and the strawberry stalk coordinates of strawberries are important to automate strawberry harvesting. We propose a model that simultaneously learning the ripeness of strawberry and the coordinates of the stalk through two-path model with semantic segmentation. Two-path model can be seen as a model that performs two tasks at the same time. This model detecting the ripeness of strawberries with 90.33% accuracy and recognizing the coordinates of the stalk with 71.15%

Keywords—Multi-path model; Semantic segmentation; Multi-task segmentation; Detecting strawberry ripeness; Recognize strawberry stalk coordinate

I. Introduction

Recently, with the development of smart farm technology using artificial intelligence robots, related research continues. Research on recognizing weeds, research on disease detection of tomatoes, and research on fruit detection are representative smart farm research using artificial intelligence [1,2]. In addition, research on controlling robots or farms through information on objects found through deep learning is also underway [3]. In this paper, we propose a system for a robot that harvests strawberries by applying recent research to strawberries.

Detecting the ripeness of strawberries is the most import factor for harvesting strawberries. Ripe strawberries are red while unripe strawberries are green color. Since only ripe strawberries should be selected and harvested, in this paper, the ripeness of strawberries is detected using a semantic segmentation model. Second, we need to find the stalk of the strawberry. This sent to the robot the coordinates of the location it should go to harvest strawberries. As a method for this, a point with a value of 255 is drew on the area corresponding to the stalk on the strawberry, and the coordinates of this point are recognized using keypoint segmentation.

Since the proposed model has to detect two information at the same time, the number of parameters of the model is increased. As a method to solve this problem in this paper, multi-path convolution is proposed differently from normal convolution. This means that when generating a feature map, instead of performing (input × output) convolution, if convolution is performed with N path, only (input/N × output/N) ×N convolutions are performed, and then N feature maps are combined. It has the effect of greatly reducing the number of parameters while having the same number of feature maps. Semantic segmentation and keypoint segmentation can be simultaneously performed using the encoder-decoder model or U-net using multi-path convolution. We got result using mIoU accuracy is over 90% for detect ripeness of strawberries and over 70% for recognize the coordinates of the stalk of strawberries.

II. Related Work

As a related work, there is a study in which classification and segmentation were simultaneously learned using the 3D breast ultrasound images in 2021[4]. "Joint Stem Detection and Crop-Weed Classification for Plant-specific Treatment in Precision Farming" presented by Philipp in 2018 [5] paper also related with our study. As shown in Fig.1, he proposed a model that simultaneously performs keypoint segmentation and semantic segmentation by connecting two decoders in the last feature map of the encoder. In order to solve the problem that the decoder is doubled to increase the overall parameters, a method the reduce the parameters using dense block is proposed.

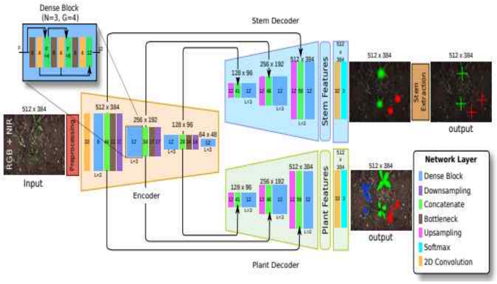


Fig.1. Reparameterization of variable autoencoders.

Research on reducing parameters in the convolution layer is also ongoing. Depth-wise convolution can be seen as a representative study for reduce parameters. If the existing convolution performed both the total number of channels and the spatial direction at once, the depth-wise convolution is to completely separate the two part. As shown in Fig.2, in the existing convolution, each input channel passes through the kernel and the merges to make one feature map, but in depth-wise convolution, the output from each input channel passes through the kernel becomes feature map, so the number of input channels it has the advantage of reducing the amount of computation and parameters.

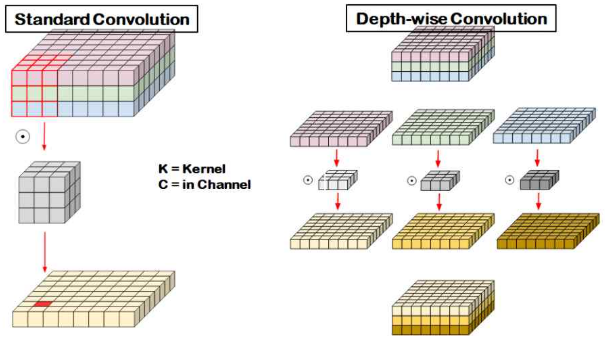


Fig. 2. Comparison of standard convolution and Depth-wise convolution.

III. Data Set

In this paper, the ripeness level is detected and the coordinates of stalk are found through dataset was taken directly in the strawberry farm. The strawberry dataset used in this paper is shown in Fig. 3.

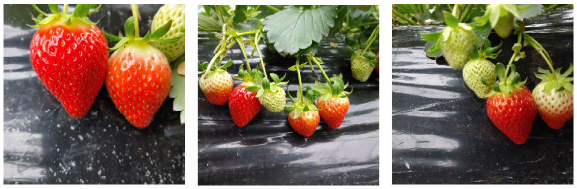


Fig. 3. Strawberries original image data.

Fig. 4 is examples of labeling for detect ripeness of strawberries. We pixel-based labeling well-ripened strawberries to brown, half-ripened strawberries to gray, and unripe strawberries to purple for semantic segmentation

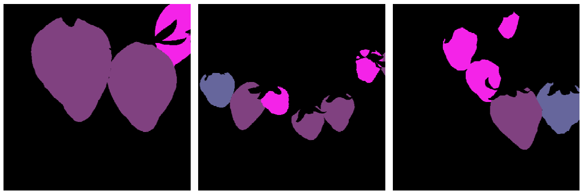


Fig. 4. Pixel-based labeling according to the degree of strawberry ripeness for semantic segmentation.

The area of the stalk of strawberry was labeled by drawing a circle with a radius of 15 pixels with the center coordinate like Fig. 5. This circle was labeled using a gaussian filter so that the pixel values of the circles decrease as the distance from the center coordinates increases. This is an efficient way to find keypoints while being used in a paper published in 2017 in a way that helps learning to focus on the center coordinates of a circle during learning [6].

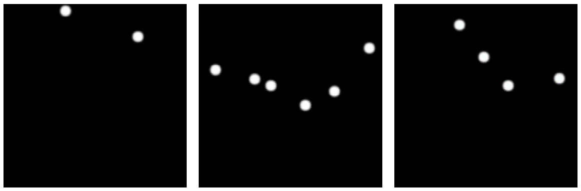


Fig. 5. Pixel-based labeling according to the points of strawberry stalk for keypoints segmentation.

IV. Model

A. Two-path convolution

In order for the model to perform two tasks in real time, it is important to reduce parameters of the model. As a method for this, two-path convolution method was proposed. If the 320×320×3 image is convolved with a 320×320×64 feature map, the number of parameters are (3×3×3+1) × 64, with a total of 1,792 parameters, and is fully restored to 320×320×64. The connected convolution is had number of parameters are (3×3×64+1) ×64, with a total of 36,928 parameters, as shown in Fig. 6.

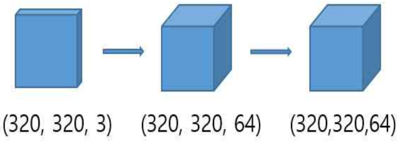


Fig. 6. Image of standard convolution method

Unlike the existing convolution, the method proposed in this paper learns weights while making a feature map through two paths instead of one path. This is shown in Fig. 7. If you make two 320×320×32 size feature maps from a 320×320×3 image, the number of parameters is (3×3×3+1) ×32 ×2, which has 1,792 parameters and is fully 320×320×32. The two convolutions that are connected are (3×3×32+1) ×32×2 with 18,496 parameters, which reduces 18,496 parameters compared to the existing convolution method, while creating a feature map with the same resolution and number of channels.

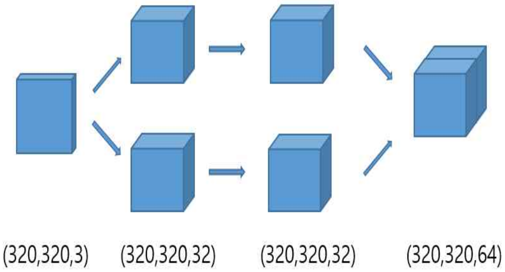


Fig. 7. Image of two-path convolution method.

B. Three-path convolution

The three-path convolution, which adds one more path to the two-path convolution, learns weights through three paths and concatenates the same feature map with 32 channels made of three. Since the size of the created feature map is (320×320×96), the number of channels is reduced to 64 through 1×1 convolution to create a feature map with same size as (320×320×64). Three-path convolution also has the effect of reducing the number of parameters while making various feature maps compared to the existing convolution. Fig.8 shows the appearance of three-path convolution.

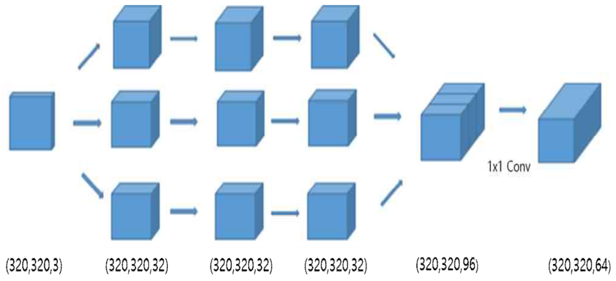


Fig. 8. Image of three-path convolution method.

C. Multi-path convolution segmentation model

The multi-path convolution model, which is a model to be made by mixing two path and three path convolution, was created based on the encoder-decoder model, which is a basic segmentation model. Two-path convolution is used in layer1 ,2, 8 and 9 of the encoder-decoder model, and three-path convolution is used in layers 3, 4, 5, 6 and 7. This is a method to maintain the general deep learning structure that generally generates more feature maps as the layer goes down. The proposed model can be seen through Fig. 9.

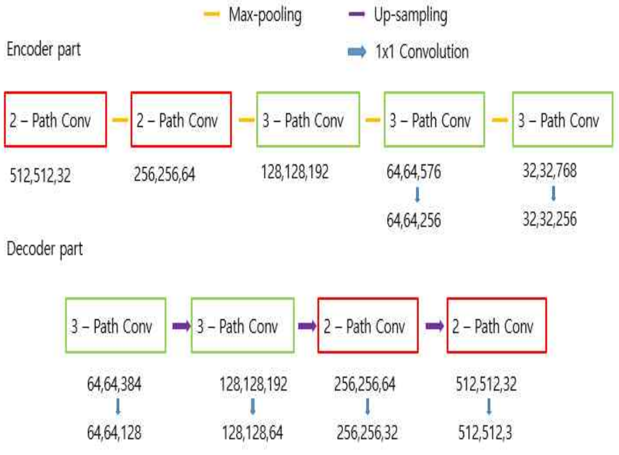


Fig. 9. Encoder-decoder segmentation model using multi-path convolution.

In encoder part, three-path convolution of the layer 3 does not use 1×1 convolution, but the feature map concatenated in three paths is used as it is, and 1×1 convolution is used in the layer4,5 for reduce channels. In the decoder part, up-sampling was performed using multi-path convolution, and 1×1 convolution was used for all layers to adjust the number of channels.

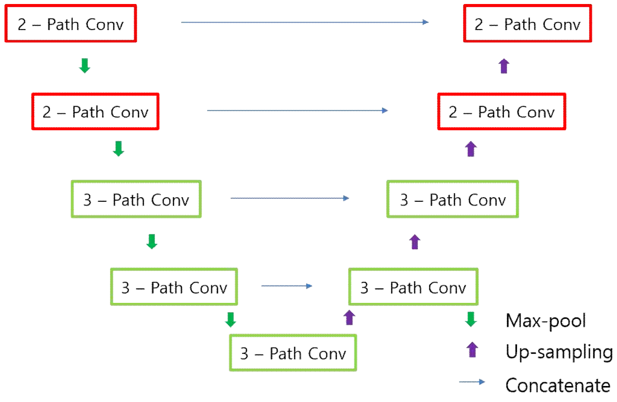


Fig 10. U-net segmentation model using multi-path convolution.

Fig10 shows the U-net model using multi-path convolution. U-net is called U-net because it has a U shape. In the basic encoder-decoder model, the feature map of the encoder and the feature map of the decoder are connected and concatenated. This shows a high accuracy improvement in edge and color detection, which are low-level information, because low-level information of the encoder can also be utilized, compared to recovering the size using only the decoder information [7].

D. Multi-branch for multi segmentation

The model proposed in this paper segments two targets simultaneously. Multi-branch was used as a method for this. The feature map obtained from the last stage of the multi-path convolution segmentation model is again divided into two paths, one performs semantic segmentation learning to detect strawberries ripeness, and the other performs keypoint segmentation to detect strawberry stalk. As a method for this, output of multi-path convolution U-net goes into 2 two-path convolution. One performs semantic segmentation for detect strawberries ripeness, other performs keypoints segmentation for detect strawberry stalk like Fig 11.

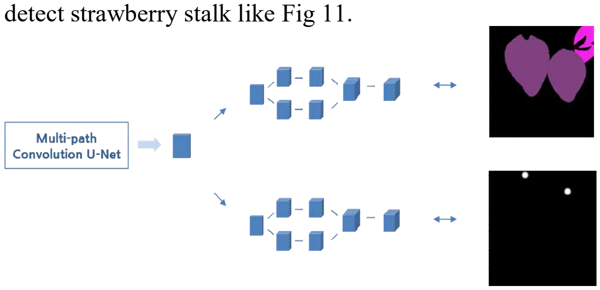


Fig 11. Image of Multi-branch for multi segmentation

E. Loss function

The proposed model simultaneously learns semantic segmentation to detect strawberry ripeness and keypoint segmentation to find point corresponding to the stalk position. As a method to learn this efficiently, semantic segmentation used the dice-coefficient loss function and keypoint segmentation used the Focal loss function. The dice-coefficient function is a loss function proposed by Milletari in 2016. It is a loss function proposed to compensate for the blurring problem in the edge part of the object of the cross-entropy loss function. It shows strength in semantic segmentation. Focal loss function is a function proposed by Tsung-Yi in 2017. When there is an extreme class imbalance problem in the training data, a small weight is given to a sample that is well classified and a large weight is given to a sample that is difficult to classify. It is a loss function that learns by focusing on this difficult sample. This shows strength in keypoint segmentation where relatively few pixel keypoint points among many background pixels given a value of 0 must be learned. Fig 12 shows the change of the loss function according to the formula of the Focal loss function and the value of the r parameters. In this paper, learning was carried out by assigning a value of 0.5 to *r*, which showed the best result as a result of conducting this experiment while changing the value of *r* [8][9].

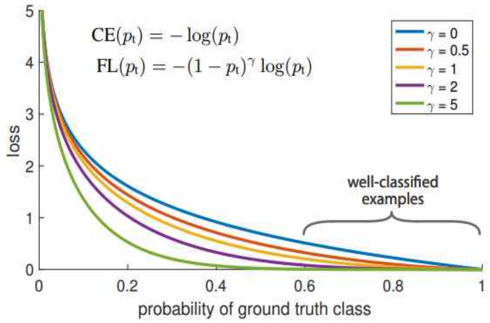


Fig 12. Focal loss output graph according to r parameters.

In this paper, the dice-coefficient loss function for semantic segmentation and the Focal loss function for keypoint segmentation were combined and used as the final loss function like Formula (1).

$$Final\,loss \;=Focal\,loss+Dice\,coefficient \qquad (1)$$

V. Experiments and Results

A. Experimental method

The training data for strawberry ripeness detection and stalk region coordinate recognition consists of images with a resolution of 2048×2048. This is a large number for learning, and the image was learned after resizing the resolution to 512×512, which is 1/4 of the size. The data for learning consists of 2456 training data, 60 validation data, and 412 test data. The learning rate was initially set to 0.01, and as the epoch increased, the learning rate was forcibly reduced to the value in Table 1, and the optimizer used the adam optimizer technique. For augmentation, flip and random rotate augmentation were used. Total of 150 epochs were learned, and the mIoU method was used as the evaluation method [10,11,12].

TABLE I. Learning rate change according to epoch.

Epoch	Learning rate
0 ~ 20	0.01
20 ~ 40	0.008
40 ~ 60	0.005
60 ~ 80	0.001
100 ~ 120	0.0008
120 ~ 150	0.0005

B. Experimental results

The final goal in this experiment was to configure the proposed system for the robot, and the robot to detect well-ripened strawberries in real time and to recognize the position of the stalk at the same time, so the number of parameters had to be reduced as much as possible. In order to compare the performance of the models, the U-net model based on VGG16, U-net based on the Multi-path convolution proposed in this paper, and the Encoder-decoder model based on Multi-path convolution, Res-net 101, were used. The parameters and accuracy of U-net based on the comparison were compared, and the results can be confirmed through Table 2 .

TABLE II. Comparision of parameter and model performance result.

Model	Parameters	mIoU1	mIoU2
VGG16 U-net	11,676,815	84.56	59.58
Multi-path U-net **(only 2multi.)**	8,467,268	85.18	62.11
Res U-net	27,497,154	89.87	69.45
Multi-path encoder-decoder	22,489,348	91.48	72.48
Multi-path U-net	16,137,678	90.33	71.15

* mIoU1 = strawberry ripeness semantic segmentation.

* mIoU2 = stalk of strawberry keypoint segmentation.

The second model in Table2 means a multi-path U-net using only two-path convolution which shows strength in both parameters and mIoU accuracy compared to U-net using VGG16. Both the encoder-decoder model using multi-path convolution and U-net showed higher accuracy than res U-net. In this study, multi-path U-net with few parameters was selected as the final model among the models that detect strawberry ripeness over 90% find stalk area over 70%, although there are few parameters to obtain quick results. And the experiment was conducted, and the results can be seen in Fig 13~15.

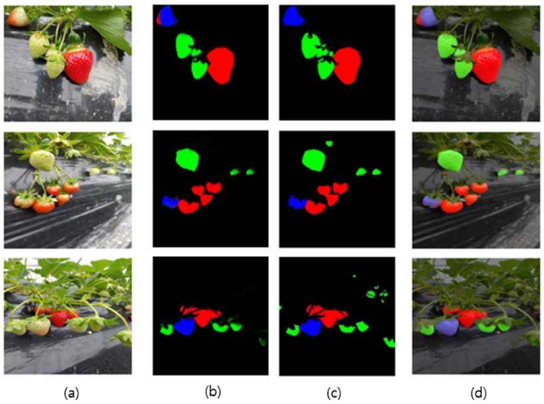


Fig 13. Result of strawberry ripeness semantic segmentation result (a) is original data, (b) is predict data, (c) is ground truth data, (d) is image that overlaps (a) and (b)

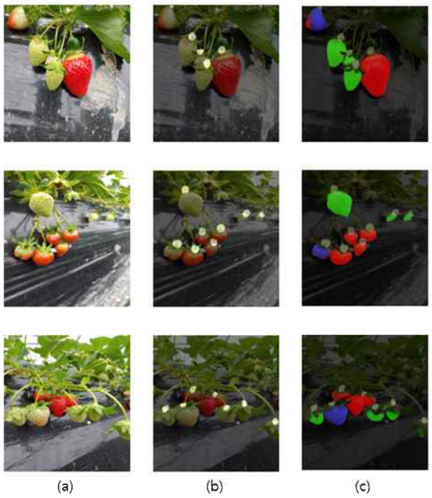
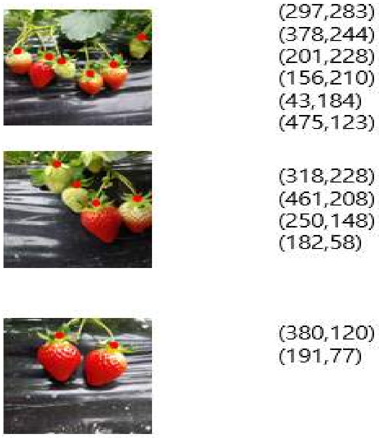


Fig 14. Result of stalk of strawberry keypoint segmentation result (a) is original data, (b) is predict data, (c) is image that overlaps (a), (b) and strawberry ripeness predict result image



```
(297,283)
(378,244)
(201,228)
(156,210)
(43,184)
(475,123)

(318,228)
(461,208)
(250,148)
(182,58)

(380,120)
(191,77)
```

Fig 15. Result of stalk prediction by displaying the center coordinates of the predicted stalk region on the right and drawing circle with a radius of 5 at the center coordinates

VI. Conclusion

Using multi-path convolution, it has the same feature map as the existing model, but by reducing the parameters, the overall parameters and learning speed of the model were reduced, and similar accuracy was maintained while finding the ripeness and location of strawberries in real time. This model predicted strawberry ripening degree more than 90% and stalk area of strawberry more than 70%. It is a study that effectively found ripe strawberries by well detecting the area and ripeness of strawberries, and it is also possible to calculate how many strawberries are in the image by using the number of coordinates of the stalk.

Acknowledgment

This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry(IPET) through Smart Farm Innovation Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA) and (421003041HD030)

References

[1] N. Chebrolu, P. Lottes, A. Schaefer, W. Winterhalter, W. Burgard, and C. Stachniss. Agricultural Robot Dataset for Plant Classification, Localization and Mapping on Sugar Beet Fields. Intl. Journal of Robotics Research (IJRR), 2017.

[2] S. Haug, A. Michaels, P. Biber, and J. Ostermann. Plant Classification System for Crop / Weed Discrimination without Segmentation. In IEEE Winter Conf. on Appl. of Computer Vision (WACV), 2014.

[3] I. Sa, Z. Chen, M. Popvic, R. Khanna, F. Liebisch, J. Nieto, and R. Siegwart. weedNet: Dense Semantic Weed Classification Using Multispectral Images and MAV for Smart Farming. IEEE Robotics and Automation Letters (RA-L), 3(1):588–595, 2018.

[4] Y.Zhou, H.Chen, Y.Li, Q.Liu, X.Xu, S.Wang, P.Yap, D.Shen, “Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images,” Medical Image Analysis, vol. 70, May 2021.

[5] P.Lottes, J.Behley, N.Chebrolu, A.Milioto, C.Stachniss, “Joint stem detection and crop-weed classification for plant-specific treatment in precision farming”,arxiv, 2018.

[6] V. Belagiannis and A. Zisserman. Recurrent human pose estimation. In FG 2017. IEEE, 2017.

[7] O.Ronneberger, P.Fischer, T.Brox, “U-net: Convolutional Networks for Biomedical Image Segmentation”, MICCAI, Springer, LNCS, vol. 9351, pp. 234-341, 2015.

[8] L.Chen, G.Papandreou, L.Kokkinos, K.Murphy, AL.Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol, no.4, pp.834-848, 2018.

[9] Lin, T.; Goyal, P.; Girshick, R. B.; He, K.; and Dollar, P. 2017b. Focal loss for dense object detection. ´ In ICCV 2017, 2999–3007.

[10] Cai, Z., and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In CVPR 2018.

[11] K.Simonyan and A.Zisserman, ″Very Deep Convolutional Networks for Large-Scale Image Recognition″, ICLR, 2015.

[12] C.Szegedy, W.Liu, Y.Jia, P.Sermanet, et al, ″Going Deeper with Convolutions″, CVPR, 2015.