JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 19, No. 12, pp. 2385-2391, Dec. 2018

Mask R-CNN을 이용한 실시간 감귤 분할 및 검출 시스템

김진원 · 이말례*
전북대학교 컴퓨터공학과

A Real-time Citrus Segmentation and Detection System using Mask R-CNN

Jin-Won Kim · Malrey Lee*

Center for Advanced Image and Information Technology, School of Electronics & Information Engineering, Chon Buk National University, 664-14, 1Ga, Deokjin-Dong, Jeonju, Chon Buk, 561-756, South Korea

[요 약]

본 논문에서는 감귤을 촬영하여 800x800으로 변환 후 200장의 사진을 수집하고, 각 사진의 감귤영역을 마스크 레이블링하여 JSON형식으로 저장하여 데이터셋을 생성하였고, 최신 알고리즘인 Mask R-CNN을 적용하여보다 높은 신뢰성을 가진 시스템을 구축하여 감귤 검출과 분할하는 시스템을 구현하였다. 학습을 진행하는데 적은 데이터셋으로 인한 과적합의 문제를 해결하기 위하여 데이터 증가 기법을 이용하여, 검출 성능은 0.97로 크게 향상 시켰다. 농가의 실질적인 욕구를 충족시키기 위하여 사진을 찍어서 1차적으로 마스크 레이블링 되고 2차적으로 추가 작업 후 바로 학습이 가능한 플랫폼을 개발할 계획이다.

[Abstract]

In this paper, 200 photograph of citrus were collected and converted to 800x800. The areas of each citrus in the photograph were mask-labeled and stored in JSON format to generate a data set. The latest algorithm, Mask R-CNN, I constructed a reliable system to detect and divide citrus fruits. In order to solve the over-fitting problem due to small data sets, the data augmentation was used and the detection performance was as high as 0.97 with small data sets. In order to meet the farmer's practical needs, I plan to develop a platform that can take photographs, label them with a mask first, and then train them immediately after doing additional mask labeling work.

색인어 : 과일 검출, 심층 콘볼루션 신경망, 빠른 학습, 실시간 성능, 자동 수확

Key word : Visual fruit detection, Deep convolutional neural network, Rapid training, Real-time performance, Auto-harvesting

http://dx.doi.org/10.9728/dcs.2018.19.12.2385

Received 06 September 2018; **Revised** 17 September 2018
Accepted 23 December 2018

***Corresponding Author; Malrey Lee**

Tel: +82-63-270-3993
E-mail: mrlee@chonbuk.ac.kr

Ⅰ. Introduction

Low birth and aging phenomenon is accelerating the decrease in agriculture population and rural aging. [1] The demand for the provision of alternative labor force or the development of a more convenient and advanced agricultural system is on the rise. As the global competition intensifies due to the proliferation of FTAs and opportunities for expanding access to export markets coexist simultaneously, researches are being actively conducted in advanced countries to innovate agricultural products and to harness future growth engines through the convergence of agriculture and robotic technology.

Recently, artificial intelligence, sensors, and big data technologies have been applied to various industrial and agricultural fields in line with the advent of the fourth industrial revolution. In particularly the powerful performance of deep-learning technology [2] in the field of image processing is making the world come alive. Examples of utilizing image processing techniques using deep running in the field of agricultural automation include yield prediction [3], pest awareness [4], and quality evaluation systems [5].

The development of an accurate fruit detection system is of strenuous work for fully automated harvesting robots, since vision with perception system needs to be exist before subsequent manipulation and grasping systems. Robots are unable to detect the differences between similar objects and fruit, hence the need for segmentation between the fruits and the background.

This paper is about technology for detecting fruits and mapping mask through object recognition technique. Our goal is to build a system that accurately detects fruits and makes mask from citrus photographs, measures the number of fruits and provides a better environmental friendly alternative for auto harvesting.

This is not only effective in terms of labor savings through unattended and automated systems, but also contributes to quality improvement challenge overseas and market competition with low prices due to the proliferation of FTA.

The composition of this paper is as follows. We will discuss some of the studies related to this problem in Chapter 2. Deep learning algorithms and methods for detecting fruits are described in Chapter 3. Chapter 4 is the experiment based on the photographs obtained from the citrus farm in Jeju Island, and the results will be discussed in Chapters 5.

Ⅱ. Related Works and Background

Various works related to fruit detection shows a significant interest across a variety of orchard for the yield mapping or autonomous harvesting [14]. Detection system is typically the task of finding the different fruit or background object in an image and classifying them.

Past few year, many researchers have mentioned the main problem of fruit detection, like the works issued in [8 – 13], the main problem of building a fast and accurate fruit detection system persists, as found by Kapach, K et al. [14] The high variation in the appearance of the fruits in field settings, including shape, size, color, texture and reflectance properties causes the problem that make hard to detect.

Most of the works presented in the past, address the problem of fruit detection as segmenting fruits from background. Ulzii et al. [15] examined the issue of citrus detection for yield prediction. They developed an image processing system to detect citrus and use the data to predict the yield estimation. the system is based on conversion of RGB image to HSV, noise removal and watershed segmentation[25]. Wang et al. [11] proposed a color and distinctive specular reflection pattern for detecting apples. By using two-camera stereo rig for image acquisition, the system was able to work at nighttime with controlled artificial lighting to reduce the variance of natural illumination. Hung et al. [13] presented a multi-class image segmentation for automate fruit segmentation. Relevant features from data were capture by unsupervised feature learning which is called Sparse Autoencoder(SAE). The classifier output which is used to learn the label association to the multiscale responses is passed into the CRF. They achieved outstanding segmentation performance, but did not present object detection.

More recently, Faster R-CNN[7] were used in the study for sweet pepper and rock melon detection in a greenhouse and showed expanded object detector which is collected from Google Image search.

Our work differs substantially from that study, because our system not only detects citrus but also segment each fruit which is essential for auto harvesting system.

Ⅲ.Materials and methods

This section presents the Mask R-CNN framework for fruit detection in orchards as well as introduces details about transfer learning and dataset of our work and data augmentation techniques, which are used for the studies experimented in section experimental results and discussions.

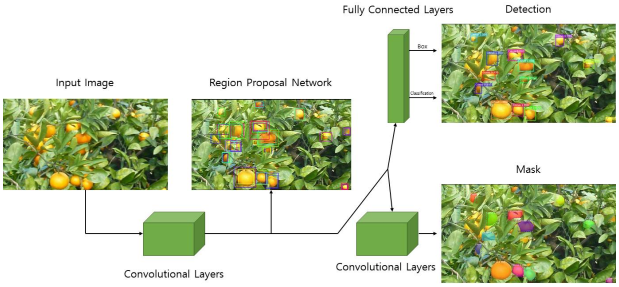

Fig. 1. The Structure of Mask R-CNN

1) Mask R-CNN

Is a simple, flexible and fast system extended from Faster R-CNN. The segmentation results surpass prior state-of-the-art results. The system is composed of two stages. First stage is a Region Proposal Network(RPN) which proposes candidate of bounding box. Second stage is extraction of features using RoIPool from output of first stage and classifies the regions and regresses a bounding box around the object. In parallel to predicting previous job, it also outputs a binary mask for each RoI.[17]

The input to the training model is a 3-channel color image (RGB) of free size (within the range of capacity of GPU memory) which are resized to 800 pixels, and annotated data of bounding boxes and mask around each object. Our paper uses ResNet-101. We train on 2 GPUs, mini-batch size of 4 for 1K iterations, with a learning rate of 0.02 and weight decay of 0.001 and a momentum of 0.9.

2) Transfer learning

It became standard in computer vision to train CNN on a large basic network then transfer the learned functions to a new target task, which typically includes fewer labels (i.e. fine tuning). MS-COCO data sets are often used as a base for 80 object categories and 200,000 million images. Using MS-COCO pre-programmed CNN features, a high-tech result for image processing tasks ranging from image classification to image caption was acquired by the CNN using the MS-COCO data weighted CNN feature [17] basically supported by the CNN. However, if the target class differs significantly from the default class during the delivery of weights to the target task, performance may be compromised. This is because the deeper layers of the CNN network learn a function rather than the neighboring task [18]. In case of few new data to train on, but similar to organizational data, due to the small amount of data, the fine-tunes (running a backpropagation on the entire model) are not subject to risk of over-fitting. The new data is like the organizational data, so in this case, only the final linear-classifier layer is learned.

3) Microsoft COCO [22]

In recent years, computer vision has been focused on image classification, object recognition, and segmentation. In 2012, AlexNet [19] had an error rate of 15% at the ILSVRC Object Classification Contest. With the recent revival of deep learning, the accuracy of the image classification has already exceeded the level of human beings by 3.5% [20], and object recognition is the next it is an assignment. There are many kinds of objects in a single image and it is more difficult to recognize the position of the object accurately. Dataset such as PASCAL VOC [21], ImageNet [22], SUN [23] and MS COCO are typical examples of data that is used in object recognition such as a data set including a label of an object instance in an image and a position of a boundary box. In this study, the study was conducted using the COCO data set of the Microsoft Research team.

MS COCO is a data set for large-scale object recognition, segmentation and captioning. It provides over 200,000 images in 80 categories for object recognition problems. Every object instance has a detailed refinement mask annotated. COCO annotations are created through a three-step pipeline of category labeling, instance positioning, and instance segmentation. Mask R-CNN's research suggests that the use of additional COCO in the learning process has resulted in significant performance improvements [17].

In the process of creating our data set for citrus recognition, we segmented the fruit photographs taken directly from the citrus farm in Jeju Island, South Korea, into appropriate sizes. Then, through the process of the above-mentioned three-stage pipeline, we annotated the type information that is necessary for our system to be as rich and accurate as the format of the COCO data set. We have drawn a separate segmentation mask for dozens of citrus instances overlapping in a single citrus picture [26]. We also fine-tuned the weighed values learned in MS COCO to our model to reduce the cost of learning time and obtain higher accuracy.

4) Data Augmentation

Data Augmentation is one of the best methods to use when you want to study with a small amount of data. You can increase the number of images by various methods such as reversing, enlarging, reducing, rotating, or adjusting the brightness from the original image of the data set. Deep learning technology itself used in the field of computer vision is very dependent on the absolute amount of data. Because it is possible to raise the accuracy of learning with a very simple trick, many studies use this method, and research on the data augmentation methodology itself is actively proceeding.

Because our research requires data that includes segmentation annotations in images, complex forms of data augmentation are not easy to use. For example, in the method of randomly cropping or scaling an image to an image, the probability of a segmented instance being damaged is very high. Therefore, only the left and right inverted data enhancement method was used for learning.

Ⅳ. Experimental results and discussions

The orchard data evaluated in this paper are collected during the day in orchards in Jeju Island, Korea. Citrus data was captured by citrus researchers. The data contains multiple rows of orchestral image data. The longer distance between the citrus trees was compensated with a higher resolution picture. In addition to the citrus image, we also collected fruit images to reduce various error between other fruits. Therefore, high resolution images were needed to express fruit well, and this image could be obtained using a handheld high pixel digital camera. The images captured at each orchard span the entire tree and were driven by key experimental objectives of effective output estimation and mapping. The fruit detection work presented in this document is therefore an important component of automatic harvesting.

This section describes the abstinence study for the detection network and evaluates the fruit detection performance related to the number of training images, the transfer learning between the orchard and data enhancement techniques. These studies were conducted using ResNet-101 because its accuracy is very high, but performance evaluation for FPN network is also presented. Finally, there is a simple technique for the proposed source network. Mask R-CNN framework allows for multi-class detection, binary problems are considered for orchard data, and new models must be trained for each fruit type. Limiting the number of classes can generally lead to better classification accuracy [18]. Orchard blocks are generally allowed in orchard applications because one fruit per block is homogeneous without mixing.

In the ResNet-101 and VGG16 networks, the sub-sampling factor is 16 in the final transformation layer, and the smallest possible object size is 16 pixels. To ensure this, all training sub-images were scaled to a shorter side of 800 pixels, which meant enlarging the citrus sub-image. The sub-image dimensions specified in the previous section are selected to rescale fruit of sufficient size later. All networks were initialized with a coco filter and trained until there is convergence of detection performance across the validation set. This was about 4k repetitions for citrus. Further investigation of the initialization and learning rate for the citrus data set will enable faster training.

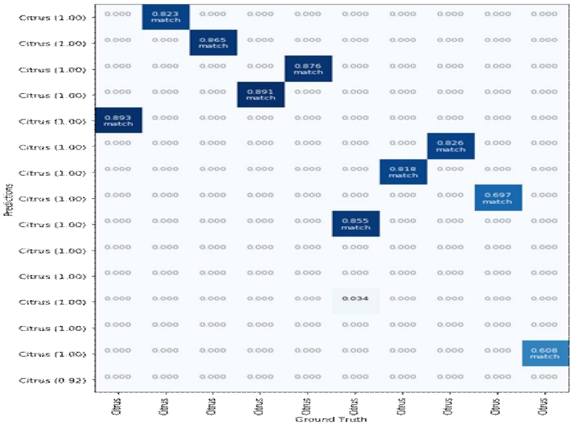

Fig. 2. Grid of ground truth objects and their predictions

Figure 2 shows a grid of ground truth objects and their predictions its prediction varies as the images of the citrus is hidden by a leaf. But as you can see in the Table 1 its precision-recall curve scored 0.971. Detection performance is reported using the average precision response to fruit, which is the area of the precision recall curve. As in [6], the result is reported using an F1 score, where the class threshold is evaluated against the set of head-out validation.

Table 1. Fruit Detection Results (As F1-scores)

Network	Citrus
ResNet-101	0.971
FPN	0.956

ResNet-101 scored the very best between two networks, 0.971 for citrus. This can be attributed to auto-harvesting platform. End to end learning for detection and segmentation. Fig. 3 and 4 show the results of citrus detection.

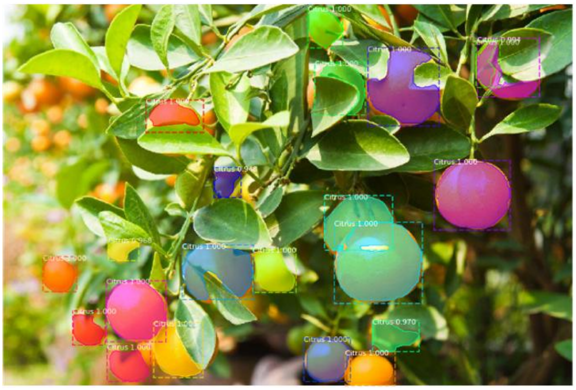

Fig. 3. Example of detections and segmentations for each citrus (close shot)

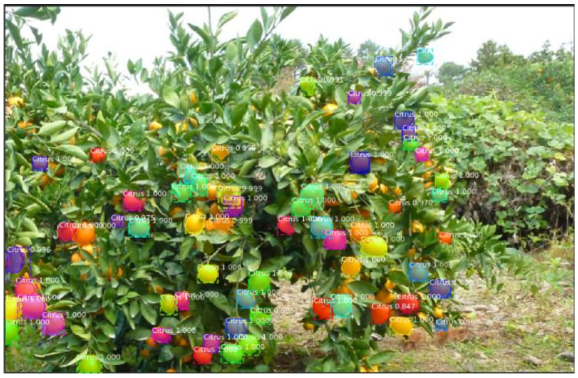

Fig. 4. Example of detections and segmentations for each citrus (full shot)

The Mask R-CNN detection framework yields state-of-the-art performance on orchard images. It creates a mask for citrus by its shape. The study of transferring learning turnout to be successful as we can see with fig 4. 400 of citrus image data was enough to perform a real-life use for auto-harvesting platform. Data augmentation was very helpful in training a few numbers of data which boosted detection performance. It efficiently reduced a number of images of training by almost 50% which eventually leads to reduce in necessity of labelled data.

The methods of detection presented in this paper can easily be extended to other orchards. The supplied labeling toolbox can be used in the harvest process, and a parallel training test process can alter the detection performance according to the increasing number of educational images. The labeling process can then be terminated according to the labeling budget and/or performance requirements. For small fruits, the image needs to be resized.

Although the minimum fruit size is less than 16 pixels, the fruit resolution can be harmful to labeling and sensing performance. The results are thought to use simple data growth techniques, such as image frying and relocation, and transfer learning between other fruits is not necessary. Transfer learning may still be important when the baseline task is very similar to the destination job. For example, a trained model in a given apple data set can still be useful for initializing the sense network for another captured apple data set using different lighting conditions and/or different sensors. However, due to these variations in the data set, further testing of how the model is applied is required [5] shows reasonable qualitative performance for a new set of data without re-learning.

Ⅴ. Conclusion

This paper introduces a fruit detection system for image data captured in orchards using the Mask R-CNN detection framework. In order to better understand the practical development of such a system, we conducted ablation studies on one of famous fruits in Korea. The study of detection performance for the number of training images showed the amount of training data required to reach convergence. Transfer learning analysis showed that the transfer weight between orchards did not result in significant performance improvements on a network that was directly initialized from the features of highly generalized coco dataset. Data expansion techniques such as flip-and-scale augmentation have improved performance on a variety of training images and achieved comparable performance in less than half the number of training images. In this study, the author's previous study obtained the best detection performance, and the citrus F1 score was 0.9. In advanced applications such as auto-harvesting, masks are extremely helpful in controlling robots. Future research will integrate the detection output from the mask R-CNN with yield mapping to provide object association between adjacent frames. Additional analysis for fruit detection is performed to understand a dataset representing the same fruit, a dataset captured under different illumination conditions, a sensor configuration, and transmission learning over a period of a year.

For future work, I plan to develop a platform that can take photographs, label them with a mask first, which will help people making mask easily and for those object that was not masked user can mask by themselves. And also by using GUI structure people will be able to train their own train set that is custom to agriculture easily.

Reference

[1]. Roh, Jae-Sun·Jung, Jin Hwa·Jeon, Ji Yeon*; Returning Farmers and the Aging of Farm Households: Prospects of Changes in Rural Population by Their Influx. *Journal of Korean Society of Rural Planning*, 19.4: 203-212, 2013.

[2]. LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. nature, 2015, 521.7553: 436.

[3]. Aggelopoulou, A. D., Bochtis, D., Fountas, S., Swain, K. C., Gemtos, T. A., & Nanos, G. D. Yield prediction in apple orchards based on image processing. *Precision Agriculture*, 12(3), 448-456, 2011.

[4]. Weizheng, S., Yachun, W., Zhanliang, C., & Hongda, W. Grading method of leaf spot disease based on image processing. In 2008 international conference on computer science and software engineering, *Wuhan*, pp. 491-494, 2008, December.

[5]. Cubero, S., Aleixos, N., Moltó, E., Gómez-Sanchis, J., & Blasco, J. Advances in machine vision applications for automatic inspection and quality evaluation of fruits and vegetables. *Food and Bioprocess Technology,* 4(4), 487-504, 2011

[6]. BARGOTI, Suchet; UNDERWOOD, James P. Image segmentation for fruit detection and yield estimation in apple orchards. Journal of Field Robotics, 34.6: 1039-1060, 2017.

[7]. Girshick, R., Donahue, J., Darrell, T., & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. *In Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 580-587, 2014

[8]. Nuske, S., Achar, S., Bates, T., Narasimhan, S., & Singh, S. Yield estimation in vineyards by visual grape detection. In Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ *International Conference on* ,pp. 2352-2358, IEEE, 2011, September.

[9]. Nuske, S., Wilshusen, K., Achar, S., Yoder, L., Narasimhan, S., & Singh, S. Automated visual yield estimation in vineyards. *Journal of Field Robotics*, 31(5), 837-860, 2014.

[10]. Yamamoto, K., Guo, W., Yoshioka, Y., & Ninomiya, S. On plant detection of intact tomato fruits using image analysis and machine learning methods. *Sensors,* 14(7), 12191-12206, 2014

[11]. Wang, Q., Nuske, S., Bergerman, M., & Singh, S. Automated crop yield estimation for apple orchards. In Experimental robotics pp. 745-758, Springer, Heidelberg. 2013.

[12]. Bac, C. W., Hemming, J., & Van Henten, E. J. Robust pixel-based classification of obstacles for robotic harvesting of sweet-pepper. *Computers and electronics in agriculture*, 96, 148-162, 2013.

[13]. Hung, C., Nieto, J., Taylor, Z., Underwood, J., & Sukkarieh, S. Orchard fruit segmentation using multi-spectral feature learning. In Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on pp. 5314-5320, IEEE, November,2013.

[14]. Kapach, K., Barnea, E., Mairon, R., Edan, Y., & Ben-Shahar, O. Computer vision for fruit harvesting robots–state of the art and challenges ahead. *International Journal of Computational Vision and Robotics,* 3(1-2), 4-34, 2012.

[16]. I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. McCool, “DeepFruits: A Fruit Detection System Using Deep Neural Networks,” *Sensors*, vol. 16, no. 8, p. 1222, 2016.

[17]. Kaiming He, Georgia Gkioxari Piotr Dollar Ross Girshick, “Mask R-CNN” Computer Society, pp.2980- 2988 ,2017

[18]. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” in *Advances in neural information processing systems*, pp. 3320–3328, 2014.

[19]. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham, September, 2014 .

[20]. KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. p. 1097-110, .2012.

[21]. HE, Kaiming, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 770-778, 2016

[22]. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. IJCV, 2010.

[23]. DENG, Jia, et al. Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, CVPR 2009. *IEEE Conference on. IEEE,* p. 248-255, 2009.

[24]. Hyuk Jin Yang, Dong Hyun Kim, Yeong Geon Seo, Noise-robust Hand Region Segmentation In RGB Color-based Real-time Image, *Journal of Digital Contents Society* Vol. 18 No. 8 pp. 1603-1613, Dec. 2017

[25]. Ki-Hong Park, Hui-seong Noh, Effective Acne Detection using Component Image a* of CIEL*a*b* Color Space, *Journal of Digital Contents Society* Vol. 19 No. 7 pp.

1397-1403, July. 2018

[26]. Jae-Young Lee, Jun-Sik Kwon, Image Annotation System for mobile Augmented Reality Environment, Journal of *Digital Contents Society* Vol. 16 No. 3 pp. 437-444, Jun. 2015

Jin-Won Kim

2018: M.S, Chonbuk National University

※Research Area : Artificial Intelligence, Big data, Game engine, VR, AR and So on

Malrey Lee

1998 : Ph.D, Chung-ang University

2003 - 2018: Professor of Chonbuk National University
※Research Area : Artificial Intelligence, Big data, Game engine, VR, AR and So on