DEEP LEARNING BASED PLANT PART DETECTION IN GREENHOUSE SETTINGS

Manya Afonso[1], Ruud Barth[2] and Aneesh Chauhan[3]

[1]Biometris, Wageningen University and Research, The Netherlands

[2]Greenhouse Horticulture, Wageningen University and Research, The Netherlands

[3]Food and Bio-based Research, Wageningen University and Research, The Netherlands

manya.afonso@wur.nl, ruud.barth@wur.nl, aneesh.chahan@wur.nl

ABSTRACT

Precision agriculture challenges such as automatic harvesting, phenotyping, and yield prediction require precise detection of plant parts such as the fruits, leaves or stems. Deep learning has emerged as the state-of-the-art technology for image segmentation and object detection in several domains, notably in self-driving vehicles and medical imaging. In recent years, deep learning methods are being increasingly adopted in vision-based applications for precision agriculture. In previous work, methods were investigated to segment the image for plant parts. However, such an approach did not yield object instances. In this work, we applied the state-of-the-art deep learning object detector, MaskRCNN, to the problem of detecting fruit and other plant parts, in the sweet pepper (capsicum annuum) plant. An extensive study was carried out where we investigated different transfer learning schemes, different convolutional neural network architectures, and varying numbers of training images. Experimentally, we found that MaskRCNN trained with the synthetic data and fine-tuned with very few empirical images is able to detect more than 95% of the sweet pepper fruit. It was also found that training on the synthetic data and then fine-tuning over a few empirical images led to a better performance in the detection of fruit, over training only on the limited set of empirical images. Furthermore, results show that the best model could successfully generalize to different imaging conditions. This work is a necessary step for applying deep learning in high-throughput robotics and phenotyping approaches and will open up many opportunities for smart farming and more efficient use of resources. Currently, training deep learning models is dependent on the knowledge and expertise of the scientists involved. The insights gained from this work should lead to more automatic training protocols, allowing widespread use in very different applications.

Keywords*:* Computer Vision, Robotics, Deep Learning, Synthetic Data, Instance Segmentation.

1. INTRODUCTION

Automation and robotic tasks in agriculture such as harvesting, pruning or localized spraying require detailed and accurate localization of plant parts from images (Bac et. al., 2014). Traditionally, computer vision solutions to detect these parts were based on hand-crafted features such as shape and texture (Kapach et. al., 2012). However due to the large variability between different varieties of the crop, specimens of the same crop and variety, and in the imaging conditions (Barth, 2018; Bac, 2015), methods from the domain of machine learning such as convolutional neural networks, are being increasingly used due to their ability to cope with large amounts of variation in training data. Deep neural networks (Lecun et. al., 2015) have emerged as the state-of-the-art in computer vision, from image level classification (Krizhevsky et. al., 2012) to pixel-wise or semantic segmentation (Long et. al.,

2015). In many applications including agriculture, object detection rather than semantic segmentation is required to discern individual instances in order to determine manipulation or application points for each object. FastRCNN (Girshick et. al., 2015) is an object detector which uses region based convolutional neural networks (R-CNN) combined with the selective search method to detect region proposals. Another fast and popular deep learning object detector is the You Only Look Once (YOLO) (Redmon et. al., 2016) which applies a single neural network to the full image which divides the image into regions and predicts bounding boxes and probabilities for each region, unlike Fast-RCNN and its variants which apply the model to an image at multiple locations and scales. These methods provide a classification with a bounding-box localization of instances of the object of interest. A variant of FastRCNN has been used to detect sweet peppers and other fruit from color and near infrared images (Sa et. al., 2016). A recent method, Mask-RCNN, extends FastRCNN by a branch for predicting an object mask in parallel with the existing branch for bounding box recognition (He et. al., 2017). In our previous work, it was shown that plant parts could be successfully semantically segmented in the image, based on synthetic data bootstrapping (Barth et. al., 2018) and fine-tuning with a small empirical dataset (Barth et. al., 2017).

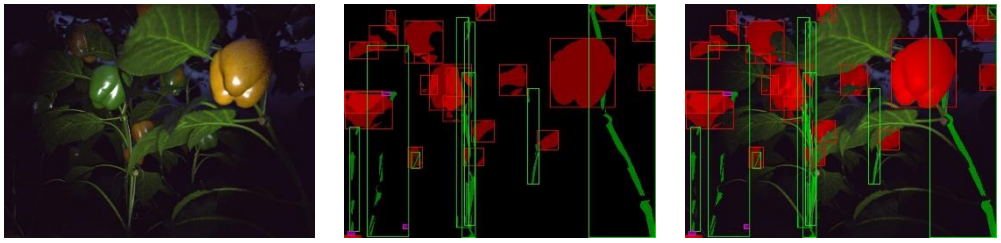
In this paper we build upon previous pixel-wise segmentation results, by using Mask-RCNN to obtain instance detection of plant parts. This paper also investigates different transfer learning techniques, with and without training on the synthetic or empirical dataset. Experimental results show that under the best training configuration (architecture and transfer learning scheme), almost 100% of the fruits, 75% of the stems, and 80% of the peduncles can be detected, with few false positives. An improvement over the Intersection over union (IOU) metric was obtained compared to previous pixel-wise segmentation. Thus, the use of MaskRCNN, resulted in a practically useful method for sweet pepper part detection with state-of-the-art performance.

2. Materials and Methods

2.1 Dataset

The image dataset consists of the first 200 synthetic images out of the dataset from (Barth et. Al., 2017 and Barth, 2018), modelled to approximate the empirical set visually and 50 empirical (real) images of a sweet-pepper crop in a commercial high-tech greenhouse. The dataset is available online. Because the ground truth of this dataset was on a per-pixel level and did not contain instances, the empirical dataset was manually relabeled for unique objects of the classes fruit, stem, and peduncle using the LabelMe tool (Russell et. al., 2008). For the synthetic images, the instance wise ground truth was obtained by re-rendering each object alone in the 3D scene, thus producing the complete shape of each object. To deal with occlusions, a post processing step was used to include only the class pixels for each instance that were also present in the original ground truth. The 200 synthetic images were used only for training. The first 30 images from the empirical dataset were used for training or fine-tuning while image numbers 31 to 50 are used for testing. An alternative empirical dataset of 20 images, that used different illumination and camera, was also acquired and labeled with the aim to verify the trained model's generalizability and robustness to new conditions in other datasets.

An example image from the synthetic dataset and its ground truth are shown in Figure 1. It can be seen that the color image is quite realistic, compared to empirical ones in Figures 2(a) and 3(a).



(a) (b) (c)

Figure 1: Example of synthetic image (left), of its corresponding ground truth instance labels (middle), and of the GT overlaid (right).

2.2 Deep Learning Software Setup and algorithms

Detectron, Facebook AI Research's software with the implementation of Mask R-CNN, was used to train convolutional neural network models with the image datasets from the previous section to perform instance segmentation of plant parts. It was installed on a workstation with an NVIDIA GeForce GTX 1080 Ti 11GB GPU with Ubuntu LTS 16.04 64-bit supported by CUDA 9.0.

Mask-RCNN uses a convolutional neural network architecture such as ResNet (He et. al., 2016) as the backbone, which extracts the feature maps. The region proposal network (RPN) applies a sliding window over this feature map and calculates region proposals, which are then pooled with the feature maps and over each, a classifier is applied resulting in a bounding box prediction corresponding to an instance of the particular class. The scheme until this point is the same as FasterRCNN. An additional convolutional network is applied on the aligned region and feature map to obtain a mask for each bounding box.

The following backbone CNN architectures were varied: the 50 and 101-layer ResNets,and ResNext (Xie et. al., 2017) configurations ResNext 101-32x8d (101 layers, cardinality 32, bottleneck width 8) and ResNext 101-64x4d (101 layers, cardinality 64, bottleneck width 4). For these architectures, models pre-trained on the ImageNet1k dataset were downloaded from the Detectron repository. Since this paper addresses the problem of sweet-pepper plant part detection, models pre-trained on ImageNet-1K offer an advantage since one of the classes in this dataset is 'bell-pepper'.

Different transfer learning schemes, training only on the synthetic dataset or only on the empirical dataset, or training on the synthetic set followed by fine-tuning on the empirical were also studied.

3. RESULTS

For the detected objects, true positives, false positives, and false negatives were determined using a criterion that a 25% in overlap of the segmented instance and the ground truth masks should result in a positive detection thereof. This level of overlap was chosen to take into account the fact that some fruit and stems may have disjoint sections due to occlusions, not all of which get detected as part of the same instance. From these measures, the precision and recall were calculated for each plant part class. For comparison with earlier pixel-wise segmentation, the Jaccard Index similarity coefficient or intersection-over-union (IOU) (Csurka et. al., 2013) was calculated by combining all pixels for a particular class and image and comparing against the pixel-wise ground truth.

Figures 2 and 3 present results of segmentation obtained with different transfer learning schemes and architectures, and Table 1 summarizes the figures of merit for the different configurations.

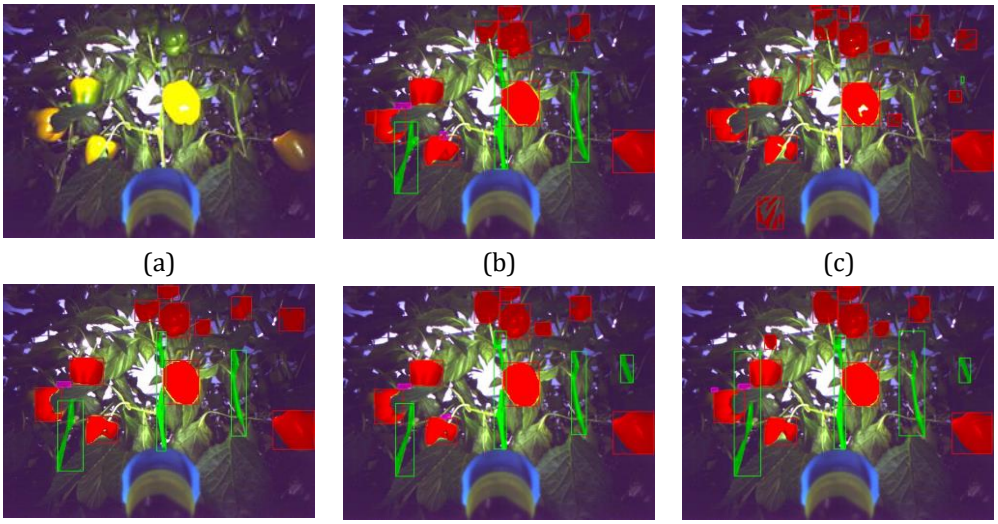


Figure 2: Detection on empirical image no. 36: (a) RGB image; (b) RGB with overlaid GT; detections with: (c) R101 trained on synthetic set; (d) R101 trained only on emp. images 1-30; (e) R101 and (f) X101 64x4d, both trained on synthetic and fine-tuned on emp. images 1-30.

Table 1: Summary of detection statistics using Mask-RCNN with different backbone architectures and training schemes. The 3 best values for each metric over the empirical test set are in bold.

Network		**Fruit**			**Stem**			**Peduncle**	
	P	**R**	**IoU**	**P**	**R**	**IoU**	**P**	**R**	**IoU**
Tested on: emp 31-50									
Pixelwise (VGG)			0.76			0.41			0.39
Experiment 1, Trained on: synth 1-200, Tested on: emp 31-50									
ResNet101	0.56	0.92	0.56	0.11	0.04	0.01	0.73	0.3	0.08
ResNext101 32x8d	0.43	**0.99**	0.54	0.17	0.14	0.09	0.56	0.24	0.09
ResNext101 64x4d	0.41	0.95	0.45	0.4	0.21	0.1	0.7	0.4	0.16
Experiment 2, Trained on: synth 1-200, Finetuned: emp 1-30, Tested on: emp 31-50									
ResNet101	**0.9**	0.96	**0.78**	0.92	0.59	**0.39**	**0.85**	0.71	**0.41**
ResNext101 32x8d	0.87	**0.97**	0.77	**1**	0.56	**0.37**	0.78	**0.84**	**0.46**
ResNext101 64x4d	0.87	0.96	**0.78**	**0.98**	**0.76**	**0.47**	0.77	**0.82**	**0.48**
Experiment 3, Trained on: emp 1-30, Tested on: emp 31-50									
ResNet101	0.87	**0.97**	0.76	0.97	0.56	0.33	0.79	0.62	0.36
ResNext101 32x8d	0.87	0.94	0.74	**1**	**0.67**	0.36	0.82	0.65	0.34
Tested on: alt. emp 1-20									
ResNet101 train syn 1-200	0.21	0.6	0.32	0	0	0.02	0.06	0.08	0.02
ResNet101 train emp 1-30	0.77	0.71	0.72	0.77	0.5	0.16	0.64	0.6	0.2
R101 train syn 1-200 ft emp 1-30	0.67	1	0.69	0.7	0.46	0.17	0.33	0.46	0.07

(a) (b) (c)

Figure 3: Detection results alternative emp. image no. 11: (a) RGB image; (b) RGB with overlaid GT; (c) detection with R101 trained on synthetic images and fine-tuned on emp. images 1-30.

4. Discussion

From the results in Table 1, it can be seen that the best transfer learning scheme was training on the synthetic dataset, followed by fine-tuning on the empirical dataset. This scheme was found to perform better than training only on the empirical dataset. For the detection of the fruits, the best architecture was found to be ResNet101, whereas ResNext101 64x4d was the best for the detection of the stems. The two did not vary much in the detection of the peduncles. With this combination of transfer learning, architectures, and the full empirical training dataset, we obtained for the fruits, a precision of above 0.90 and a recall of close to 1.0, for the stems, a precision of close to 1.0 and a recall above 0.75, and for the peduncles a precision and recall of close to 0.80.

Overall, the intersection over union values for each class were higher than the respective ones reported in previous work on pixel-wise segmentation. Thus, state-of-the-art performance was achieved in detection of these plant parts using Mask-RCNN whilst determining the optimal transfer learning strategy, network architecture and dataset size.

5. Conclusions

In this work, deep learning instance detection was applied to the problem of detecting pepper plant parts. Experimental results show that this approach works well for the detection of fruit, which is useful for practical applications such as harvesting and yield estimation. Our results also show that the synthetic pepper dataset is useful for initial training, to detect peppers in real life images, thus reducing the need for a large number of manually annotated images.

In the experiments, the best numerical results were obtained for the detection of the fruit. The detection of other plant parts such as stems and peduncles is a harder problem likely due to the similarity in color to leaves, twigs, and other unlabeled parts. The segmentation of plant parts could be useful for other plants grown in greenhouses and similar environments, for example tomatoes regarding yield estimation or stem thickness as an important trait for phenotyping.

References

Bac, C. W. (2015). Improving obstacle awareness for robotic harvesting of sweet-pepper. PhD thesis, Wageningen University and Research.

Bac, C. W., van Henten, E. J., Hemming, J., and Edan, Y. (2014). Harvesting robots for high-value crops: State-of-the-art review and challenges ahead. Journal of Field Robotics, 31(6):888–911.

Barth, R. (2018). Vision principles for harvest robotics : sowing artificial intelligence in agriculture. PhD thesis, Wageningen University and Research.

Barth, R., IJsselmuiden, J., Hemming, J., and E.J. Van Henten (2017). Synthetic bootstrapping of convolutional neural networks for semantic plant part segmentation. Computers and Electronics in Agriculture.

Barth, R., IJsselmuiden, J., Hemming, J., and E.J. Van Henten (2018). Data synthesis methods for semantic segmentation in agriculture: A capsicum annuum dataset. Computers and Electronics in Agriculture, 144:284 – 296.

Gabriela Csurka, Diane Larlus, F. P. (2013). What is a good evaluation measure for semantic segmentation? In Proceedings of the British Machine Vision Conference. BMVA Press.

Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 1440–1448.

Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., and He, K. (2018). Detectron. https://github.com/facebookresearch/detectron.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In Computer Vision (ICCV), 2017 IEEE International Conference on, pages 2980–2988. IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778.

Kapach, K., Barnea, E., Mairon, R., Edan, Y., and Ben-Shahar, O. (2012). Computer vision for fruit harvesting robots – state of the art and challenges ahead. Int. J. Comput. Vision Robot., 3(1/2):4–34.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105.

LeCun, Y., Bengio, Y., and Hinton, G. (2015), Deep learning, nature, 521(7553):436.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 779–788.

Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. (2008). Labelme: A database and web-based tool for image annotation. International Journal of Computer Vision, 77(1):157–173.

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., and McCool, C. (2016). Deepfruits: A fruit detection system using deep neural networks. Sensors (Basel), 16(8):1222. 27527168[pmid].

Uijlings, J. R., Van De Sande, K. E., Gevers, T., and Smeulders, A. W. (2013). Selective search for object recognition. International journal of computer vision, 104(2):154–171.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 5987– 5995. IEEE.