Real-Time Vegetables Recognition System based on Deep Learning Network for Agricultural Robots

Yang-yang Zheng[1], Jian-lei Kong[1,*], Xue-bo Jin[1], Ting-li Su[1], Ming-jun Nie[2] and Yu-ting Bai[3]

[1]School of Computer and Information Engineering, Beijing Technology and Business University, Beijing,100048, China
[2]Beijing Agricultural Information Technology Research Center, 100097
[3]School of Automation, Beijing Institute of Technology, 100081, Beijing
[*]Correspondence: kongjianlei@btbu.edu.cn; Tel.: +86-138-1029-8315

***Abstract*—Automation is a major challenge in the application of agricultural robots. Therefore, in order for agricultural robots to automatically classify and detect vegetables, an effective identification system should be established, which will increase production efficiency.In this paper, we propose a real-time vegetables recognition system based on deep learning network to detect and classify vegetables. On the basis of the large vegetable dataset obtained by our pinking robot, our goal is to find the appropriate framework for pinking mission with high accuracy and fast speed. Therefore, we select several one-stage and two-stage networks as alternative detectors including Faster R-CNN, SSD, RFB Net,YOLOv2, and YOLOv3. With the promotion of data augmentation, we demonstrate the recognition network based on YOLOv3 combing the depth meta-framework and feature extractor achieve the well performance in terms of both accuracy and speed by reducing the number of false positives in training phase. Comparative experiments indicate that our recognition system can effectively identify seven different kinds of vegetables (the mean AP can reach 87.89%, and the detection speed is up to 38FPS), which would handle with real-time detection task and improve pinking capability of our agricultural robot.**

***Index Terms*—convolutional neural networks, vegetable, real-time detection**

I. Introduction

With the development of intelligent agriculture, automatic agricultural machines and Agricultural automation is the main trend in the field of agricultural development. The existing agricultural robots include picking robots, grafting robots and splashing robots, which are intelligent and precise to gradually replace manual labor in the foreseeable future. However, those robots presented in automatic or semiautonomous ways cannot accomplish the complex operations including vegetable picking, pesticide splashing, etc. In order to achieve agricultural automation, agricultural robots currently have many problems to solve, such as how to detect the location and classification of vegetables. In order to enable the picking robot to automatically pick vegetables and fruits, an automatic picking system should be constructed, which can improve the efficiency of agricultural picking. With the rapid development of deep learning and high-computational hardware technology, picking robots usually take advantage of convolutional neural networks (CNN) to form the detection and classification method for vegetable, which are mainly divided into two-stage object detectors and single-stage object detectors. For two-stage detection network, a series of candidate object boxes is first generated, then perform classification and regression operations. The basic region-based convolutional neural network (R-CNN) [1] is the earliest application of CNN features to construct the detection system with well performance. Then a fast region-based convolutional neural network (Fast R-CNN) [2] is proposed to combine the target classification with bounding box regression to solve multi-task detection. Further, the regional proposal network (RPN) is proposed by the faster region-based network (Faster R-CNN) [3] to generate amounts of anchors, which are richer proposals to improve accuracy slightly. While Faster R-CNN has been a milestone of the two-stage detectors, these algorithms have some disadvantages in dealing with resources and time consumption for large datasets, which is not suitable to subsequent realistic application of agricultural robots.

Therefore, single-stage object detection networks with faster speed are trained by regular and dense sampling over locations, scales and aspect ratios in an end-to-end flow. The main advantage of single-stage method is its high improvement in computational efficiency greatly, which is suitable for realistic tasks. The representative methods such as You only look once v2(YOLOv2) [4] and YOLOv3 [5] simplify object detection as a regression problem. However, the single-stage detection accuracy is lower than the two-level detection accuracy, mainly caused by the class imbalance problem. The latest method in the one-stage approach aims to solve this problem. Single shot multibox detector (SSD) [6] improves accuracy performance by producing different scale predictions and fusing feature maps of different layers. Similarly, receptive field block network (RFB Net) [7], inspired by the receptive fields of human vision, proposes a novel RFB block module to significantly reduce the search space of objects. Further, RetinaNet [8]

This work is partially supported by National Key R&D Program of China No. 2017YFC1600605, Research Foundation for Youth Scholars of Beijing Technology and Business University No. QNJJ2017-15 and QNJJ2016-13, NSFC under Grant No. 61673002, Beijing Municipal Education Commission No. KM201810011005, Construction of technological innovation and service capability - Basic scientific research service fee-innovation platform No. PXM2018_014213_000033.

978-1-7281-1312-8/18/$31.00 ©2018 IEEE

design a FPN-based single-level detector involving Focal-Loss to reduce false positives resulting from class imbalance caused by extreme foreground-to-background ratios.

In our opinion, the current two-stage and single-stage methods build the framework tone of the target detection together. The essential difference between two detectors is the tradeoff between the recall and localization, which fundamentally determine the accuracy and detection time. The single-stage detector has a higher recall at cost of low localization. Instead, the two-stage detector has a higher positioning capability, but the recall is lower, since the refine of the box's precision could kill some positive samples by mistake. Both of above methods have achieved top performances on several challenging benchmarks, including PASCAL VOC and MS COCO. However, the detection effect is still unclear in face of special inspection tasks such as vegetable pinking.

In this paper, we proposed a new vegetable dataset and designed an object detection framework to identify various vegetables by the application of depth meta-architectures [9] and feature extractor. Then an automatic pinking system was developed successfully to collect the different types of vegetables in real agriculture scenes. In addition, the system can handle with complex tasks that the fruit is partially blocked by leaves, complex campus environments, etc. By comparing different deep networks, confirming the classification and detection model with faster speed and high accuracy is the main purpose for the detection system, which improve the efficiency of realistic vegetables picking and other agriculture operation.

The content of this paper is divided into the following parts: Section 2 introduces the vegetables dataset. Section 3 describes the deep convolutional neural network used. Section 4 presents our experiments to results on the vegetable dataset. Finally, we make a conclusion and discuss the further research in Section 5 presents.

II. Vegetable Dataset

A. Data Collection

The vegetables dataset was collected by an automatic robot for picking vegetable in the vegetable greenhouse as showed in Figure1. In the greenhouse the tables of vegetables are arranged in rows, where the height of the plants is about 1.5 meters. The robot system takes photos along these lines. Vegetable monitoring is done by installing two cameras above the horizontal line. The robot takes pictures of vegetables at intervals of 0.1 meters with different angles and positions, making the dataset more generalized.

B. Data Annotation

On the basis of the vegetable dataset, we pre-process the image as a uniform size of 800*800, then manually annotate the images. The annotation process is to mark the vegetables in the picture with a bounding box and output the corresponding vegetable class. Therefore, the vegetables dataset is divided into the following seven categories: tomato, unripe tomato, tomato early-blossom, tomato full-blossom, cucumber, cucumber blossom, unripe cucumber. Figure2 shows sample of seven classes vegetables.

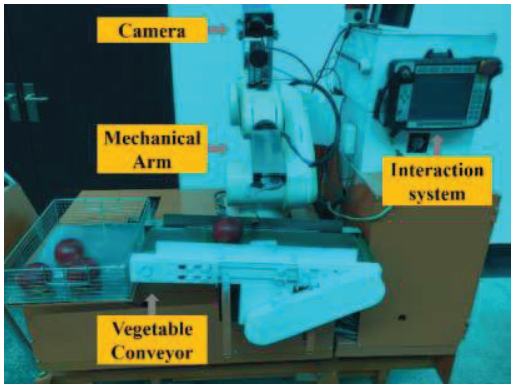


Fig. 1. Automatic picking robot of vegetable.

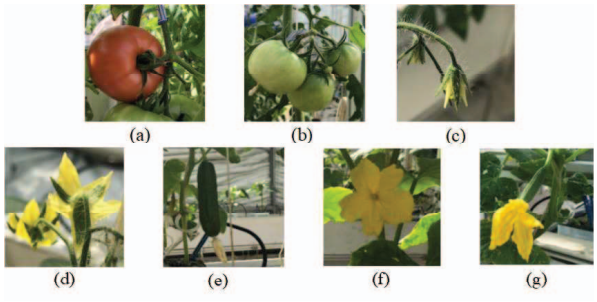


Fig. 2. Figure 2 Representative of the vegetables dataset. (a) tomato, (b) unripe tomato, (c) tomato early-blossom, (d) tomato full-blossom, (e) cucumber, (f) cucumber blossom, (g) unripe cucumber.

C. Dataset structure

Our vegetables dataset includes approximately 7,083 images collected from the Vegetable Base at Beijing Agricultural Information Technology Research Center. These images were taken under a greenhouse. They include four stages of tomato from flowering to fruiting; three stages of cucumber from flowering to fruiting. The categories and quantities of annotated samples used in the system can be seen in TableI. Each image contains multiple annotated samples, so the number of samples for each class is different.

III. Deep Learning Detector

A. Real-time Vegetables Detection System

With the continuous improvement of hardware technology, the development of deep convolutional networks with better performance has been promoted.We introduced a series of inspection networks in front, so we mainly focus on the following five architectures: Faster R-CNN, SSD, RFB Net, YOLOv2, and YOLOv3. Our task is to detect and classify vegetables in the image. An overview of the real-time vegetable detection system is shown in Figure3.

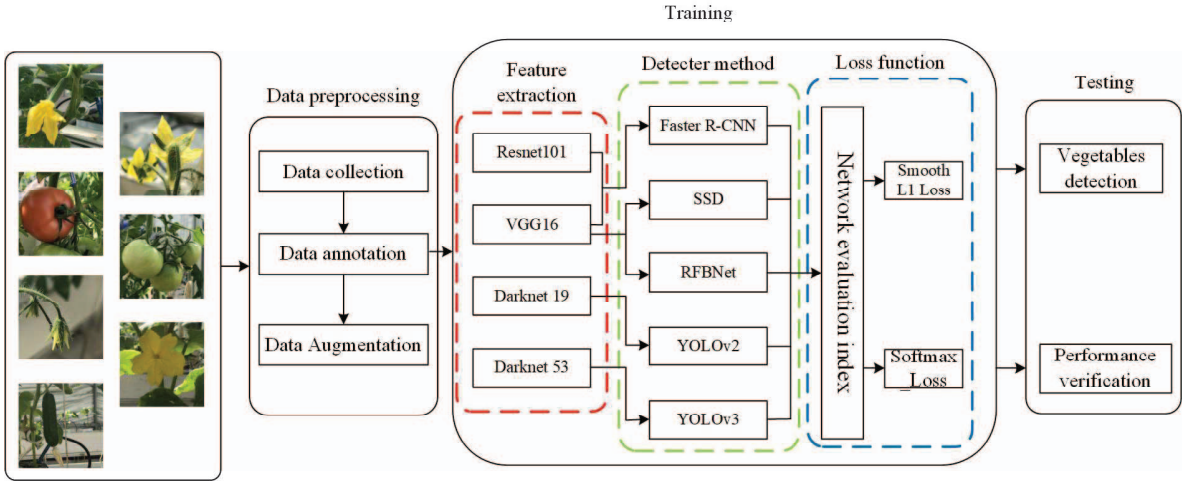


Fig. 3. The structure flow of the vegetable recognition system based on deep learning. The architecture of the system includes input image preprocessing, training networks, detection and classification of vegetables in the image.

TABLE I
CATEGORIES LIST OF OUR VEGETABLE DATASET AND ACCORDING ANNOTATED SAMPLES

Class	Number of Images in the Dataset	Number of Annotated Samples (Bounding Box)	Percentage of Bounding Box Samples(%)
Tomato	1021	3219	12.99
Unripe tomato	898	2987	12.06
Tomato early-blossom	1153	4568	18.44
Tomato full-blossom	1089	4134	16.68
Cucumber	872	2567	10.36
Unripe cucumber	1138	4098	16.54
Cucumber blossom	912	3204	12.93
Total	7083	24777	100

B. Faster Region-based Network

In the Faster R-CNN, the detection process takes place in two phases. In the first phase, a Regional Proposal Network (RPN) processes the image through a feature extractor and produces a score for each proposal.In the second phase, these uniformly sized features are sent to the classification and regression layers to predict the category and bounding box to which each regional suggestion bounding box belongs. This results in a regional proposal with almost no cost. Since Faster R-CNN has a relatively high degree of accuracy in object identification and classification, several networks have been developed on the basis of it.

C. Single Shot Detector Network

The SSD Network solves the object recognition problem by using a feed-forward convolution network. This feedforward convolution network produces a fixed bounding box size and predicts the score of the object in the box.The network predicts objects of different sizes by using feature maps of different resolutions. Moreover, the SSD encapsulation process is an end-to-end network, so it can avoid generating solutions and thereby save computation time.

D. Receptive Field Block Network

Based on the receiving field (RF) structure in the human visual system, a novel RF module (RFB) module is proposed. This module considers the relationship between RF size and eccentricity to enhance the identity of the feature. The RFB module is assembled by further assembling the RFB module to the top of the SSD using a lightweight CNN model to further assemble the RFB module. RFB Net enables the accuracy of advanced deep backbone network probes while maintaining real-time speed.

E. YOLOv2 Network

Although YOLOv1 [10] has a fast detection speed, it is not as accurate as the Faster R-CNN detection method in terms of detection accuracy. YOLOv1 is not accurate enough for object localization and has a low recall rate. After that, YOLOv2 is proposed to improve YOLO in several aspects, i.e., add batch normalization on all convolution layers, use high resolution classifier, use convolution layers with anchor boxes to predict bounding boxes instead of the fully connected layers, etc. YOLOv2 follows a principle in the improvement: maintaining detection speed, which is also a big advantage of the YOLO model.

F. YOLOv3 Network

YOLOv3 is proposed to improve YOLOv2 in several aspects. i.e., (1) YOLOv3 predicts an object score for each bounding box using logistic regression. (2) Each box uses a multi-label classification to predict which classes the bounding box might contain. YOLOv3 uses binary cross entropy loss instead of Softmax for class prediction in training. (3) YOLOv3 predicts three different scales of boxes. Our system extracts feature from these scales using concepts similar to pyramid networks. (4) The new network is used to perform feature extraction. It has 53 convolutional layers, so it is called Darknet-53.

The goal of our training is to reduce the error between the true and predicted values by adjusting the parameters in the network, thus minimizing the loss function. TableII list the loss function, train picture size, and parameter quantities used in this work.

TABLE II
DEEP LEARNING META-ARCHITECTURES AND FEATURE EXTRACTORS OF EACH NETWORK

Meta-Architecture	Feature Extractor	Train Size	Loss Function	parameter
Faster R-CNN	VGG-16	800*800	SmoothL1	573M
Faster R-CNN	ResNet-101	800*800	SmoothL1	769M
SSD	VGG-16	800*800	SmoothL1	98M
RFB Net	VGG-16	800*800	SmoothL1	130M
YOLOv2	Darknet-19	800*800	Softmax	203M
YOLOv3	Darknet-53	800*800	Softmax	246M

IV. RESULT AND DISCUSSION

A. Experimental Implementation

We experimented with the vegetable dataset, which included seven categories of annotated vegetables. Since the image data of the vegetable dataset is still small, so as to avoid over-fitting of the network, a large amount of data is added using the data augmentation method. We used a method to enhance the image dataset. These data enhancement techniques include geometric transformations such as resizing images, random cropping, rotation, and horizontal flipping. Intensity transformations such as enhanced contrast and brightness, color changes, increased noise, etc.During the experiments, our data set was divided into 80% training set, 10% verification set and 10% testing set. The vegetables dataset was trained and tested on an Intel Core i7 3.6GHz processor with four NVIDIA Tesla p40 GPUs and 256G RAM.

B. Test Result

Our system can detect the class and location of the vegetables in the picture, as shown in Figure4. We use IoU > 0.5 as a reference to compare estimates result and ground-truth. Each class is independent of each other because they exhibit different characteristics in the network. Using the meta-architecture and the depth feature extractor, the system shows several advantages when dealing with objects of various sizes, shapes, colors, etc., compared to previous conventional methods.

C. Accuracy Evaluation

Our proposed system handles complex vegetable datasets in images by using different depth networks. Thereby enabling detection and identification. The performance of our system was first evaluated based on the Intersection-over-Union (IoU) and Average Precision (AP) introduced in the Pascal VOC Challenge.

$$IoU(A,B) = \left| \frac{A \bigcap B}{A \bigcup B} \right| \tag{1}$$

where A represents the ground-truth box collected in the annotation, and B represents the predicted result of the network.

The Average Precision is the area under the Precision-Recall curve of the detection task. Like the Pascal VOC Challenge, the AP is calculated by averaging the precision of a set of intervals recall levels $[0, 0.1, ..., 1]$, which is the AP calculated for all categories in our task.

$$AP = \frac{1}{11} \sum_{r \in \{0, 0.1, ..., 1\}} p_{\text{interp}}(r) \tag{2}$$

$$p_{\text{interp}}(r) = \max_{\tilde{r}:\tilde{r} \geq r} \max p(\tilde{r}) \tag{3}$$

where $p(\tilde{r})$ is the measure precision at recall $\tilde{r}$. Next, we compute the mAP averaged with the IoU = 0.5(due to the complexity of the scenarios). The detection results are shown in TableIII.

Vegetable test results show that shallow networks perform better than the deeper network in our task. Such as the case of Faster R-CNN with VGG-16 obtains the total mean AP at 81.08%, which improve the result about 13% than the same meta-architecture with ResNet-101 achieving 67.73%. This indicates that the detection accuracy is affected by the feature extraction. Since the type of the dataset is too small to meet the requirements of the deep network, the low- quality feature would lead to the lower accuracy. When using different meta-architectures, the mean AP of Faster R-CNN, YOLOv2, and YOLOv3 can reach 80% or even more. The mean AP of YOLOv3 reached 87.89%, indicating that the YOLOv3 meta-architecture showed better detection performance for the vegetable dataset.

D. Speed Evaluation

To further analyzing the performance of each network on the detection speed, this system needs to be operated on a vegetable picking robot, which requires not only accuracy but also real-time performance. We need system to perform real-time detection on the picking robot, thus improving the efficiency of the picking robot. The speed is evaluated on a machine with NVIDIA P40, CUDA 8.0 and cuDNN v6 which Speed detection evaluation indicator using Frames Per Second (FPS). The result of the detection time is shown in Figure5.

The result of speed detection shows that Faster R-CNN is too slow to achieve the effect of real-time detection and It can only reach 7FPS. SSD and RFB Net are approximate 19FPS that it can achieve real-time detection. However, its detection accuracy is too low, which it does not have good performance for vegetable dataset detection. YOLOv2 and YOLOv3 have the fastest detection speed and It can reach approximate 40FPS. In particular, YOLOv3, with the speed of real-time detection, has a detection accuracy of up to 87.89%, which is the best for the detection of vegetables dataset.

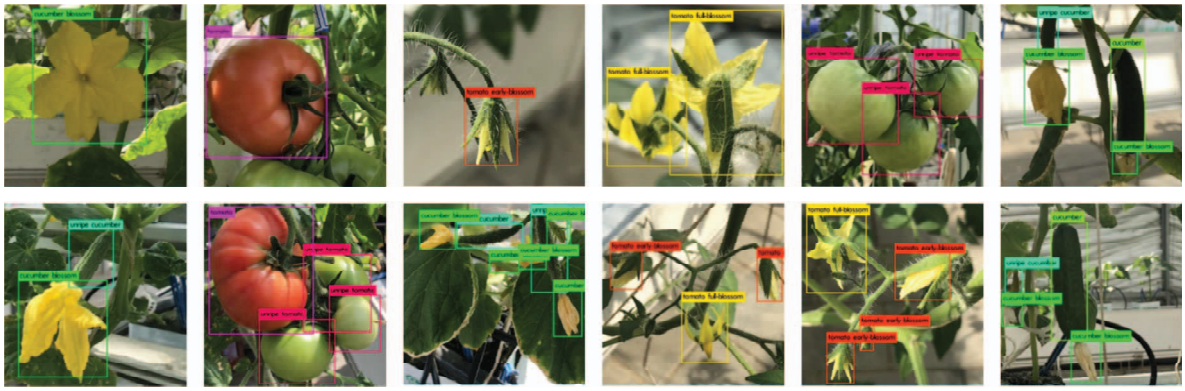


Fig. 4. Recognition result for vegetable dataset.

TABLE III

DETECTION RESULTS OF OUR PROPOSED SYSTEM USING DEEP-LEARNING META-ARCHITECTURES AND FEATURE EXTRACTORS

	Meta-Architectures					
	Faster R-CNN	Faster R-CNN	SSD	RFB Net	YOLOv2	YOLOv3
Class/Feature Extractor	VGG-16	ResNet-101	VGG-16	VGG-16	Darknet-19	Darknet-53
Tomato	0.8842	0.6932	0.8267	0.8835	0.8808	0.9201
Unripe tomato	0.8580	0.7015	0.8172	0.8632	0.8909	0.9188
Tomato early-blossom	0.8219	0.7317	0.7856	0.8121	0.8597	0.8724
Tomato full-blossom	0.7748	0.6838	0.7423	0.8369	0.8746	0.8960
Cucumber	0.8066	0.6572	0.7791	0.8549	0.8950	0.8888
Unripe cucumber	0.7821	0.6489	0.7197	0.8628	0.7699	0.8831
Cucumber blossom	0.7482	0.6249	0.7534	0.7157	0.7896	0.7730
Total mean AP	**0.8108**	0.6773	0.7749	**0.8326**	**0.8515**	**0.8789**

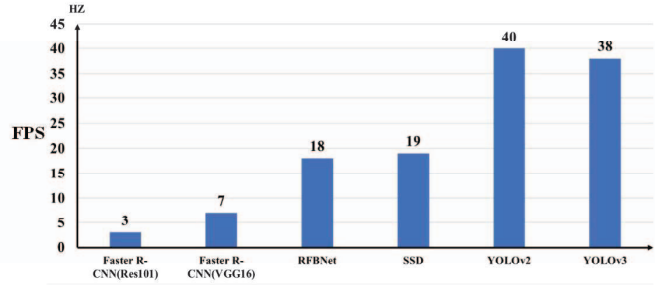


Fig. 5. Detection speed of deep networks selected

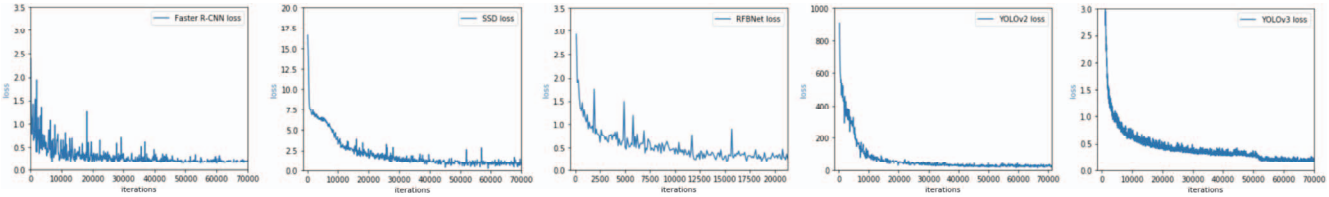
E. *Loss function*

The loss function is used to estimate the difference between the predicted value $f(x)$ of the model and the true value Y. It is a non-negative real-valued function, usually represented by $L(Y, f(x))$. The smaller the loss function, the better the robustness of the model. The structural risk function of the model includes empirical risk terms and general terminology. Usually can be expressed as follows:

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i; \theta)) + \lambda\phi(\theta)] \qquad (4)$$

where the previous mean function represents the empirical risk function, L represents the loss function, and the latter θ is a regularization or a penalty term, which can be $L1$ or $L2$, or other regular functions. The whole expression means to find the value of θ when the objective function is minimized. Figure6 shows the loss function diagram for the five meta-architectures.

The comparative results of loss function show that in our task, the trend of the loss function of five meta-architectures is reduced. It stabilized at about 50,000 Iterations, which the predicted value of the network was closer to the true value. Among them, the loss function of the three meta-architectures of Faster R-CNN, SSD and RFB Net has large fluctuations in the process of descent, while the loss function of YOLOv2 and YOLOv3 is relatively smooth. In comparison, the loss function of YOLOv2 and YOLOv3 decreases faster, and the learning ability of the network is stronger.



(a) Faster R-CNN (b) SSD (c) RFB Net (d) YOLOv2 (e) YOLOv3

Fig. 6. Loss function graphs of each network.

F. *Confusion Matrix*

Because the patterns displayed in each category have different complexity, especially in terms of vegetable categories and backgrounds, the system does not accurately distinguish between several categories, resulting in reduced accuracy. In Figure7, we calculated the confusion matrix of the vegetable dataset test results. By analyzing the confusion matrix, we can directly evaluate the performance of the network. In addition, the confusion matrix helps us further analyze the program to avoid re-confusion between these different classes. For instance, it can be seen from the figure that the prediction and true correlation of cucumber flowers are lower than others. It indicates that some confusion between cucumber blossom and tomato full-blossom and tomato early-blossom, which leads to the decrease of detection precision of cucumber blossom.

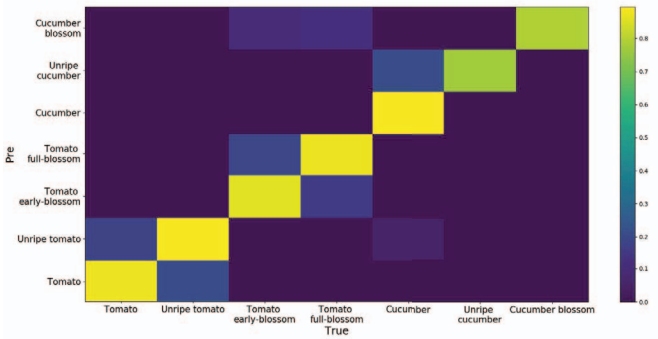


Fig. 7. Confusion matrix of the vegetables detection results

V. Conclusion

Agricultural picking robots need to identify and classify vegetables, which it is not only necessary to improve the accuracy of detecting vegetables, but more importantly, the speed of detection. In this work, we focus on finding a more appropriate deep learning framework which can real-time identify and classify vegetables to accomplish the function of picking robot. Experiments show that with the promotion of data augmentation, the YOLOv3 network has higher accuracy of 87.89%, indicating that YOLOv3 combing the depth meta-framework and feature extractor achieve the well performance in terms of both accuracy and speed by reducing the number of false positives in training phase. Moreover, this method has real-time performance that the detection speed is up to 38FPS. Thereby, the work efficiency of the picking robot can be improved, and the automation of the agricultural equipment can be greatly promoted. In the future work, we will improve the current results and improve the accuracy based on the speed. It is hoped that the system will be extended to the identification of agricultural pests and diseases, improve the efficiency of agriculture, and make agriculture more intelligent.

References

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

[2] R. Girshick, "Fast R-CNN," In Proceedings of the IEEE Inter- national Conference on Computer Vision, 2015, pp. 1440-1448.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," In Advances in neural information processing systems, 2015, pp. 91-99.

[4] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA, 2017, pp. 6517-6525.

[5] Redmon J, Farhadi A, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.- Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox Detector," In European Conference on Computer Vision, 2016, pp. 21-37.

[7] Liu S, Huang D, Wang Y, "Receptive Field Block Net for Accurate and Fast Object Detection," preprint arXiv:1711.07767, 2017.

[8] T. Lin, P. Goyal, R. Girshick, K. He and P. Dollar, "Focal Loss for Dense Object Detection," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2018, pp. 2999-3007.

[9] J. Huang et al., "Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA, 2018, pp. 3296-3297.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," In Pro-ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779-788.