Peer-reviewed version available at *Agriculture* **2018**, *8*, 196; doi:10.3390/agriculture8120196

Article

Detection of key organs in tomato based on deep migration learning in complex background

Sun Jun*, He Xiaofei , Ge Xiao , Wu Xiaohong , Shen Jifeng and Song Yingying

School of Electrical and Information Engineering of Jiangsu University, Zhenjiang 212013, China; sun2000jun@sina.com (J.S.); 769598959@qq.com (X.H.); 2685326848@qq.com (X.G.) wxh419@ujs.edu.cn (X.W.); shenjifeng@ujs.edu.cn (J.S.); 810541544@qq.com (Y.S.)
*Correspondence: sun2000jun@sina.com; Tel.: +13775544650

Abstract: In the current natural environment, due to the complexity of the background and the high similarity of the color between immature green tomato and plant, the occlusion of the key organs (flower and fruit) by the leaves and stems will lead to low recognition rate and poor generalization of the detection model. Therefore, an improved tomato organ detection method based on convolutional neural network has been proposed in this paper. Based on the original Faster R-CNN algorithm, Resnet-50 with residual blocks was used to replace the traditional vgg16 feature extraction network, and K-means clustering method was used to adjust more appropriate anchor size than manual setting to improve detection accuracy. A variety of data augmentation techniques were used to train the network. The test results showed that compared with the traditional Faster R-CNN model, the mean average precision (mAP) of the optimal model was improved from 85.2% to 90.7%, the memory requirement decreased from 546.9MB to 115.9 MB, and the average detection time was shortened to 0.073S/sheet. As the performance greatly improved, the training model can be transplanted to the embedded system, which lays a theoretical foundation for the development of precise targeting pesticide application system and automatic picking device.

Key words: object detection; tomato organ; K-means clustering; Soft-NMS; migration learning; convolutional neural network; deep learning

1. Introduction

Tomato is native to South America and has a cultivation history of more than 100 years in China. It is one of the most popular fruits and vegetables, which has health effects of lowering blood pressure, slowing down aging, slimming and supplementing vitamins[1].

Due to the wide planting area, high yield and short maturity, tomato is difficult to be stored. The picking of tomato is a labor-intensive and long-consuming operation. Therefore, the development of the automatic picking robot for tomato is particularly important. The technology of object detection based on image processing is the most basic and important link in the research of picking robot. It is a typical object recognition problem that detecting the key organs (flower and fruit) of tomato from the image of plants. Data from the internet display that immature green tomato contains alkaloids, which may cause poisoning after eating, while mature red tomato does not contain alkaloids. Therefore, it is significantly important to detect the key organs (tomato flower, immature green tomato and mature red tomato) accurately and immediately. It will provide a theoretical basis for the detection of pest and disease, targeting pesticide application system, and development of picking robot [2]. Traditional tomato recognition methods mostly extract the information of color, shape and some shallow features, then use classifier to detect and recognize

tomatoes. Zhao Yuanshen et al.[3] used the combination of Harr features and Adaboost classifier to identify tomatoes, and the recognition rate of mature tomato in the test set was 93.3%; Jiang Huanyu et al.[4]utilized the difference of color characteristics between mature tomatoes and background to identify red mature tomatoes by threshold segmentation method. However, the recognition rate of immature green tomato similar to the color of the leaf is not high; Zhang Ruihe et al.[5] transformed the red tomato image to extract the edge features of background region, then segmented the tomato from the background by fitting curve, and obtained the three-dimensional coordinates of the target by using the principle of stereo vision imaging; K Yamamoto et al. [6] extracted the color characteristics of leaves, stems and backgrounds, then constructed a decision tree through regression tree classifier and extracted the pixels of tomato fruits by blob pixel segmentation to identify tomato. But this method has a poor identification rate under the complex background. At present, most traditional object recognition and detection methods are based on the shallow feature extraction to detect and identify tomato organs. The overlapping and occlusion problems between tomatoes under the complex background cannot be well solved, and the time cost for feature extraction is relatively expensive and the applicability is not strong.

In recent years, with the emergence of deep learning, convolutional neural networks (CNN) can extract hierarchical features by using unsupervised or semi-supervised feature learning, which has stronger generalization than artificial features. It can not only learn shallow semantic information, but also learn deep abstract information. In the field of generalized recognition (such as object class [8], object detection [9] and object segmentation [10]), Faster R-CNN [11]and YOLO [12] models have been widely used and achieved good results. Suchet Bargoti et al. [13] constructed an image segmentation processing framework by using orchard image data. And multi-scale multilayer perceptron (MLP) and convolutional neural network were include in the framework. The image data captured from network was extended by context information, as well as some of the appearance changes and category distributions observed in the data. Finally, Watershed Segmentation (WS) and Circular Hough Transform (CHT) processing algorithms were used for fruit detection and counting; Inkyu Sa et al. [14] constructed the Faster R-CNN model by pixel level superposition of RGB and near infrared images, then the model was used for fruit detection by migration learning, but the model parameter was too large; Zhou Yuncheng [15] proposed a Fast R-CNN model based on double convolutional chains. VGG Net-based feature extraction network was trained by merging RGB and grayscale feature images of tomatoes, and then Fast R-CNN was initialized with the parameters to identify flowers, fruits, and stem of the tomato in plant images. The above literature proves that convolutional neural network can not only extract the shallow texture and color of key organs of tomato automatically, but also learn deeper abstract features. This will improve the detection accuracy of tomato flowers and fruits, and reduce the cost of feature extraction, which is more robust to the detection and identification in complex environment. So it is feasible to use deep learning methods to detect and identify tomato organs. However, the accuracy of these models is not high under the conditions of overlapping and occlusion. And the anchor size used in the above method was set artificially by the target in the VOC 2007 data set, without updating according to the actual size of the key organs of tomato, which will have a certain impact on the recognition accuracy.

Aiming at the problem of low recognition accuracy of key organ detection in tomato with fixed size, overlap and occlusion, an improved Faster R-CNN model was proposed in this paper. Firstly, K-means clustering was used to obtain the anchor size that suitable for key organs of tomato.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2018 doi:10.20944/preprints201810.0524.v1

Peer-reviewed version available at Agriculture 2018, 8, 196; doi:10.3390/agriculture8120196

Secondly, residual blocks were used to replace the basic feature extraction network of the original model in order to improve the detection and recognition accuracy of tomato key organs. Finally, Soft-NMS was used to attenuate the bounding boxes of the tomato organ instead of completely removing them, which can solve the above problems to a certain extent.

2. Materials and Methods

2.1. Data sources

Figure1. Examples of key organ of tomato

A total of 5624 RGB images of tomato flowers, mature red tomatoes and immature green tomatoes were collected from the agricultural digital greenhouse of Jiangsu University with high-definition camera. In order to prevent the poor performance of model caused by insufficient diversity of training samples, following measures were taken during the process of image acquisition: considering the difference in imaging results caused by different greenhouse ambient light conditions, images were collected in sunny weather and cloudy days respectively; in the process of sampling, different forms and occlusions of tomato organs were taken into account, and fruits with different maturity were photographed from multiple angles to increase the diversity of samples. In this paper, python script was used to augment a small number of sample maps, including random flip (horizontal, vertical), transform angle (0 ~ 180), random scaling of the original image scaling factor (1 ~ 1.5) etc. [16]. The total number of expanded samples was 8929, and tags were produced according to the standard voc2007 dataset format, then the expanded set was randomly divided into 4:1 ratios between the training set and the test set.

2.2. Structure of the detection model of tomato organs

Convolutional neural network (CNN) includes convolutional layer, pooling layer and fully connected layer. The convolutional layer uses semi-supervised feature learning and hierarchical feature extraction efficient algorithm to extract image abstract features. It can automatically extract and reduce the dimension of input images, and has stronger generalization than the human-set features. The fully connected layer mainly performs image classification based on the extracted features. The neurons in the convolutional layer extract the primary visual features of the image by using local receptive field and reduce the network parameters by sharing the weights. The pooling layer not only reduces the dimension of the features, but also realizes the invariance of displacement,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2018 doi:10.20944/preprints201810.0524.v1

Peer-reviewed version available at Agriculture 2018, 8, 196; doi:10.3390/agriculture8120196

scaling and distortion. The convolutional layer and pooling layer in CNN usually appear alternately, and the activation unit is set to realize the nonlinear transformation, which accelerates the convergence rate of the network.

The traditional Faster R-CNN object detection algorithm can be divided into two parts. Vgg16 was selected as the basic feature extraction network for image feature extraction and classification. The network consists of eight convolutional layers, five maximum pooling layers and three fully connected layers, and Re LU (Rectified Linear Unit) was used as the activation function. In region proposal network (RPN), an arbitrary scale image is taken as input and outputs a series of object proposals, and each proposal has an object score. In order to generate region proposals, slide on the feature map of the last convolutional layer of vgg16, and multiple region proposals are predicted at the same time in each sliding window position [17].

2.3. Improved feature extraction network

Deep convolutional neural network makes a great progress in image recognition rate compared with traditional methods. As the number of network layers deepens, the convolutional layer can learn deeper abstract features, and the related research also shows that the extracted features can be enriched by increasing network layers. Simon yan et al. [18] demonstrated that the accuracy of recognition rises with the increase of the depth of network, but the simple stacked convolutional layer cannot train the network smoothly due to the gradient explosion when propagating backward. Although relevant literature has shown that the deep network can be trained through batch normalization [19] and dropout [20], the problem of precision decline after a certain iteration still exists. In order to break through the problems of precision degradation of deep network and the limitation of network depth, He et al. proposed a deep residual resnet model by using the concept of identity mapping. This method solved the degradation problem by fitting a residual map with a multi-layer network [21]. For a stacked layer structure (stacked by several layers), if the residual is zero, then the stacked layer only makes an identity map, the network performance will not decrease. But in fact the residual is not zero, which can enable the stacked layers to learn new features based on the input characteristics, thus having better performance. The basic idea is as follows (as shown in Figure 2): Since direct mapping is difficult to learn, the basic mapping relationship from X to $H(x)$ is no longer being learned, but the difference between the two is learned, that is the residual, then in order to calculate $H(x)$, just add this residual to the input.

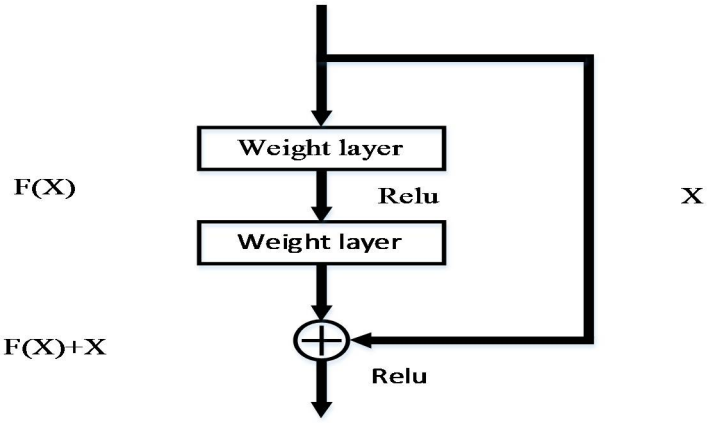


Figure2. Residual learning module

At this point, the original optimal solution mapping $H(x)$ can be equivalent to$F(X) + X$, that is, the fast connection implementation in the feed-forward network shown in Figure 2. The method of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2018 doi:10.20944/preprints201810.0524.v1

Peer-reviewed version available at Agriculture 2018, 8, 196; doi:10.3390/agriculture8120196

quick connection can be expressed by formula (1):

$$Y = F(X, \{W_i\}) + W_s X \quad (1)$$

Among the formula(1), X represents the input vector of the module, Y represents the output vector of the module, W_i represents the weighting layer parameters, and a linear projection W_s is needed to match the dimensions to ensure the consistency of the input and output dimensions.

ResNet-50 is consist of 50 modules with the same structure as shown in Figure 2. In this paper, the vgg16 feature extraction model was replaced by the 50-layer residual network ResNet-50 to improve the classical Faster R-CNN depth learning model, and the improved detection framework was displayed in Figure 3.

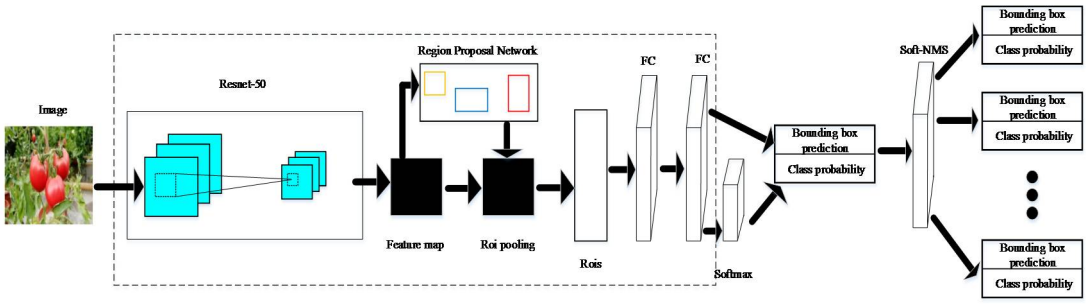


Figure3. Improved framework for the Faster R-CNN model

2.4. Use K-means to cluster the appropriate anchor size

The purpose of the clustering algorithm (K-Means) is to divide n objects into k different clusters according to their respective attributes, so that the similarity of each object in the cluster is as high as possible, and the similarity between clusters is as small as possible. The criterion function used for evaluating similarity is the sum of squared errors. Since the convolutional neural network has translation invariance and the position of the anchor boxes is fixed by each grid, it is only necessary to calculate the width and height of the anchor boxes by k-means. Due to the use of euclidean distance will cause larger bounding boxes to produce more errors than smaller bounding boxes, a new distance formula is defined for this purpose:

$$d(box, centroid) = 1 - \mathrm{IOU}(box, centroid) \quad (2)$$

When the anchor boxes are calculated, the x and y coordinates of all boxes' center points will be set to zero so all the boxes are in the same position, which is convenient for calculating the similarity between boxes by the new distance formula [22].

2.5. Region proposal network

The core idea of RPN is to directly generate region proposals by using convolutional neural network, which is essentially a sliding window [23]. At the center of the sliding window, a total of nine kinds of anchors are generated, corresponding to three scales (8*8, 16*16, 32*32) and three aspect ratios (1:1, 1:2, 2:1) of the input image. Then the predicted region proposals are sent to two fully connected layers: cls layer and reg layer, which are used for classification and box regression respectively. And finally, the top 300 region proposals are selected as input of Fast R-CNN after sorting according to the score of region proposals.

2.6. Soft-NMS

Non-maximum suppression (NMS), as the name implies is to suppress elements that are not the maxima, searching for local maxima. This local area represents a neighborhood with two

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2018 doi:10.20944/preprints201810.0524.v1

Peer-reviewed version available at Agriculture 2018, 8, 196; doi:10.3390/agriculture8120196

variable parameters: one is the dimension of the neighborhood, and the other is the size of the neighborhood. Non-maximum suppression is an important part of the object detection process, which generates the detection box based on the object detection score. The detection box with the highest score is selected, while other detection boxes with obvious overlap with the selected detection box are suppressed. This process is continuously recursively applied to the remaining detection frames [24]. For example, In pedestrian detection, each window will get a score after feature extraction and classifier recognition, but sliding window will also cause many windows to overlap with other windows, so NMS is needed to select the detection box with the highest score in the neighborhood, and suppress those boxes with low scores.

Soft-NMS has the same algorithmic complexity as traditional NMS. It only needs to make simple changes to the traditional NMS algorithm without adding additional parameters and training, and can be easily integrated into any object detection process with high efficiency and easy implementation.

The score reset function of the traditional NMS (rescoring function) is calculated as follows:

$$s_i = \begin{cases} s_i, & iou(M, b_i) < N_t \\ 0, & iou(M, b_i) \geq N_t \end{cases} \tag{3}$$

In the above formula (3), a threshold is used in NMS to determine whether adjacent detection frames are reserved. Among them, M is the bounding box with the highest current score, b_i is the adjacent detection box, S_i is the score of detection box, and N_t is the threshold.

The score reset function of the Soft-NMS (rescoring function) is calculated as follows:

$$s_i = \begin{cases} s_i, & iou(M, b_i) < N_t \\ s_i(1 - iou(M, b_i)), & iou(M, b_i) \geq N_t \end{cases} \tag{4}$$

The score of the adjacent detection box that overlaps with the detection frame M by attenuation is an effective improvement to the NMS algorithm. The higher the overlap of M is, the more serious the fractional attenuation may be. When the degree of overlap between the adjacent detection frame and M exceeds the threshold N_t, the detection score of the detection box is linearly attenuated. In this case, the algorithm attenuates the detection score of the non-maximum detection frame instead of completely removing it, and does not attenuate the original detection score of the detection frame without overlapping. Therefore, the non-maximum detection box of the tomato under overlapping and occlusion is retained, and the detection precision is improved.

3. Model training

3.1. Test platform

The test operating platform is the Ubuntu 16.04 system, which uses the Tensorflow as deep learning framework. The computer memory is 32GB, equipped with Intel® Core™ i7-7700K CPU@4.00GHz ×8 processors, GPU uses NVIDIA GTX1080Ti, adopts 16nm production process, memory type is GDDR5, and capacity is 11GB.

3.2. Test parameter setting

The mini-batch stochastic gradient descent (SGD) method with momentum factor was used to train the network [25]. The number of mini-batch was 256, and the momentum factor was set to 0.9. Since the initialization of weights affects the convergence speed of network, the gaussian distribution (mean value was zero and the standard deviation was 0.01) was used to initialize the weights of all layers of the network randomly in this paper. All bias of the convolutional layers and

fully connected layers were initialized to zero. The same learning rate was adopted for all layers in the network, and the initial learning rate was set to 0.001. During the training process, the current learning rate was reduced by 1/10 step by step, and the regularization coefficient was set to 0.0005.

4. Results and analysis

4.1. The effect of different anchor sizes on mAP

Table 1 shows the model parameter settings and mean average precision (mAP). Since anchors were used to predict the bounding boxes, three sizes 8 * 8, 16 * 16, 32 * 32 and three conversion ratios 1:1, 1:2, 2:1(a total of 9 kinds of anchors) were applied to the original Faster R-CNN. Because the target size of the original voc dataset was very different, if the tomato dataset was used to train the detection target, some anchors were unreasonable. Therefore, the appropriate anchor size by K-means clustering method was calculated in this paper. The original size was updated to 4*4, 16*16, 64*64, and the transformation ratio remained unchanged in order to improve the detection rate of the bounding box. Generally, the threshold of mAP is 0.5, which means if the coincidence degree is greater than 0.5, the key organs of the tomato are correctly detected. From Table 1, it can be seen that the mAP of model 2, 4, 6 and 7 with anchor size 4 * 4, 16 * 16, 64 * 64 is 0.7, 1.6, 0.7 and 1.6 percentage points higher than model 1, 3, 5 and 8 with original size 8 * 8, 16 * 16, 32 * 32, respectively. The above situation indicated that the updated anchor size is more suitable for the data set of this paper, and the recognition rate of key organs of tomato is improved.

Table 1 Model parameter settings and mAP

Model NO.	Basic feature extraction network	Model parameters		Detection box retention algorithm	mAP (%)
		Basic anchor size	Pooling type of basic feature extraction network		
0	VGG16	8*8, 16*16 , 32*32	max-pooling	NMS	85.2
1	ResNet-50	8*8, 16*16 , 32*32	average-pooling	NMS	86.6
2	ResNet-50	4*4, 16*16 , 64*64	average -pooling	NMS	87.3
3	ResNet-50	8*8, 16*16 , 32*32	max-pooling	NMS	87.1
4	ResNet-50	4*4, 16*16 , 64*64	max-pooling	NMS	88.7
5	ResNet-50	8*8, 16*16 , 32*32	average -pooling	Soft-NMS	88.9
6	ResNet-50	4*4, 16*16 , 64*64	average -pooling	Soft-NMS	89.6
7	ResNet-50	4*4, 16*16 , 64*64	max-pooling	Soft-NMS	90.7
8	ResNet-50	8*8, 16*16 , 32*32	max-pooling	Soft-NMS	89.1

4.2. The effect of different pooling types of feature extraction convolutional layers on mAP

There are usually two types of pooling: maximum pooling and average pooling. In Table 1, model 3, 4, 7, 8 and model 1, 2, 5, and 6 respectively adopts the maximum and average pooling types. The results show that the effect of maximum pooling is better than the average pooling when other conditions remain unchanged. Combining the literature [27] with internet data, it can be explained that the effect of average pooling is to average all the values of the entire feature map. It

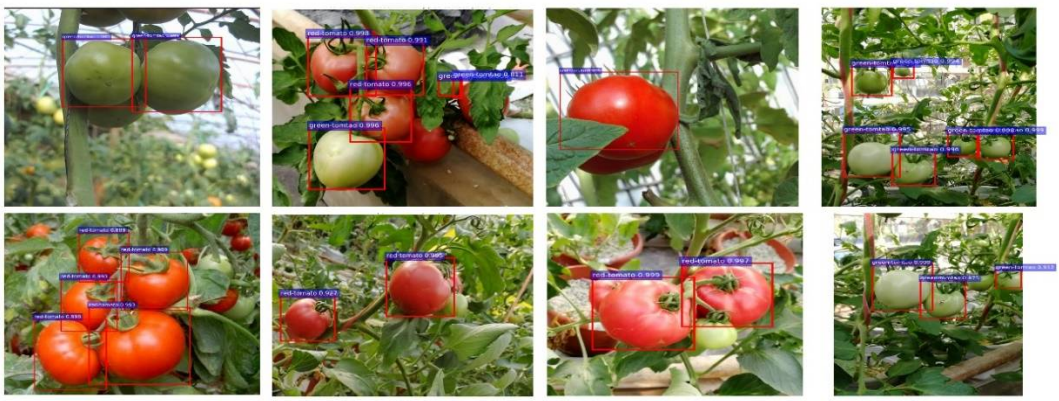
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2018 doi:10.20944/preprints201810.0524.v1

Peer-reviewed version available at Agriculture 2018, 8, 196; doi:10.3390/agriculture8120196

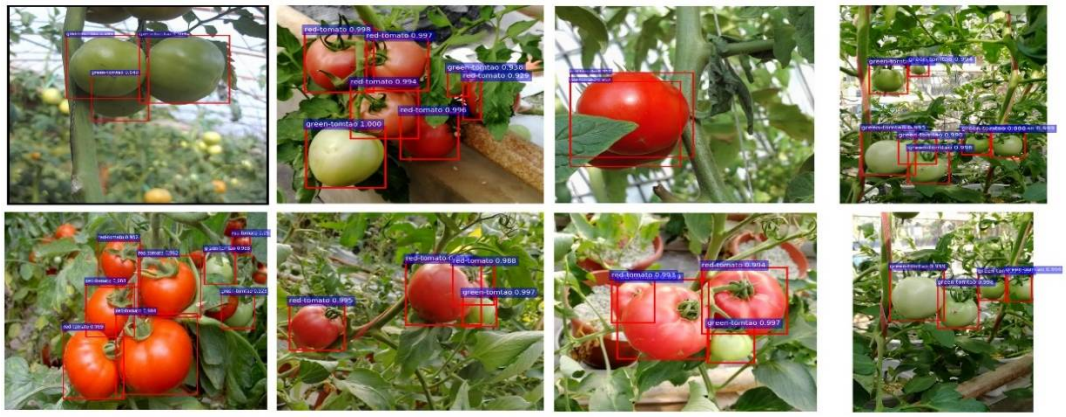
can reduce the error caused by the increase of estimation variance due to the limited size of the neighborhood, and retains more background information of the image. However, maximum pooling can reduce the deviation of estimation mean caused by the error of convolutional layer parameters. So using maximum pooling in the middle layer of convolution can preserve texture features more and discard redundant features, which is beneficial to extracting the deep key features of target in the image and improving the recognition accuracy.

4.3. The effect of Soft-NMS on mAP

It can be seen from Table 1 that the model 5, model 6, model 8 and model 7 retained by the Soft-NMS method are 2.3, 2.3, 2 and 2 percentage points higher than those of the models 1, 2, 3 and 4 retained by the NMS method, respectively. In order to better demonstrate the effect of the Soft-NMS algorithm on the detection of tomato organs in the case of overlapping occlusion, the detection effect was visualized. As shown from Figure 4, the optimal model 7 using Soft-NMS is based on the area of overlapping portion to set an attenuation function for the adjacent detection frame instead of completely zeroing its score. This operation can effectively solve the problem of low detection and recognition rate of key organs of tomato under overlapping and occlusion.



(a) Model 4 using NMS



(b) Model 7 using Soft-NMS

Figure4. Visualization of detection effects in case of overlapping and occlusion. (a) Detection of key tomato organs using a retention algorithm with non-maximum suppression; (b) Detection of key tomato organs using a retention algorithm with Soft-NMS.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2018 doi:10.20944/preprints201810.0524.v1

Peer-reviewed version available at Agriculture 2018, 8, 196; doi:10.3390/agriculture8120196

5. Indicators for model performance evaluation

The purpose of object detection is to find the target in a given image, classify the target, and locate the target in the image. Target detection model is usually trained on a set of fixed classes, so the model can only locate and classify the categories contained in the image. In addition, the location of target is usually in the form of a boundary matrix, so target detection involves the location information of target in the image and the classification of target. Mean average precision is an algorithm that is particularly well-suited for the prediction of target category and its location, which is very useful for the assessment of performance in the target detection model. Besides, there are several important evaluation parameters: accuracy, recall and average precision [28].

5.1. Precision & Recall

Precision is just the accuracy. In the field of information retrieval, precision and recall appear together. For a query, a series of goals returned, and the correct rate refers to the proportion of the relevant targets in the returned results. Precision is defined as:

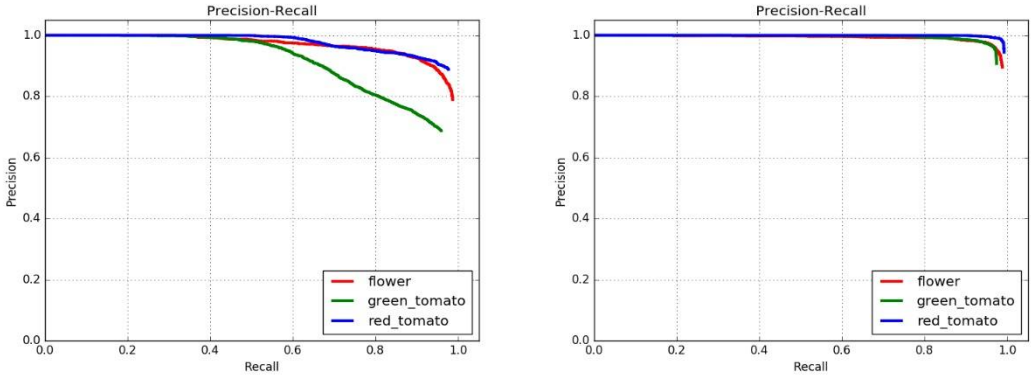
$$Presision = \frac{A}{B} \tag{5}$$

The recall rate is the ratio of relevant targets in the returned results to all relevant targets, and it is defined as:

$$Recall = \frac{A}{C} \tag{6}$$

Among these formulas (5,6), A represents the number of related targets in the return results, B indicates the number of returned results and C represents the number of all relevant targets.

Suppose true positive (TP): predict the number of positive classes as a positive class, true negative (TN): predict the number of negative classes as negative classes, false positive (FP): predicts negative classes as positive ones (false rate), false negative (FN): predicts the positive class as the number of negative classes (missing rate). Then $Precision = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, and $Recall = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$. The performance of object detection model is usually evaluated in the paper using the recall and precision score curves with the threshold, that is, the R-P curve [29]. Figure 5 shows the P-R curve of the original Faster R-CNN and the improved model 7. It can be seen from the P-R curve that the detection accuracy of model 7 is higher and the performance is better.



(a) Original Faster RCNN (b) Improved model 7

Figure5. Precision and recall curve of the model. (a) Precision-recall curve of original Faster R-CNN model; (b) Precision-recall curve of improved model 7.

5.2. Average Precision & Mean Average Precision

Compared to the graph, in some cases, the specific values can more clearly show the performance of test model. Average recision (AP) is usually used as a metric, and the calculation formula is:

$$\mathrm{AP} = \int_0^1 p\ (r)\ d\ (r) \tag{7}$$

Among this formula(7), p represents $Precision$, r represents $Recall$, p is a function of r, so average precision is equal to the area under the P-R curve, and mean average precision (mAP) equals the average of all categories of average precision. It can be seen from Table 2 that the AP of the flower, immature green tomato and the mature red tomato in improved model 7 is 13.7, 2.4 and 0.5 percentage points higher than that of the original Faster R-CNN, respectively. And the mAP is 5.5 percentage points higher. It indicates that the improved model 7 is better than the original model for detection and recognition of tomato key organs.

Table 2 Model average precision and mean average precision

Model	Average precision（%）			mAP（%）
	Flower	Green tomato	Red tomato	
Original Faster R-CNN	76.8	88.4	90.4	85.2
Improved model 7	90.5	90.8	90.9	90.7

5.3. Memory requirements and detection time of models

It can be seen from Table 3 that compared with the original Faster R-CNN model, the memory requirement and detection time of the improved model 7 (using Resnet-50) as the basic feature extraction network are reduced by about 79% and 23%, respectively, and the detection accuracy has a large improvement. It indicates that the improved model reduced the model parameters and improved the precision of model detection on the basis of guaranteeing no additional test time, which lays a theoretical foundation for the subsequent implantation of the model into the embedded system and the development of portable devices for accurate and real-time detection of tomato key organs.

Table 3 Model parameter quantity and detection time

Model	Model memory requirement (MB)	Average testing time(s)	FPS
Original Faster R-CNN	546.9	0.095	10.5
Improved model 7	115.9	0.073	13.7

5.4. Detection effect of model

In order to verify the actual field prediction effect of the optimal model in this paper, the key organs of tomato were tested at 9:00 am on a sunny morning in the greenhouse. It can be seen from Table 1 and Figure 6 that the average precision of the improved model 7 for the detection of key organs of tomato is higher, which can reach 90.7%. And the confidence of each object is basically over 0.99. The above performance shows that the model has a good detection effect on tomato flowers, immature green tomatoes and mature red tomatoes in the actual field background.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 October 2018 doi:10.20944/preprints201810.0524.v1

Peer-reviewed version available at Agriculture 2018, 8, 196; doi:10.3390/agriculture8120196

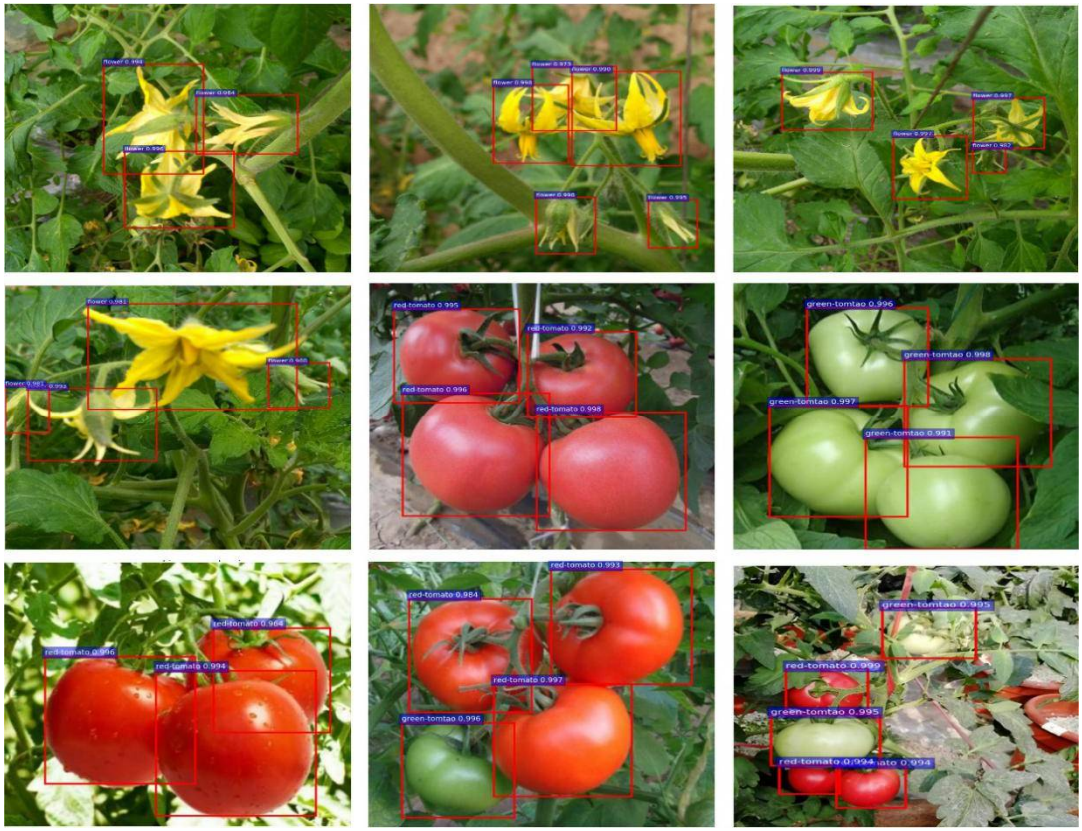


Figure6. Examples of improved model detection

6. Conclusion

In this paper, convolutional neural network was used to detect and identify the key organs of tomato. The original Faster R-CNN model based on vgg16 feature extraction network was improved, the resnet-50 with residual structure was used to replace the vgg16 network. And K-means clustering was adopted to fit the anchor size of the data set in this paper, avoiding the problem of recognition accuracy reduction caused by artificial setting. Compared with the original Faster R-CNN model, this model used Soft-NMS to retain the generated detection frame, which solved the problem of low detection accuracy of key tomato organs under overlapping and occlusion in a certain extent.

The results shows that the APs of tomato flowers, immature green tomatoes and mature red tomatoes were increased from 76.8%, 88.4%, 90.4% to 90.5%, 90.8%, 90.9%, respectively. And the memory required by the model was reduced from 546.9 MB to 115.9 MB. The average detection time was shortened from 0.095 S/sheet to 0.073 S/sheet. The mean average precision was increased from 85.2% to 90.7%, and the performance of the model was greatly improved.

The training model can be transplanted to the embedded system in the future, which lays a theoretical foundation for the development of precise targeting pesticide application system and automatic picking robot device of tomato.

Author Contributions: Sun Jun and He Xiaofei contributed to the development of the systems, including farm site data collection and the manuscript writing. He Xiaofei provided significant suggestions on the development, and contributed to performance evaluation. Ge Xiao contributed to the grammar modification. Sun Jun, He Xiaofei, Ge Xiao and Song Yingying analysed the results. All authors wrote the manuscript together.

Funding: This work is partially supported by National natural science funds projects (31471413，61875089), A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), Six Talent Peaks Project in Jiangsu Province (ZBZZ-019), Science and Technology Support Project of Changzhou (Social Development) (CE20175042).

Conflicts of Interest: The authors declare no conflict of interest

References

1. Paran, E.; Engelhard, Y. P-333: Effect of tomato's lycopene on blood pressure, serum lipoproteins, plasma homocysteine and oxidative sress markers in grade I hypertensive patients. *Am. J Hypertens.* **2001,** 14, (4), A141-A141.
2. Geng, C.X.; Zhang, J.X.; Cao, Z.Y.;Li, W. Cucumber Disease Toward-target Agrochemical Application Robot in Greenhouse. *Trans. CSAM.* **2011**, 42(1):177-180.
3. Zhao, Y.S.; Gong, L.; Huang, Y.; Niu, Q. Object Recognition Algorithm of Tomato Harvesting Robot Using Non-color Coding Approach. *Trans. CSAM.* **2016**, 47(7):1-7.
4. Jiang, H.Y.; Peng, Y.H.; Shen, H.; Ying, Y. Recognizing and locating ripe tomatoes based on binocular stereo vision technology. *Trans. CSAE.* **2008**, 24(8):279-283.
5. Zhang, R.H.; Ji, C.Y.; Shen, M.X.; Cao, K. Application of Com puter V ision to Tomato Harvesting. *Trans. CSAM.* **2001**,32(5):50-52.
6. Yamamoto, K.; Guo, W.; Yoshioka, Y.; Ninomiya, S. On plant detection of intact tomato fruits using image analysis and machine learning methods. *Sensors* **2014,** 14, (7), 12191-12206.
7. Farabet, C.; Couprie, C.; Najman, L.; Lecun, Y. Learning Hierarchical Features for Scene Labeling. *IEEE T. Pattern.Anal.***2013,** 35, (8), 1915-1929.
8. Sun, J.; Tan, W.J; Mao, H.P.; Wu, X.H.; Chen, Y.; Wang, L. Recognition of multiple plant leaf diseases based on improved convolutional neural network. *Trans. CSAE.* **2017**, 33(19):209-215.
9. Xiang, X.; Lv, N.; Guo, X.; Wang, S.; Saddik, A. E. Engineering Vehicles Detection Based on Modified Faster R-CNN for Power Grid Surveillance. *Sensors* **2018,** 18, (7), 2258.
10. Philipsen, M.P.; Dueholm, J.V.; Jørgensen, A.; Escalera, S.; Moeslund, T.B. Organ Segmentation in Poultry Viscera Using RGB-D. *Sensors* **2018,** 18, (1), 117.
11. Zhang, L.; Lin, L.; Liang, X.; He, K. Is Faster R-CNN Doing Well for Pedestrian Detection? In European Conference on Computer Vision(ECCV2016), Springer: Amsterdam, Netherlands, 11-14October2016; pp. 443-457.
12. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection, In Computer Vision and Pattern Recognition(CVPR2016), Las Vegas, USA, 27-30 June2016; pp. 779-788.
13. Bargoti, S.; Underwood, J. Deep fruit detection in orchards, In IEEE International Conference on Robotics and Automation(ICRA2017), Marina Bay Sands, Singapore,29May-3June2017.
14. Inkyu, S.; Ge, Z.; Feras, D.; Ben, U.; Tristan, P.; Chris, M. C., DeepFruits: A Fruit Detection System Using Deep Neural Networks. *Sensors* **2016,** 16, (8), 1222.
15. Zhou, Y.C.; Xu, T.Y; Zhen, W.; Deng, H.B. Classification and recognition approaches of tomato main organs based on DCNN. *Trans. CSAE.* **2017**, 33(15):219-226.
16. Sun, J.; He, X.F; Tan, W.J.; Wu, X.H.; Shen, J.F.; Lu, H. Recognition of crop seedling and weed recognition based on dilated convolution and global pooling in CNN. *Trans. CSAE.***2018**,

34(11):159-165.

17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks, In International Conference on Neural Information Processing Systems(NIPS2015), Montreal,Canada, 7-12December2015; pp. 91-99.
18. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computer Science.* **2014,** 14, (7), 123.
19. Ioffe, S.; Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift, In International Conference on International Conference on Machine Learning(ICML2015), Lille, France, 6-11July2015; pp. 448-456.
20. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks, In International Conference on Neural Information Processing Systems (NIPS2012), Lake Tahoe, NV, USA, 3-6December2012; pp. 1097-1105.
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks, In European Conference on Computer Vision(ECCV2014), Springer: Zurich, Switzerland, 6-12 September 2014; pp. 630-645.
22. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. **2018**.
23. Liu, J.; Li, P. F. A Mask R-CNN Model with Improved Region Proposal Network for Medical Ultrasound Image, In International Conference on Intelligent Computing(ICIC2018), Wuhan, China, 15-18August2018; pp. 26-33.
24. Neubeck, A.; Gool, L.V. Efficient Non-Maximum Suppression, In International Conference on Pattern Recognition(ICPR2006), Hong Kong, China, 20-24August2006; pp. 850-855.
25. Wang, L.; Zang, J.; Zhang, Q.; Niu, Z.; Hua, G.; Zheng, N. Action Recognition by an Attention-Aware Temporal Weighted Convolutional Neural Network. *Sensors* **2018,** 18, (7).
26. Barbedo, J.G.A. Factors influencing the use of deep learning for plant disease recognition. *Biosystems Engineering* **2018,** 172, 84-91.
27. Boureau, Y. L.; Bach, F.; Lecun, Y.; Ponce, J. Learning mid-level features for recognition, In Computer Vision and Pattern Recognition(CVPR2010), San Francisco, CA, USA, 13-18 June 2010; pp. 2559-2566.
28. Everingham, M.; Gool, L.V.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int J Comput Vision.* **2010,** 88, (2), 303-338.
29. Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves, In Proceedings of the International Conference on Machine Learning(ICML2006), Pittsburgh, PA, USA, 25-29June2006; pp. 233-240.
30. Li, K.; Huang, Z.; Cheng, Y.C.; Lee, C.H. A maximal figure-of-merit learning approach to maximizing mean average precision with deep neural network based classifiers, In IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP2014), Florence, Italy, 4-9May2014; pp. 4503-4507.