Fruit image recognition and classification method based on improved single shot multi-box detector

Jinfan Cai¹, Jiaxin Tao¹, Yahong Ma¹, Xiaojiao Fan¹, Lin Cheng²

¹College of Information Engineering, Xijing University, Xi’an, China
²State Grid Shannxi Electric Power Research Institute, Xi’an, China

Correspondingauthor and e-mail: Yahong Ma, mayahong@xijing.edu.cn

Abstract. In recent years, pattern recognition has been gradually applied to the field of agriculture, especially fruit classification and rating based on image. However, the target fruit is affected by the interference factors such as light change, uneven brightness, similar background, branches and leaves, shadow coverage, etc. In order to solve the problems of low recognition rate and low generalization of fruits in the natural environment, we proposed an improved SSD (Single Shot Multi-box Detector) to classify apple, persimmon, nectarine and pear. The original network is optimized and the more advanced soft NMS (Non-Maximum Suppression) is used to obtain the anchor boxes, so as to provide a better initial value for the prediction of the bounding box. At the same time, under the premise of keeping good performance independent of the network architecture, batch normalization is used to initialize the de-random training model, which brings stable and predictable gradient of the detector. Through the identification test of four kinds of fruits collected in the natural background, the experiment shows that the improved SSD fruit detection model is effective. The fruit recognition speed is the fastest and the detection accuracy is the highest compared with other methods. The average detection accuracy in various environments has reached 92.4%. It can be concluded that the proposed method will realize the accurate detection of many kinds of fruits, and provide a new scheme for the fruit recognition and detection problems in agricultural automatic picking.

1. Introduction
In recent years, the research of target fruit detection occupies more and more proportion in the field of machine algorithm. In the natural environment, fruit recognition and detection mainly use computer vision technology to obtain fruit target location information [1]. Fruit target recognition is the basis of automatic picking. Many researchers mainly study fruit target recognition methods in limited environment or natural environment from the perspective of color, texture, edge and other features [2-4]. By using a variety of classification and clustering algorithms to design the target recognition model, a better target detection effect is achieved. Traditional target detection methods mostly extract the shallow features of the target object, such as hog features [5], SIFT features [6], LBP [7], color features, local binary features and so on. Finally, the detection results are classified according to the features of the data. These characteristics of artificial design only apply to some specific scenes, and the performance of complex scenes is not satisfactory, which makes the target recognition model
constructed difficult to meet the needs of complex field scenes, and the detection effect is very unstable.

At present, the three main target detection algorithms are Faster-RCNN [7] (Faster Regions with CNN), YOLO [8] (You Only Look Once) and SSD (Single Shot Multi-box Detector) [9]. Taking fruit recognition as an example, its principle is to extract the characteristics of the input quantity by using the convolution network, and then judge the category and original coordinates of the fruit in the preselected area through the frame classification and frame regression, and realize the target location of the fruit, which improves the accuracy of fruit recognition. However, the basis of the above method is to obtain the image features from the characteristics of the fruit itself. When there are many common interference factors in the natural picking environment, such as light change, shadow coverage, uneven coloring, branches and leaves occlusion and fruit overlap, the characteristics of the fruit change significantly, which makes the characteristics used to describe the fruit also appear obvious differences. Therefore, the fruit recognition method based on image features is not ideal in the natural environment.

In this paper, apple, persimmon, nectarine and pear are taken as the research objects, and the recognition and detection technology of the fruit images collected in the natural environment is studied. Based on the SSD network structure, the more advanced soft NMS is used to obtain the anchor boxes, so as to provide a better initial value for the prediction of the bounding box. The SPP method is used to replace the original pooling layer in the pooling layer to improve the image scale and reduce over fitting. In order to simplify the process of feature extraction and improve the accuracy and robustness of fruit recognition, the experiment of fruit recognition is compared, which provides a reference for fruit picking technology in natural environment.

2. Data set production

The fruit images used in this paper are collected by ourselves. When collecting, Canon EOS 250D camera is used with a pixel value of 24.1 million. In order to facilitate image acquisition, all the fruits are common in life. The collected fruit images mainly include four kinds of fruits: apple, persimmon, nectarine and pear. Each fruit image is collected 300 pieces, compressed into 300 × 300 and stored in JPEG format.

In the research of image preprocessing, labeling image annotation tool is used to annotate the image manually. In the process of neural network model training, there are often over fitting problems. In order to prevent over fitting problems, it is necessary to expand the data set of manually labeled image. The expansion methods include rotation, translation and so on. After data expansion, each of the four fruit images has 500 pieces, and the fruit image data set has 2000 pieces in total. The data set is allocated according to 4:1 ratio, among which 1500 pieces are training sets and 500 pieces are test sets.

3. Improved SSD Multi-Types fruit detection method

3.1. Soft NMS algorithm

In SSD network, the accuracy of anchor boxes dimension directly affects whether the network can learn more accurate prediction methods. For this problem, the traditional SSD network model uses soft NMS instead of the original NMS, and retains the anchor boxes with lower confidence. The scale parameters of anchor boxes are defined as eq. (1):

$$S_i = S_{\text{min}} + \left( \frac{S_{\text{max}} - S_{\text{min}}}{N - 1} \right) (i - 1), i \in [1, N]$$  \hspace{1cm} (1)$$

After the anchor boxes are generated, the features of different scales are selected and sent to the prediction network for prediction calculation. In this paper, we use linear weighting non maximum suppression improvement to change the confidence to the function of IOU, F(IOU), and retain the anchor boxes with lower confidence value, shown in eq. (2).
3.2. Improved SSD network structure

SSD is one of the most popular deep learning object detection models currently. SSD uses VGG16 as the basic model to obtain the feature map for detection. VGG 16 is used as the feature map of direct action of the basic network to predict the multi-objective categories and peripheral frames.

The accuracy of recognition increases with the depth of the network. However, the gradient explosion in the back propagation leads to a simple stacking and convolution layer which cannot train the network smoothly. Some researchers have made the limited deep-seated training possible by batch normalization and dropout technology, but after a certain training iteration, the problem of accuracy saturation leading to accuracy degradation still exists. The improved SSD target detection algorithm uses multi-layer network to fit a residual mapping to solve the degradation problem. Suppose \( H(x) \) is the target optimal solution mapping, and another mapping \( f(x) \) is fitted with a stacked nonlinear layer (as shown in Figure 1), which can be represented by eq. (3).

\[
F(X) = H(X) + X
\]

At this time, the original optimal solution mapping \( H(x) \) can be equivalent to \( f(x) + X \), that is, the fast connection implementation in the feedforward network shown in Figure 1.

The way of fast connection can be expressed in eq. (4).

\[
Y = F(X, \{W_i\}) + W_iX
\]

Where \( X \) represents the input vector of the module, \( Y \) represents the output vector of the module, \( W_i \) represents the weight layer parameter. When the input and output dimensions are consistent, a linear projection \( W_i \) needs to be added to match the dimensions.

3.3. Batchnorm algorithm

For VGG, low-level relu4. The scale of the feature graph out of 3 will be much smaller than that of the later high-level features. The average weight of conv4 is about 10.6 times of that of the latter conv7. The larger scale layer will play a leading role in updating, which is not conducive to training. Therefore, batch standardization is needed. The difference between the later layers is not so big, so batch normalization is needed.

A d-Dimension input \( a = (a^{(1)}, a^{(2)}, ..., a^{(n)}) \), batchnorm will normalize each dimension feature:

\[
\hat{a}^{(k)} = \frac{a^{(k)} - E[a^{(k)}]}{\sqrt{Var[a^{(k)}]}}
\]

Where \( E[a^{(k)}] \) represents mathematical expectation and \( Var[a^{(k)}] \) variance. However, if it is simply normalized, it will affect the learning characteristics of the network layer. In order to solve this problem, two parameters \( \gamma^{(k)}, \beta^{(k)} \) are introduced to scale and shift the normalized value.

\[
y^{(k)} = \gamma^{(k)} \hat{a}^{(k)} + \beta^{(k)}
\]
These two parameters are learned together with the original model parameters. When $\gamma = \text{Var}(a^{(i)})$, $\beta = E[a^{(i)}]$ can restore the features that the original network needs to learn.

In the process of the random gradient training of the prediction block after each detection layer, it participates in the back propagation process with the gradient, allowing more search space and faster convergence.

### 3.4. Improved SSD model structure

In this paper, the SSD network model is modified, as shown in Figure 2. The red part is the software network management, aiming to improve the testing accuracy of the bounding box. The introduction of batchnorm restrains the over fitting in the data training process, improves the generalization ability of the training model, and speeds up the network training.

![Figure 2. Improved SSD model detection framework.](image)

### 4. Experimental results

#### 4.1. Experimental environment and parameter setting

We used Python 3.7 as the programming language, NVIDIA Titan rtx1070 GPU, windows 10, Intel i7 as the training platform, and tensorflow as the open framework for deep learning.

In the training of the improved SSD network, the data input size is $300 \times 300$, and the loss function is optimized by batch random gradient descent method, with a total of 30000 iterations. Using default parameters, learning rate is 0.001, weight attenuation $\text{val}\_\text{Loss}$ is 0.0005, batch_ Size is set to 16. Use the callback function reducerlonplateau to monitor Val during model training_ Loss, patience value is set to 10, when val_ Learning when loss does not decline after 10 epochs_ Rate reduced 10 times to 0.0001.

#### 4.2. Experimental results and performance comparison

In this study, the average precision (AP) and the total average precision (mAP) of a single category are used, and the calculation formula is as follows eq. (7):

$$P = \frac{TP}{TP + FP}, \quad AP(s) = \frac{P_t}{N_t}, \quad mAP = \frac{1}{s} \sum_{s=1}^{s=1} \frac{AP(s)}{N_t}$$

Among them, $P$ is the accuracy, $TP$ is the correct number of fruit positive samples, $FP$ is the wrong number of fruit positive samples, $t$ is the fruit category, $NT$ is the number of all samples including the...
category T, AP (s) is the average accuracy of the category s fruit, K is the total number of fruit categories, the higher the map value, the more accurate the fruit detection on the data set.

In this paper, four kinds of fruits are used for experiment and comparative study, and different target detection algorithms are used to train and test these four kinds of fruits. Figure 3 shows the test effect on the test set.

![Test sample](image)

**Figure 3.** The results of fruit testing.

The test results are shown in Table 1, which shows the comparison results between improved SSD proposed in this paper and common target detection algorithms CNN, Faster-RCNN, YOLO-v3 and SSD-300.

|                | Apple  | Persimmon | Nectarine | Pear   | mAP    | Recognitionspeed |
|----------------|--------|-----------|-----------|--------|--------|------------------|
| CNN            | 80.6%  | 75.2%     | 81.7%     | 80.7%  | 79.5%  | 7 f/s            |
| Faster-RCNN    | 88.9%  | 83.2%     | 88.4%     | 84.3%  | 86.2%  | 13 f/s           |
| YOLO-v3        | 83.2%  | 81.7%     | 86.2%     | 83.1%  | 83.6%  | 19 f/s           |
| SSD-300        | 88.5%  | 78.3%     | 83.7%     | 75.1%  | 81.4%  | 34 f/s           |
| Improved SSD   | 90.3%  | 92.9%     | 93.0%     | 93.2%  | 92.4%  | 36 f/s           |

5. Conclusions
Through the analysis and comparison of the experimental results, the detection accuracy of the improved SSD algorithm for apple, persimmon, nectarine and pear is 1.4% - 9.7% higher than that of the faster-RCNN, YOLO-v3 algorithm and SSD-300 algorithm, and it has obvious advantages in the speed detection of fruit recognition. In particular, compared with the original SSD algorithm, the map of the improved SSD algorithm is 92.4%, 11.0% higher than the original SSD algorithm, and the recognition speed is increased by 2f / s. It can be seen that the model studied in this paper not only has better test accuracy, but also has better test speed. In the research of fruit classification and recognition, it has better recognition ability.

Acknowledgement
This work is supported by the Shaanxi province innovation capacity support program (Grant Nos. 2018KJCK-095), as well as the Xijing university special talent research fund (Grant Nos. XJ17T03). This work is also supported by the Graduate Innovation Fund of Xijing University (2020YC-08). At the same time, I would like to thank Professor Ma Yahong for her guidance and help me finish this work successfully.

References
[1] Edan Y, Han S and Kondo N. Automation in Agriculture[M]. New York: Springer Publishing Company, 2009.
[2] BEHROOZI-KHAZAEI Nand MALEKI M R A robust algorithm based on color features for grape cluster segmentation[J]. Computers and Electronics in Agriculture,2017,142:41-49
[3] CHAIVIVATRAKUL Sand DAILEY M N. Texture-based fruit detection[J]. Precision Agriculture, 2014, 15(6): 662-683.

[4] LU J and SANG N. Detecting citrus fruits and occlusion recovery under natural illumination conditions[J]. Computers and Electronics in Agriculture, 2015, 110: 121-130.

[5] Singla N. Motion detection based on frame difference method [J]. International Journal of Information & Computation Technology, 2014, 4(15): 1559-1565.

[6] Kong T, Yao A, Chen Y, et al. HyperNet: towards accurate region proposal generation and joint object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 845-853.

[7] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448.

[8] Redmon J, Divvala S and Girshick R. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[9] Guo chuanlei and He Jia. Improved single-stage multi frame target detection method based on transposition convolution operation [J]. Computer application, 2018, 38 (10): 2833-2838. (In Chinese)