Citrus recognition based on YOLOv4 neural network

Wang Changhong$^{1,a}$, Luo Qiang$^{1,b,*}$, Chen Xin$^{1,c}$, Yi Bingliang$^{2,d}$, Wang Haibao$^{2,e}$

$^1$Intelligent Mountain Agricultural Machinery Technology Research Center of Chongqing three Gorges university, Wan Zhou, Chong Qing, China;

$^2$Heifei Institute of Intelligent Machinery, Chinese Academy of Sciences, Hei Fei, An Hui, China

$a$email:1147810696@qq.com, $^c$email:cx48699684@163.com, $^d$email:1751663188@qq.com, $^e$email:58973481@qq.com

*corresponding author: $^b$email: qluosxy@bjtu.edu.cn

Abstract: In order to realize the accurate and rapid recognition of citrus fruit by autonomous picking robot in natural environment, a citrus recognition method based on YOLOv4 neural network is proposed. This algorithm improves the YOLOv4 model and uses the Kmeans++ algorithm to obtain the prior frame in the model to enhance the scale adaptability. Take citrus pictures independently and expand the data, use the LabelImage platform to mark the data, and train the network model under the Darknet framework. According to the results, the citrus recognition model has good robustness and real-time performance to the common interference factors and their superposition in natural picking environment. The average recognition accuracy is 89.23 and the average detection speed is 60 ms. The average recognition accuracy is 89.23. It meets the requirements of real-time image recognition speed and accuracy of citrus picking robot.

1. Introduction

In recent years, China's fruit production has greatly increased to become one of the world's important fruit producers, especially apple, citrus and other fruits in the planting area and yield has steadily ranked first in the world, citrus production accounted for 1/4 of the world's citrus production. During the whole process of citrus production, the labor cost is relatively high, accounting for 50%~70% of the whole cost. The artificial operation of citrus picking in China is still the mainstream. Automatic picking is of great significance for solving labor shortage, picking fruit in time, reducing cost and improving fruit picking quality$^{[1]}$.

In the study of fruit recognition, the method of fruit target recognition in limited environment is studied from the point of view of color, texture, edge and so on. The target recognition model is designed by using various classification and clustering algorithms, and good target detection effect is obtained. But the basis of the above method is to obtain image features from the characteristics of the fruit itself$^{[2]}$.When there are many common interference factors in natural picking environment, such as light change, shadow cover, uneven coloring, foliage occlusion and fruit overlap, the fruit characteristics change obviously, which makes the characteristics used to describe the fruit obviously different. Therefore, the citrus recognition method based on image features is not ideal in natural environment. Convolutional neural network is robust to target feature learning and extracting key information. Recently, the target detection algorithm based on convolution neural network mainly
relies on two ideas, one is based on the idea of target candidate frame two-stage, the other is based on the one-stage. of regression idea Two-stage first extract the target candidate box and then train the detection model, such as FastR-CNN, FasterR-CNN, on its basis [3-4]. The one-stage has no target candidate box extraction operation, directly using the detection network to generate target category and location information, with higher detection speed, such as SSD, YOLOv3, YOLOv4, etc [3-6]. YOLOv4 is a recently proposed deep convolutional neural network with higher detection efficiency. To ensure the efficiency of citrus detection and recognition model, this paper designs a citrus recognition method based on YOLOv4 depth learning network, which can make citrus data images of various forms by adding noise and rotating citrus images, using YOLOv4 network model to learn and train the data set, and using industrial CCD camera to collect and recognize citrus images.

2. YOLO algorithm

YOLO (You Only Look Once) [7] Network is a kind of target detection algorithm based on regression, which has fast detection speed and has achieved good results in many target detection tasks [8-9]. Compared with other algorithms, YOLO multi-scale prediction algorithm can detect targets more effectively. At the same time YOLO the real-time performance is outstanding, which can meet the needs of citrus real-time picking identification and detection. YOLOv4 algorithm introduces some optimization methods from the aspects of YOLOv3 data processing, backbone network, network training, activation function, loss function and so on, which makes the model achieve the optimal matching in the detection speed and accuracy so far [6].

2.1 YOLOv4 algorithmic framework

YOLOv4 backbone network is CSPDarknet53 the core of the algorithm, Used to extract target features, SPP as Neck additional module, PANet as a Neck feature fusion module, YOLOv3 as Head. Darknet53 contains five large fragments, the number of small residual units contained in these five large residual blocks is 1, 2, 8, 8 and 4, respectively. CSPDarknet53 added CSP Net to each large residual block of the Darknet53 (Cross Stage Part Network), Integration into feature maps by gradient changes, Divide the feature map into two parts, part of the convolution, the other part is combined with the last convolution result. In target detection, CSP can effectively improve CNN learning ability, at the same time reduce the amount of calculation. PANet (Path Aggregation Network) makes full use of feature fusion, YOLOv4 also changed the fusion method from addition to multiplication, So that the network can get more accurate target detection ability. YOLOv4 overall network framework is shown in Figure 1:

![YOLOv4 Network framework](image)

YOLOv4 Mosaic data enhancement strategy is used to increase the input image variability and enrich the image feature information. The proposed target detection model can achieve higher robustness. At the same time, the network training process is optimized by using label smoothing, learning rate cosine annealing attenuation and other techniques [6].
2.2 YOLOv4 Loss function

The YOLOv4 loss function in network training is composed of three parts: confidence loss function, boundary loss function and classified loss function. If there is no objective in a certain boundary, only confidence loss is calculated, and if there is an objective, three kinds of losses are calculated. Loss function expression:

\[
L_{\text{conf}} = \sum_{i=0}^{g} \sum_{j=0}^{B} I_{ij}^{\text{obj}} \left[ 1 - \text{IoU} + \frac{S^2 \left( b_i, b^j \right)}{c^2} + av \right]
\]

\[
L_{\text{class}} = \sum_{i=0}^{g} \sum_{j=0}^{B} I_{ij}^{\text{noobj}} \left[ \log(C_i^j) + (1 - C_i^j) \log(1 - C_i^j) - \lambda_{\text{noobj}} \sum_{i=0}^{g} \sum_{j=0}^{B} I_{ij}^{\text{obj}} C_i^j \log(C_i^j) + (1 - C_i^j) \log(1 - C_i^j) \right]
\]

\[
L_{\text{bbox}} = \sum_{i=0}^{g} \sum_{j=0}^{B} \sum_{c=1}^{C_i^j} \left[ \rho \left( P_i^j - \hat{P}_i^j \right) + \left( 1 - \rho \right) \log(1 - P_i^j) \right]
\]

\[
\text{LOSS} = L_{\text{conf}} + L_{\text{class}} + L_{\text{bbox}}
\]

In the formula:
- \( S \): feature map scale and prior frame;
- \( \lambda_{\text{noobj}} \): weight factor;
- \( I_{ij}^{\text{obj}} \) and \( I_{ij}^{\text{noobj}} \): if there is a target at the first box of the grid, take 1 and 0 respectively, and if there is no target, take 0 and 1 respectively;
- \( \rho () \): Euclidean distance;
- \( c \): the diagonal distance between the predicted box and the actual box closure area;
- \( b \): central coordinates and width of the prediction box;
- \( b^j \): center coordinates and width height of the actual frame;
- \( C_i^j \) and \( \hat{C}_i^j \): confidence levels of the prediction and tagging boxes;
- \( P_i^j \) and \( \hat{P}_i^j \): category probability of the prediction box and annotation box.

The confidence loss and classification loss are calculated by cross-entropy method, and the boundary box regression loss is calculated by CIoU loss function. Compared with the traditional mean square error loss function, the problem of sensitivity to the scale of the target object is CIoU effectively avoided.

3. YOLOv4 Algorithm Optimization

3.1 K-means++ Clustering

Because K-means clustering algorithm is sensitive to initial point selection, it is necessary to obtain a better solution by clustering multiple times. K-means++ algorithm is used in this paper. By calculating clustering, K-means++ algorithm is improved at the initial point selection, the distance between cluster centers can be as far as possible. K-means++ basic steps of the algorithm are as follows:

1. Random sampling of a sample from the data set as the initial cluster center \( u_i \)
2. Calculate the distance of \( X \) between each \( x_i \) sample and the nearest cluster center \( D(x_i) \)
\[ D(x_i) = \arg \min_i |x_i - u_i|, \quad (2) \]

In the formula, \( j = 1, 2, ..., k \)

3. Calculate the probability of each sample being selected as the next cluster center and select the next cluster center \( \frac{D(x)^2}{\sum_{x \in X} D(x)^2} \)

(4) Repeat steps (2) and (3) until \( k \) cluster centers are selected.

The larger 76×76 feature extraction convolution layer uses three smaller priori boxes, has the largest receptive field, and the medium 38×38 feature extraction convolution layer uses three intermediate priori boxes. The smaller 19×19 feature extraction convolution layer uses three larger priori boxes with the smallest receptive field. Has used less random k-means++ instead of the original clustering algorithm, the clustering algorithm can effectively reduce the clustering bias caused by the original algorithm at the initial clustering points poor. A better size Anchor obtained and matched to the corresponding feature map can improve the detection accuracy and recall rate.

3.2 Residual network

The network depth is very important to the performance of the network. The deep network can extract the high efficiency features for recognition, but the network performance rapidly reaches saturation with the increase of the network depth, and even begins to decline rapidly, which is called degradation problem. In order to solve the problem of performance degradation caused by the increase of convolution layer number, he Kaiming et al.\cite{12} A deep residual learning framework is proposed. The basic idea is to form a basic residual learning unit by adding shortcut connection (shortcut connections) branch to construct convolution neural network, and to use stacked nonlinear convolution layer to fit a residual mapping. Its core is residual unit, as shown in figure 2.

![Figure 2. Basic modules of the residual network](image)

The \( x \) in the figure is the network input and the \( G(x) \) is the expected output. The ResNet is only the difference between the learning output and the input \( G(x) - x \), that is, the residual \( F(x) \). By performing identity mapping, the latter layer can directly learn the residuals through branch \( x \), and back propagation is easier to return the gradient. When the network reaches the optimum, the module will be set 0, transfer the features from the identity map to the bottom and keep the network in the optimal state without being affected by too many layers.

The application of residual learning unit to deep convolution neural network can effectively alleviate the problem of gradient disappearance in back propagation during network model training, and then solve the problem of difficult training and performance degradation of deep network.
4. Experimental analysis

4.1 Data set readiness

The data set used in the training of citrus network detection model is a self-made data set, which is taken by camera at citrus base, 2000, and 1680 selected. In order to improve the ability of network model and avoid network over fitting, the original image is rotated at random by 30° and 30°, the original image is flipped randomly, horizontally, vertically, and the data set is expanded by cutting and scaling; The data is enhanced by adjusting the saturation and hue, histogram equalization, median filtering and other image processing techniques. Considering that the data enhancement will lead to the change of image shape and the serious change of quality in the picture, 3216 images were obtained by random amplification of each picture.

Label image tool is used to mark the detection target. As shown in figure 3, the minimum external rectangle is used to mark citrus with an area of more than 50% exposed. Randomly 80% of labeled images were used as training set and 20% as test set.

![Image of labeling tool](image)

Figure 3. Picture caption

4.2 Evaluation indicators

The precision P and recall R are selected in the field of target detection. The formulas are as follows:

\[
P = \frac{TP}{TP + FP} \quad (3)
\]

\[
R = \frac{TP}{TP + FN}
\]

The TP represents the number of targets detected correctly by the model, the FP represents the number of targets detected by the model, and the FN represents the number of targets missed by the model.

4.3 Experimental configuration training

And this experiment is Windows10 system, Graphics cards are RTX2080Ti 11G independent graphics cards produced by Nvidia, CUDA10.2, of installation cuDNN7.6.5, python3.7, VS2019, OpenCV3.4.2, Darknet-53 is the network framework His training parameters are: Input image size 608×608, The total number of images entered per iteration is 64, In 16 batches, momentum is 0.9, The weight attenuation coefficient is 0.0005, The maximum number of iterations is 10000, the initial
The learning rate was 0.001, choose mosaic data enhancement strategy. Iterated to 8000, at 9000 steps, the learning rate begins to decay. The loss curve during network training is shown in figure 4:

![Figure 4. Network iteration loss change chart](image)

4.4 Comparative analysis

After training and testing the citrus data set using the YOLOv4 model, the configuration information of the training platform remains unchanged, and the configuration information of the training platform is used to train Faster RCNN, YOLOv2, YOLOv3, and YOLOv4 neural network learning model on this data set. The detection accuracy $\text{MAP}$, accuracy $\text{P}$, and recall rate of different models are compared in Table 1:

| Network category | Detection accuracy $\text{MAP} / \%$ | Accuracy $\text{P} / \%$ | Recall rate $\text{R} / \%$ |
|------------------|-----------------------------------|-----------------|-----------------|
| Faster RCNN      | 84.58                             | 85.23           | 83.25           |
| YOLOv2           | 88.56                             | 90.13           | 87.36           |
| YOLOv3           | 87.22                             | 88.76           | 86.63           |
| YOLOv4 of this article | 89.23                             | 93.52           | 88.59           |

As can be seen from Table 1, and Faster RCNN, YOLOv2, YOLOv3 comparison of the YOLOv4 model shows that the detection accuracy $\text{MAP}$, accuracy $\text{P}$ and recall rate of the model are improved, and the detection accuracy is 89.23, which meets the precision requirements of citrus recognition by picking robot. The detection time is 60 ms, so the model also has good real-time performance and can meet the target recognition speed requirement of automatic picking.

4.5 Actual testing results

In order to detect the practical application ability of the model, the algorithm is used to detect in the actual scene. The detection results are shown in figure 5. There is slight occlusion in figure 5 a, serious occlusion of leaves and branches in figure 5 b, overlapping occlusion between citrus c in figure 5, excessive number of citrus, uneven size, complex background and serious occlusion. The detection accuracy is 96.8 and the average detection time is 60 ms, which meets the requirement of real-time image recognition speed of citrus picking robot. This method can effectively deal with the
interference factors under various outdoor picking conditions.

![Fig 5. Test results of actual environment citrus](image)

5. Conclusion
This paper proposes a method for citrus quality detection based on YOLOv4 depth learning network. The method uses K-means++ algorithm to cluster to obtain a priori box and integrate into the residual network. It is better to collect and label citrus pictures, make data sets and test models after training Faster RCNN, YOLOv2, YOLOv3 and other networks, the accuracy of the model for citrus fruits in complex environments is 93.52 and the detection time is 34 ms. Compared with other networks, the model takes into account the requirements of recognition accuracy and speed, and the detection and positioning accuracy is the highest and the comprehensive performance is the best. The recognition model is robust to the common interference factors and their superposition in the actual picking environment, such as light change, uneven brightness, mutual occlusion of fruit and branches and leaves, and shadow coverage, which can provide a reference for the design of citrus picking robot.

Acknowledgments
This paper is a key project of Chongqing Education Commission (Research and Application of Intelligent Citrus picking Technology and equipment) (KJZD-K201901202) and The Natural Science Foundation of Chongqing Science and Technology Commission (Study on Mechanical damage Mechanism and Hyperspectral Imaging Nondestructive testing of Citrus picked by Robot)(cstc2020jcyj-msxmX0143) is one of the phased achievements.
References

[1] Wang Dandan, Song Huibo, He Dongjian. Advances in the Visual System of Apple Picking Robot [J]. Research Journal of Agricultural Engineering (10):59-69.
[2] Xu Yue, Li Yinghui, Song Huibo, He Dongjian. Target Segmentation Method of Double Fruit Overlapping Apple Based on Snake Model and Kernel Point Detection [J]; and Journal of Agricultural Engineering 31(01):196-203.
[3] Zhang Fukai, Yang Feng, Li Tze. Rapid Vehicle Detection Method [J]. Based on Improved YOLOv3 Computer Engineering and Applications ,2019,55(02):12-20.
[4] Liu Jizhan. Development and Analysis of Greenhouse Picking Robot Technology Journal of Agricultural Machinery 48(12):1-18.
[5] Redmon J,Farhadi A.YOLOv3:An Incremental Improvement[J] 2018.arXiv preprint arXiv:1804.02767,20.
[6] Bochkovskiy A ,Wang C Y , Liao H Y M.YOLOv4:Optimal Speed and Accuracy of Object Detection[J]. 2020.
[7] Redmon J,Divvala S,Girshick R,et al.You only look once:Unified,real-time object detection[C] Proceedings of CVPR,2015:779-788.
[8] Zhang Fukai, Yang Feng, Li Tze. Rapid Vehicle Detection Method [J]. Based on Improved YOLOv3 Computer Engineering and Applications ,2019,55(02):12-20.
[9] Han Wen, Wei Chaoyu, Liu Huijun. Target Detection Method of Green Citrus in Field [J]. Based on Tiny-YOLOv3 Journal of China University of Metrology (03):349-356+392.
[10] Wang Shuqing, Huang Jianfeng, Zhang Pengfei, Wang Juan. Method of crayfish quality detection [J/OL]. based on YOLOv4 network Food and Machinery :1-9[2020-12-09]. http://kns.cnki.net/kcms/detail/43.1183.TS.20201204.1312.009.html.
[11] ARTHUR D ,V ASSILVITSKII S.K-Means++: The Advantages of Careful Seeding[C] Eighteenth Acmsiam Symposium on Discrete Algorithms. New Orleans:Society for Industrial and Applied Mathematics,2007:1027-1035.
[12] He K M,Zhang X Y,Ren S Q,et al.Deep residual learning for image recognition[C] Proceedings of the 29th IEEE
[13] Li Na, Jiang Zhi, Wang Jun, Dong Xingfa. Faster R-CNN - based instrument identification method [J]. Liquid Crystal and Display ,2020,35(12):1291-1298.
[14] Xue Yueju, Huang Ning, et al. Improved YOLOv2 Identification method of immature Mango [J]. Journal of Agricultural Engineering 34(07): 173-179.
[15] Wu Xing, Qi Zeyu, et al. Apple Detection method [J]. Based on lightweight YOLOv3 convolution Neural Network Journal of Agricultural Machinery, 2020,51(08): 17-25.