

Research on Apple Target Recognition and Counting Method Based on YOLOv5.

Kun Chen *, Wei Zhou, Quanjie Xu, Jiajiang Shen, Kaiyang Zhao

School of Zhangjian, Nantong University, Nantong, China, 226019

* Corresponding Author Email: 3013507370@qq.com

Abstract. In this study, a YOLOv5 target recognition model was established to extract apple feature information and predict the number and location of apples in 200 pictures. First, the model is trained using the training data set, and the validation data set is validated to obtain the best model weight and parameter information. Secondly, this paper specifies that the training times are 50 times, the batch size of one read is 16 times, and the training is carried out through data enhancement operations to enrich the training data set. Finally, the weight and parameter information with the highest average verification accuracy and recall rate were selected for the 20th time to predict 200 pictures, and then the number of apples in each picture was counted.

Keywords: YOLOv5, Apple Recognition, Object detection, image processing, Deep Learning.

1. Introduction

With the continuous development of science and technology, artificial intelligence has gradually penetrated into all aspects of our lives. Among many artificial intelligence technologies, deep learning, as a key technology, has achieved remarkable results in many fields. Deep learning is a kind of learning method that simulates the structure of human brain neural network. It learns and analyzes data through large-scale neural network, so as to recognize and predict the target. In the field of agriculture, deep learning technology has been successfully applied to crop disease identification, fruit maturity detection and other aspects. This paper will discuss the use of deep learning for apple recognition research.

As one of the most popular fruits in the world, the detection of apple quality and yield is of great significance to both farmers and consumers. Traditional apple recognition methods mainly rely on manual observation and experience judgment, which is not only time-consuming and labor-intensive, but also subjective, difficult to achieve large-scale automatic detection. With the development of deep learning technology, we can realize the automatic identification and classification of apples by building a neural network model with strong learning ability.

In order to indicate the origin of problems, the following background is worth mentioning. We need to build an efficient and accurate apple image recognition model to solve the problem of insufficient recognition of apple pickers and robots. Through feature analysis and extraction of marked fruit images, the model can identify the maturity, quality and variety of apples. This paper aims to solve the following problems: The features of these images are extracted, and the corresponding mathematical model is built to count the number of apples in each image. The result is a histogram of the number of apples in the data set.

2. Research Methods

By using the YOLOv5 algorithm to solve the problem of apple number statistics based on the harvested apple image dataset, we need to carry out the following steps:

Data preparation: The image data set is divided into a training set and a test set, ensuring that each image has a corresponding number of apples labeled (groundtruth).

Model selection: YOLOv5 is an object detection algorithm that solves problems by locating and identifying objects in an image. Since our goal is to count the number of apples, YOLOv5 is a suitable choice.

Model training: Using the image data of the training set and the corresponding apple number annotation, we can train the YOLOv5 model. The training process will apply a backpropagation algorithm to optimize the model parameters so that it can accurately detect and locate apples in the image.[1]

Model evaluation: Use the test set to evaluate the model, calculate the accuracy of the model in detecting the number of apples, recall rate and other indicators, and evaluate the performance of the model.

Counting the number of apples: Using the trained YOLOv5 model, target detection was carried out on the image to identify and locate apples. Then, by counting the number of apples detected in each image, the distribution of apples throughout the dataset can be obtained.

Draw a histogram: According to the statistics of the number of apples, draw a histogram of the distribution of all apples to visually show the distribution of apples.

It should be noted that in practical applications, steps such as data preprocessing, hyperparameter tuning, and model optimization also need to be considered to obtain better results (Figure 1).

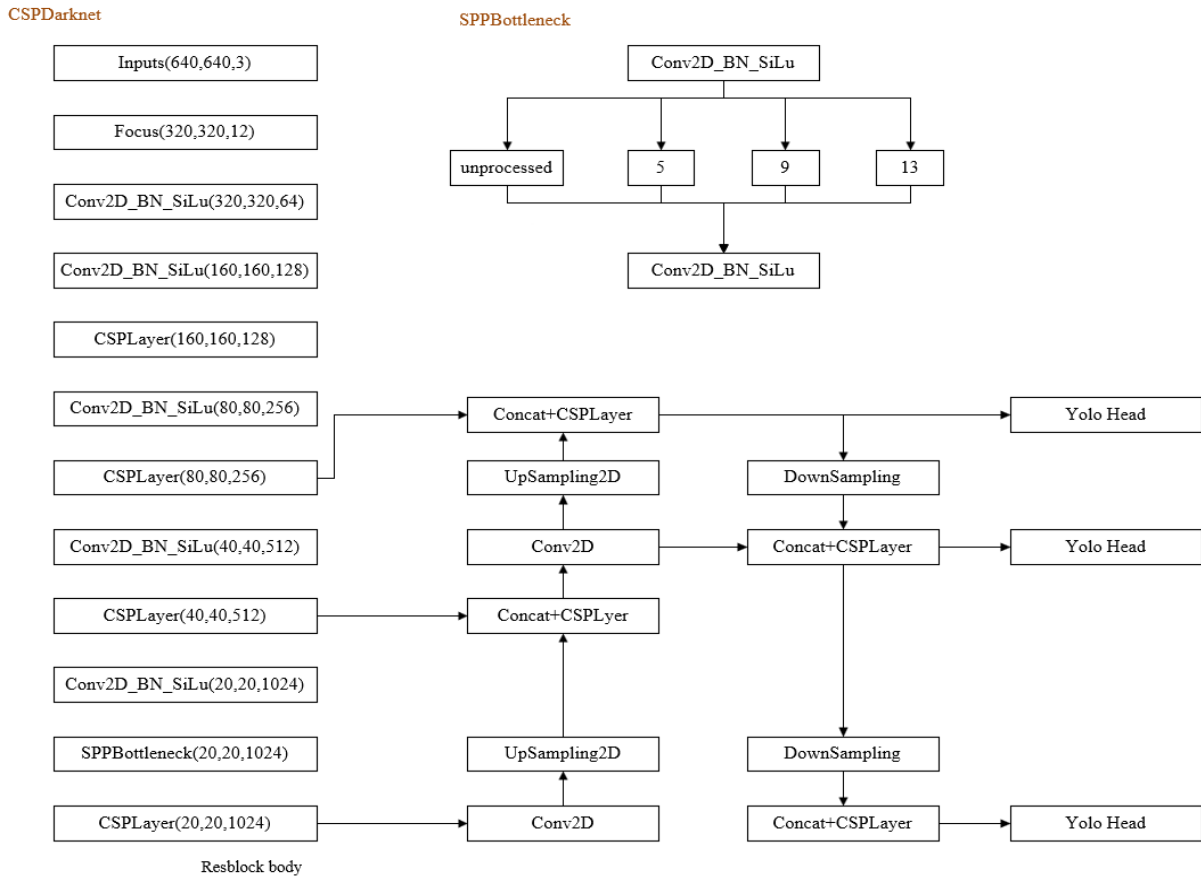


Figure. 1 YOLOv5 Network Structure

This paper uses Precision (P), Recall (R), AveragePrecision (AP), and average precision mean, Mean Detection Time (MDT), and Parameters are used to evaluate the performance of the model. Parameters measure the consumption of memory resources by the model in megabytes. MDT is the average of the time it takes the model to detect a single test set image, in milliseconds. When robot picking is carried out in a natural environment, in order to avoid apple damage, the picking action is generally slow, so the detection speed does not need to be high. Therefore, in this paper, the average precision mean (mAP) is used as the main evaluation index to evaluate the improved detection model. The calculation formulas of evaluation indicators are as follows:

$$P = \frac{TP}{TP + FP} \quad (1)$$

$$R = \frac{TP}{TP + FN} \quad (2)$$

$$AP = \int_0^1 PRdR \quad (3)$$

$$mAP = \frac{1}{n} \sum_{i=1}^n AP_i \quad (4)$$

3. Model Building And Solutions

3.1. Data Preprocessing

Since the number of Apple pictures provided is only 200, which is far from meeting the requirements of training data set, we obtained the data set named ApplesDetection from the official Baidu AISTudio. The data set is very similar to the conditions such as the shooting Angle, surrounding environment, blurring degree, brightness degree and so on of the Apple picture, and the training set, test set and verification set have been divided in this folder. At the same time, LabelImg annotation tool was used to annotate all the 200 pictures and draw the location and category information of the target. The working space is as follows (Figure 2).



Figure. 2 Labelimg's workspace

After the annotation is complete, we save the annotation result as an.xml or.txt file, and ensure that the file name is the same as the image file name. Then, we divided the marked data set into training set, verification set and test set, and merged the training set, test set and verification set into the ApplesDetection folder for model training and evaluation.^[2]

3.2. Build YOLOv5 Target Detection Model

3.2.1 Input data augmentation and Adaptive anchor frame calculation

First, since there are only 200 data sets, Mosaic data augmentation was adopted in this paper, that is, four images were joined together into a large image by random scaling, random cropping and random arrangement. Meanwhile, data augmentation methods such as upside-down flipping, random scaling and color dithering were also adopted to enrich the data set. Improve the detection performance of the model for objects of different scales and shapes. Secondly, we use adaptive anchor

frame calculation, which is a technique used to automatically adjust the size of anchor frame in object detection algorithm. In the training process, the adaptive anchor frame calculation will automatically adjust the size and position of the anchor frame according to the size and shape of the target object to better match the target object, so as to improve the accuracy and efficiency of the target detection.

3.2.2 Building a Backbone Network

Backbone network is a kind of deep neural network, which is mainly used to extract the features of input data and provide effective information for subsequent tasks. The main structures of Backbone constructed in this paper include Conv module, C3 module and SPPF module. The Conv module consists of a Conv2d convolution operation layer^[3],

$$Y = (\sum_{i=1}^n \alpha_{\omega i} \odot \alpha_{ci} \odot \alpha_{fi} \odot \alpha_{si} \odot W_i) * X \quad (5)$$

A BatchNorm2d batch normalization operation layer and activation function. In this module, we use SiLu activation function to improve the generalization ability of the model.

$$Leaky ReLU(x) = \begin{cases} x, & x > 0 \\ \alpha x, & x \leq 0 \end{cases} \quad (6)$$

$$SiLU(x) = \frac{x}{1 + e^{-x}} \quad (7)$$

The C3 module is composed of three Conv modules and a Bottleneck module. We introduce a Bottleneck module for residual connection and add the input into the second Conv module to the output of the third Conv module for matrix addition, in order to avoid the backpropagation of updating weight information when optimizing. The problem of disappearing gradients arises. Finally, the SPPF module belongs to a special pooling layer, SPPF adopts a multi-scale spatial pyramid pooling structure, which can divide the input feature map into multi-scale grids and carry out pooling operations in each grid to generate fixed-length feature vectors. This enables SPPF to capture target information of different scales and improves the accuracy of target detection.

3.2.3 Building the Neck Network

The Neck network structure is a key component of the model. It is mainly responsible for the fusion and transmission of features extracted by the Backbone network to generate high-quality feature maps for the subsequent Head network for target detection. The Neck network constructed in this paper is composed of SPP module and PANet module. The SPP module pools the feature maps of different scales to obtain a fixed length feature vector, which solves the problem of target detection with different scales and aspect ratio. The PANet structure consists of two parts: feature upward fusion and feature downward fusion. Feature downward fusion starts from the top feature map and fuses the top feature with the bottom feature through up-sampling operation and horizontal connection to obtain a series of multi-scale feature maps. These feature maps not only contain rich semantic information, but also have high resolution, which is conducive to the detection of small targets. Feature upward fusion starts from the feature map of the lowest layer and fuses low-level features with high-level features through downsampling operations and horizontal connections to further enhance the transmission and fusion of features. The output feature maps of the two paths are stitched together to form a more comprehensive feature representation.

3.2.4 Output Construction

At the output end, we use loss function and non-maximum suppression to improve the performance of the model. We adopt the FocalLoss loss function. On the basis of balancing the cross-entropy loss function, FocalLoss adds a regulatory factor that can reduce the weight of easily classified samples and focus on the training of difficult samples.

$$FL(p) \begin{cases} -\alpha(1-p)^\gamma \log(p) & \text{if } (y == 1) \\ -\alpha(1-p)p^\gamma \log(1-p) & \text{otherwise} \end{cases} \quad (8)$$

$$FL(P) = -\alpha(1-p)^\gamma \times y \log(p) - (1-\alpha)p^\gamma \times (1-y) \log(1-p) \quad (9)$$

In other words, compared with the cross-entropy loss, FocalLoss can reduce the weight of samples with inaccurate classification. The losses do not change, and for accurately classified samples, the losses will be smaller. Therefore, as a whole, it is equivalent to increasing the weight of inaccurately classified samples in the loss function. Therefore, in general, this method can alleviate the problem of class imbalance in target detection and improve the performance of the model. Finally, non-maximum suppression can remove the redundant detection frame and obtain the final detection result [4].

3.3. Build optimizer algorithm

In this model, we decided to use the Adam optimizer model, which belongs to the optimizer model of stochastic gradient descent algorithm [5]. It works by calculating the gradient of the FocalLoss loss function and then updating the parameter information of different layers in the model by backpropagation. Compared with the traditional stochastic gradient descent model, such as SGD optimizer and BGD optimizer, the optimizer selected in this paper introduces the learning rate factor as a hyperparameter to adapt the learning rate. At the same time, the optimizer also introduces a momentum term, that is, taking into account the direction of the gradient in the previous step, making parameter updates smoother, helping to accelerate convergence and reduce oscillations. So at the beginning of training, the model can converge quickly; At the later stage of training, the model can be fine-tuned at a smaller learning rate, making the model perform better.

3.4. Test analysis

Firstly, all the 200 pictures were annotated using labeling software to make the data set, and then divided into the training data set, verification data set and test data set according to 8:1:1, and then merged into the corresponding folder in the ApplesDetection data set officially obtained by Baidu AIStudio. Then, the training data set and test data set are imported into the YOLOv5 target detection model for loop training. We specify 50 training times and 16 batch size for one-time reading [6]. After several hours of model training, we get the average accuracy and recall rate of verification set for each training under the condition of 30 training times. The line chart is as follows (Figure 3).

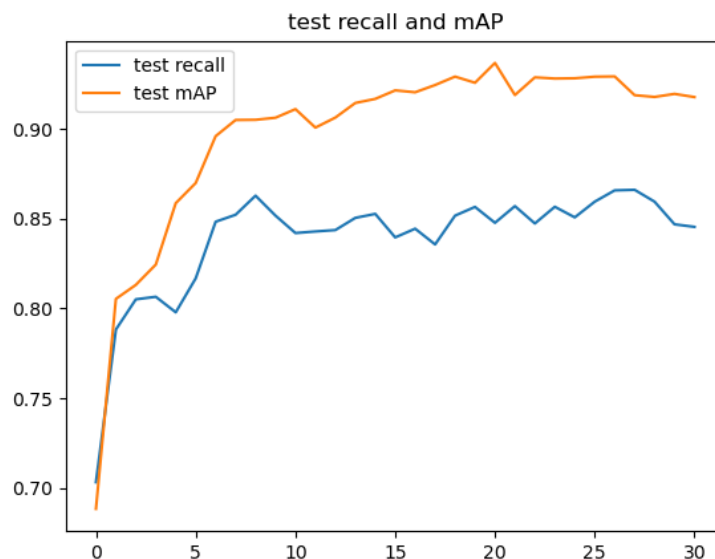


Figure. 3 Test recall and mAP

We found that the average accuracy of the verification set was the highest and the recall rate was the largest in the 20th training, while the average accuracy of the verification set and the recall rate decreased from the 20th training, so we ended the training at the 31st cycle and chose to use the weights, parameters and other information of the 20th training. Finally, we used this information and the YOLOv5 model to identify and detect the 200 pictures one by one, and counted the apples in the pictures when the model marked them, that is, a number was recorded for each apple marked and then continuously added up^[7], and the number information of apples in each picture was stored. Therefore, we finally draw the distribution histogram of all apples according to the quantity information of each apple stored, as shown below (Figure 4 and 5).

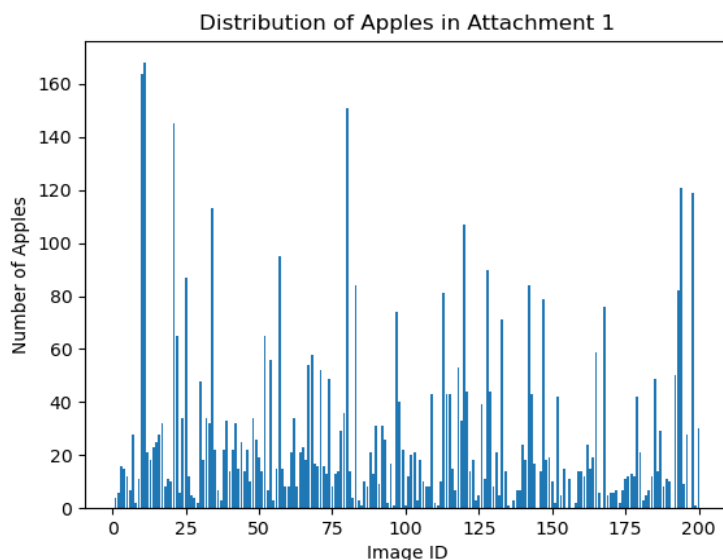


Figure. 4 The distribution histogram of the number of apples (Method 1)

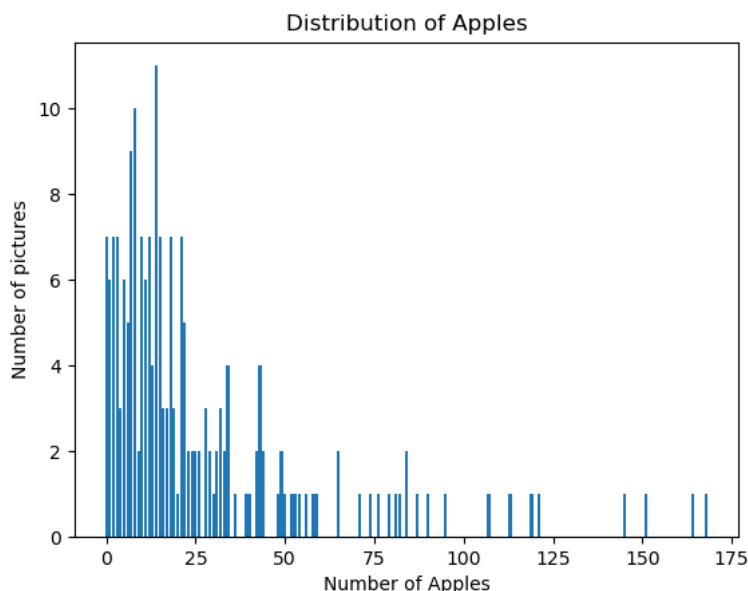


Figure. 5 The distribution histogram of the number of apples (Method 2)

4. Conclusions

In this paper, by establishing the YOLOv5 target detection model, the accurate identification and counting of apples are realized. First, 200 apple pictures were collected as a data set, and Labellmg software was used to label apples in each picture to obtain the quantity and location information of apples. Then, a YOLOv5 target detection model was constructed, data enhancement method was used

to expand the data set, adaptive anchor framework was used to calculate and improve the detection performance, and Focal Loss function and Adam optimizer were used for model training. Finally, the model performance was evaluated on the verification set, the model with the highest average accuracy and recall rate was selected for apple identification and counting, and the distribution histogram of the number of apples in the data set was counted. The experimental results show that the proposed method can improve the recognition accuracy and counting efficiency of apples and provide support for apple picking robots. In general, by constructing the YOLOv5 target detection model, this paper realized the accurate identification and counting of apple targets^[8] and provided effective support for apple picking robots.

References

- [1] ZHONG Shihao. Research on target recognition and localization algorithm of clustered Chaotianpepper based on deep learning [D]. Guizhou normal university, 2023.
- [2] Li Changlu. Research on recognition and localization of apple picking robot based on machine vision [D]. Lanzhou university of technology, 2022. D
- [3] Gao Meng. Research on sentence matching method based on Fused residual network attention mechanism [D]. Chongqing university of posts and telecommunications, 2021.
- [4] Guo X. Research on fruit detection algorithm for apple picking robot in complex environment [D]. North central university, 2023.
- [5] Manoj J H K. An improved method for text detection using Adam optimization algorithm[J]. Global Transitions Proceedings,2022,3(1):230-234.
- [6] Mohd A Z, Asraf M M, Adhi H S, Siti R A, et al. Automated Apple Recognition System Using Semantic Segmentation Networks with Group and Shuffle Operators[J], AGRICULTURE-BASEL, 2022, 12(6).
- [7] Zhuoqun Zhao, Jiang Wang, Hui Zhao. Research on Apple Recognition Algorithm in Complex Orchard Environment Based on Deep Learning. [J], Sensors (Basel, Switzerland), 2023, 23(12): 5425-5425.
- [8] Yong Zhong, Ming Zhao. Research on deep learning in apple leaf disease recognition. [J], Computers and Electronics in Agriculture, 2020, 168.