WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 3, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n3.26
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Application of Image Restoration and Object Detection Techniques in Fruit-Picking Robots

Qiong Yang*, Jiana Meng, Yuhai Yu, Siwei Han, Guijun Luo

College of Computer Science and Engineering, Dalian Minzu University, Liaoning, China
*Corresponding Author: hnsyq69@163.com

ABSTRACT

This paper focuses on the problem that existing apple picking robots are unable to accurately recognize targets in complex and unstructured orchard environments. The target detection algorithm of the robot is optimized based on YOLOv5 and ResNet neural network model. First, considering the low resolution of the robot's own binocular camera, this paper adopts the RealSR model to process the image with super-resolution. Then, the YOLOv5 model is chosen as the optimized target detection algorithm, which can obtain the number, location, ripeness and quality information of apples in the image through training. Secondly, by comparing the VGG and ResNet models, which are widely used in the field of image classification, the ResNet model is finally chosen as the main body for recognizing and classifying different fruits, and its correct rate can reach more than 97%.

KEYWORDS

RealSR, YOLOv5, VGG, ResNet.

1. INTRODUCTION

In today's society, the development of orchard automation technology is of great significance in solving labor shortages and improving agricultural productivity. However, the application of apple picking robots in complex orchard environments around the world is currently subject to many challenges, such as the "leaf shade", "branch shade", "fruit shade", etc. The application of apple picking robots in complex orchard environments is also subject to many challenges, such as the "leaf shade", "branch shade", and "fruit shade". The problem of accurately recognizing obstacles, as well as the difficulty of classifying and recognizing different fruits. These challenges directly affect the picking efficiency and fruit quality, bringing potential economic losses and safety risks [1]. In order to cope with this problem, this paper is dedicated to the study of the application of machine learning and computer vision techniques in the optimization of apple picking robots. Through in-depth analysis of various occlusion obstacles existing in complex orchard environments, we combine RealSR technology with super-resolution processing of low-resolution images to improve image clarity and accuracy. In terms of target detection, we compared several advanced models and chose the YOLOv5 model for optimization based on the consideration of the robot hardware facilities. By processing and data enhancement of 200 low-resolution images, we successfully improved the image recognition effect, which provides reliable support for the practical application of the robot. In addition, for the problem of recognizing and classifying different fruits, we carried out in-depth research and comparison, and finally chose the ResNet model as the main body, whose accuracy can reach more than 97%. This research not only provides key technical support for the production efficiency and quality of the apple industry, but also provides useful exploration and practice for the further development of orchard automation technology. By analyzing the stability and objectively

evaluating the deep learning models used, we provide an important reference for future research and application, helping the orchard intelligent technology to move towards a smarter and more efficient future.

2. SUPER-RESOLUTION MODEL

Super resolution is the process of restoring low quality compressed images into high resolution images. With the rapid development of mobile Internet, smart devices are gradually popularized to every corner of life. Along with it comes a large amount of real image data, the quality of these pictures will be compressed due to the need of storage and transmission, in order to enable users to obtain a higher quality visual experience, image restoration/super-resolution algorithms came into being. The importance of super-resolution as an underlying visual task is indisputable, and the most intuitive effect is the enhancement of the sensory quality of the human eye.

2.1. SRGAN(SRResNet)

SRGAN (Super-Resolution Generative Adversarial Network) is a deep learning model for image super-resolution reconstruction. It generates more realistic and high-quality super-resolution images by utilizing the adversarial loss function and the perceptual loss function's works as follows: the G-net generates a high-resolution image from a low-resolution image, and it is up to the D-net to determine whether the image obtained is generated by the G-net or is the original image in the database. When the G-net can successfully fool the D-net, then super resolution can be accomplished by this GAN.

2.2. CARN

CADN (Contextual Attention-based Deep Network) is a deep learning model for image restoration. It restores missing or damaged regions by utilizing contextual information in the image, making the restored image more realistic and accurate.

2.3. RealSR

RealSR is a deep learning model for super-resolution reconstruction, which aims to convert low-resolution images into high-resolution images to enhance image quality and detail richness [2]. The RealSR model adopts the training of real photo pairs and generates a low-resolution image and the corresponding high-resolution image by random operations such as cropping and rotating, in order to improve the generalization ability and robustness of the model. Compared with the existing super-resolution methods, the innovation of RealSR is mainly in three aspects:

RealSR adopts a new picture degradation method designed independently to simulate the degradation process of real pictures by analyzing the blur and noise in real pictures.

Pairs of training data are not required, and training can be performed by utilizing unlabeled data.

It can deal with the problem of blurring and noise in low resolution images and get clearer and cleaner high-resolution results.

2.4. Super-resolution reconstruction

We deployed three models in Anaconda's virtual environment, respectively, with the help of already trained models for migration learning and training on 200 images, respectively, after 200 rounds, 120 rounds and 40 rounds of training, according to the generated images it is known that RealSR works best, and RealSR is chosen for super-resolution reconstruction of the dataset. The results are shown in Figure 1.

Figure 1. Comparison of the effects of the four super-resolution models

We super-resolved 200 images to get 200 high-definition resolution apple images, and then we data-enhanced the 200 high-definition resolution apple images by doing the following for each of the 200. The enhanced result is shown in Figure 2.

Figure 2. Data enhancements

4000 high resolution images of apples were obtained through data enhancement. We scientifically randomly selected a portion of the images and manually labeled them using the Labeling tool to obtain the location and area of all apples in the image.

3. CONSTRUCTION AND SOLUTION OF APPLE DETECTION MODEL BASED ON YOLOV5

YOLOv5 is the latest development in the YOLO target detection series introduced by Alexey Bochkovskiy and others at the University of California, Berkeley. Compared to previous versions, YOLOv5 offers significant improvements in inspection accuracy, speed, and model compactness. The algorithm uses the most advanced computer vision technology and neural network architecture, which is characterized by speed, accuracy and efficiency [3-4].

3.1. Model training

YOLOv5 is divided into five models, YOLOv5n, YOLOv5s, YOLOv5m, YOLOv51, and YOLOv5x, and after the comparison test and the size of the data volume, we finally chose the YOLOv5s model.

Firstly, the manually labeled images are inputted into the input side of YOLOv5s for Mosaic data enhancement, image size processing and adaptive anchor frame calculation, etc. Mosaic data enhancement combines the four images to achieve the effect of enriching the background of the images; image size processing adaptively adds the least black edges to the original images of different lengths and widths, and unifies them to be scaled to the standard size; and adaptive anchor frame calculation compares the output predicted frame with the real frame on the basis of the initial anchor frame, calculates the gap, and then updates the parameters in reverse to obtain the most suitable anchor frame value. Anchor frame calculation on the basis of the initial anchor frame, the output predicted frame is compared with the real frame, the gap is calculated and then updated in reverse, and the parameters are continuously iterated to obtain the most suitable anchor frame value.

In the feature extraction stage of Backbone, the basic structure of Conv, C3, and SPPF is used to extract features from the input image; Conv is used to down sample the input (a total of 5 times down sampling); C3 is used to extract features from the input, fusion, and enrichment of semantic information of the features, and Bottleneck is used to reduce the number of parameters and the amount of computation in the process. Then in the processing feature stage, shallow features are fused for the three scales of feature maps to be used for target detection (shallow features are beneficial for detection). Drawing on PANet to fuse shallow features to the extracted feature maps, the feature maps are rich in semantic information as well as accurate location information of the objects.

Finally, Head outputs a vector with the category probability of the target object, the object score and the location of the bounding box of the object. The detection network consists of three detection layers, where feature maps of different sizes are used to detect target objects of different sizes. Each detection layer outputs the corresponding vector and finally the predicted bounding box and category of the target in the original image is generated and labeled.

We conducted 200 rounds of training on the YOLOv5s model and got the best result to generate the best.pt file, and used the trained best.pt to deduce the remaining images after super resolution processing. Finally, the model reasoning identified the apples in all 200 images and generated 200 txt files, each containing the location and area of all apples in the corresponding image, and marked the maturity of all apples. The inference obtained from randomly selected images from the test set is shown in Figure 3.

3.2. Model evaluation

When using neural network models for target detection algorithms, some mathematical metrics are needed to evaluate the performance of a network model, which are:

(1) True Positives (TP): Denotes the number of positive samples for which the target result can be correctly identified.

(2) True Negatives (TN): Denotes the number of negative samples for which the target result can be correctly identified.

(3) False Positives (FP): Denotes the number of positive samples of target results that were incorrectly identified.

(4) False Negatives (FN): Indicates the number of negative samples of incorrectly identified target results.

(5) Precision: Accuracy rate, which indicates the proportion of the results detected by the algorithm's recognition that the target is truly present.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{1}$$

(6) Recall: which indicates the ratio of correct targets recognized among the results of image recognition detection to the total number of targets in the class.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2}$$

Figure 4 shows the changes in the values of loss, precision and accuracy during the 200 epochs of model training. The size of the loss value represents the accuracy of the model, the smaller the loss means the higher the prediction accuracy of the model. From the figure, it can be seen that the model stabilizes after 100 epochs of training, the loss value reaches a smaller level, and the precision rate and recall rate are both close to 0.9.

Figure 3. Model reasoning effect diagram

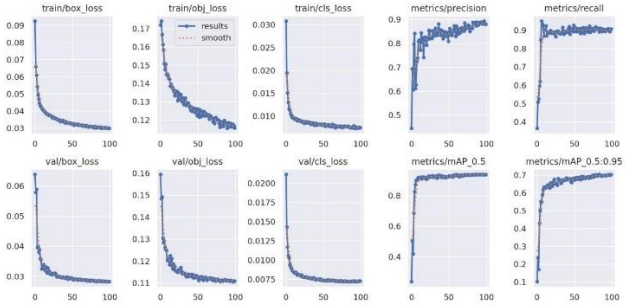

Figure 4. Training/validation loss

3.3. Model solving

3.3.1. Estimating the number of apples

Based on the trained YOLO model, we can know how many apples there are in a picture. Each line of the output contains the category of apples, the position of the block diagram (normalized), so the number of rows in the label corresponds to the number of apples in the image. Figure 5 is the distribution of the number of apples in the 200 images shown in the statistical histogram.

3.3.2. Estimating the location of apples

The center point of each apple can be located according to the position and area information of the apple, and the center point of all apples can be compiled into an array, and the two-dimensional scatter diagram of the geometric coordinates of all apples can be drawn according to the data in the array. The position of the apple is shown in the Figure 6.

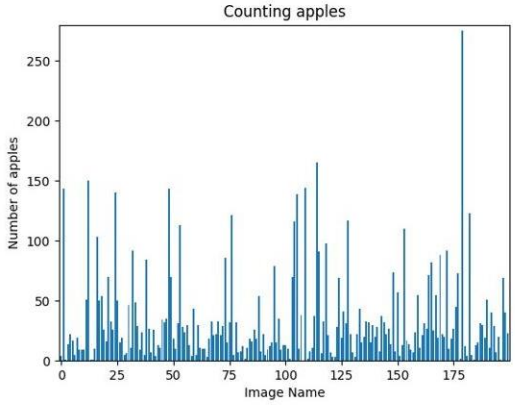

Figure 5. Histogram of the distribution of the number of apples

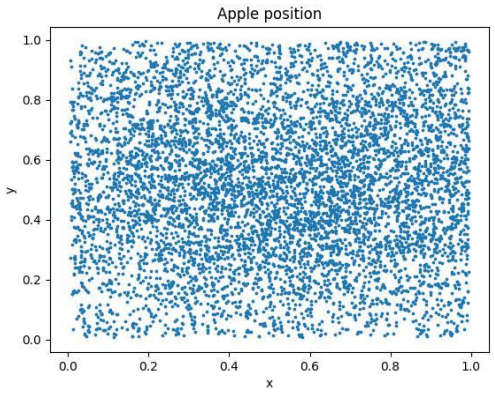

Figure 6. Scatter plot of all Apple locations

3.3.3. Estimating the maturity state of apples

All apples reasoned using the trained YOLOv5s model are automatically labeled by the model, and the apples are labeled by us into four categories, i.e., fully ripe stage, semi-ripe stage, immature stage, and flowering stage. The ripeness of the apple is shown in Figure 7.

3.3.4. Estimating the masses of apples

Since the output results contain the position information and area information of all apples, based on the area information, we can deduce the projected area S of a two-dimensional plane image. And according to a lot of experiments to find a suitable coefficient K', so that K'multiplied by S close to the actual surface area of the apple. Finally, approximate surface area is used to represent the size of the apple. The mass of a normal ripe apple is about 250g. We take one apple as the standard, dynamically calculate the mass of each apple based on the approximate surface area, put the quality information of all apples into an array, and then plot a histogram of the mass distribution of all apples, as shown in Figure 8.

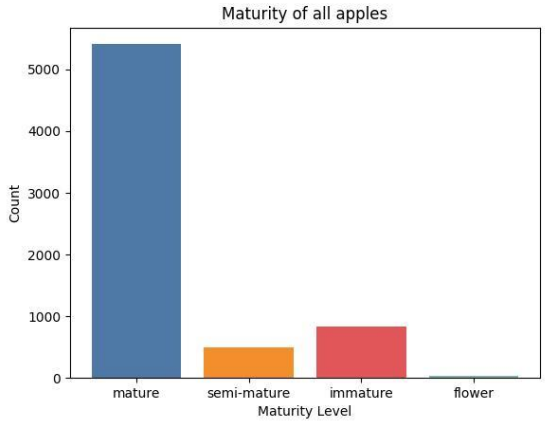

Figure 7. Histogram of ripening status of all apples

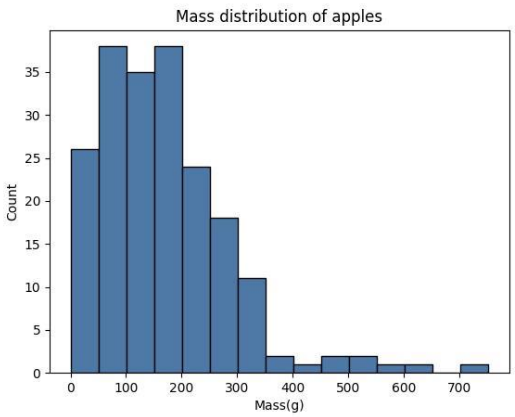

Figure 8. Histogram of mass distribution of all apples

4. APPLE CLASSIFICATION MODEL BASED ON RESNET

4.1. Selection of image classification model

VGG: VGG (Visual Geometry Group) is a deep convolutional neural network architecture proposed by a visual geometry group in 2014. VGG network uses consecutive small convolutional kernels (3×3) and pooling layers to build deep neural networks, which can be as deep as 16 or 19 layers, of which VGG16 [5] and VGG19 [6] are the most famous. Compared to previous network structures, VGG16 uses deeper network depth and small-sized 3x3 convolutional kernels, which improves the network's nonlinear modeling ability and reduces the number of parameters, making the model more trainable and efficient. The convolutional layers are nonlinearly mapped to each other using the ReLU activation function, while the fully connected layers are followed by a Dropout technique to prevent overfitting. However, VGG consumes more computational resources and uses more parameters, resulting in a larger memory footprint (140M). The vast majority of these parameters are from the first fully connected layer.

ResNet: ResNet (Residual Network) is a deep neural network model for image classification tasks [7]. Compared to traditional convolutional neural networks, ResNet introduces residual blocks in the network structure, which allows the network to better handle deep features, thus improving the classification performance. The residual block solves the problem of gradient vanishing in the optimization process by adding "shortcut connections", which enables efficient training of deep

networks. In ResNet, each residual block is actually a function $f(x)$ consisting of two or more convolutional layers. Thus, the ResNet model achieves connections between the main path and the shortcuts, which improves training in the deeper layers by adding "shortcut connections" connections between the main path and shortcuts, which improves the expressive power and generalization performance of the network.

4.2. Comparison of image classification model results

Expand the dataset consisting of 200 images and divide them randomly at a ratio of 7:3. The image size in the data set is uniformly cut to 224*224. The enhanced image data obtained after random Angle rotation and random horizontal flip are passed into two deep learning models, VGG16 and ResNet18, which pass through multiple convolutional layers and residual blocks respectively, and finally output the categories of five kinds of fruits through the fully connected layer. And the experimental parameters are set as Table.1:

Table 1. Experimental parameter table

Parameter	Value	Meaning
Epoch	20	Training rounds
Batch Size	16	Images number per batch
Loss	CrossEntropyLoss	Loss function
Optimizer	adam	Optimization function
Learning Rate	0.001	Speed of parameter change
Dropout	0.5	Percentage of parameters to discard

We also compared the precision, recall and accuracy of the two models, and the changes of several evaluation indexes with epochs on the verification set are as shown in Figure 9 and Figure 10. By comparing the evaluation graphs of the two models, it can be seen that although the two models are very effective, the comprehensive comparison of F1 values shows that ResNet18 is still better, and its accuracy rate can reach more than 97%. So, we choose ResNet18 as the final classification model.

More than 11,000 apple images were used as a test set for classification training, which was passed into yolov5s model for inference, and the specific number of apples in each apple image was obtained. In this paper, the ID number is set as a group of 500, and the number of apples in each group of apple images is counted respectively. The histogram of ID number distribution of all apple images is finally obtained, as shown in Figure 11.

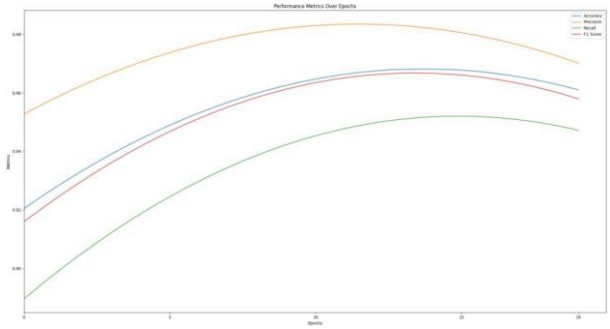

Figure 9. VGG16 training result graph

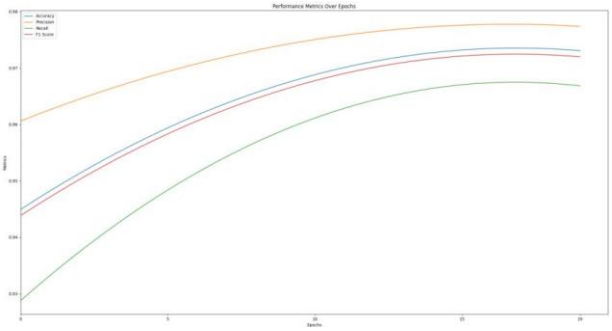

Figure 10. ResNet 18 training result graph

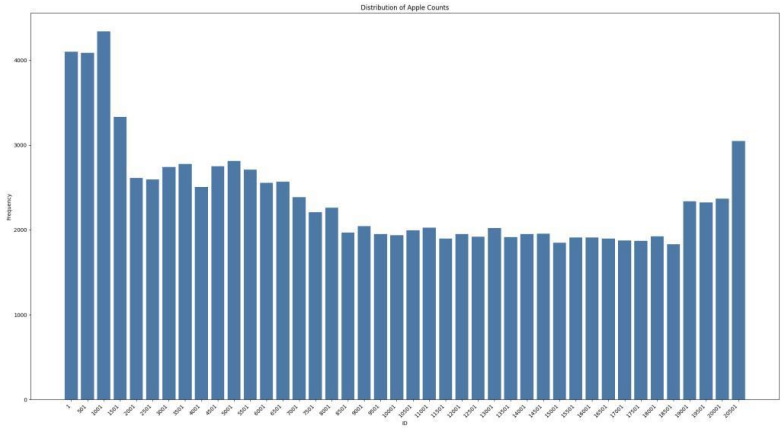

Figure 11. Histogram of the distribution of all apple image ID

5. CONCLUSIONS

In this study, we aim to address the recognition barriers faced by apple picking robots in orchard environments in order to improve picking efficiency and fruit quality. By using RealSR super-resolution processing and optimized target detection algorithm YOLOv5, we successfully improved the robot's ability to recognize obstacles such as "leaf shade", "branch shade", "fruit shade", etc., thus reducing the risk of accidental damage and injury. We successfully improve the robot's ability to recognize obstacles such as "leaf shade", "branch shade", "fruit shade", etc., which reduces the risk of accidental damage and injury, and improves the harvesting efficiency.

We further explored the problem of recognizing and classifying different fruits and chose the ResNet model as the main body for experimental validation and achieved an accuracy rate of more than 97%. This lays the foundation for automatic recognition and classification of diverse fruits in orchards, which helps to realize intelligent harvesting and handling processes and improve overall agricultural production efficiency.

By comparatively analyzing the stability and performance of different deep learning models, we have objectively assessed the applicability of the selected model and verified its reliability and effectiveness in complex environments. This provides useful reference and guidance for future research and practical applications.

In summary, this study provides a useful idea and practical basis for the improvement of apple picking robots in orchard environments through the combined use of image processing techniques and deep learning models. Our work has not only made some breakthroughs in technology, but also provided useful exploration and reference for the development of agricultural intelligence and modernized production. The future research direction can focus on further optimizing the performance of the algorithm, expanding the scope of application, and transforming the research results into actual productivity to promote the innovation and development of the agricultural field.

REFERENCES

[1] Sivakumar, Dharini, Yuming Jiang, and Elhadi M. Yahia. "Maintaining mango (Mangifera indica L.) fruit quality during the export chain." Food Research International 44.5 (2011): 1254-1263.

[2] Cai Jianrui, et al. "Toward real-world single image super-resolution: A new benchmark and a new model." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[3] Dou Zhan; Zhou Hang; Liu Zhe; Hu Yuanhao; Wang Pengchao; Zhang Jianwen; Wang Qianlin; Chen Liangchao; Diao Xu; Li Jinghai. An Improved YOLOv5s Fire Detection Model.[J]. Fire Technology, 2023: 1-32.

[4] Yaodi Li; Jianxin Xue; Mingyue Zhang; Junyi Yin; Yang Liu; Xindan Qiao; Decong Zheng; Zezhen Li. YOLOv5-ASFF: A Multistage Strawberry Detection Algorithm Based on Improved YOLOv5[J]. Agronomy, 2023, Vol.13(1901): 1901.

[5] Bansal, Kanishk | Singh, Amar. Development of VGG-16 transfer learning framework for geographical landmark recognition[J]. Intelligent Decision Technologies,2023, Vol.17(3): 1-12.

[6] Weiqiang Fan; Xiaoyu Li; Zhongchao Liu. Fusion of visible and infrared images using GE-WA model and VGG-19 network[J]. Scientific Reports,2023,Vol.13(1): 190.

[7] Omar El Ariss; Kaoning Hu. ResNet-Based Parkinson's Disease Classification[J]. IEEE Transactions on Artificial Intelligence, 2023, Vol.4(5): 1258-1268.