Research on Spatial Positioning System of Fruits to be Picked in Field Based on Binocular Vision and SSD Model

Xing Zhang 1,2, Qiaoming Gao *,1,2, Dong Pan 1,2, Peng cheng Cao 1,2, Dong hui Huang 1,2

1 Department of Mechanical and Transportation Engineering, Guangxi University of Science and Technology, Guangxi Liuzhou 545616, China
2 Guangxi Automotive Parts and Vehicle Technology Key Laboratory, Guangxi University of Science and Technology, Liuzhou, 545616, China;

Email of all the authors:
* Corresponding Author: Xing Zhang; email: 221032988@stdmail.gxust.edu.cn; phone:13135623863.
Qiaoming Gao; email: 100001864@gxust.edu.cn; phone: 19943065682.
Dong Pan; email: 767051614@qq.com; phone: 15738532260.
Peng cheng Cao; email: 1123075936@qq.com; phone: 18409166451.
Donghui Huang; email: 401370126@qq.com; phone: 18877239965.

Abstract: The accurate fruit recognition in the field was one of the key technologies of fruit picking agricultural robots. An improved Single Shot Multi-Box Detector (SSD) model based on the color and morphological characteristics of fruit was proposed in this paper when aimed at the large collection workload and low secondary transfer efficiency of fruit such as palm fruit, durian, pineapple and other fruits grown in a complex field environment. A binocular depth camera RealSense D435i was used to collect images of the fruit to be picked in the field. Meanwhile, the MobileNet was replaced with the VGG16 basic network based on the TensorFlow deep learning framework to reduce the amount of convolution operations for extracting image features in the SSD model, and a spatial positioning system for pineapple fruit was designed. Furtherly, experiments showed that the improved SSD depth detection model had a smaller size and it was more convenient to be deployed on the mobile end of agricultural robots, which the model had a high accuracy in the effective recognition of the fruits to be picked under the weed occlusion and overlapping scenes. The frame rate of the video reading and detection for the binocular depth camera reached 16.74 Frames Per Second (FPS), which had good robustness and real-time, and a good solution for the automatic picking of agricultural picking robots could be provided in the field.

1. Introduction

With the development of intelligence agriculture, a series of applications such as machine learning, deep learning, image processing, and robotic have been a popular issue in current agricultural information technology research [1]. Some larger fruit such as oil-palm fruits, durians and pineapples that grow in complex fields with rough surfaces and burrs has large harvesting workload and low secondary transport efficiency. It is of great significance to study how to locate the fruit space position, and to
develop a fruit picking robot to autonomously collect and transport the fruits to be picked on the roadside.

Researchers have successively proposed a number of recognition and localization methods for fruits based on image processing in natural environments. In traditional recognition algorithm, Wu et al. [2] proposed a peach detection method based on point cloud clustering and spherical segmentation, which achieved a recognition accuracy of 88.68%. Sun et al. [3] proposed a green apple detection method based on GrabCut model and Neut algorithm, which achieved a recognition accuracy of 94.12%. Fu Lanhuai et al. [4] proposed a banana detection method based on fruit texture features and Support Vector Machine (SVM) classifiers, which achieved a recognition accuracy of 92.55%. He Zhi-Liang [5] proposed a method of green litchi recognition based on improved linear discriminant analysis (LDA) classifier and hough transform fitting circle, which achieved a recognition accuracy of 80.4%. Zhao et al. [6] proposed a green citrus detection method based on fruit color difference model and SVM classifier, which achieved a recognition accuracy of 83%. Zhao Yuanshen et al.[7] proposed a tomato detection method for threshold segmentation of fruit images after wavelet transformation in Luminance In-phase Quadrature-phase (YIQ) color space, which achieved a recognition accuracy of 93%. However, those methods have low robustness and accuracy caused by factors such as fruit color, texture, and light intensity in a complex environment.

Compared with the traditional methods above, the Deep convolutional neural network (DCNN) method have been widely used in agriculture due to their strong feature extraction ability and autonomous learning ability. Liu Guoxu et al. [8], Liang Cuixiao et al. [9]and Tian et al. [10] respectively proposed the tomato, litchi and apple detection method based on YOLO-v3 model, which achieved a recognition accuracy of 94.58%, 96.52% and 96% respectively. Santos Thiago T al. [11]and Yu Y et al. [12] respectively proposed grape and strawberry detection method based on Mask-RCNN model, which achieved a recognition accuracy of 91% and 95.78% respectively. Villacrés Juan Fernando et al. [13] and Song Zhenzhen et al. [14] respectively proposed the cherries and Kiwifruit detection method based on Faster-RCNN model, which achieved a recognition accuracy of 85% and 87.61% respectively. Shanjun L et al. [15] and Hongxing P et al. [16] proposed the citrus detection method based on SSD, which achieved a recognition accuracy of 87.89% and 89.53% respectively. Therefore, a detection method for the fruit to be laid aside roadside after manual picking in the complex environment of the field based on MobileNet-SSD model was proposed in this paper. The main objectives of this research are as follows:

- (1) Using binocular depth camera realSenseD435i to collect images for recognition research.
- (2) Proposing an end-to-end fruit target detection framework based on MobileNet-SSD model.
- (3) Designing a space positioning system for roadside fruits to be picked in a complex environment to lay the technical foundation for the later development of agricultural picking robots.

2. System design and research methods

2.1. System design composition

Figure 1 shows a field fruit spatial positioning system based on binocular vision and SSD model after studying the actual environment of picking robots in the field. It is mainly composed of a fruit recognition part and a spatial coordinate acquisition part. The binocular depth camera realSenseD435i collects images of fruits which are manually picked and placed on the roadside and completes the depth extraction through camera calibration. The SSD convolutional neural network recognizes the fruits, and finally obtains the actual spatial coordinates through coordinate transformation.
2.2. Fruit recognition method

The current deep convolutional neural network object models are mainly divided into two categories. One is the detection algorithm based on region suggestions, such as RCNN[17], Fast-RCNN[18], Faster RCNN[19], Mask R-CNN[20], another directly converts the problem of target frame positioning into regression problem without generating candidate frames, such as RetinaNet[21], SSD[22], YOLO[23], etc. Figure 2 shows a lightweight deep neural network called MobileNet, which the standard convolution can be divided into deep convolution and point-wise convolution, where the deep convolution performs convolution on each input channel separately, and the point-wise convolution is linearly connected to the output of the deep convolution. If M is the number of image input channels, N is the number of output channels, and Dk is the size of the convolution kernel of the depth separable convolution, the ratio of the calculation amount of the separable convolution to the calculation amount of the standard convolution can be derived as in Equations (1):

$$\frac{D_k \times D_k \times D_k \times M}{D_k \times D_k \times D_k \times D_k} = \frac{1}{N} + \frac{1}{D_k \times D_k}$$

(1)

Figure 2. Schematic diagram of the deep separable convolution process of the MobileNet network
Therefore, it can be seen that MobileNet network can effectively reduce the amount of convolution operation for extracting image features, shorten the model training time, and ensure that the model can still have better recognition accuracy. Therefore, figure 3 shows a MobileNet-SSD model in this paper, which is mainly divided into two parts. One part is located at the front end, which the MobileNet network replaces the VGG16 basic network in the SSD for the preliminary extraction of fruit targets. Another is the multi-scale feature detection network at the back end starting from the 11th convolutional layer to extract the features of the front-end network under different scale conditions to obtain six feature layers to predict the position and category of the fruit. Finally, the non-maximum suppression module is used to filter out the repeated prediction targets. In general, a model was shown in this paper can accelerate the calculation speed and effectively ensure the real-time performance while reducing the model size and the performance requirements of the deployment equipment, laying a technical foundation for the later deployment of agricultural picking robots.

![Figure 3. Schematic diagram of the MobileNet-SSD network model structure](image)

2.3. Location method of fruits to be picked based on binocular vision

The traditional fruit spatial positioning research is mainly based on the calibration of the two ordinary cameras on the left and right, the three-dimensional matching of the fruit picking points in the left and right images is performed to calculate the matching point parallax and the three-dimensional coordinates according to binocular vision triangulation [24]. Ling Xiao et al. [25] and Xiang Rong et al. [26] successfully used binocular stereo vision to obtain the spatial coordinates of tomatoes and applied them to tomato identification and picking. Whereas the positioning stability and accuracy of ordinary binocular cameras were affected by the structure of binocular cameras, meanwhile, its real-time performance was not as good as binocular depth cameras. Therefore, figure 4 shows a binocular depth camera RealSense D435i produced by Intel, whose working principle is that structured light is emitted through an infrared transmitter, meanwhile, two infrared cameras triangulate the structured light to generate a computer depth image, and cooperate with the color image to generate a point cloud image with three-dimensional characteristics. The specific method steps are described as follows.

(1) Through the development kit RealSense SDK2.0 and dynamic calibration application programming interface (API) provided by Intel, the camera is calibrated in real time, and the depth map is aligned with the color map to obtain the matching color frame and depth frame.

(2) Establishing the mapping relationship among the RGB image, depth image pixel coordinate system, and the camera coordinate system based on the internal and external parameter matrix of the depth camera. Furtherly, the internal parameter matrix $K$ of the camera displaying RGB images in the binocular depth camera intelRealsense D435i can be derived as in Equations (2), where the focal length $(f_x, f_y)$ and main spot $(u_0, v_0)$ can be derived as in Equations (3) and (4):

$$
K = \begin{bmatrix}
649.58 & 924.13 \\
0 & 355.84 \\
0 & 0 & 1
\end{bmatrix}
$$
(3) Obtaining the three-dimensional coordinates of the fruit through coordinate transformation. A linear relationship between the fruit's three-dimensional space coordinates and the image coordinates based on the principle of binocular vision triangulation. If the fruit pixel coordinate system coordinates in the RGB image is \( p(x, y) \), and the depth coordinates \( z \) is obtained from the depth image, the fruit's spatial coordinates \( p'(x', y', z') \) in the camera coordinate system can be calculated as in Equation (5):

\[
\begin{align*}
    x' &= \frac{z}{f_x} (x - u_0) = \frac{z}{649.58} (x - 924.13) \\
    y' &= \frac{z}{f_y} (y - v_0) = \frac{z}{355.84} (x - 924.20) \\
    z' &= z
\end{align*}
\]

Figure 4. Target 3D positioning diagram based on RealSense D435i

3. Image acquisition and model training

3.1. Image acquisition and data set production

As shown in Figure 5, the field positioning system for fruits in this paper is mainly for palm fruits, durians, pineapples, etc., which need to be picked and placed on the roadside and secondly transport. Whereas consider palm fruits and durians are no large-scale planting in china, we used pineapple fruits with similar shapes and collection methods as a research object in this paper. By studying the working environment of picking robots picking roadside fruits in the field, manually setting up a complex picking environment, the image resolution of binocular depth camera Realsense D435i is 1280×720 pixels to collect 500 image samples. Meanwhile, the data set was expanded through geometric transformation methods such as translation, flip, zoom, and cropping, and data enhancement was performed...
by adjusting the brightness, contrast, and Gaussian filtering of the image. Then 1400 images were screened to improve the generalization ability of the training model, while kept 200 images as the verification set and configure the training data set and test set for the rest of the images according to 3:1. Finally, the data set was labeled to rectangles for the fruits in the image by labelImg tool. As shown in Figure 6, the labeling information is stored in the format of the PASCAL VOC data set. Each set of data contains the category sequence of the target fruit, the abscissa and ordinate of the center point of the target fruit, and the width and height.

![Figure 5](image1.png)

Figure 5. Sample labeling of the data set: (a) Palm fruit, (b) Durian, (c) Pineapple

![Figure 6](image2.png)

Figure 6. Fruits to be picked and identified in the field by the system

3.2. Model training and evaluation indicators

The total loss used in the training of the improved SSD model include positioning loss (bounding box regression loss) and confidence loss (target classification loss), which can be derived as in Equations (6). Meanwhile, the accuracy of recognition used to evaluate the performance of the model in detecting single-category fruits to be picked in the field, which can be derived as in Equations (7):

$$L(\chi, c, l, g) = \frac{1}{N} \left( L_{\text{conf}}(\chi, c) + \alpha L_{\text{loc}}(\chi, l, g) \right)$$  

$$P = \frac{TP}{TP + FP}$$

where N represents the number of default matching boxes, if N=0, set the loss to 0. $\chi$ represents whether the prediction box matches the real label box (1 if it matches, 0 if it does not match). c represents the confidence of the softmax function for each category. l represents the predicted bounding box, g represents the real label box, $\alpha$ represents the weighting coefficient of confidence loss and positioning loss. True positive (TP) represents a positive sample with the number of fruits correctly identified by the model, and false positive (FP) represents a negative sample with the number of fruits identified by the model in interference environment.

Finally, the experiments were conducted on a computer that has Intel(R) Xeon(R) CPU E5-2673v3, 64-bit 2.40 GHz, and a NVIDIA GeForce GTX 1080Ti GPU. Figure 7 showed the change curve of the average loss value with 2000 iterations during the training process, while the batch stochastic gradient descent algorithm was selected under the TensorFlow framework, the batch size was 16, the weight decay was 0.0005, the momentum was 0.9, the initial learning rate was 0.001, and the
non-maximum suppression was 0.5. It could be seen from the figure that when the number of iterations is about 1750, the loss value decreases to close to 1.1 and tends to be stable, the last training model was saved and used as the target detection model for the fruit.

![Figure 7. Curve diagram of loss value and epochs](image)

4. System experiment results and analysis

4.1. System interface test

In this paper, the main operating interface of the field positioning system for fruits based on binocular vision and SSD was shown in Figure 8, which detection model was available for the recognition accuracy of the fruits offline or online. And the emergency shutdown system button was designed to prevent failure. Determine the system crash problem and display recognized fruit space coordinates on the system interface in real time, which could lay the technical foundation for the next deployment of agricultural robots to achieve autonomous picking, collection and transfer of fruits.

![Figure 8. The interface of the pineapple field positioning system](image)

4.2. Experiment results and analysis

The system has been tested and found that in the actual field identification and positioning of the fruits to be picked. It was easy to recognize when the image of the fruits was clear and contained one or a small amount in the image, but there were differences in the recognition rate of the target detection model for the fruits when the fruits under dense condition fruits, weeds covering fruits, and overlap-
ping fruits. Therefore, the three scenes of dense, weed occlusion, and overlapping fruits were set, and recognition results of the single-stage target detection algorithm such as RetinaNet, YOLOv3, SSD and the improved algorithm of this paper were shown in Figure 9. It could be seen that these models can basically recognize weed occlusion and overlapping fruits, whereas a big difference in the accuracy of fruit recognition in dense scenes would be the focus of this paper in the future.

Finally, this paper took the pineapple fruit as an example of the fruit to be picked by an agricultural robot. The accuracy of the pineapple fruit images in the three scenes in the test set was calculated according to Equation (7) to evaluate the performance of the model, and the average value was repeatedly taken to reduce the result error. The specific experimental results were shown in Table 1. It could be seen that although the recognition rate of the improved algorithm in this paper was not very high, the improved lightweight network model in this paper could partially improve the real-time performance of fruit detection, and it could recognize weeds and overlapping fruit scenes. The accuracy rate reached 99.0% and 90.0%, and the recognition rate under single-pair dense conditions is not high. This would be the key research optimization direction of the next step of the model in this paper. And the model size was greatly reduced, it was easier to move the fruit picking robot. End-to-end deployment, overall the research in this article has certain research and engineering application value.

Figure 9. The order from left to right is the images of three scenes: dense, weeds, and overlapping fruits. And the recognition effect of four algorithms on fruit picking: (a). RetinaNet, (b). YOLOv3, (c). SSD, (d). Algorithm in this paper
Table 1: The results of four algorithms on the fruit of pick-up pineapple in different scenarios

| Algorithm        | Fruits in dense scenes | Fruit covered by weeds | Overlapping fruits | Model size (MB) | FPS | Average detection time of a single image (s) |
|------------------|------------------------|------------------------|--------------------|-----------------|----|--------------------------------------------|
| RetinaNet        | 49.7                   | 92.0                   | 96.0               | 139             | 10.37 | 0.212                                      |
| YOLOv3           | 66.2                   | 99.0                   | 79.5               | 235             | 15.94 | 0.116                                      |
| SSD              | 78.7                   | 99.0                   | 88.0               | 90.7            | 14.91 | 0.133                                      |
| This paper       | 50.5                   | 99.0                   | 90.0               | 24.2            | 16.74 | 0.128                                      |

5. Conclusions

This study proposed the use of a field fruit spatial positioning system based on binocular vision and MobileNet-SSD model in a natural field environment, and the main conclusion were as follows.

(1) In order to solve the problem of insufficient robustness and real-time performance of agricultural picking robots in fruit recognition under complex field environments. The deeply separable lightweight network MobileNet was proposed to replace the VGG network in SSD target detection model, which the fruits to be picked in the weed occlusion and overlapping scenes could be effectively recognized. Whereas the accuracy of fruit recognition in dense fruit scenes is not high, which would be the key research direction of this paper in the next step.

(2) Using the binocular vision camera RealSense D435i to collect images and locate the spatial coordinates of the fruits to be picked, which avoided the cumbersome steps of camera calibration, image segmentation, stereo matching and three-dimensional distance measurement of traditional binocular cameras. Meanwhile, it was easier to adapt to light intensity and the delay of the mutual transmission of binocular camera images was lower.

(3) In order to verify the feasibility of the improved target detection network proposed in this paper, the accuracy and real-time performance of fruit recognition in a complex field environment as indicators were used. The improved lightweight network model in this paper greatly reduced the model size, kept overall recognition accuracy, and improved the real-time performance while compared it with RetinaNet, Faster-RCNN, YOLOv3, and SSD models. Meanwhile, it was easier to deploy in the mobile terminal of the fruit picking robot. The research in this paper had certain engineering practical applications.

Author Contributions

X.Z., Q.G., D.P., and W.Z. conceptualized and designed the experiments. P.C., D.H., and J.Z. performed the experiments. X.Z., Q.G., and D.P. analyzed the data and prepared the manuscript. P.C., D.H., and J.Z. prepared the experimental materials and contributed to the data analysis. All authors have read and agreed to the published version of the manuscript.

Funding

The authors acknowledge support from the Project of National Natural Science Foundation of China (No.51805312); Guangxi University of Science and Technology Graduate Education Innovation Project (No.GKYC202004)

Conflicts of Interest: The authors declare no conflicts of interest.
Abbreviations
The following abbreviations are used in this manuscript:
SVM Support Vector Machine
LDA Linear Discriminant Analysis
YIQ Luminance In-phase Quadrature-phase
DCNN Deep Convolutional Neural Network
YOLOv3 You Only Look Once version 3
SSD Single Shot MultiBox Detector
Faster-RCNN Faster Regions with CNN features
TP True Positive
FP False Positive

References
[1] Pathan Misbah, Patel Nivedita, Yagnik Hiteshri, et al. (2020) Artificial cognition for applications in smart agriculture: Artificial Intelligence in Agriculture, 4:81-95.
[2] Wu Gang, Zhu Qibing, Huang Min, et al. (2019) Automatic recognition of juicy peaches on trees based on 3D contour features and colour data. Biosystems Engineering, 188:1-13.
[3] Sun Sashuang, Jiang Mei, He Dongjian, et al. (2019) Recognition of green apples in an orchard environment by combining the GrabCut model and Ncut algorithm. Biosystems Engineering, 187:201-213.
[4] Fu Lanhui, Duan Jieli, Zou Xiangjun, et al. (2019) Banana detection based on color and texture features in the natural environment. Computers and Electronics in Agriculture, 167:105057.
[5] He Zhi-Liang, Xiong Jun-Tao, Lin Rui, et al. (2017) A method of green litchi recognition in natural environment based on improved LDA classifier. Computers and Electronics in Agriculture, 140:159-167.
[6] Zhao Chuanyuan, Lee Won Suk, He Dongjian. (2016) Immature green citrus detection based on colour feature and sum of absolute transformed difference (SATD) using colour images in the citrus grove. Computers and Electronics in Agriculture, 124:243-253.
[7] Zhao Yuanshen, Gong Liang, Huang Yixiang, et al. (2016) Robust Tomato Recognition for Robotic Harvesting Using Feature Images Fusion. Sensors, 16(2):173.
[8] Liu Guoxu, Nouaze Joseph Christian, Touko Mboembe Philippe Lyonel, et al. (2020) YOLO-Tomato: A Robust Algorithm for Tomato Detection Based on YOLOv3. Sensors, 20(7):2145.
[9] Liang Cuixiao, Xiong Juntao, Zheng Zhenhui, et al. (2020) A visual detection method for nighttime litchi fruits and fruiting stems. Computers and Electronics in Agriculture, 169:105192.
[10] Tian Yunong, Yang Guodong, Wang Zhe, et al. (2019) Apple detection during different growth stages in orchards using the improved YOLO-V3 model. Computers and Electronics in Agriculture, 157:417-426.
[11] Santos Thiago T., de Souza Leonardo L., Dos Santos Andreza A., et al. (2020) Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. Computers and Electronics in Agriculture, 170:105247.
[12] Yu Y.Zhang K.Yang L,et al. (2019) Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. Computers and Electronics in Agriculture. 163:104846.
[13] Villacrèis Juan Fernando, Auat Cheein Fernando. (2020) Detection and Characterization of Cherries: A Deep Learning Usability Case Study in Chile. Agronomy, 10(6):835.
[14] Song Zhenzhen, Fu Longsheng, Wu Jingzhu, et al. (2019) Kiwifruit detection in field images using Faster R-CNN with VGG16. IFAC-PapersOnLine, 52(30):76-81.
[15] Li Shanjun, Hu Dingyi, Gao Shumin, et al. (2019) Real-time classification and detection of citrus based on improved single short multibox detector. Transactions of the Chinese Society of Agricultural Engineering, 35(24):307 – 313. (in Chinese with English abstract) https://doi.org/10.11975/j.issn.1002-6819.2019.24.036
[16] Peng Hongxing, Huang Bo, Shao Yuanyuan., et al. (2018) General improved SSD model for picking object recognition of multiple fruits in natural environment. Transactions of the Chinese Society of Agricultural Engineering 34(16): 155 – 162. (in Chinese with English abstract) https://doi.org/10.11975/j.issn.1002-6819.2018.16.020

[17] GIRSHICKR, DONAHUE J, DARRELL T, et al, (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 580 – 587

[18] GIRSHICK R, (2015) Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision, 1440 –1448.

[19] REN S, HE K, GIRSHICK R, et al. (2015) Faster R-CNN: towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 91 –99.

[20] HE K, GKIOXARI G, DOLLAR P, et al. (2017) Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 99:2961 –2969.

[21] LIU W, ANGUELOV D, ERHAN D, et al. (2016) SSD: single shot multibox detector. European Conference on Computer Vision, 21 –37.

[22] LIN T Y, GOYAL P, GIRSHICK R, et al. (2017) Focal loss for dense object detection. IEEE Transactions on Pattern Analysis & Machine Intelligence, 99:2999 –3007.

[23] Redmon J, Divvala S, Girshick R, et al. (2016) You only look once:Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 779–788.

[24] Rovira-Más Francisco, Wang Qi, Zhang Qin. (2010) Design parameters for adjusting the visual field of binocular stereo cameras. Biosystems Engineering, 105(1):59-70.

[25] Ling Xiao, Zhao Yuanshen, Gong Liang, et al. (2019) Dual-arm cooperation and implementing for robotic harvesting tomato using binocular vision. Robotics and Autonomous Systems, 114:134-143.

[26] Xiang Rong, Jiang Huanyu, Ying Yibin. (2014) Recognition of clustered tomatoes based on binocular stereo vision. Computers and Electronics in Agriculture, 106:75-90.