IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 15, 2019, accepted June 20, 2019, date of publication July 1, 2019, date of current version August 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2925812

Window Zooming–Based Localization Algorithm of Fruit and Vegetable for Harvesting Robot

CHENGLIN WANG[1], TIANHONG LUO[1], LIJUN ZHAO[1], YUNCHAO TANG[2], AND XIANGJUN ZOU[3]

[1]College of Mechanical and Electrical Engineering, Chongqing University of Arts and Sciences, Chongqing 402160, China
[2]College of Urban and Rural Construction, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China
[3]College of Engineering, South China Agricultural University, Guangzhou 510642, China

Corresponding author: Lijun Zhao (20190005@cqwu.edu.cn)

This work was supported in part by a grant from Key Projects of National Key Research and Development Plan under Grant 2018YFB2001400, in part by the Chongqing University of Arts and Sciences Foundation for Major Scientific Research Projects under Grant P2018JD11, and in part by the Chongqing University of Arts and Sciences Introducing Talents Project under Grant R2018SJD17.

ABSTRACT Localization of fruit and vegetable is of great significance to fruit and vegetable harvesting robots and even harvesting industries. However, uncontrollable factors, such as varying illumination, random occlusion, and various surface color and texture, constrain the localization of fruit and vegetable using the vision imaging technology under unconstructed environment. Our previous studies have developed various methods (illumination normalization, features-based classification, etc.) to localize a certain kind of fruit or vegetable using the binocular stereo vision. However, the localization of the multiple fruit and vegetable still faces challenges due to the uncontrollable factors. In order to address this issue, this study proposed an intelligent localization method of targets in fruit and vegetable images acquired by the two charge-coupled device (CCD) color cameras under unstructured environment. The method utilized the Faster region-based convolutional neural network (R-CNN) model to recognize the fruit and vegetable. Based on the recognition results, a window zooming method was proposed for the matching of the recognized target. Finally, the localization of the target was completed by calculating the three-dimensional coordinates of the matched target using the triangular measurement principle. The experimental results showed that the proposed method could be robust against the influences of varying illumination and occlusion, and the average accurate recognition rate was 96.33% under six different conditions. About 93.44% of 1036 pairs of tested targets from unoccluded and partially occluded conditions were successfully matched. Localization errors had no significant difference and they were less than 7.5 mm when the measuring distance was between 300 and 1600 mm under varying illumination and partially occluded conditions.

INDEX TERMS Fruit and vegetable localization, vision imaging technology, binocular stereo vision, unstructured environment.

I. INTRODUCTION

In recent years, numerous kinds of fruit and vegetable harvesting robots have been developed by researchers with an aim of rapid, automatic and effective implementing harvesting mission [1]–[7]. Localization of fruit and vegetable including recognition and matching is the critical step of the automatic-based harvesting [8]. The localization information of fruit and vegetable under the natural environment can be captured by using the vision imaging system of the harvesting robot. Then, the robot can harvest fruit and vegetable by following the guide of the information.

However, the vision imaging system is easily affected by the growth environment of fruit and vegetable such as intensity changes of illumination, random occlusion of surface and so on, which seriously affects the accuracy of localization of fruit and vegetable. Many approaches based on the vision imaging technology have been proposed using

The associate editor coordinating the review of this manuscript and approving it for publication was Nan Liu.

different sensors and algorithms for the recognition and matching of fruit and vegetable. However, there has not been an ideal method yet. Xu et al. [9] proposed a method to detect citrus using infrared thermal imaging. Jiménez et al. [10] developed a robust system to detect orange fruits on an artificial orange tree using an infrared laser. A rate of 80-90% of detection was reported under varying illumination, shadows and background objects conditions. In [11], a hyperspectral camera of 369–1042 nm was used to capture green citrus images. 70–85% of the fruit were correctly identified using a developed image processing method under partial occlusion and highly contrasted fruits condition. In [12], hyperspectral data was analyzed by using a proposed algorithm for apple recognition. 88.1% of the recognition accuracy of apple fruit was reported. The above methods adopting advanced vision sensors to detect fruit achieved high correct detection rates. However, due to time-consuming of data acquisition, high-consuming of devices and high-complexity of image analysis, these hyperspectral and thermal image analysis-based methods would be limited for developing a fruit harvesting robot [13].

The analysis of fruit color images acquired using a red–green–blue (RGB) color camera has advantages of intact information reservation of fruit surface and cheap devices. Researchers have developed numerous RGB color image analysis-based detection algorithms of fruit. In these approaches, the algorithms developed by using a charge-coupled device (CCD) camera have become research hotspots. In [14], three charge-coupled device (CCD) cameras were used to capture the strawberry images. The stereo depth of the object from the camera could be calculated by the two cameras. Another camera was used to detect the target strawberry. 76.55% of strawberries in a real greenhouse could be successfully harvested based on the proposed method. Linker et al. [15] developed an apple detection algorithm, which used four steps to determine the number of green apples in the orchard. Wang et al. [16] proposed an RGB model for cotton recognition using the R-B feature. Similarly, Hanan et al. [17] used the R/(R+G+B) feature of the orange color image to recognize orange fruit. Thus, orange fruits on the tree could be picked using the harvesting robot cooperating the developed vision algorithm. However, varying illumination under natural environment seriously affects the fruit recognition accuracy obtained by using the above methods. In order to reduce the impact of lighting conditions, Payne et al. [18] used a method combining multiple color spaces analysis to detect mango fruit. However, fruit recognition results were not still ideal under varying illumination conditions. A method detecting pomegranate fruits on trees was proposed by Akin et al. [19]. However, the accuracy of detection rate was still affected by lighting and occlusion.

At the aspect of the matching of fruit and vegetable, various methods have been proposed based on the binocular color images. Song et al. [20] used SIFT (Scale Invariant Feature Transform) characteristics as a feature for fruit matching and applied Euclidean distance as the similarity metrics of image matching. They claimed the algorithm can effectively solve the problem of fruits recognition and matching with translation, rotation, scaling, and partial occlusion. Yao et al. [21] developed an image feature extraction and matching method based on the SIFT algorithm, which acquired image feature by using the multi-view of fruit. The results showed that their algorithm presented a better detection result by comparing with speeded-up robust features (SURF) algorithm. More methods for matching fruit and vegetables can be found in [22]–[24], in which the general idea of the methods is to find the matching point based on the most similar feature region in two images. However, these methods have not been validated for matching fruit or vegetable under the unstructured environment.

Especially, the research group of the authors of this paper has developed several algorithms for detecting and matching fruits or vegetables with the aim of reducing the impact of varying illumination and partially occlusion on CCD color camera. Wang et al. [25] segmented litchi illumination-normalized color image using color-based K-means algorithm. Thus, the mature litchi fruits could be obtained and the highest average recognition rate for unoccluded and partially occluded litchi was 98.8% and 97.5%, respectively. Luo et al. [26] used an AdaBoost-based classifier to recognize grape fruits under unstructured environment. The classification accuracy reached up to 96.56%. Xiong et al. [27] proposed a color model for night-time green grape detection through analysis of color features of grape images under daytime natural light and night-time artificial lighting. The accuracy of green grape fruit detection was 91.67%. Wang et al. [13] used four basic classifiers to detect clustered mature litchi fruits. The highest and lowest average recognition rates were 94.17% and 92.00% under sunny back-lighting and partial occlusion, and sunny front-lighting and non-occlusion conditions, respectively. The corresponding fruit harvesting robot could be developed by combining the above methods with the manipulator and the binocular vision system reported in [28]. However, although the manipulator could adjust to multiple fruit and vegetable harvesting, the algorithms only could meet the recognition of a certain kind of fruit or vegetable. Meanwhile, the recognition accuracy still was not ideal under the varying illumination, random occlusion, and various fruit surface color and texture.

Recently, deep learning technology has been applied in many different research fields. Especially in object detection, some excellent artificial neutral networks contributed ideal detection results such as a convolutional neural network (CNN), region-based convolutional neural network (R-CNN), fast region-based convolutional neural network (Fast R-CNN), faster region-based convolutional neural network (Faster R-CNN) and their improvement networks and so on. Chen et al. [29] developed a method to detect vehicles in satellite images using deep convolutional neural networks. Park et al. [30] reported an approach for wild image object detection using pretrained convolutional neural network. The experiment showed their proposed method

could obtain good detection results. In [31], traffic signs were automatically detected by using a CNN-based method. Although there was much noise in the experiment, the classification accuracy was able to achieve to 99.55%. In [32], [33], the methods based on Faster R-CNN and CNN were applied in the detection of human pose acquired using a CCD color camera or video. The detection results reported by these papers were satisfactory, especially short in time-consuming.

From the above, we see that CCD color camera-based methods were suitable for the fruit or vegetable recognition under natural environment. However, most developed methods did not give ideal detection results due to varying illumination and random occlusion. Moreover, the detection time-consuming of the developed method was still a factor to constrain in-real time detection. Inspired by object detection based on deep learning neural networks and the fruit or vegetable recognition methods, this paper mainly focused on the combination of the two methods to find a robust and rapid method for recognition and matching of fruit and vegetable. In order to eliminate the influences caused by unstructured environments and improve the detection rate of fruits, we proposed a method for the recognition of multiple fruit and vegetable by using a Faster R-CNN. In order to match the recognized fruit or vegetable, a window zooming-based matching method was proposed. Compared to previous works, this study is novel in the following two ways:

1) The localization method for multiple fruit and vegetable was developed by only using a CCD color camera-based vision imaging system.
2) Window zooming-based method was proposed for matching the recognized fruit and vegetables under the unstructured environment.

II. MATERIALS AND METHODS

A. AN OVERVIEW OF THE PROPOSED APPROACH

As shown in Fig. 1, the proposed localization flowchart of fruit and vegetable consists of capturing images, resizing the images, the target recognition, and the target matching. The image of fruit or vegetable firstly was acquired by a binocular stereo vision system. Then, the left and right images were both resized into 224 × 224 pixels image by using a bilinear interpolation method for computational efficiency. The resized image is used as the inputting image of the Faster R-CNN for the recognition of fruit or vegetable. The recognized fruit or vegetable was used for matching by using the proposed window zooming method. Last, the fruit or vegetable was localized by calculating its special coordinates. Faster R-CNN-based recognition and window zooming-based matching of fruit or vegetable will perform in the following Sections in detail.

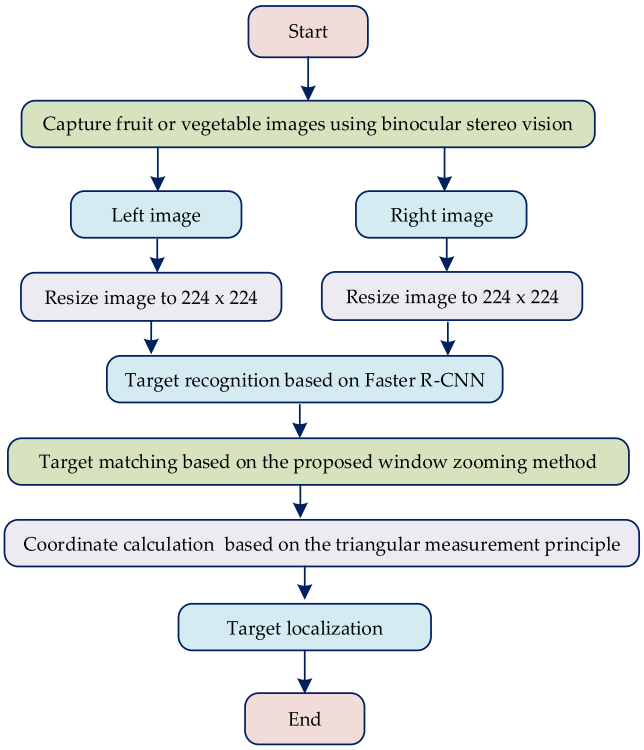


FIGURE 1. Flowchart of the proposed method.

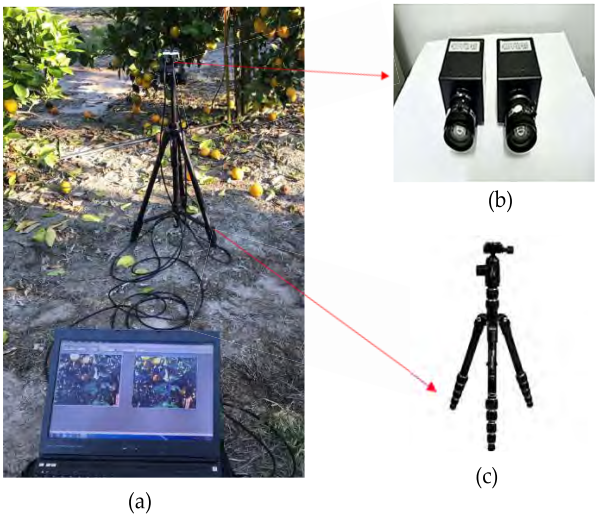


FIGURE 2. The experiment system figure: (a) image processing system; (b) Two CCD color cameras; (c) a tripod.

B. EXPERIMENT MATERIALS AND EQUIPMENT

As shown in Fig. 2, the experiment system of fruit or vegetable localization mainly consisted of an image processing system, two CCD color cameras, and a tripod. The cameras (model MV-VD120SC, supplied by Microvision company in Xi'an, China), as shown in Fig. 2b, which integrated CCD vision sensors, had a digital video output of 1280 by 960 effective pixels, and were parallel mounted on the tripod as shown in Fig. 2c. The distance between the centers of the two camera lenses was fixed at 200 mm in this study. The focal length of the cameras was selected as 6 mm. The two cables have two USB2.0 interfaces, which can be used for the two cameras connecting to the two interfaces of A personal computer (PC), respectively. The PC as shown in Fig. 2a

was the hardware platform of the image processing system, which had 8 GB RAM, an Intel Core i5-4590 CPU, and a Windows 7 operating system. The software system of the image processing system running in the PC was OpenCV 3.0, TensorFlow, and Matlab 8.3. OpenCV 3.0 (supplied by Intel Corporation at Santa Clara, CA, USA) was used for cameras calibration. The camera calibration process in detail can be obtained in our previous study reported in [13]. TensorFlow 1.12.0 was used for Faster RCNN-based recognition of fruit and vegetable with configuration environment including IDE Pycharm 2017, Anaconda 3, GPU 1.4.0, CUDA 8.0, and cuDNN 6.0. And window zooming-based matching method was developed in Matlab 8.3 (supplied by MathWorks Corporation at Nedik, MA, USA). Color images of citrus, litchi, pepper, and eggplant were captured using the above binocular vision system. Image acquisition time was October 2017 for citrus, June 2016 for litchi, September 2018 for pepper and eggplant, respectively.

C. FEATURE EXTRACTION BASED ON VGG16

CNN has excellent advantages on the feature extraction of images. Due to the weight sharing network structure, CNN reduces the complexity of the network model and the numbers of weight, which is similar to a biological neural network. Compared with LeNet, AlexNet, and ZFNet, the structure of VGG16 is deeper and can extract features better. VGG structure has five convolutions, and each convolution is followed by the maximum pooling layer. The depth of the configuration increases from the left to the right. The input of the network is 224 × 224 image, and the output is the result of image classification. A convolution core of 3 × 3 sizes is used to capture changes in horizontal, vertical and diagonal pixels of an inputting image. Three layers of full connection (FC) layer, which is FC-4096, FC-4096, and FC-1000, are used in the last layers of VGG network, and a soft-max layer is used as the end of VGG network. We chose VGG16 and resized the acquired fruit or vegetable images into 224 × 224 pixels as the input of the VGG16 network for the feature extraction.

D. REGION PROPOSAL NETWORK (RPN)

Faster R-CNN-based object detection mainly applies principles based on the Region Proposal Network (RPN) and the Fast R-CNN. The RPN is a Fully Convolutional Network (FCN), which is used for producing regional suggestion boxes of high quality. The structure of RPN is shown in Fig. 4, where 512 features can be obtained by sliding a 3 × 3 convolution kernel on a convolution feature map. The feature map is just the result obtained by using VGG16 stated in section 2.3. Then, the 512 features are delivered into two parallel full link layers, which are the box-classification layer (cls) and the box-regression layer (reg). Thus, the classification information and the localization information can be obtained. There are k anchor points at the center of every sliding window. As shown in Fig. 3, every anchor point corresponds to one box that has one certain size and ratio of length to width. The RPN uses three kinds of sizes and ratios of length to

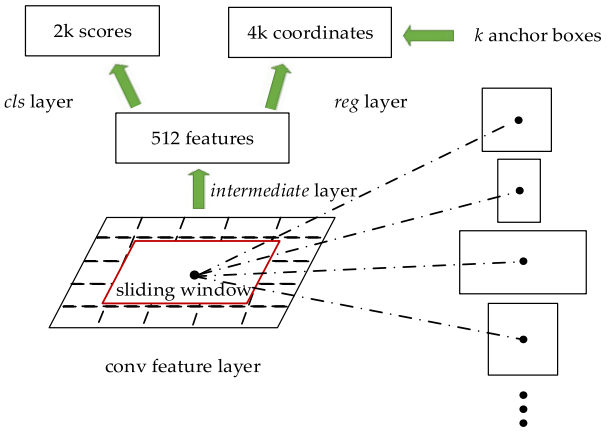


FIGURE 3. The network structure of RPN.

width, thus there 9 anchor points at the center of every sliding window. Therefore, there are 9 region proposals at every sliding window position, there are 4 × 9 outputs representing coordinate coding of the 9 region proposals boxes in the reg, and there are 2 × 9 outputs representing the probability of every region proposal box being object or non-object in the cls.

E. OBJECT RECOGNITION BASED ON FASTER R-CNN

After producing proposal regions by using the RPN, the Faster R-CNN-based detection is implemented by using the Fast R-CNN. The Fast R-CNN shares the convolution features extracted by the VGG16 with the RPN. The detail training procedure is as follows.

- Initialization of the RPN training is implemented by using a model M_0obtained by using ImageNet. Then, a model $M_1$is obtained after finishing the RPN training. A proposal region $P_1$is produced by using the model M_1.
- A model M_2 is obtained by using the Fast R-CNN training by using the proposal region P_1producing from the RPN of Step 1. And the Fast R-CNN model is initialized by using M_0.
- A model M_3 is obtained by initializing the RPN training using the model M_2. Then a proposal region P_2 is produced after finishing the RPN training.
- A final model M_4 is obtained by using the model M_3 and P_2 to train the Fast R-CNN.

F. TARGET LOCALIZATION BASED ON WINDOW ZOOMING

After Faster RCNN-based recognition of fruit and vegetable, the recognized targets from the left and right images needed to be matched for localization. However, the position of the Faster RCNN output box is the target approximate position. As can be seen in Fig. 4a, the output box is not tangent to the target contour, which will affect matching if using the output box as the matching label. Therefore, in this study, a window zooming algorithm was proposed for matching the recognized targets. Before matching the targets, we changed the original code of the Faster RCNN for deleting the probability

(a) (b)

FIGURE 4. The recognized targets for matching: (a) output result of Faster RCNN; (b) Simplified output result.

values of the outputting results and removing non-targets to reduce interference for matching as shown in Fig. 4b. Then we used the minimum region proposal box as a basic window and recorded the whole pixels of the window. Move the window to the interior of other boxes so that the geometric center of the window coincides with the geometric center of the box. Calculate the ratio R_1 of the target pixel to the whole pixels of the basis window and the ratio R_2 of the non-target pixel to the whole pixels of the basic window, respectively. If R_1 is greater than R_2, the basis window is considered as the matching label. And if R_1 is smaller than R_2, the basis window will be enlarged according to the relationship between height and width of the external box. When the height is greater than the width, the basis window will be enlarged in the vertical direction. The enlargement will stop until the target pixel is more than the non-target pixel in the basis window. At this moment, the enlarged basis window is considered as the matching label. If there are no more target pixels than non-target pixels in the process of window enlargement, the enlargement process will stop when it coincides with the outer box edge length. The matching label will be the outer box. Similarly, when the width is greater than the height, the basis window will be enlarged in the horizontal direction. The enlargement will stop until the target pixel is more than the non-target pixel in the basis window. And the enlarged basis window is considered as the matching label. If there are no more target pixels than non-target pixels in the process of window enlargement, the enlargement process will stop when it coincides with the outer box edge length. The outer box will be considered as the matching label.

Assumptions and Terminologies:

BasW: Basic Window, which is the minimum region proposal box.

BasWP: Pixels of Basic Window, which is the whole pixels of the basic window.

OthW: Other Window, which is another window that is another region proposal box except for the basic window.

GeoC: Geometric center, that is the geometric center of the region proposal box.

TarP: Target Pixels, that are the pixels of the fruit or vegetable included in the window.

Non-TarP: Non-Target Pixels, that are the pixels of the non-fruit or non-vegetable included in the window.

WinH: Window Height, which is the height of the window.

WinW: Window Width, which is the width of the window.

N: N is the total number of the region proposal boxes.

ML: Matching label, which is defined as the matching label of fruit or vegetable for matching the recognized targets in both right and left images.

Algorithm 1 Window Zooming Algorithm

```
Input: BasW, BasWP, GeoC of BasW
fori=1; i ≤ N; i++
  GeoC of BasW = GeoC of OthW
 R1 = TarP/ BasWP
 R2= Non-TarP/ BasWP
 ifR1>R2 then
   ML = BasW
 else
   if WinH of BasW > WinW of BasW then
     While( WinH of BasW < WinH of OthW)
       WinH = WinH + 1
       ifR1>R2 then
         ML = BasW
         break
   else
     While( WinW of BasW < WinW of OthW)
       WinH = WinH + 1
       ifR1>R2 then
         ML = BasW
         break
   end if
 end if
end for
```

When the matching is implemented using the above label method, some matching constraints need to be stated. For example, by the ordering constraint, if fruit is on the left side of another in the left image, it should also be on the left side of it in the right image. If there were no fruits on the left side of the fruit in the right image, the corresponding position of the fruit could not be determined. Under this condition, this kind of fruit was ignored. If some portion of a fruit needed to be matched was in the left image, and the whole part or some part of the fruit was in the right image, the preset threshold constrained the correct matching. The epipolar constraint stated that for the mapping point on an image, its match point must fall on another image of epipolar online. In a binocular stereo vision system, this epipolar constraint may hypothesis that the matching points were in the same horizontal line, that was values of coordinates of the matching points in the two images were equal. So, matching in the different rows should be discarded. Furthermore, the length between the centers of the two camera lens, defined as a baseline, should be adjusted to a suitable distance. By triangulation, if the baseline is shorter, both the position of litchi in left and right image will be more consistent, which makes the matching of features be easy. The shorter baseline will also make distance

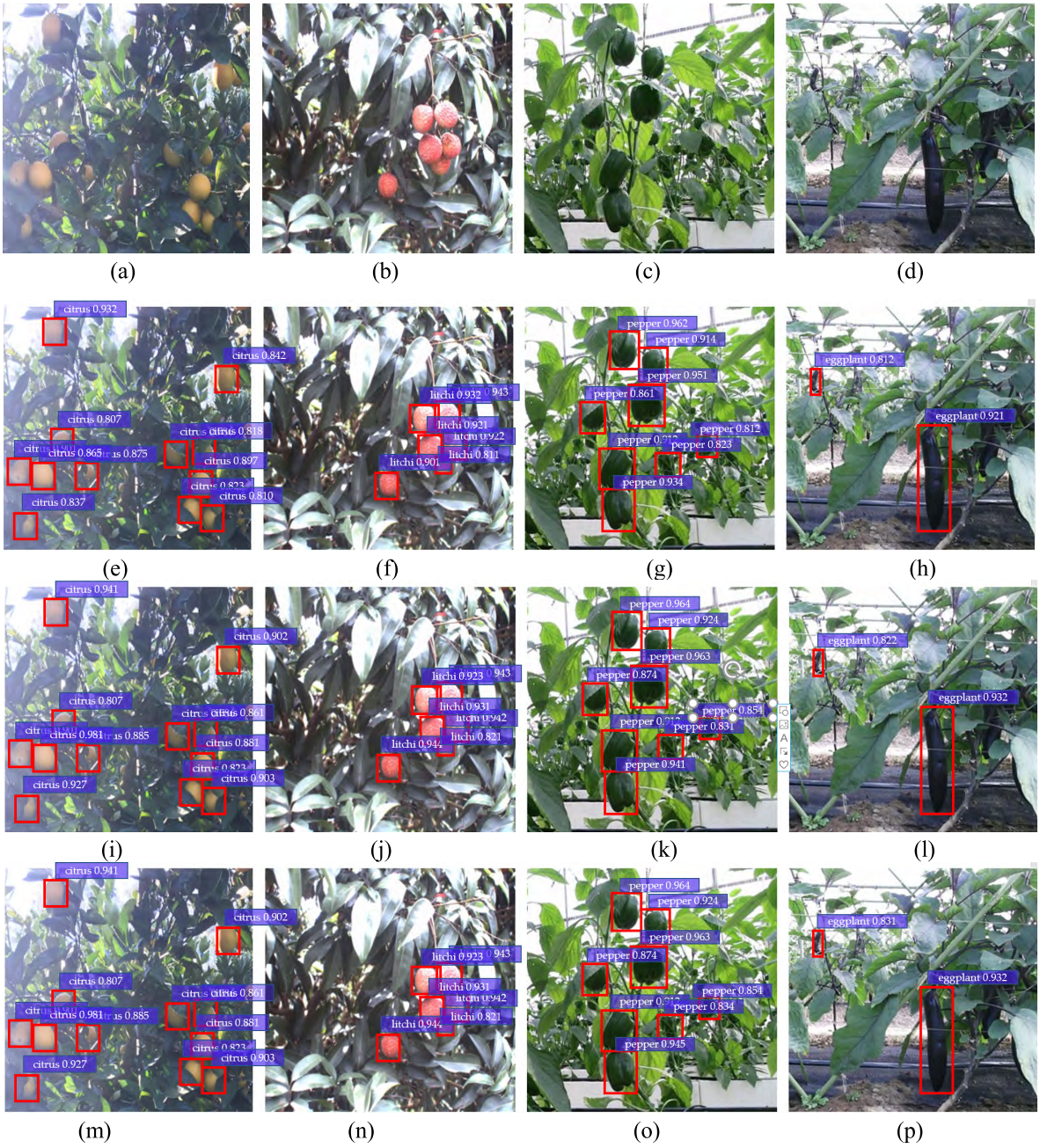


FIGURE 5. The recognition results based on Faster RCNN. (a)-(d) Original fruit and vegetable color images; (e)-(h) Results after 3000. iterations; (i)-(l) Results after 4000 iterations; (m)-(p) Results after 5000 iterations.

measurement errors become larger. In this study, the baseline was fixed at 200 mm.

Therefore, the matching label of the left image can be considered as the template and it is traversed into the right image along the epipolar line to find the most similar window, the diagonal intersection point of the found window is the matching point of a fruit feature point of the left image. The similarity measure method was selected the normalized cross-correlation (NCC) based on the grey value matching given in equation (1), as shown at the bottom of the next page, which was invariant to all linear illumination changes, and hence was suitable for unstructured environments.

Here, (u, v) was the diagonal intersection point coordinates of fruit label template in the left image, the size of the template was $M \times N$. $I_1(u+i, v+j)$ was the grey value of the point $(u+i, v+j)$, and $I_1(u, v)$ was the average value of grey values of label template. The coordinates $(u+ i\text{-}d, v + j)$ was the translation result of the coordinates $(u+i, v+j)$ in right image, and similarly, $I_2(u+ i\text{-}d, v+j)$ was the grey value of the point $(u+ i\text{-}d, v+j)$. $I_2(u\text{-}d, v)$ was the average value of grey values of an $M \times N$ size window that its diagonal intersection point coordinates were $(u\text{-}d, v)$. By solving equation (1), obtain the solution d which makes $NCC(d)$ reach the maximum value. Thus, the disparity in d was calculated, the matching point

coordinates of (u, v) was obtained and the coordinates $O(x_1, y_1, z_1)$ of fruit feature point were determined by equation (2).

$$\begin{cases} x_1 = \dfrac{b(u_1 - u_0)}{(u_1 - u_2)} \\ y_1 = \dfrac{ba_x(v_1 - v_0)}{a_y(u_1 - u_2)} \\ z_1 = \dfrac{ba_x}{(u_1 - u_2)} \end{cases} \tag{2}$$

where u_0, v_0, a_x, a_y are the camera internal parameters and they were obtained by cameras calibration, b is the value of baseline, 200mm. Thus, if $u_1 - u_2$ is determined, can be calculated. The value of $u_1 - u_2$ is defined as disparity d and determined by the matching procedure.

III. RESULTS

A. RECOGNITION OF FRUIT AND VEGETABLE UNDER NATURAL ENVIRONMENT CONDITIONS

With the aim of illustrating the performance of the proposed recognition method, original color images of different kinds of fruit and vegetable were captured under natural environment were shown in Fig. 6a–d, respectively. The results were seen in Fig. 6e-p using the Faster RCNN-based method, respectively, where Fig. 6e-h showed the recognition results after 3000 iterations, Fig. 6i-l was the recognition results after 4000 iterations, and Fig. 6m-p showed the recognition results after 5000 iterations, respectively.

The above results indicated that Faster RCNN performed well in fruit and vegetable recognition. And with iteration number increasing from 3000 to 4000, recognition probability of fruit and vegetable increased comparing Fig. 6e-h and Fig. 6i-l. While the recognition results have hardly changed comparing Fig. 6i-l and Fig. 6m-p, which also could be confirmed in Fig. 7. As seen from Fig. 7, the loss curve fluctuated within a fixed range and remains at a fixed value, which implied the convergence of the Faster RCNN model after 4000 iterations.

Since the aim of this paper is to test the algorithm rather than to target a certain kind of target, four kinds of targets were put together for statistics. True positive rates, false positive rates, false negative rates, precision, recall and F1 of the total four kinds of targets, which were obtained using the Faster RCNN model under the condition that the output probability of Faster RCNN was set to 0.8, were recorded in Table 1 compared with the corresponding values obtained by our previous method stated in [13], [25]. The true positives rates, which were the rates of correct fruit and vegetable recognition, were 96.33% and 92.66% obtained by using the proposed method and our previous method, respectively.

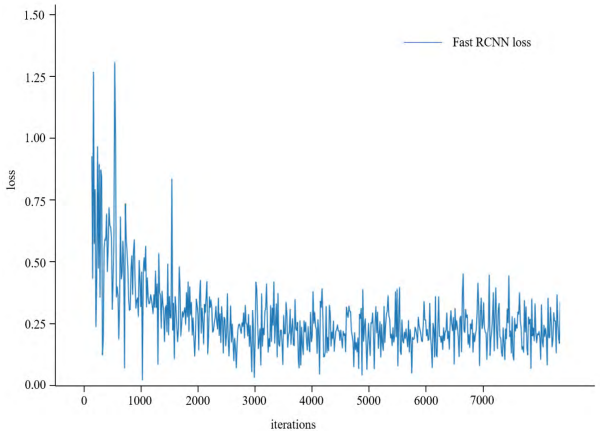


FIGURE 6. Loss curve with the number of iterations.

(a) (b)

FIGURE 7. The matching result using our previous method. (a) left image; (b) right image.

The false positives rates, which were the rates of incorrect identifying non-fruits into fruits, were 1.82% and 11.00% under six different conditions, respectively. The false negatives rates, which were the rates of incorrect identifying fruits into non-fruits, were 3.67% and 7.34% under six different conditions, respectively. The precision, recall, and F1 were 98.15%, 96.33%, 97.23% and 89.39%, 92.66%, 90.99% obtained by using the proposed method and our previous method under six different conditions, respectively. The highest true positives rates were 96.69% and 94.17%, achieved under cloudy day and partial occlusion conditions and sunny back-lighting and partial occlusion conditions, respectively. The lowest true positives rates were 94.56% and 91.07% under sunny front-lighting and non-occlusion conditions and sunny front-lighting and partial occlusion conditions. The proposed algorithm received the lowest false positive rate, 1.04%, under sunny back-lighting and non-occlusion conditions, which was much lower than that 6.67% obtained by our previous method. However, the highest false positive rate was 2.18%, happened under sunny front-lighting and partial occlusion conditions, which was still much lower than that 28.13% obtained by the previous method under sunny

$$NCC(d) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}[I_1(u+i, v+j) - \bar{I}_1(u, v)][I_2(u+i-d, v+j) - \bar{I}_2(u-d, v)]}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}[I_1(u+i, v+j) - \bar{I}_1(u, v)]^2}\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}[I_2(u+i-d, v+j) - \bar{I}_2(u-d, v)]^2}} \tag{1}$$

TABLE 1. The recognition results of the proposed method and the previous method.

Illumination Conditions	Target	True Positives Rate %		False positives Rate %		False Negatives Rate %		Precision %		Recall %		F1 %	
		Pro	Pre	Pro	Pre	Pro	Pre	Pro	Pre	Pro	Pre	Pro	Pre
SFP	126	95.68	91.07	2.18	10.53	4.32	8.93	97.77	89.64	95.68	91.07	96.71	90.35
SFN	35	94.56	92.00	2.04	28.13	5.44	8.00	97.89	76.58	94.56	92.00	96.19	83.58
SBP	118	96.87	94.17	1.37	9.49	3.13	5.83	98.61	90.85	96.87	94.17	97.73	92.48
SBN	48	95.98	92.11	1.04	12.50	4.02	7.89	98.93	88.05	95.98	92.11	97.43	90.03
CP	136	96.69	93.58	1.98	9.73	3.31	6.42	97.99	90.58	96.69	93.58	97.34	92.06
CN	55	96.54	93.33	2.15	6.67	3.46	6.67	97.82	93.33	96.54	93.33	97.18	93.33
Total	518	96.33	92.66	1.82	11.00	3.67	7.34	98.15	89.39	96.33	92.66	97.23	90.99

[a] Pro = Proposed method; Pre = Previous method; SFP = Sunny front-lighting and partial occlusion; SFN = Sunny front-lighting and non-occlusion; SBP = Sunny back-lighting and partial occlusion; SBN = Sunny back-lighting and non-occlusion; CP = Cloudy day and partial occlusion; CN = Cloudy day and non-occlusion; True positives rate = Amount of true positives/(amount of true positives +amount of false negatives) × 100%; False positives rate = Amount of false positives/(amount of false positives + amount of true positives) × 100%; False negatives rate = Amount of false negatives/(amount of false negatives + amount of true positives) × 100%; Precision = Amount of true positives/(amount of false positives + amount of true positives) × 100%; Recall = Amount of true positives/(amount of false negatives + amount of true positives) × 100%; F1 = 2 × Precision × Recall/(Precision + Recall) × 100%.

front-lighting and non-occlusion conditions. Both of the lowest false negatives rates were 3.13% and 5.83% under sunny back-lighting and partial occlusion conditions. The highest false negatives rates were 5.44% and 8.93% under sunny front-lighting and non-occlusion conditions and sunny front-lighting and partial occlusions, respectively.

B. MATCHING PERFORMANCE BASED ON WINDOW ZOOMING

The matching result of the experiment based on our previous method from Fig. 7 showed some mistaken matching. There were no matching citrus fruits in the right image for citrus A and B. The reason was that citrus A and B seriously partially occluded each other in the left image, and there were no similar template windows with citrus fruits in the right image after searching on the same rows. By comparison, citrus A and B obtained the correct matching citrus fruits in the right image showed from Fig. 8. It could be seen that the previously proposed template matching method did not perform well under severe occlusion condition. After the matching experiment, the matching success rates of the fruit and vegetable targets under different illumination and occlusion conditions using previous method and the proposed method were recorded in Table 2. 93.44% of the total numbers of the targets were successfully matched under six different conditions by using the proposed method, which was higher than 85.71% obtained by using our previous method.

By using the proposed method, the highest matching success rate was 97.03% under sunny back-lighting and partial occlusion conditions. The lowest matching success rate was 79.69% under sunny front-lighting and non-occlusion conditions. The matching success rates were 94.93%, 88.18%, 96.40% and 81.25% under other conditions. However, by using the template matching method, 85.71% of the total numbers of the targets were successfully matched under six different conditions by using our previous method. The highest matching success rate was 91.75% under sunny back-lighting and partial occlusion conditions. The lowest

(a) (b)

FIGURE 8. The matching result using the proposed method. (a) left image; (b) right image.

TABLE 2. The matching results of the proposed method and the previous method.

Illumination Conditions	Pair	Correct Matching Rate of the Proposed Method		Correct Matching Rate of the Previous Method	
		Amount	%	Amount	%
SFP	217	206	94.93	186	85.71
SFN	64	51	79.69	48	75.00
SBP	303	294	97.03	278	91.75
SBN	110	97	88.18	87	79.09
CP	278	268	96.40	243	87.41
CN	64	52	81.25	46	71.88
Total	1036	968	93.44	888	85.71

[a] SFP = Sunny front–lighting and partial occlusion; SFN = Sunny front–lighting and non–occlusion; SBP = Sunny back–lighting and partial occlusion; SBN = Sunny back–lighting and non–occlusion; CP = Cloudy day and partial occlusion; CN = Cloudy day and non–occlusion.

matching success rate was 71.88% under cloudy day and non-occlusion conditions. The matching success rates were 85.71%, 75.00%, 79.09% and 87.41% under other conditions.

C. LOCALIZATION ERROR OF THE PROPOSED ALGORITHM

Localization experiment was implemented under four conditions. Localization errors of 15 random targets per 100mm from 300mm to 1600mm were calculated with respect to the true value. The average value of the errors at each measurement point all fell within ±15 mm as shown in Fig. 9. Four sets of data were divided into six pairs of data. T-test was

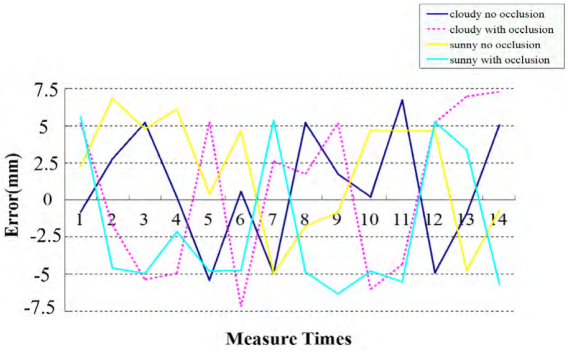


FIGURE 9. Distance measurement average error result.

implemented by using SPSS 19.0 (supplied by SPSS Corporation in Chicago, USA) for analyzing these six pairs of data. The corresponding values of Sig., 0.981, 0.423, 0.218, 0.621, 0.080 and 0.058 were all larger than 0.05, and there was no significant difference for all conditions. Using the distance measurement equation (2), the value of defined as disparity d was determined by the matching procedure. The above result indicated that the proposed algorithm was sufficiently robust under variable illumination and partially occluded conditions.

D. REAL-TIME PERFORMANCE OF THE PROPOSED ALGORITHM

Although the time consuming for Faster RCNN model training was about 5 hours, the time for recognition of fruit and vegetable targets was short by using the model. The targets matching occupied the majority of the whole time consumed by the proposed method. Choosing the minimum region proposal box to move it into other proposal boxes for finding the matching label cost time; then traversing the matching label of the target of the left image into the right image to search for the optimal matching window based on the normalized cross-correlation similarity measure function cost time; the average time consuming for matching was 2042 ms based on the window zooming method. However, the proposed method operating on localization calculation consumed few time.

IV. DISCUSSION

To locate fruit and vegetable under natural environment, the Faster RCNN model firstly was used in the target recognition. By increasing iteration numbers to 4000 and 5000, the recognition results tend to be stable. The fruits and vegetables under different lighting conditions were tested by using the Faster RCNN model. The experimental results demonstrated that the method could well identify the fruit and vegetable from the background (e.g., leaves, branches and sky). The performance tests of the fruit and vegetable recognition method indicated that the method could partly account for the robustness against the influence of the varying illumination and occlusion. Based on the successful fruit and vegetable recognition, a matching method based on window zooming was proposed, which could not only utilize the regular graph advantages of the region proposal box of the Faster RCNN model in matching, but also avoid mistaken matching disadvantages of the target under occlusion conditions. The experimental results implied that the proposed method could improve the accurate matching result than the result obtained by our previous temple matching method. Based on the correct matching, the localization of the target was implemented by using the triangular measurement principle. The localization experiment showed localization errors had no significant difference and they were less than 7.5 mm when the measuring distance was between 300 mm and 1600 mm under varying illumination and partially occluded conditions. From the interactive performance of the developed approach in the tests, the average processing time from the recognition of target to target localization was 2042 ms. Based on the discussions above, it can be seen that the proposed method can successfully and robustly recognize and locate fruit and vegetable under the real environment. However, there were still some shortcomings of the developed method. Although the training samples of the Faster RCNN have included different illumination and occlusion conditions, it could not cover all conditions, which would affect the correction of the recognition. And when the fruit and vegetable targets were severely occluded, some targets may not be correctly matched. Therefore, additional research is still needed to improve the recognition and matching rates for satisfying more varied natural environments.

V. CONCLUSION AND FUTURE WORK

An approach was developed in this paper for localization of fruit and vegetable under varying illumination and random occlusion conditions using vision imaging technology. The approach included two parts, recognition, and matching. In the first part, Faster RCNN was applied in fruit and vegetable recognition. The original color images of fruit and vegetable were converted into the training format of Faster RCNN. Then, the Faster RCNN was trained by using the converted format images. Last, the fruit and vegetable targets could be effectively recognized by using the trained Faster RCNN. In the second part, based on the successful recognition of fruit and vegetable, the matching was completed using the zooming window in the left image to be a template to traverse the right image to search for an optimal window based on the similarity measure method of normalized cross-correlation. The experiments were implemented by using the fruit and vegetable images from the real natural environment, and the results were quantitatively assessed and compared with other corresponding methods. The main results are as follows.

- The proposed method using the Faster RCNN model to identify the fruit and vegetable under the real environment received stable and satisfying recognition results when the iteration number was greater than 4000.
- The recognition method was able to automatically separate fruit and vegetable from the background, and the accuracy rate of the recognition could achieve 96.33% under six different conditions, which was more accurate than the result obtained by our previous method.

- The highest and lowest matching success rates of fruit and vegetable of the proposed method were 97.03% under sunny back-lighting and partial occlusion conditions and 79.69% under sunny front-lighting and non-occlusion conditions, respectively, which were superior to the method based on the template matching.
- Localization errors had no significant difference and they were less than 7.5 mm when the measuring distance was between 300 mm and 1600 mm under varying illumination and partially occluded conditions.
- The interactive performance of the proposed algorithm was investigated, and the average time consuming was about 2042 ms based on the window zooming method.

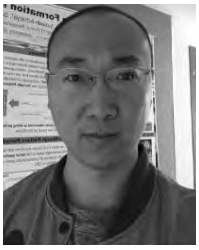
In conclusion, the developed localization approach can effectively recognize and match the fruit and vegetable targets under a complex real natural environment. However, the accurate localization of the fruit and vegetable under sever varying illumination conditions are still issues that need to be solved and will require further research. In future research, we will consider localization methods of fruit and vegetable under a dynamic environment.

NOMENCLATURE

RGB	red, green and blue
CCD	charge-coupled device
CNN	convolutional neural network
R-CNN	region-based convolutional neural network
SIFT	scale invariant feature transform
SURF	speeded-up robust features
NCC	normalized cross-correlation
Fast R-CNN	fast region-based convolutional neural network
Faster R-CNN	faster region-based convolutional neural network

REFERENCES

[1] Z. De-An, L. Jidong, J. Wei, Z. Ying, and C. Yu, "Design and control of an apple harvesting robot," *Biosyst. Eng.*, vol. 110, no. 2, pp. 112–122, Oct. 2011.
[2] X. J. Zou, H. X. Zou, and J. Lu, "Virtual manipulator-based binocular stereo vision positioning system and errors modelling," *Mach. Vis. Appl.*, vol. 23, no. 1, pp. 43–63, Jan. 2012.
[3] C. W. Bac, J. Hemming, and E. J. van Henten, "Robust pixel-based classification of obstacles for robotic harvesting of sweet-pepper," *Comput. Electron. Agricult.*, vol. 96, pp. 148–162, Aug. 2013.
[4] F. Qingchun, C. Wei, Z. Jianjun, and W. Xiu, "Design of structured-light vision system for tomato harvesting robot," *Int. J. Agricult. Biol. Eng.*, vol. 7, no. 2, pp. 19–26, Apr. 2014.
[5] S. S. Mehta and T. F. Burks, "Vision-based control of robotic manipulator for citrus harvesting," *Comput. Electron. Agricult.*, vol. 102, pp. 146–158, Mar. 2014.
[6] F. Dimeas, D. V. Sako, V. C. Moulianitis, and N. A. Aspragathos, "Design and fuzzy control of a robotic gripper for efficient strawberry harvesting," *Robotica*, vol. 33, no. 12, pp. 1085–1098, Jun. 2015.
[7] A. Silwal, J. R. Davidson, M. Karkee, C. Mo, Q. Zhang, and K. Lewis, "Design, integration, and field evaluation of a robotic apple harvester," *J. Field Robot.*, vol. 34, no. 1, pp. 1140–1159, Sep. 2017.
[8] A. Gongal, S. Amatya, M. Karkee, Q. Zhang, and K. Lewis, "Sensors and systems for fruit detection and localization: A review," *Comput. Electron. Agricult.*, vol. 116, pp. 8–19, Aug. 2015.
[9] H. Xu and Y. Ying, "Detecting citrus in a tree canopy using infrared thermal imaging," *Proc. SPIE*, vol. 30, no. 3, pp. 321–327, Mar. 2004.
[10] A. R. Jimenez, R. Ceres, and J. L. Pons, "A survey of computer vision methods for locating fruit on trees," *Trans. ASAE*, vol. 43, no. 8, pp. 1911–1920, Aug. 2000.
[11] H. Okamoto and W. S. Lee, "Green citrus detection using hyperspectral imaging," *Comput. Electron. Agricult.*, vol. 66, no. 11, pp. 201–208, May 2009.
[12] O. Safren, V. Alchanatis, V. Ostrovsky, and O. Levi, "Detection of green apples in hyperspectral images of apple-tree foliage using machine vision," *Trans. ASABE*, vol. 50, no. 6, pp. 2303–2313, Nov. 2007.
[13] C. Wang, Y. Tang, X. Zou, L. Luo, and X. Chen, "Recognition and matching of clustered mature litchi fruits using binocular charge-coupled device (CCD) color cameras," *Sensors*, vol. 17, no. 2, p. 2564, Aug. 2017.
[14] P. Rajendra, N. Kondo, K. Ninomiya, J. Kamata, M. Kurita, T. Shiigi, S. Hayashi, H. Yoshida, and Y. Kohno, "Machine vision algorithm for robots to harvest strawberries in tabletop culture greenhouses," *Eng. Agricult., Environ. Food*, vol. 2, no. 5, pp. 24–30, Dec. 2008.
[15] R. Linker, O. Cohen, and A. Naor, "Determination of the number of green apples in RGB images recorded in orchards," *Comput. Electron. Agricult.*, vol. 81, nos. 5–6, pp. 45–57, Feb. 2012.
[16] Y. Wang, X. Zhu, and C. Ji. "Machine vision based cotton recognition for cotton harvesting robot," *Comput. Technol. Agricult.*, vol. 259, no. 4, pp. 1421–1425, Aug. 2012.
[17] M. W. Hannan, T. F. Burks, and D. M. Bulanon, "A machine vision algorithm combining adaptive segmentation and shape analysis for orange fruit detection," *Agricult. Eng. Int.*, vol. 5, pp. 1–41, Jan. 2010.
[18] W.-G. Song, H.-X. Guo, and Y. Wang, "A method of fruits recognition based on SIFT characteristics matching," in *Proc. Int. Conf. Artif. Intell. Comput. Intell.*, Nov. 2009, pp. 119–122.
[19] L. Yao, G. Zhou, Z. Ni, P. Zhang, and S. Zhu, "Matching method for fruit surface image based on scale invariant feature transform algorithm," *Trans. Chin. Soc. Agricult. Eng.*, vol. 31, no. 9, pp. 161–166, May 2015.
[20] J. Lü, D. A. Zhao, and W. Ji, "Research on matching recognition method of oscillating fruit for apple harvesting robot," *Trans. Chin. Soc. Agricult. Eng.*, vol. 29, no. 20, pp. 32–39, Jan. 2013.
[21] M. Huang, "Apple fruit recognition based on template matching," *Comput. Appl. Softw.*, vol. 5, no. 5, pp. 223–228, Dec. 2010.
[22] H. J. Pan, X. X. Li, G. W. Wang, and C. S. Qi, "The shape of fruit recognition using matching round," *Adv. Mater. Res.*, vol. 645, no. 4, pp. 251–254, Dec. 2013.
[23] S. Oke, M. Ookado, and Y. Nakamura, "Recognition of fruits by image processing-application of template matching," *Control Appl. Post-Harvest Process. Technol.*, vol. 1, pp. 129–134, Jan. 1995.
[24] W.-H. Chang, S. Chen, S.-C. Lin, P.-F. Huang, and Y.-Y. Chen, "Vision based fruit sorting system using measures of fuzziness and degree of matching," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Oct. 1994, pp. 2600–2604.
[25] C. Wang, X. Zou, Y. Tang, L. Luo, and W. Feng, "Localisation of litchi in an unstructured environment using binocular stereo vision," *Biosyst. Eng.*, vol. 145, pp. 39–51, May 2016.
[26] L. Luo, Y. Tang, X. Zou, C. Wang, P. Zhang, and W. Feng, "Robust grape cluster detection in a vineyard by combining the AdaBoost framework and multiple color components," *Sensors*, vol. 16, no. 3, pp. 1–20, Oct. 2016.
[27] J. Xiong, Z. Liu, R. Lin, R. Bu, Z. He, Z. Yang, and C. Liang, "Green grape detection and picking-point calculation in a night-time natural environment using a charge-coupled device (CCD) vision sensor with artificial illumination," *Sensors*, vol. 18, no. 2, p. 969, Dec. 2018.
[28] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, Oct. 2014.
[29] S. Park and Y. S. Moon, "Wild image object detection using a pretrained convolutional neural network," *IEIE Trans. Smart Process. Comput.*, vol. 3, no. 6, pp. 366–371, Dec. 2014.
[30] H. H. Aghdam, E. J. Heravi, and D. Puig, "A practical approach for detection and classification of traffic signs using convolutional neural networks," *Robot. Auton. Syst.*, vol. 84, pp. 97–112, Oct. 2016.
[31] X. Zhao, W. Li, Y. Zhang, T. A. Gulliver, S. Chang, and Z. Feng, "A faster RCNN-based pedestrian detection system," in *Proc. IEEE 84th Veh. Technol. Conf.*, Sep. 2016, pp. 1–5.
[32] K. Wei and X. Zhao, "Multiple-branches faster RCNN for human parts detection and pose estimation," in *Proc. Asian Conf. Comput. Vis.*, Mar. 2017, pp. 453–462.



CHENGLIN WANG received the B.E. degree in mathematics and the M.E. degree in CIS from Qiqihar University, Qiqihar, China, in 2009 and 2013, respectively, and the Ph.D. degree in AME from South China Agricultural University, Guangzhou, China, in 2018. He has a total of ten-year experience in academic profession. His research interests include the computer vision, image processing, and development of harvesting robot.

TIANHONG LUO received the B.E. and M.E. degrees in mechanical design theory and the Ph.D. degree in mechanical design theory and automation from Chongqing University, Chongqing, China, in 1998 and 2002, and 2005, respectively. His research interests include kinematics and system dynamics of robotic mechanisms, and new functional components of robots.

LIJUN ZHAO received the B.E. degree in mechanical design theory and automation and the M.E. and Ph.D. degrees in agricultural mechanization engineering from Northeast Agricultural University, Harbin, China, in 2005, 2013, and 2016, respectively. His research interests include the computer vision, image processing, and agricultural machinery.

YUNCHAO TANG, photograph and biography not available at the time of publication.

XIANGJUN ZOU, photograph and biography not available at the time of publication.