Fruit maturity and location identification of beef tomato using R-CNN and binocular imaging technology

Kuang-Wen Hsieh^{1,2} · Bo-Yu Huang² · Kai-Ze Hsiao² · Yu-Hao Tuan² · Fu-Pang Shih² · Li-Cheng Hsieh² · Suming Chen³ · I-Chang Yang⁴

Received: 6 November 2020 / Accepted: 15 July 2021 / Published online: 2 August 2021
© The Author(s) 2021

Abstract

The objective of this study was to identify the maturity and position of tomatoes in greenhouse. Three parts have been included in this study: building the model of image capturing and object detection, position identification of mature fruits and prediction of the size of the mature fruits. For the first part, image capturing in different time and object detection will be conducted in the greenhouse for identification of mature fruits. For the second part, the relative 3D position of the mature fruits calculated by the binocular vision was compared with the actual measured position. For the third part, the size of the bounding box from the object detection was compared with the actual size of the mature fruit, and the correlation was calculated in order to pre-adjust the width of the gripper for plucking operation in the future. The precision and the recall of the mature fruits of this study are over 95%. The average error of the 3D position is 0.5 cm. The actual size of the fruits and the R-squared of the size of the bounding box are over 0.9.

Keywords Beef tomato · Object detection · Binocular vision · R-CNN · 3D position

Introduction

The investment of protected cultivation of fruits and vegetables is getting relatively high in the agriculture in many countries. The facilities should be able to adjust the environmental variables including light source, temperature and humidity, in order to ensure the quality of the cultivated fruits and vegetables. Therefore, the price and quality of fruits and vegetables from protected cultivation would be higher than those from outdoor, and the incomes of facility growers would also be higher. At present, the insufficient and aging of labor in rural areas are very common. Due to

the existing structure of the large fixed facilities, it could be equipped with different mechanized devices to reduce the repetitive operations and heavy manual labor. This could save labor and improve the comfort of the work.

The image pre-processing for judging ripe fruits during fruit harvesting uses HIS [1, 2] or color balance [3] and other methods to separate the fruit from the background. An edge algorithm is used to interpret the appearance of fruits [4, 5]. The obtained interpretation data is then used in support vector machine (SVM) or artificial neural network (Artificial Neural Network, ANN) for non-linear operations [2, 6–8]. In recent years, many studies focused on the new deep learning algorithm with very good success rate in image recognition and to identify the positioning and maturity of fruits [9–12]. There have been many studies on fruit harvesting robots in the past, using 3D cameras for image acquisition and harvesting [13, 14].

A clustered tomato recognition method based on depth map was developed to be acquired by a binocular stereo camera [15]. The recognition accuracy of clustered tomatoes was 87.9% at an acquisition distance of 300–500 mm. A low-cost camera was utilized to recognize the tomato in the tree canopy [16]. A robust tomato recognition algorithm based on multiple feature images and image fusion was developed.

✉ I-Chang Yang
allen2004@gmail.com

¹ Agricultural Automation Center, National Chung Hsing University, Taichung City 40227, Taiwan

² Department of Bio-Industrial Mechatronics Engineering, National Chung Hsing University, Taichung City 40227, Taiwan

³ Department of Biomechatronics Engineering, National Taiwan University, Taipei 10617, Taiwan

⁴ Taiwan Agricultural Mechanization Research and Development Center, Taipei 11051, Taiwan

The final segmentation result was processed by morphology operation to reduce a small amount of noise. In the detection tests, 93% target tomatoes were recognized out of 200 overall samples. A stereo camera was employed to detect and locate mature apples in tree canopies [17]. The stereo camera was mounted on the slide bar in parallel with a distance of 200 mm between the centers of the two camera lens. The author reported that over 89.5% of apples were successfully recognized and the errors were less than 20 mm when the measuring distance was between 400 and 1500 mm. A binocular vision sensor was applied to develop a dual-arm cooperative approach for a tomato harvesting robot [18]. A tomato detection algorithm combining AdaBoost classifier and color analysis was proposed and employed by the harvesting robot. Over 96% of target tomatoes were correctly detected with the speed of about 10 fps.

There are many devices for image acquisition in different aspects of image capture [6], such as color cameras, spectrum cameras, thermal imaging cameras, etc. In object detection field, the accuracy of the proposed algorithm is getting higher and higher, and the recognition speed is getting faster. In terms of speed, YOLO (you only look once) proposed by Redmon et al. [19] was ahead of other algorithms. The author also updated the architecture year by year so it can have better recognition results without losing speed. In terms of judgment accuracy, the regional network-based CNN architecture (region-based CNN, R-CNN) series was more advanced. The backbone on top of AlexNet was proposed to increase the R-CNN with the region proposal function, and achieve 53.3% mAP (mean average precision) in PASCAL VOC [20, 21]. Therefore, a binocular vision camera with R-CNN algorithm was employed in this research. The study aims to develop the imaging system to evaluate the maturity of the fruit as an index and its 3D position could be detected as well. The information would be applied to the grabbing system to harvest the fruit automatically.

Materials and methods

Crop and facility

The crop, beef tomato (*Solanum lycopersicum*, cv.), was the target plant grown in the greenhouse in Taiwan Agriculture Research Institute. The species of beef tomato applied in the project was Chuan-Fu 993 (All Lucky Seed Co., Ltd, Taiwan).

Depth camera, system structure, and calibration

In order to harvest fruits, the robot arm does not only move in one dimension but in three dimension space. The position on a plane could be calculated from a general RGB camera.

However, the depth between the camera and the object is not easy to be calculated by single camera. If single camera moves with the slide rail for linear movement, the error of calibration of the motor and the slide rail equipment must be considered.

The study applied ZED mini binocular vision camera. This camera connects with the computer via the USB 3.0 type-C interface. The resolution of the image is 1280×720 pixels. The detected distance of the three dimensional depth would be 0.15–12 m.

For the calculation of the depth, this study applied the inferred bounding box from Mask R-CNN [22] for which the ZED official function library was applied for the calculation of three dimensional positions. For the three dimensional spatial information of the depth camera, as shown in Fig. 1, X direction is the center of the left lens pointing to the right lens. Y direction is the X direction perpendicular to the left lens extended downwards, and Z direction is the XY plane perpendicular to the left lens extended downwards.

A tomato harvesting equipment in this study was designed to be installed on a transport vehicle, including a depth camera and a central control host. The research combined technologies, such as image capture, object detection, and binocular stereo vision. After capturing a good image, the object detection model would be applied to analyze the maturity. The depth camera was then applied to calculate the coordinates of the three dimensional space.

There are two coordinate systems in this research, namely the depth camera and the robotic arm coordinate system. The three-dimensional information of the depth camera is defined here. Referring to Fig. 2a, for the coordinate system, the X direction is the direction from the center of the left lens to the right lens; the Y direction is perpendicular to the left lens and the X direction extends downward; the Z

Fig. 1 Mask R-CNN saves the data. X,y represents the position of that tomato in the image. Width, height represent the frame and height of the bounding box. Real_x, real_y, real, depth represent the calculated three dimensional distance from the depth camera

Fig. 2 **a** Three dimensional spatial information of the depth camera. **b** Binocular vision positioning correction architecture

direction is perpendicular to the XY plane of the left lens extends forward.

This experiment will use the same tomato to calculate the deviation between the calculated position and the actual position of the depth camera at different distances and different positions. Considering that the distance between the actual harvested tomatoes and the camera is about 50–60 cm, the depth part will be divided into three parts of 50, 55, and 60 cm.

Because this experiment will perform positioning detection at three depths, 25 detection points are distinguished by the shortest distance, that is, under 50 cm, the image range that can be captured, as shown in Fig. 2b. At this distance, the horizontal distance of the tomato is 20 cm, the vertical distance is 12 cm, and the depth camera origin (left lens center) extends forward through the center of the 25 detection points, that is, the horizontal position of these 25 points is $-40, -20, 0, 20, 40$, the vertical position is $-24, -12, 0, 12, 24$, the depth is 50, 55, 60, and the unit is cm. The three-dimensional position calculated by the depth camera under this architecture is the location of the testing point erected.

In this experiment, three-dimensional position calculations will be performed at each point, and regression analysis will be performed on all the predicted three-dimensional positions and actual positions.

Building up the training dataset

In the image, the sample labeling of the object to be detected would be done manually. Then the image and the labeled data would be input to the neural network for training. The Mask R-CNN was applied to calculate the error from the predicted result and the actual result. Based on the error value, it was in turn to modify the inner parameters of the algorithm. However, the machine learning is greatly affected by the input data. The wrong detection or failing to detect the object would easily occur for the untrained image. In order to evaluate the training effectiveness of a model, the collected data will be divided into training dataset and validation dataset.

The depth camera captured 255 images for training and validation classification. There were 205 images for training dataset and 50 images for validation dataset. None of them was duplicated. Besides, in order to increase the amount of data, the data would be flipped horizontally before it was entered to the neural network. This could double the amount of data.

Dataset labeling

VGA Image Annotator (VIA) [23] was applied for mature sample labeling. VIA is a lightly but powerful manual labeling software. The labeling task can easily be done on web browser, and different data formats can be exported for users to use.

The export labeling format of VIA is *.json (JavaScript Object Notation). The complete labeled data includes image file name, image size, other information of the labeled frame and labeled type. Authors converted the complete labeled data by VIA to Microsoft common objects in context (MS COCO) format to enter the neural network.

During the growth process of the tomatoes, the actual growing conditions of them were different as they may be pushed by other fruits, covered by the net, or covered by the branches. Not all the mature tomatoes could be plucked. Therefore, this study classified the object into three types: (1) mature tomato without cover (Full), (2) mature tomato but covered (Covered), and (3) peduncle.

As shown in Fig. 3, the lower and middle part of the mature tomato is covered by the unripe green tomato so it has been classified as Covered. Besides, its peduncle was also labeled. Lower left is the mature tomato without cover. Its peduncle was not labeled. When the actual position of the

Fig. 3 VGA image annotator operation interface

Table 1 Training stage

Stage	Epoch	LR	Training layers
1	1–40	0.001	Heads
2	40–120	0.001	ResNet Stage 4 and up
3	120–160	0.0001	All

object could not be identified by human eyes, it would not be labeled by the system.

Setting the parameter of Mask R-CNN

The model of this study was pre-trained model that was trained base COCO of ResNet-101 for transfer learning. COCO dataset included 80 classifications and tens of the thousands of images. Pre-trained model applied this dataset for pre-training. Since the COCO dataset was big and diverse, the pre-trained model could include the information of different classifications. In addition to the own prepared situation for final training, this model could be applied to different research domains, and it could be converted into the required specific model. This could save the time for training complete model in different domains.

This study included three stages as shown in Table 1. The first stage applied 0.001 learning rate and only the RPN (region proposal network), classification and mask level were trained. The second stage applied 0.001 learning rate and the fourth and latter part of ResNet was trained. The third stage applied 0.001 learning rate to train the whole network. The training process would record the overall error,

RPN error, and mask layer error of the training dataset and validation dataset. This provides one of the indicators for the final selection of the model.

Estimate the actual size of the fruit from the image

In general, the harvesting claw will open to the maximum in plucking the tomato fruits. In fact, the sizes of the tomatoes are different and it is not necessary for the claw opening to the maximum for each harvesting action. Therefore, this study recorded size of the tomato species of this research under different position of the image with different depth. The information was recorded by comparing the bounding box of the tomatoes detected by Mask R-CNN and information from depth camera with the actual size.

In this experiment, the depth camera was applied for three dimensional space calculating and compared with the error of the actual position. Then the tomatoes with the width and height of 80×70 mm, 60×50 mm, 68×58 mm, and 55×50 mm were put to the nine positions of the image to record the data. Besides, with the width and height of 40–100 cm, the steps above repeated once for every 10 cm.

Results and discussions

Object detection

The maturity and position image of the target tomato calculated by using Mask R-CNN was shown in Fig. 4. This study applied Mask R-CNN and the backbone was COCO Pre-trained model from ResNet101. The learning rate was set to

Fig. 4 The maturity and position image of the target tomato calculated by using Mask R-CNN

0.001. The stochastic gradient decent (SGD) was applied. The size of the input image was 1024×704 pixels and the training was 160 epochs. There were totally 255 images, and 205 of them randomly were selected as the training samples. The rest of 50 images were regarded as the validation set. The VIA was applied for polygonal edge labeling. There were 1538 marked tomatoes in all images. The intersection over union (IoU) would be the threshold with 0.5 to determine the position of the border frame correct or not. The inference border frame would do the IoU calculation with the ground truth, and the IoU value larger than 50% would be inference positive. Therefore, the IoU over and included 90% were calculated and discussed in this research.

In order to evaluate the effectiveness of the model and the choice of threshold, the following formula was applied for calculation:

$$\text{Precision} = \frac{(\text{True Positive})}{(\text{True Positive} + \text{False Positive})} \quad (1)$$

$$\text{Recall} = \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})} \quad (2)$$

$$\text{Intersection over Union (IoU)} = \frac{\text{Intersection}}{\text{Union}} \quad (3)$$

The objects for Mask R-CNN training included: full tomato, covered tomato and peduncle. As shown in Fig. 5, under the validation of IoU@99%, if IoU is larger than 0.9, it is judged correct. If it is less than 0.9, it is judged wrong. When mAP is 1.0, it fails only under IoU@95%. For the training result the Precision is 98.6% and Recall is 98.6%. Figure 5 shows three categories of APs when IoU@95%.

In order to apply in harvesting in the future, this study classified the objects into full, covered and peduncle. The full

tomato is red frame, the covered tomato is blue frame and peduncle is green frame. From the image, it was clearly found that Mask R-CNN could effectively distinguish whether the tomatoes were blocked. By using the robot arm to harvest tomatoes, the size of the tomatoes, the overlapping of the fruits, and the harvesting claw not easy to bypass the front row of tomatoes, peduncle and net should be considered. After the Mask R-CNN identified the position and shape of the peduncle, the position of the cutting operation of the peduncle could be determined according to the shape of the mask.

According to the experimental field, the brightness of the tomato surface is between 4000 and 20,000 Lux. As shown in Fig. 6, Mask R-CNN can be used for identification, and the brightness is between 6000 and 10,000 Lux in the daily lighting environment. Therefore, the model trained in this study can have a good recognition success rate in a general light environment.

Maturity training of tomatoes in the model

Since the target tomatoes labeled in this study are tomatoes that experts judge to be harvestable, the tomatoes identified by Mask R-CNN only judge whether they are mature or not. According to the aforementioned training results, both Precision and Recall reach 98%, which means that the ripe tomatoes deduced by Mask R-CNN and the ripe tomatoes judged by experts have high reliability.

There were 363 images grabbed in this study including (mature tomatoes, semi-mature and immature tomatoes), and proposed the identified tomato regions in the images, as shown in Fig. 7, and converted them into HSV color space to calculate the color of mature tomatoes, as shown in Fig. 8a–c. According to the converted HSV value, the average and standard deviation of each value can be counted,

Fig. 5 Mask R-CNN performance **(a)** Precision-recall curves of two IoUs **(b)** Precision-recall curves of three classes

and the upper and lower limits can be calculated with three times the standard deviation, as shown in Table 2. This table can define that the Hue (H) of mature tomatoes is between 0 and 16.47, Saturation (S) is between 157.27 and 252.09, Value (V) is between 119.05 and 255; Hue of green

tomatoes (immature) is between 23.15 and 33.48, Saturation is between 137.48 and 207.65, Value is between 120.86 and 211.06; semi-mature tomatoes have Hue between 12.05 and 23.22, Saturation between 160.76 and 233.69, and Value between 145.96 and 238.37.

Fig. 6 Light intensity from the surface of the tomato in the greenhouse

Fig. 7 The extracted image of the mature tomato

Actual fruit size estimation

Position calculation by depth camera

ZED mini binocular vision depth camera was employed in the project and played an important role to determine the depth of the object. The depth can be calculated from the range of 0.15–12 m. The field camera in this study was installed on a self-propelled vehicle. The distance XY direction was 1.14 cm while the average error was 0.48 cm. The maximum error of depth Z direction was 1.86 cm while average error was 0.67 cm.

Fruit size prediction

According to the comparison results, there was a certain error for ZED mini depth camera. But the calculated

three-dimensional spatial information was credible. Therefore, this study would classify the mature tomato fruits according to their size in the image. By calculating the bounding box and the actual size under different depth ratio by Mask R-CNN, the size of the fruit could be estimated. Figure 9 shows the tomato images in four different width and height distributed in nine different positions. It could be observed that under the same depth, although ratio of width and height with different size would not be focused on the farther location, it was very close within the required range of this study.

The formula of the ratio calculation is shown in Eq. (4):

$$\text{Factor} = \frac{\text{Real size}}{\text{Bounding Box size}} \left(\frac{\text{mm}}{\text{pixel}} \right) \quad (4)$$

where Factor represents the ratio of the bounding box and actual size.

According to Fig. 9 the width and height ratio did not vary obviously under the same depth with different size in nine different positions. This means the image did not distort much. After integrating all the data of width and height, and calculated by curvilinear regression Eqs. (5) and (6), the result is shown in Fig. 10:

$$\text{Factor}_w = 0.0161 \times \text{depth}^{0.9384} \quad (5)$$

$$\text{Factor}_h = 0.0191 \times \text{depth}^{0.8985} \quad (6)$$

The R^2 of width ratio (Factor_w) and height ratio (Factor_h) are 0.9041 and 0.9344, respectively. That statistical analysis has considerable accuracy. Therefore, the image could be applicable for object size estimation. Finally, the width and height of the bounding box inferred by Mask R-CNN multiplied by formula (5) and (6), and the size could be estimated.

The size estimation formula is as following:

$$\text{Width}_{\text{real}} = \text{bbox}_w \times \text{Factor}_w \quad (7)$$

$$\text{Height}_{\text{real}} = \text{bbox}_h \times \text{Factor}_h \quad (8)$$

where $\text{Width}_{\text{real}}$ and $\text{Height}_{\text{real}}$ are the estimated size (mm) of the fruit and bbox is the bounding box.

Identification algorithms comparison and discussion

Tomato recognition algorithm based on multiple feature images and image fusion [16]. A novel Faster R-CNN model was established by using the RGB and near infrared images to develop a precision and fast fruit inspection system [24]. In 2017, machine vision and artificial neural network system (ANN) were applied to build up a mango maturity inspection system [8]. The current results calculated by the developed

Fig. 8 Hues (a), saturations (b), and values (c) of mature tomatoes in HSV color index

Table 2 HSV statistic results of different mature level tomatoes

Color index	Average	Standard deviation	Upper limit	Lower limit
Immature tomato (green, N: 23)				
Hue	27.38	2.54	33.48	23.15
Saturation	166.53	19.05	207.65	137.48
Value	162.80	23.87	211.06	120.86
Semi-mature tomato (N: 260)				
Hue	17.26	2.83	23.22	12.05
Saturation	196.28	16.25	233.69	160.76
Value	194.77	21.35	238.37	145.96
Mature tomato (N: 1113)				
Hue	7.04	3.14	16.47	0
Saturation	204.68	15.81	252.09	157.27
Value	198.56	26.51	255	119.05

Mask R-CNN algorithm in this study could reach the resolution accuracy of tomatoes recognition up to 93%. When the distance is 50 and 60 cm, compared with the algorithm used above, the sweet pepper with the resolution accuracy of 80.7% and the mango with the resolution accuracy of 94% are compared. There are higher precision and recall, and the results are both 100%.

A fast three-dimensional scene reconstruction method in the simulation environment could be obtained by using the stereo camera with the point cloud [18]. There were several error occurrences in using Xbox 360 Kinect Model 1414 (Microsoft, Inc., USA) with ToF (Time of Flight) technology. The positioning error was one of it, and the error is less than 10 mm [3]. Compared with these studies, the positioning errors of our cameras are all less than 10 mm after being calibrated. The similar results may be due to the light factor in the environment caused the image brightness to be unstable during capturing. It would lead to larger errors for calculating the feature points of the camera. In Table 3, most

of the differences between prediction position and measured position were less than 10 mm.

Conclusions

A fruit maturity detection system has been developed in this research using the Mask R-CNN object detection algorithm. The Transfer learning using ResNet-101's COCO pre-trained model was applied in backbone. After evaluating the validation set, the IoU threshold value was set to 0.9, and both precision and recall can reach 100%. The model trained in this experiment had had good recognition ability in unrecognized scenes. The system to estimate the actual size of ripe tomato fruits could obtain very high correlations in width and height regressions. The detection system could be applied to judge the spatial location of the fruit to provide the information to the crop grabbing robotic system in the future.

The object detection network and depth camera were integrated to get the position of the mature tomatoes and estimated the size. Object detection network was applied in precision of Mask R-CNN and ResNet to evaluate the mature fruits. The pixel of inspection ability was good enough to estimate the covered objects. There is good opportunity to apply in the agricultural field. The results could be applied within the robotic grabbing system to harvest solanaceae crops and cucurbitaceae crops in the greenhouse.

The positions and locations of the mature tomato fruits could be determined by using the information of the depth and position data. According to the test result of the estimated size from the image, different positions in the image did not have obvious distortion. Besides, the ratios of different size of tomatoes were very close. Therefore, the width and height could be grouped to a set of data, which could be processed with regression analysis to get the estimated size of the fruits.

Fig. 9 Factor for different size and depth

Fig. 10 Factor and depth data

Table 3 Comparison between three dimensional spaces calculated from depth camera (prediction) and actual position (measurement)

Position	Prediction			Measurement			Differences		
	X	Y	Z	X	Y	Z	X	Y	Z
1	– 15.2*	– 131.4	480.8	– 20.0	– 120.0	480.0	4.8	11.4	0.8
2	– 51.7	– 114.6	560.7	– 55.0	– 110.0	560.0	3.3	4.6	0.7
3	84.6	– 79.3	498.0	80.0	– 80.0	505.0	4.6	0.7	– 7
4	146.4	– 105.7	643.5	140.0	– 105.0	625.0	6.4	– 0.7	18.5

*The unit of the depth is mm

Acknowledgements This research was financially supported by the Ministry of Science and Technology, Taiwan. Project number: MOST 108-2321-B-002-026.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. S. Hayashi, K. Shigematsu, S. Yamamoto, K. Kobayashi, Y. Kohno, J. Kamata, M. Kurita, Evaluation of a strawberry-harvesting robot in a field test. *Biosyst. Eng.* **105**(2), 160–171 (2010)
2. W. Ji, D. Zhao, F. Cheng, B. Xu, Y. Zhang, J. Wang, Automatic recognition vision system guided for apple harvesting robot. *Comput. Electr. Eng.* **38**(5), 1186–1195 (2012)
3. T.T. Nguyen, K. Vandevoorde, N. Wouters, E. Kayacan, J.G. De Baerdemaeker, W. Saeys, Detection of red and bicoloured apples on tree with an RGB-D camera. *Biosyst. Eng.* **146**, 33–44 (2016)
4. D.M. Bulanon, T. Kataoka, Y. Ota, T. Hiroma, AE—automation and emerging technologies: a segmentation algorithm for the automatic recognition of Fuji apples at harvest. *Biosyst. Eng.* **83**(4), 405–412 (2002)
5. C. Wang, X. Zou, Y. Tang, L. Luo, W. Feng, Localisation of litchi in an unstructured environment using binocular stereo vision. *Biosyst. Eng.* **145**, 39–51 (2016)
6. A. Gongal, S. Amatya, M. Karkee, Q. Zhang, K. Lewis, Sensors and systems for fruit detection and localization: a review. *Comput. Electron. Agric.* **116**, 8–19 (2015)
7. J. Xiong, R. Lin, Z. Liu, Z. He, L. Tang, Z. Yang, X. Zou, The recognition of litchi clusters and the calculation of picking point in a nocturnal natural environment. *Biosyst. Eng.* **166**, 44–57 (2018)
8. E. Yossy, J. Pranata, T. Wijaya, H. Hermawan, W. Budiharto, Mango fruit sortation system using neural network and computer vision. *Procedia Comput. Sci.* **116**, 596–603 (2017)
9. S. Bargoti, J.P. Underwood, Image segmentation for fruit detection and yield estimation in apple orchards. *J. Field Robot.* **34**(6), 1039–1060 (2017)

10. M. Rahneemoonfar, C. Sheppard, Deep count: fruit counting based on deep simulated learning. *Sensors* **17**(4), 905 (2017)
11. Z. Wang, K.B. Walsh, B. Verma, On-tree mango fruit size estimation using RGB-D images. *Sensors* **17**(12), 2738 (2017)
12. Y. Yu, K. Zhang, L. Yang, D. Zhang, Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Comput. Electron. Agric.* **163**, 104846 (2019)
13. Q. Quan, T. Lanlan, Q. Xiaojun, J. Kai, F. Qingchun, in *Selecting candidate regions of clustered tomato fruits under complex greenhouse scenes using RGB-D data*. 2017 3rd International Conference on Control, Automation and Robotics (ICCAR) (2017), pp. 389–393
14. K. Tanigaki, T. Fujiura, A. Akase, J. Imagawa, Cherry-harvesting robot. *Comput. Electron. Agric.* **63**(1), 65–72 (2008)
15. R. Xiang, H. Jiang, Y. Ying, Recognition of clustered tomatoes based on binocular stereo vision. *Comput. Electron. Agric.* **106**, 75–90 (2014)
16. Y. Zhao, L. Gong, Y. Huang, C. Liu, Robust tomato recognition for robotic harvesting using feature images fusion. *Sensors* **16**(2), 173 (2016)
17. Y. Si, G. Liu, J. Feng, Location of apples in trees using stereoscopic vision. *Comput. Electron. Agric.* **112**, 68–74 (2015)
18. L. Xiao, Y. Zhao, L. Gong, C. Liu, T. Wang, Dual-arm cooperation and implementing for robotic harvesting tomato using binocular vision. *Robot. Auton. Syst.* **114**, 134–143 (2019)
19. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, in *You only look once: unified, real-time object detection*. Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (2016), pp. 779–788
20. R. Girshick, J. Donahue, T. Darrell, J. Malik, in *Rich feature hierarchies for accurate object detection and semantic segmentation*. 2014 IEEE Conference on Computer Vision & Pattern Recognition (2014), pp. 580–587
21. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process Syst.* **25**, 1097–1105 (2012)
22. K. He, G. Gkioxari, P. Dollár, R. Girshick, in *Mask r-cnn*. Proceedings of the IEEE International Conference on Computer Vision (2017), pp. 2961–2969
23. A. Dutta, A. Zisserman, in *The VIA annotation software for images, audio and video*. Proceedings of the 27th ACM International Conference on Multimedia (2019), pp. 2276–2279
24. I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, C. McCool, Deep-fruits: a fruit detection system using deep neural networks. *Sensors* **16**(8), 1222 (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.