Recognition and localization of strawberries from 3D binocular cameras for a strawberry picking robot using coupled YOLO/Mask R-CNN

Heming Hu[1], Yutaka Kaizu[1], Hongduo Zhang[1], Yongwei Xu[1], Kenji Imou[1], Ming Li[2], Jingjing Huang[2], Sihui Dai[3*]

(1. *Graduate School of Agricultural and Life Sciences, the University of Tokyo, Tokyo 113-0033, Japan*;
2. *Hunan Agricultural Equipment Research Institute, Changsha, Hunan 410125, China*;
3. *College of Horticulture, Hunan Agricultural University, Changsha, Hunan 410128, China*)

Abstract: To solve the problem of high labour costs in the strawberry picking process, the approach of a strawberry picking robot to identify and find strawberries is suggested in this study. First, 1000 images including mature, immature, single, multiple, and occluded strawberries were collected, and a two-stage detection Mask R-CNN instance segmentation network and a one-stage detection YOLOv3 target detection network were used to train a strawberry identification model which classified strawberries into two categories: mature and immature. The accuracy ratings for YOLOv3 and Mask R-CNN were 93.4% and 94.5%, respectively. Second, the ZED stereo camera, triangulation, and a neural network were used to locate the strawberry in three dimensions. YOLOv3 identification accuracy was 3.1 mm, compared to Mask R-CNN of 3.9 mm. The strawberry detection and positioning method proposed in this study may effectively be used to supply the picking robot with a precise location of the ripe strawberry.

Keywords: strawberry detection, 3D point cloud, mean-shift, clustering method
DOI: 10.25165/j.ijabe.20221506.7306

Citation: Hu H M, Kaizu Y, Zhang H D, Xu Y W, Imou K, Li M, et al. Recognition and localization of strawberries from 3D binocular cameras for a strawberry picking robot using coupled YOLO/Mask R-CNN. Int J Agric & Biol Eng, 2022; 15(6): 175–179.

1 Introduction

Strawberry is one of the most popular organic products in Japan[1], but its production processes are labor intensive. Strawberry has a long period of harvesting every year, and strawberry harvesting accounts for a fourth of the cost of its production[2]. Therefore, to improve productivity and efficiency, it is necessary to develop a strawberry harvesting robot[3].

Since the development of a prototype by Takenaga[4], strawberry-reaping robots have been effectively explored. Hayashi et al.[5] proposed a strawberry harvesting robot with a suction device. Han et al.[6] designed a strawberry harvesting robot with a gripping and cutting device. Xiong et al.[7,8] proposed a method to bypass obstacles to grab strawberries and designed an end device with internal containers, reducing the number of round trips from the target to the storage box. Feng et al.[9] proposed a non-destructive end effector for sucking fruit, gripping, and cutting fruit stems, which was designed to prevent peel damage and disease infection. De Preter et al.[10] proposed a strawberry-picking robot with a 3D-printed flexible manipulator. This design can effectively distribute the pressure evenly to the strawberry surface, reducing damage to the strawberry field. For fruit-picking robots with different structures, the vision system not only needs to give the recognition results of strawberries in the image but also needs to give the coordinates of strawberries in three-dimensional space[11,12].

For strawberry recognition, by studying the physiological and biochemical properties of strawberries, Li et al.[13] found that strawberries can be recognized by the R value in the RGB variant space and the a^2+b^2 values in the Lab variant space. Zhang et al.[14] proposed a method for dynamic threshold segmentation of RGB images using the normalized color difference (R-G)/(R+G) method and the OTSU algorithm. However, in actual use, these methods will be affected by changes in lighting conditions which affects the recognition rate[15,16]. In recent years, the neural network has effectively overcome the shortcomings of traditional image recognition technology that are greatly affected by environmental factors[17]. Among them, the one-stage detection of the YOLO series neural network[18] and the two-stage detection of Mask R-CNN instance segmentation[19] have received great attention. Using a neural network to locate fruits can effectively improve the detection rate of overlapping fruits and reduce the influence of illumination[20].

Through the triangulation concept, the ZED stereo camera can use its two passive RGB sensors to determine the separation between the object and the imaging system[21,22]. It is portable, affordable, and has steady imaging, making it a good choice for

Received date: 2021-12-29 **Accepted date:** 2022-09-14
Biographies: Heming Hu, PhD candidate, research interest: agricultural aerial spraying, image processing, Email: hu-heming950211@g.ecc.u-tokyo.ac.jp; **Yutaka Kaizu**, PhD, Associate Professor, research interest: machine vision, 3D mapping, field robotics, navigation, GPS, Email: kaizu@g.ecc.u-tokyo.ac.jp; **Hongduo Zhang**, PhD candidate, research interest: agricultural robot, Email: zhdhnhr@outlook.com; **Yongwei Xu**, PhD, research interest: machine vision, 3D mapping, Email: askxiongwei@gmail.com; **Kenji Imou**, PhD, Professor, research interest: renewable energy, bio fuel, navigation sensor, autonomous vehicle, image processing, Email: k-imou@g.ecc.u-tokyo.ac.jp; **Ming Li**, PhD, Professor, research interest: crops sensing information acquisition and agricultural robots, Email: liming@hunau.net; **Jingjing Huang**, MS, research interest: agricultural intelligent equipment, Email: huangjingjing2016@gmail.com.
***Corresponding author: Sihui Dai**, PhD, Associate Professor, research interest: horticultural crops, information application. College of Horticulture, Hunan Agricultural University, No.1 Nongda Road, Furong District, Changsha 410128, China. Tel: +86-15974298578, Email: daisihui@126.com.

agricultural production[23,24]. The position of the relevant target point can be found using the ZED camera after the target image coordinates are identified using the color image[25]. To build a strawberry target recognition system with high accuracy and high positioning precision to locate strawberries, YOLOv3 and Mask R-CNN were coupled with the ZED camera, which is robust enough for further use in strawberry picking robots.

2 Material and methods

This study built a 3D strawberry localization system using a neural network and a ZED binocular camera. First, 1000 strawberry images were gathered, and the YOLO and Mask-RCNN networks were used to train the strawberry identification model. After that, a 3D point cloud image was created using the left and right cameras' triangular positioning. The strawberry was then detected using the YOLO/Mask R-CNN recognition model, and the positioning was finished by reading the depth value of the positioning frame center point to determine the strawberry center point coordinates in relation to the left camera center point.

2.1 Binocular camera system architecture

The binocular stereovision system comprises ZED camera (resolution 1080×1920) and a personal computer (2.40 GHz Intel Core i5-9300H CPU and 8GB RAM). The image-processing program was developed using Python3.7 on the Microsoft Windows 10 operating system. The basic image processing functions of the program, i.e. camera calibration, image acquisition, display, spatial filtering, line fitting, and stereo matching, were implemented using ZED sdk v3.1.0.

Figure 1 presents a schematic illustration of the binocular camera system developed in this study. The origin of the camera coordinates is the center of the left camera lens, and the left and right image coordinate systems have their origins at the upper left in the images. (X, Y, Z) are the 3D coordinates of the object point ***P*** in the 3D coordinate system.

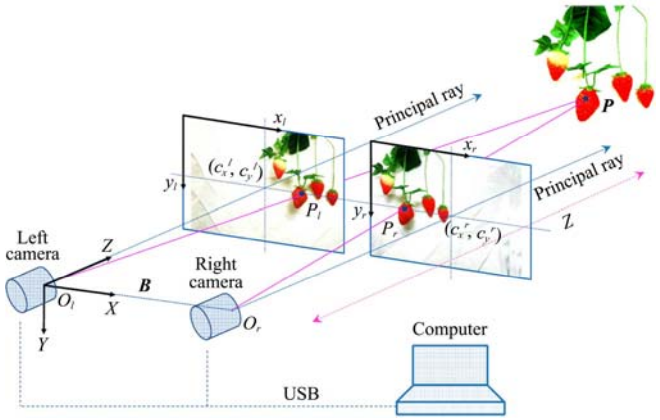


Figure 1 Schematic illustration and stereo coordinate system used in this study

The distance between the left and right cameras is ***B***. (x_l, y_l) and (x_r, y_r) are the coordinates of the object point ***P*** in the left and right image coordinate system, and (c_x^l, c_y^l) and (c_x^r, c_y^r) are the coordinates of the principal point in the left and right image coordinate system, respectively. The cameras was calibrated based on the method of Zhang[26] and the OpenCV 3.4.0 was used to run the Python program. After the calibration and mathematical rectification, the two cameras had exactly parallel optical axes (principal ray). Taking x_l and x_r to be the horizontal positions of the points in the left and right cameras, respectively, the depth is inversely proportional to the disparity between these views, where the disparity is defined simply by Equation (1).

$$disparity = x_l - x_r \tag{1}$$

The points Z, X, and Y coordinates were derived using similar triangles as follows.

$$\begin{aligned} Z &= \frac{Bf}{disparity} \\ Y &= \frac{By_l}{disparity} + c_y^l \\ X &= \frac{Bx_l}{disparity} + c_x^l \end{aligned} \tag{2}$$

2.2 Training of strawberry recognition model

2.2.1 Dataset construction

The Strawberry Digital Images Data Set contained a total of 1000 pictures, which were chosen randomly. Each image is 1008×756 in size. The dataset was labeled using the Labelimg and Labelme programs. As shown in Figure 2, Labelimg utilizes rectangular labels to train the YOLO target detection model, as shown in Figure 3, and Labelme uses polygon labels to train the Mask R-CNN image segmentation model. All photos were divided into the training set and the test set, with 950 images used for training and 50 images used for testing. The labels were classified into two categories: mature and immature.

Figure 2 Schematic diagram of Labelimg

Figure 3 Schematic diagram of Labelme

2.2.2 YOLOv3

YOLOv3 is a one-stage detection network that determines the location and kind of objects on a single image at the same time. The residual network structure is used by the feature extraction network. The network topology of YOLOv3 is mostly made up of five residual modules, as shown in Figure 4. Each residual module is made up of two DBL units and residual operations. The properties of three scales, 13×13, 26×26, and 52×52, are also included in YOLOv3, which significantly enhances the network capacity to recognize objects of various sizes.

YOLOv3 also formulates a new loss function for the target detection task. By calculating the target center point coordinate error, confidence error, and category error, the error between prediction and labeling is calculated to determine the next training direction.

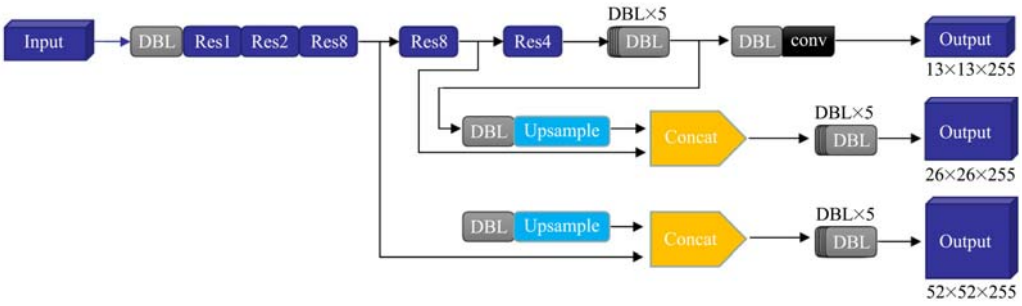


Figure 4 YOLOv3 network

1) Center coordinate error:

$$E_1 = \lambda_{cood}\sum_{i=0}^{s^2}\sum_{j=0}^{B} l_{ij}^{obj}[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2]+ \lambda_{cood}\sum_{i=0}^{s^2}\sum_{j=0}^{B} l_{ij}^{obj}[(\sqrt{w_i}-\sqrt{\hat{w}_i})^2+(\sqrt{h_i}-\sqrt{\hat{h}_i})^2] \quad (3)$$

2) Confidence error:

$$E_2 = \sum_{i=0}^{s^2}\sum_{j=0}^{B} l_{ij}^{obj}(C_i-\hat{C}_i)^2 + \lambda_{noobj}\sum_{i=0}^{s^2}\sum_{j=0}^{B} l_{ij}^{noobj}(C_i-\hat{C}_i)^2 \quad (4)$$

3) Class error:

$$E_3 = \sum_{i=0}^{s^2} l_{ij}^{obj}\sum_{c\in classes}^{sB^2}(p_i(C)-\hat{p}_i(C))^2 \quad (5)$$

where, S^2 represents the grid size, and three different scales for YOLOv3 (i.e., 13×13, 26×26, 52×52). B stands for the target box. l_{ij}^{obj} Represents that there is a target at image i, j and the value is 1, and there is no target, the value is 0. l_{ij}^{noobj} represents that if the box at image i, j has no target, its value is 1, otherwise, it is 0.

2.2.3 Mask R-CNN

Unlike YOLOv3, Mask R-CNN is a two-stage network detection method. As shown in Figure 5, the first stage of the process involves scanning the image to identify regions that are likely to contain an item. These regions are then classified in the second stage to create bounding boxes and masks. The foundation network of Mask R-CNN, an extension of Faster R-CNN that is used to extract picture information, is a convolutional neural network ResNet50. The resulting feature maps are used by the Region Proposal Network (RPN) to decide which sub-networks to use. RPN uses a sliding window to quickly scan the image and identify the target area. The Mask branch classifies each target area at the pixel level to produce an image mask. Additionally, Mask R-CNN and YOLOv3 both employ the Non-maximum suppress (NMS) to eliminate unnecessary frames and improve the accuracy of the output target frame in the common scenario in image processing where an object is detected utilizing several detection frames.

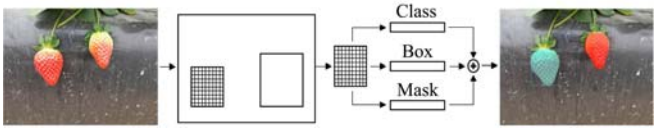


Figure 5 Mask-RCNN network

The loss function formula of Mask R-CNN is:

$$L_{all}=L_{cls}+L_{box}+L_{mask} \quad (6)$$

where, L_{all} represents the overall error, L_{cls} represents the classification error, L_{box} represents the error between the predicted box and the marked box, and L_{mask} represents the error between the predicted value of the mask and the marked value.

2.2.4 Strawberry 3D coordinate detection process

Figure 6 shows the main process of strawberry 3D coordinate detection. First, turn on the ZED camera and get its left and right views. Then the left view was passed to the YOLOv3/Mask R-CNN network to detect the weed category and its image coordinates, and the left and right views jointly calculated the depth information. Finally, the z-coordinate value was obtained by reading the depth information of the weed target frame, so as to complete the acquisition of the three-dimensional position information of the strawberry target.

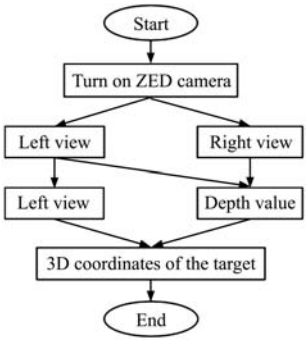
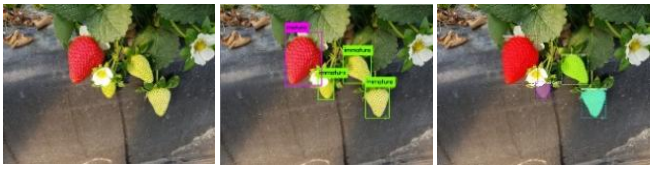


Figure 6 3D coordinate detection program flow chart

3 Results

3.1 Strawberry recognition result

The training platform computer included the CUDA10.0 computing architecture, darknet, Intel Core i5-9300H CPU, NVIDIA GeForce GTX 1650 graphics card, and Tensorflow deep learning framework. The learning rate was set to 0.001 under the darknet framework. After 2000 iterations, the loss function reduced to roughly 0.3, allowing training to stop and the YOLO target recognition model to be obtained. This enhances the model capacity to adapt to various illumination conditions. After 100 training iterations using the Tensorflow framework, the Mask R-CNN instance segmentation model was produced. The learning rate was set as 0.001. In order to measure the performance of the two models in detecting strawberries, 50 images in test sets described in 2.2.1 (which include complex situations such as large numbers, occlusions, and many background leaves) were used to identify errors, missed identifications, and classification errors. Statistical analysis and accuracy calculations are three common faults. Three immature strawberries and one ripe strawberry, one of which is hidden by blooms, are shown in Figure 7a of a testing photograph. The detection outcomes of YOLO and Mask R-CNN are shown in Figures 7b and 7c, respectively. Both networks can effectively identify the location and class of this strawberry even when it is covered by blooms.



a. b. c.

Figure 7 Strawberry testing test samples and test results

Table 1 shows that the Mask R-CNN accuracy on this dataset is 94.5 percent which is greater than YOLOv3 accuracy of 93.4 percent. Among them, YOLOv3 has more wrong detections than Mask R-CNN, misclassifying immature and ripe strawberries as well as backdrop leaves. Mask R-CNN, however, does not identify strawberries as well as YOLOv3. YOLO is a one-stage detection that simultaneously infers the position and kind of objects from the image. Since it uses a quick detection algorithm, its running speed is much faster than Mask R-CNN, which also contributes to some false detections. To lessen the likelihood of the backdrop being wrongly classified as a strawberry, the confidence threshold can be suitably raised, meaning the target frame with low confidence is not displayed. A two-stage detection is the Mask R-CNN. To classify items in accordance with the target frame, it first roughly guesses the potential targets. Since there are more unrecognized cases than in YOLOv3, its operating speed will be somewhat slower, but there will be fewer misidentifications and a higher classification accuracy rate than in YOLOv3.

Table 1 Detection model performance

Model	Total	Wrong detection	Not detected	Misclassification	Accuracy
YOLOv3	181	3	6	3	93.4%
Mask-RCNN	181	1	8	1	94.5%

Based on analysis of the main reasons for the errors, it was found that the classification errors, mainly occur on some strawberries that are about to ripen, and their tops are not fully ripe, so they are mistakenly identified as ripe, and further programs are needed to determine whether they are ripe, such as calculating the red part accounting for the proportion of the strawberry shape. The undetected error mainly occurs on the strawberry covered by the background leaves. The image size is too small, and the training sample is insufficient, so it is not detected. The training sample of this scene can be appropriately increased to reduce the error. The simple identification errors mainly occur when some leaves are identified as strawberries, and the confidence level is about 0.7, so the confidence level can be increased to avoid this problem.

3.2 Strawberry Positioning results

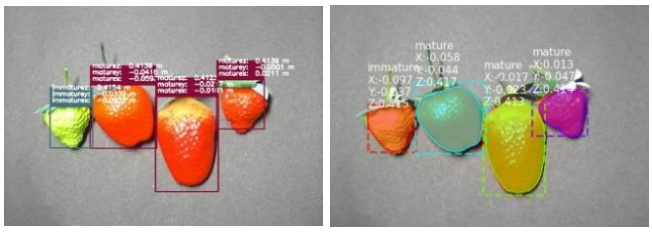
Two colored strawberry models, red strawberries to represent ripe strawberries and cyan strawberries to represent immature strawberries, are utilized to test the strawberry positioning accuracy of this system. Calculate the difference between the actual value and the positioning value between the strawberry center point and the left camera center point. The calculation method for the distance between two points was used in addition to counting the errors in the *x*, *y*, and *z* directions to measure the total error level of the system in a more intuitive manner (Equation (7)).

$$L = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} \quad (7)$$

where, x_i,y_i,z_i represent the values provided by the positioning system in the three directions, and *x*, *y*, *z* represent the actual distances in the three directions.

Figures 8a and 8b illustrate the results of measuring the three-dimensional coordinates of strawberries in the same picture using YOLOv3 and Mask R-CNN, respectively. The figure correctly distinguishes the adult and immature strawberry models, and the three-dimensional coordinates *x*, *y*, and *z* of each strawberry were obtained.

Both networks are capable of producing improved results, as shown in Table 2. YOLOv3 has a higher error than Mask R-CNN in the x direction, and Mask R-CNN has a greater error than YOLOv3 in *y* and *z* directions. The comprehensive inaccuracy of 3.1 mm with YOLOv3 is slightly better than that of 3.9 mm with Mask-RCNN. From the results in Table 2, it can be seen that the convergence ability of the target frame of Mask R-CNN seems to be weaker than that of YOLOv3. It was observed that the target frame cannot completely fit the detection results of the strawberry shape, and found that the error of Mask R-CNN comes from the detection frame being larger than the strawberry shape, while that of YOLOv3 comes from the detection box being smaller than the target shape. This difference led to a difference in the coordinate positioning points, which further increased the error in the *Z* direction. If the shape center of the Mask is used to measure, it will effectively decrease the error of Mask R-CNN.



a. b.

Figure 8 Strawberry 3D coordinate test sample

Table 2 Positioning error

Model	*X*/mm	*Y*/mm	*Z*/mm	*L*/mm
YOLOv3	2.0	1.5	1.8	3.1
Mask R-CNN	1.5	2.8	2.3	3.9

4 Conclusions

This study built a three-dimensional coordinate positioning system for strawberries which realized the functions of three-dimensional coordinate measurement and strawberry maturity classification using the YOLOv3 one-stage object detection network, the Mask R-CNN two-stage instance segmentation network, and the ZED binocular camera. The YOLOv3 network achieved a recognition accuracy of 93.4% and a placement inaccuracy of 3.1 mm. Mask R-CNN achieved a recognition accuracy of 94.5% and a localization inaccuracy of 3.9 mm. Due to its specialized candidate frame extraction network and classification network, the two-stage Mask R-CNN network had a high classification accuracy, but the one-stage YOLOv3 network had multi-scale fusion structural characteristics and had fewer missed detection situations. The system detection accuracy will be further increased by increasing the number of training samples with intricate backgrounds and severe occlusion, as well as by suitably tweaking detection parameters like confidence thresholds. According to the positioning experiment, the error of Mask R-CNN was marginally higher than YOLOv3, and its detection frame ability to converge was less effective than YOLOv3. Because in this experiment, the target frame center point was used to calculate the three-dimensional coordinates, if the mask shape center was utilized as the reference point measurement inaccuracy, the Mask R-CNN performance will be improved.

[References]

[1] Hikawa-Endo M. Improvement in the shelf-life of Japanese strawberry fruits by breeding and post-harvest techniques. The Horticulture Journal, 2020; 89(2): 115–123.

[2] Nishizawa T. Current status and future prospect of strawberry production in East Asia and Southeast Asia. In: Proceedings of the IX International

Strawberry Symposium, 2021; pp.395–402.
[3] Yoshida T, Fukao T, Hasegawa T. Fast detection of tomato peduncle using point cloud with a harvesting robot. Journal of Robotics and Mechatronics, 2018; 30(2): 180–186.
[4] Takenaga. Strawberry harvesting robot for greenhouses. Japan Strawberry Seminar 1998 and Added Information. Tokyo, Japan: The Chemical Daily, 1998; pp.6–11. (in Japanese)
[5] Hayashi S, Takahashi K, Yamamoto S, Saito S, Komeda T. Gentle handling of strawberries using a suction device. Biosystems Engineering, 2011; 109(4): 348–356.
[6] Han K S, Kim S C, Lee Y B, Kim S C, Im D H, Choi H K, et al. Strawberry harvesting robot for bench-type cultivation. Journal of Biosystems Engineering, 2012; 37(1): 65–74.
[7] Xiong Y, Ge Y, Grimstad L, From P J. An autonomous strawberry‐harvesting robot: Design, development, integration, and field evaluation. Journal of Field Robotics, 2020; 37(2): 202–224.
[8] Xiong Y, Peng C, Grimstad L, From P J, Isler V. Development and field evaluation of a strawberry harvesting robot with a cable-driven gripper. Computers and Electronics in Agriculture, 2019; 157: 392–402.
[9] Feng Q C, Wang X, Zheng W G, Qiu Q, Jiang K. A new strawberry harvesting robot for elevated-trough culture. Int J Agric & Biol Eng, 2012; 5(2): 1–8.
[10] De Preter A, Anthonis J, De Baerdemaeker J. Development of a robot for harvesting strawberries. IFAC-Papers OnLine, 2018; 51(17): 14-19.
[11] Cui Y, Gejima Y, Kobayashi T, Hiyoshi K, Nagata M. Study on cartesian-type strawberry-harvesting robot. Sensor Letters, 2013; 11(6-7): 1223–1228.
[12] Yu Y, Zhang K, Liu H, Yang L, Zhang D. Real-time visual localization of the picking points for a ridge-planting strawberry harvesting robot. IEEE Access, 2020; 8: 116556–116568.
[13] Xu L M, Zhang T Z. Influence of light intensity on extracted colour feature values of different maturity in strawberry. New Zealand Journal of Agricultural Research, 2007; 50(5): 559–565.
[14] Zhang L, Ma X, Liu G, Zhou W, Zhang M. Recognition and positioning of strawberry fruits for harvesting robot based on convex hull. In: 2014 Montreal, Quebec Canada July 13–16, ASABE, 2014. doi: 10.13031/aim.20141902612.
[15] Lei H, Huang K, Jiao Z, Tang Y, Zhong Z, Cai Y. Bayberry segmentation in a complex environment based on a multi-module convolutional neural network. Applied Soft Computing, 2022; 119: 108556. doi: 10.1016/j.asoc.2022.108556.
[16] Kai H, Huan L, Zeyu J, Tianlun H, Zaili C, Nan W. Bayberry maturity estimation algorithm based on multi-feature fusion. In: 2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 2021; pp.514–518. doi: 10.1109/ICAICA52286.2021.9498084.
[17] Liu J, Wang X. Plant diseases and pests detection based on deep learning: a review. Plant Methods, 2021; 17(1): 1–18.
[18] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.
[19] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, 2017; pp.2961–2969.
[20] Yu Y, Zhang K, Yang L, Zhang D. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. Computers and Electronics in Agriculture, 2019; 163: 104846. doi: 10.1016/jcompag.2019.06.001.
[21] Kirsten E, Inocencio L C, Veronez M R, da Silveira L G, Bordin F, Marson F P. 3D data acquisition using stereo camera. In IEEE International Geoscience and Remote Sensing Symposium, 2018; pp.9214–9217. doi: 10.1109/igarss.2018.8519568.
[22] Ortiz L E, Cabrera E V, Gonçalves L M. Depth data error modeling of the ZED 3D vision sensor from stereolabs. ELCVIA: Electronic Letters on Computer Vision and Image Analysis, 2018; 17(1): 1–15. doi: 10.5565/rev/elcvia.1084.
[23] Gupta T, Li H. Indoor mapping for smart cities—an affordable approach: Using kinect sensor and ZED stereo camera. In 2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 2017; pp.1–8.
[24] Tagarakis A C, Kalaitzidis D, Filippou E, Benos L, Bochtis D. 3D scenery construction of agricultural environments for robotics awareness. Information and Communication Technologies for Agriculture—Theme III: Decision. Cham: Springer, 2022; pp. 125–142. doi: 10.1007/978-3-030-84152-2_6.
[25] Rahul Y, Nair B B. Camera-based object detection, identification and distance estimation. 2nd International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE), 2018; pp.203–205. doi: 10.1109/icmete.2018.00052.
[26] Zhang Z. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000; 22(11): 1330–1334.