Computers and Electronics in Agriculture 187 (2021) 106237

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

3D global mapping of large-scale unstructured orchard integrating eye-in-hand stereo vision and SLAM

Mingyou Chen [a,c], Yunchao Tang [b,c,*], Xiangjun Zou [a,c,*], Zhaofeng Huang [a], Hao Zhou [a], Siyu Chen [a]

[a] Key Laboratory of Key Technology on Agricultural Machine and Equipment, College of Engineering, South China Agricultural University, Guangzhou 510642, China
[b] College of Urban and Rural Construction, Zhongkai University of Agriculture and Engineering, Guangzhou 510006, China
[c] Foshan-Zhongke Innovation Research Institute of Intelligent Agriculture and Robotics, China

ARTICLE INFO

Keywords:
Fruit-picking robot
3D mapping
Stereo vision
SLAM
Stereo matching

ABSTRACT

Large-scale, high-accuracy, and adaptive three-dimensional (3D) perception are the basic technical requirements for constructing a practical and stable fruit-picking robot. The latest vision-based fruit-picking robots have been able to adapt to the complex background, uneven lighting and low color contrast of the orchard environment. However, most of them have, until now, been limited to a small field of view or rigid sampling manners. Although the simultaneous localization and mapping (SLAM) methods have the potential to realize large scale sensing, it was critically revealed in this study that the classic SLAM pipeline is not completely adapted to orchard picking tasks. In this study, the eye-in-hand stereo vision and SLAM system were integrated to provide detailed global map supporting long-term, flexible and large-scale orchard picking. To be specific, a mobile robot based on eye-in-hand vision was built and an effective hand-eye calibration method was proposed; a state-of-the-art object detection network was trained and used to establish a dynamic stereo matching method adapted well to complex orchard environments; a SLAM system was deployed and combined with the above eye-in-hand stereo vision system to obtain a detailed, wide 3D orchard map. The main contribution of this work is to build a new global mapping framework compatible to the nature of orchard picking tasks. Compared with the existing studies, this work pays more attention to the structural details of the orchard. Experimental results indicated that the constructed global map achieved both large-scale and high-resolution. This is an exploratory work providing theoretical and technical references for the future research on more stable, accurate and practical mobile fruit picking robots.

1. Introduction

A fruit-picking robot equipped with a stereo vision system improves the efficiency and quality of orchard harvesting. At present, most of the stereo vision systems, such as binocular vision system, RGB-D vision system and multi-vision system, are able to adapt to the complex background, uneven lighting and low color contrast of the unstructured orchard (Tang et al., 2020).

The key point of further researches about fruit picking is how to build a compact, coordinated and practical fruit picking robot based on existing high-performance stereo vision systems. There have been some representative cases in this regard. Xiong et al. (2018) constructed a fruit picking robot using two CCD cameras. The main contribution of this work is to provide a method for calculating the oscillation angle of the fruit under natural disturbance. Ge et al. (2020; 2019) demonstrated a strawberry picking robot using RGB-D vision. They proposed a method that can estimate the 3D shape of the fruit from a small number of images. At the same time, they calculated the safe working area of the robot and constructed a complete strawberry picking process. Lin et al. (2019) constructed a robot for guava picking. The innovation of this work is to propose an efficient stereo vision method reconstructing obstacles on trees. Li et al. (2020) trained a high-performance semantic segmentation network, and obtained the spatial position of the tiny fruit stems directly from the original sampled images, which realized the end-to-end perception of the picking point. Silwal et al. (2017) combined agronomic methods to develop a picking robot that can be used in the orchard with ideal fruit distribution. This work is practical and constructive because it gives feasible solutions to achieve a complete

* Corresponding authors.
E-mail addresses: ryan.twain@zhku.edu.cn (Y. Tang), xjzou1@163.com (X. Zou).

https://doi.org/10.1016/j.compag.2021.106237
Received 16 February 2021; Received in revised form 30 April 2021; Accepted 31 May 2021
Available online 7 June 2021

picking process. Wibowo et al. (2017) developed an end-to-end autonomous coconut harvesting robot that can automatically climb coconut trees and detect fruits through a vision system. They designed a novel fixing and cutting mechanism, which is small in size and easy to carry. It provided a reference for the design of other similar lightweight harvesting robots. Zhang et al. (2020) introduced a robot for harvesting fruit with pedicels. They combined the deep instance segmentation network and spatial geometric methods to quickly solve the location of the picking point. They had also designed an end effector that can effectively protect the pulp from damage. This system was successfully verified in a laboratory environment. Williams et al. (2019) proposed a kiwifruit picking robot composed of four robotic arms. The vision system combined deep learning and stereo vision methods to perceive kiwifruit under natural lighting. Their dynamic fruit scheduling system successfully coordinated the various modules. The robot had passed the stability test in the orchard environment and was a complete and usable product.

The above cases had successfully applied a vision-based fruit picking robot in the unstructured environment. However, most of them are limited to the studies of basic issues or employed under local scenes. In fact, this is far from meeting the needs of fruit picking tasks targeting long time span and large scale area. The behavior of picking robots in a dynamic global scale is very complicated, and much more factors must be considered, such as the correlation of mobile platform and robotic arm, the efficiency and completeness of image sampling, the changes in the terrain and obstacles and the differences in distance between different targets, etc. This requires highly integrated hardware, stable 3D vision algorithms, and compact system framework. In addition, more stability tests under large scale orchards are necessary.

In recent years, more and more attention and efforts to visual picking robots have been shifted from local to global scale. Establishing an intelligent, robust, and large-scale picking system has gradually become a consensus in the field of harvesting robots (Tang et al., 2020). Visual SLAM technology can accurately construct a large-scale scene, track the position of the vision system in the global map, and estimate the motion trajectory; therefore, it is a potential solution for completing a large-scale picking task. There have already been some exploratory studies on this topic. For example, Dong et al. (2020) used a RGB-D camera to collect the point clouds of the trees and fruits in an orchard environment and, subsequently, utilized the ORB-SLAM2 framework for motion tracking and mapping. Habibie et al. (2018) collected simulated agricultural data through RGB cameras and lasers and, subsequently, generate a global grid map of the fruits and trees. Shalal et al. (2015) fused camera and laser scanner data to detect tree trunks and construct local orchard maps. The developed algorithm relied only on the onboard sensors of the mobile robot, without adding any artificial landmarks in the orchard. However, this work focused on the distribution of trees on the map, but lacked detailed information about the structure of the fruits. Underwood et al. (2016) proposed a 3D mobile scanning system for almond orchards to reconstruct the distribution map of flowers and fruits and estimate the yield of a single tree. They used a vehicle equipped with lidar and cameras to scan the orchard and construct a forest canopy model database over a long span. Fan et al. (2018) developed a portable and flexible RGB-D SLAM system to estimate the position, height and specific geometric parameters of trees in a large area of forest. As the point cloud resolution of the RGB-D camera dropped sharply in a large range, the point cloud fitting was relatively rough. Nellithimaru and Kantor (2019) proposed a visual SLAM system for fast counting of fruits in vineyards. It combined the classic 3D reconstruction technology with a robust instance segmentation network to obtain a clear 3D structure of the grapes. The structure of this SLAM system was relatively simple, so there may be risks of error accumulation. Ivanov et al. (2020) presented a complete set of technical solutions for outdoor mobile robots including visual perception, navigation and movement. While satisfying the basic functions, the solution guaranteed the stable exchange of internal data of the robot. It had been successfully applied to different complex scenarios and was of great importance for the behavioral logic design of future mobile robots. In addition to the above cases, more different types of agricultural SLAM systems regarding GPS, ultrasonic sensors or inertial measurement units (IMU) had also been proposed and verified (Chen et al., 2018; Gan et al., 2017; Katikaridis et al., 2019).

The SLAM technologies involved in the above research have provided possibilities for building a vision system suitable for large scale environments. However, it should to be pointed out that relevant works are, until now, in their infancy, and there are still a lot of works to be done to realize stable and practical orchard picking application. Capua et al. (2018) pointed out that the performance of the existing SLAM systems cannot fully meet the needs of large-scale agricultural tasks. They believed that by building a new form of SLAM system, optimizing the target tracking algorithm and increasing the type and number of sensors may be able to solve some of the problems. It had also been pointed out in some studies that the original SLAM system may encounter stability problems when applied to agricultural environments (Chebrolu et al., 2017; Gao et al., 2018; Pierzchała et al., 2018). In addition, the huge demand for computing resources of SLAM on a large scale is one of the most important factors restricting its application in the agricultural field (Aguiar et al., 2020; Zhao et al., 2020).

In fact, although SLAM achieves localization and mapping at the same time, it can be figured out that most of the SLAM applications regarding autonomous platforms (such as self-driving, high-altitude cruising, logistics scheduling, etc.) pay more attention to the real-time localization of the mobile platform, rather than the 3D detail of the constructed map. In order to ensure real-time performance, dense spatial features are always ignored, and only a sparse map is constructed. Even if a dense map is constructed in individual cases, it is only used for ensuring better navigations. However, on the contrary, picking in large-scale orchards require high local 3D reconstruction accuracy to better determine the shape, size, maturity and 3D structure of each observable fruits, and then generate a yield map. The more complete the output yield map, the more it can help to estimate the suitable picking area and form a better picking plan for the robot. Therefore, a global mapping system fully compatible to the nature of orchard picking task is urgently needed. Compared with the SLAM system, the stereo vision system pays more attention to the understanding of local structural details, so it can be utilized to improve the resolution of SLAM to construct a higher-performance orchard mapping system.

In this study, a new form of mobile picking robot and a mapping framework integrating stereo vision and SLAM was established. Specifically, an eye-in-hand stereo vision system was built to sample images and generate local maps, and a SLAM system was constructed to estimating the global trajectory of the cameras. Finally, the local maps were stitched according to the global trajectory to form a detailed, wide, 3D orchard map. The main contribution of this work is to combine high-accuracy stereo vision and large-scale SLAM technology, to construct a high-performance global mapping framework compatible to the fruit picking tasks. All the technologies mentioned in this study are necessities of this mapping framework, and they are mutually cooperative.

2. Methodology

2.1. Overall framework

An eye-in-hand mobile robot based on binocular vision and 6-DOF robotic arm was constructed. The system comprises four modules, namely sampling, object detection, local dense mapping, and global trajectory estimation, as shown in Fig. 1.

The sampling module was implemented based on the eye-in-hand system. An efficient hand-eye calibration method was proposed to determine the coordinate correlation between the robotic arm and the camera.

The object detection module and the local dense mapping module

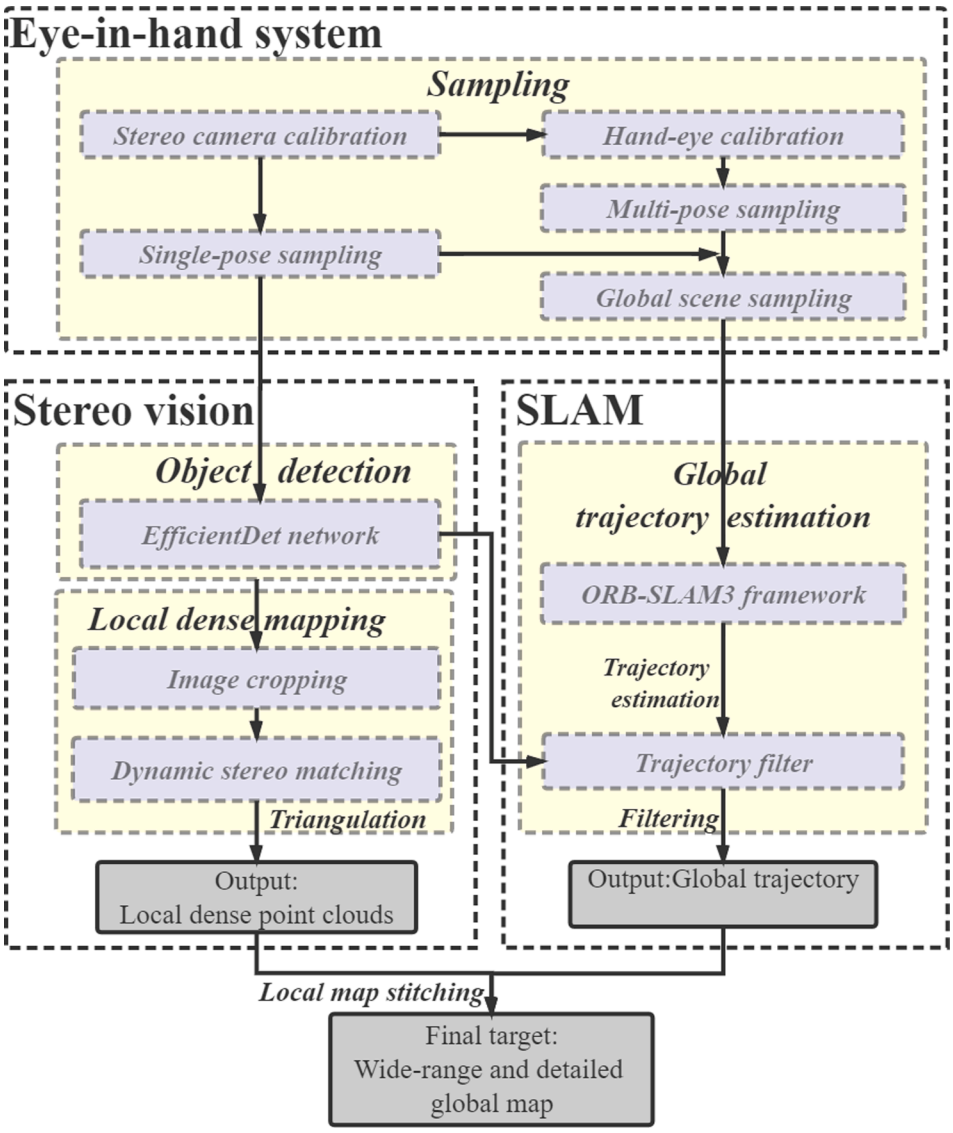


Fig. 1. Schematic of the framework.

together realized high-resolution stereo vision of local scenes. They processed each pair of binocular images acquired by the sampling module one by one. The former is an image preprocessing module, while the latter is the key to stereo vision. A dynamic stereo matching method adapted to complex unstructured orchard was proposed.

The global trajectory estimation module was implemented through a state-of-the-art SLAM system. A trajectory filter was proposed to remove the redundant and invalid trajectories.

At the end, all local dense point clouds were stitched based on the global trajectory to form a global 3D orchard map. It needs to be pointed out that although the SLAM method was used in this study, it only played a role in estimating the global trajectory of the camera, rather than performing real-time localization. In fact, the global map was constructed offline. The global map contains a wealth of ripeness, spatial location and geometry of the fruits, which has the potential to help evaluate the most suitable areas for picking and form an efficient work plan for the robot. Further operations regarding path planning, navigation or picking based on this map will be our future work. Therefore, the platform moved on a predefined route in this study.

2.2. Sampling

Camera calibration and hand-eye calibration were accomplished respectively before conducting image sampling. The former was implemented through classic Zhang's method (2000). There have been many excellent studies on improving the quality of camera calibration (Jia et al., 2015; Ramírez-Hernández et al., 2020). The latter was the focus of this section, that is, how to obtain the relative positional relationship between the flange and camera.

A checkerboard was placed in front of the robot to assist in the hand-eye calibration. The relationship of each coordinate system is shown in Fig. 2; {B} is the checkerboard coordinate system, {G} is the tool coordinate system, {F} is the flange coordinate system, {R} is the robot coordinate system, and {C} is the left-camera coordinate system. According to the basic principles of the coordinate transformation, the coordinate transformation chain was built as follows:

$$ {}^{R}_{F}T^{(i)}\,{}^{F}_{C}T\,{}^{C}_{B}T^{(i)}\,{}^{B}_{R}T = E \tag{1} $$

where E is the unit matrix, and the hand-eye relationship can be described as

$$ {}^{F}_{C}T = \left({}^{C}_{B}T^{(i)}\,{}^{B}_{R}T\,{}^{R}_{F}T^{(i)}\right)^{-1} \tag{2} $$

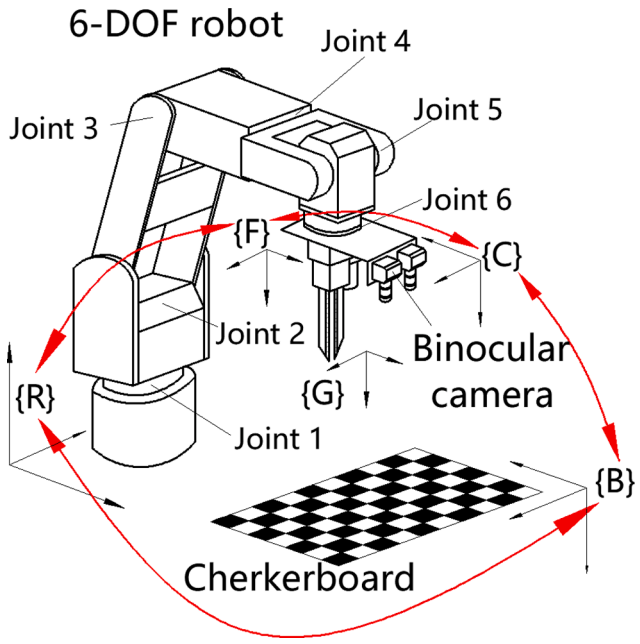


Fig. 2. Coordinate transformation chain in the eye-in-hand system.

where ${}^{B}_{R}T$ is a fixed transformation matrix of {R} relative to {B} that can be calibrated by controlling the origin of {G} to approach the origin, and the X-axis and Y-axis of {B}, respectively. ${}^{C}_{B}T^{(i)}$ is the transformation matrix of {B} relative to {C} corresponding to the *i*-th position of the robot that can be obtained by the least square method (Tang et al., 2019). ${}^{R}_{F}T^{(i)}$ is the transformation matrix of {F} relative to {R}, corresponding to the *i*-th position of the robot that can be simply obtained by forward kinematics.

A robust and reliable result can be estimated from the data corresponding to multiple positions:

$$\widehat{{}^{F}_{C}T} = \frac{1}{N_C}\sum_{i=1}^{N_C}\left({}^{C}_{B}T^{(i)}\,{}^{B}_{R}T\,{}^{R}_{F}T^{(i)}\right)^{-1} \tag{3}$$

where N_C is the number of different poses. In this study, $N_C = 15$.

After determining the hand-eye relationship, the eye-in-hand system constantly changed the field of view to scan wider scenes. Many frames scanned from different views were fused according to hand-eye relationship to form a larger scene (though it should still be called "local sample" because the mobile platform was limited to a local area when scanning). To be specific, the point cloud relative to {C}, calculated for any position of the robot, can be transformed to {R}:

$${}^{R}P^{(i)} = {}^{R}_{F}T^{(i)}\,\widehat{{}^{F}_{C}T}\,{}^{C}P^{(i)} \tag{4}$$

where ${}^{C}P^{(i)}$ is the coordinate of the point cloud obtained from the *i*-th position, relative to {C}, and ${}^{R}P^{(i)}$ is the coordinate of the point cloud obtained from the *i*-th position, relative to {R}.

As the platform moved, the local samples at each location in the orchard finally constituted a global sample.

2.3. Object detection

This is a preprocessing module of stereo vision. The binocular images were processed by a robust deep object detection network to remove the complex backgrounds and obtain the bounding boxes of the fruits in the images.

EfficientDet (Tan et al., 2020) is a new-generation object detection network developed based on the EfficientNet (Tan and Le, 2019), as shown in Fig. 3. EfficientNet is the structural backbone of EfficientDet. The bidirectional feature pyramid network (BiFPN) streamlines the structure of the PANet (Liu et al., 2018) and utilizes shortcuts to cooperate with the learnable weights to improve the efficiency of the multi-feature fusion processes. The network uses the compound scaling method to automatically determine the optimal scale of the backbone, BiFPN, and box/class prediction network, thereby achieving high speed and accuracy under the condition of the low resource supply. This is highly advantageous in mobile orchard picking tasks with complex environments and limited computing resources. The EfficientDet network is mainly divided into seven versions from D1 to D7, whose resource consumption and performance increase sequentially. In this study, EfficientDet-D3, which effectively accommodates the trade-off between resource consumption and performance, has been selected.

Fig. 4 shows the performance of the trained EfficientDet network on the sampled images, where the pixels outside the bounding boxes are all assigned a value of zero.

2.4. Local dense mapping

In this section, the 3D local orchard map with high resolution and rich details was obtained.

The input of this module is the binocular image with background removed by the above EfficientDet network. In fact, the bounding box described in section 2.3 only contained a rectangle area of the fruit. However, in order to obtain the dense map, stereo matching must be conducted, that is, realizing accurate matching between all pixels in the

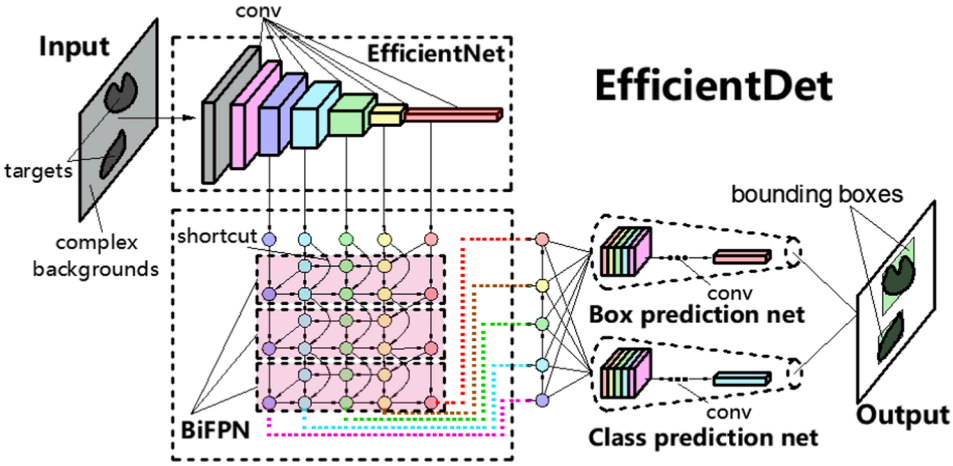


Fig. 3. Structure of the EfficientDet network.

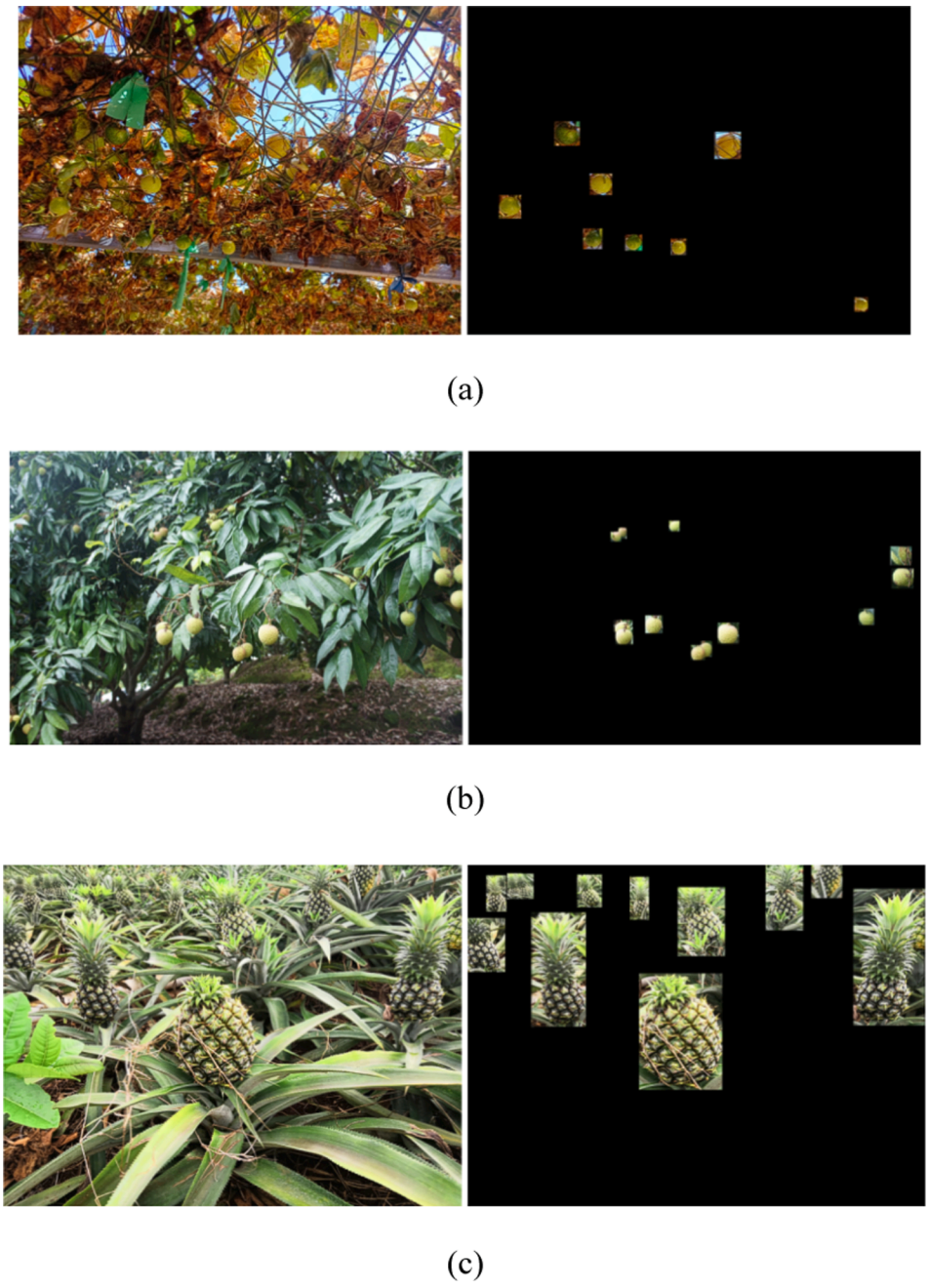


Fig. 4. Schematic of the performance of the trained EfficientDet-D3: (a) Passion fruit; (b) Litchi; (c) Pineapple.

left and right bounding boxes. A dynamic stereo matching method was proposed in this section. The process of the local dense mapping module is detailed in Fig. 5.

In our previous research (Chen et al., 2020), the difficulty of stereo matching in an unstructured banana orchard was demonstrated, and an adaptive stereo matching algorithm was designed to solve these problems. However, unlike the fixed scenario, for completing the tasks using mobile robots, increased number of uncertainties need to be considered: for example, severe mechanical vibrations, movements of targets, and complex ambiguities during stereo matching. This implies higher requirements for the speed, accuracy, and robustness of the stereo matching processes.

Therefore, a dynamic stereo matching method for multiple dynamic targets was proposed. The positional information of the bounding boxes obtained using the object detection network was utilized for image cropping and determining the searching parameters of the local stereo matching, as shown in Fig. 6. The bounding boxes were numbered from top to bottom and from left to right. For bounding boxes with the same number in the binocular images, the smallest interval that contains two boxes is obtained:

$$[v_{TX}, v_{BX}] \tag{5}$$

where v_{TX} is the minimum of v_{LT} and v_{RT}, and v_{BX} is the maximum of v_{LB} and v_{RB}. At the same time, v_{LT}, v_{LB}, v_{RT}, and v_{RB} are the row numbers of the top and bottom edges of the left and right bounding boxes, respectively; u_l, u_r, v_l and v_r are the abscissa label and ordinate label of the left and right image coordinate system, respectively.

The above interval determined the bar-shaped subimages containing single objects, as shown in Fig. 6. Stereo matching was subsequently applied to these subimages separately to obtain local disparity maps. If the coordinates of the center of the left and right boxes are $C_L(u_L,v_L)$ and

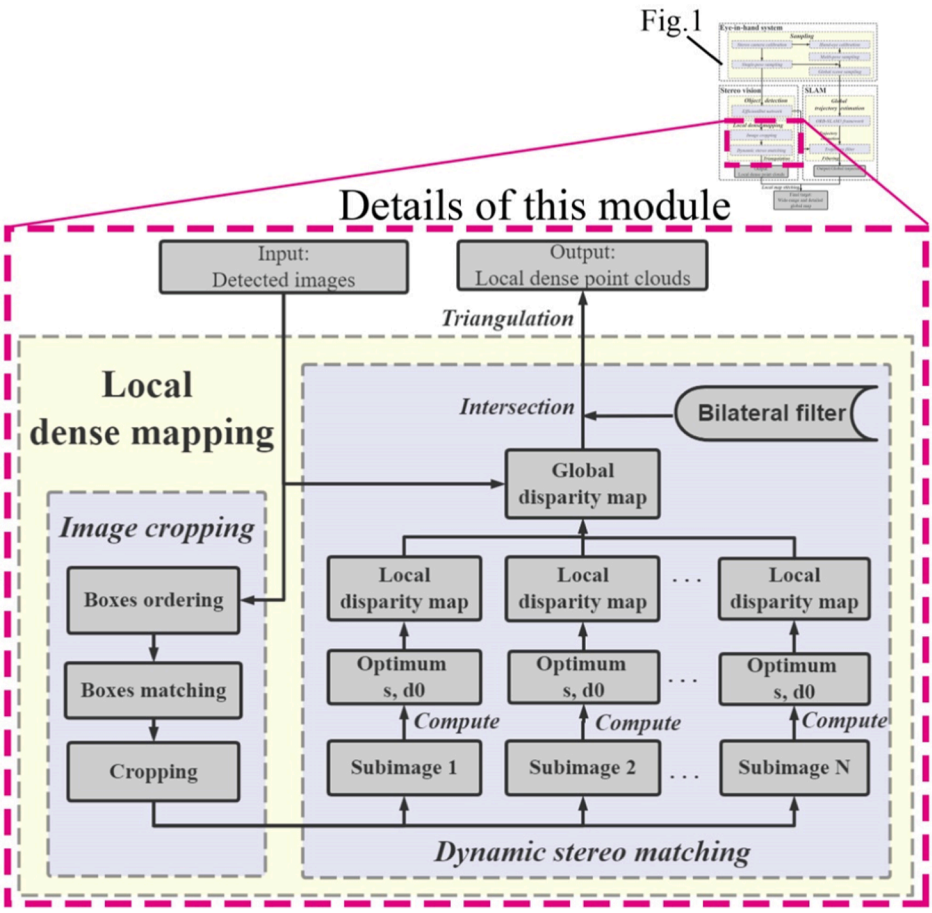


Fig. 5. Local dense mapping module.

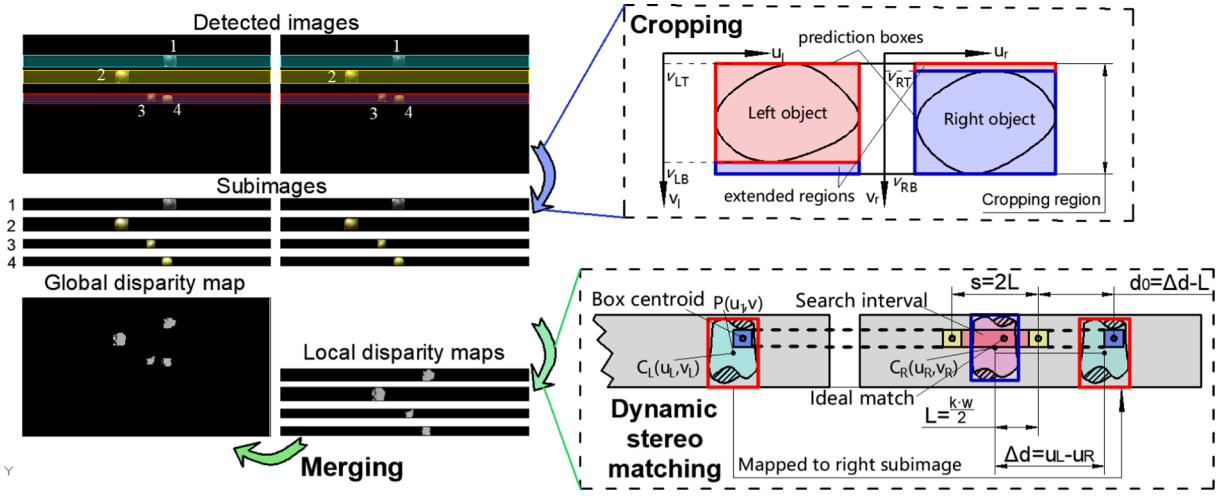


Fig. 6. Schematic of dynamic stereo matching.

$C_R(u_R,v_R)$, respectively, the optimal disparity should be close to

$$\Delta d = u_L - u_R \tag{6}$$

Therefore, for point $P(u_1,v)$ on the left image, its ideal matching point on the right image should be located near $P_0(u_1\text{-}\Delta d, v)$. P_0 was regarded as the midpoint of the search interval based on the above analysis, and the width w of the right box was considered as the reference length of the search interval. The search parameters of stereo matching are described as follows:

$$d_0 = \Delta d - k\cdot\frac{w}{2} \tag{7}$$

$$s = kw \tag{8}$$

where d_0 is the minimum for searching a disparity, s is the length of the search interval, L is half the length of the search interval, and k is a threshold that controls the width of the search interval. In this study, $k = 1.0$. These searching parameters ensure that the search interval can be

automatically adjusted according to the 2D positions of the targets.

After performing stereo matching on all the subimages, the local disparity maps were reverted to the initial position (prior to cropping) to obtain a global disparity map; this was thereafter intersected with the input left single-color background images to remove the noise caused by any mismatch. Finally, a bilateral filter was applied to smoothen the aliasing. After obtaining the disparity map, the local point cloud was generated simply by triangulation operations.

2.5. Global trajectory estimation

The global trajectory estimation module processed all the frames collected by the mobile platform during the entire sampling process and generated the global camera trajectory based on the ORB-SLAM3 framework. According to this trajectory, all local dense point clouds calculated by stereo vision system were stitched into a whole, that is, a global 3D map.

The global trajectory estimation module is detailed in Fig. 7.

The process pipeline of the ORB-SLAM3 consists of tracking, local mapping, loop closure, and map merging, which are all connected by a map module called ATLAS. ATLAS contains a wealth of visual vocabulary, a recognition database, and a large number of sparse maps that are generated during the mapping process. These data will be repeatedly optimized and combined during the mapping process to obtain a larger trajectory range. ORB-SLAM3 is able to achieve high-speed, robust, and accurate trajectory estimation in an unstructured environment, and it is currently one of the most robust and efficient SLAM systems.

With the global sample as the input, the SLAM system outputs the global camera trajectory, containing the position node of the camera relative to the global coordinate system when each frame of binocular images was captured. Each position node in the output camera trajectory has a one-to-one correspondence with the binocular image sequence, such that the dense point clouds obtained by local triangulation can be incorporated as a detailed global map.

The ratio of the area of the largest bounding box to the area of the original image was calculated using the following equation:

$$\lambda = \frac{s_b}{w_i \cdot h_i} \times 100\% \tag{9}$$

where s_b is the area of the largest bounding box, and w_i and h_i are the width and height of the image, respectively. If $0 < \lambda \leq 5\%$, the area of the fruit in the image is relatively small, and the accuracy of the fruit point cloud obtained by local mapping is unsatisfactory. At this time, the current frame and its corresponding pose in the trajectory is deleted to obtain a filtered and simplified trajectory.

The filtered trajectory was noted as a valid trajectory, and the number of pose nodes contained therein was denoted as N_K. If the local point cloud corresponding to the i-th frame is ${}^{C}P_{Ki}$ and the corresponding pose in the trajectory is $T_i \in SE(3)$, the ultimate global dense point cloud can be described by the following equation:

$${}^{G}P = T_1 {}^{C}P_{K1} \cup T_2 {}^{C}P_{K2} \cup \cdots \cup T_{N_K} {}^{C}P_{KN} \tag{10}$$

3. Experiment

This section is divided into two parts: local experiments and global experiments. In the local experiments, the performance of the sampling module, object detection module and local dense mapping module was verified under the local scenes of the orchard, respectively. In the global

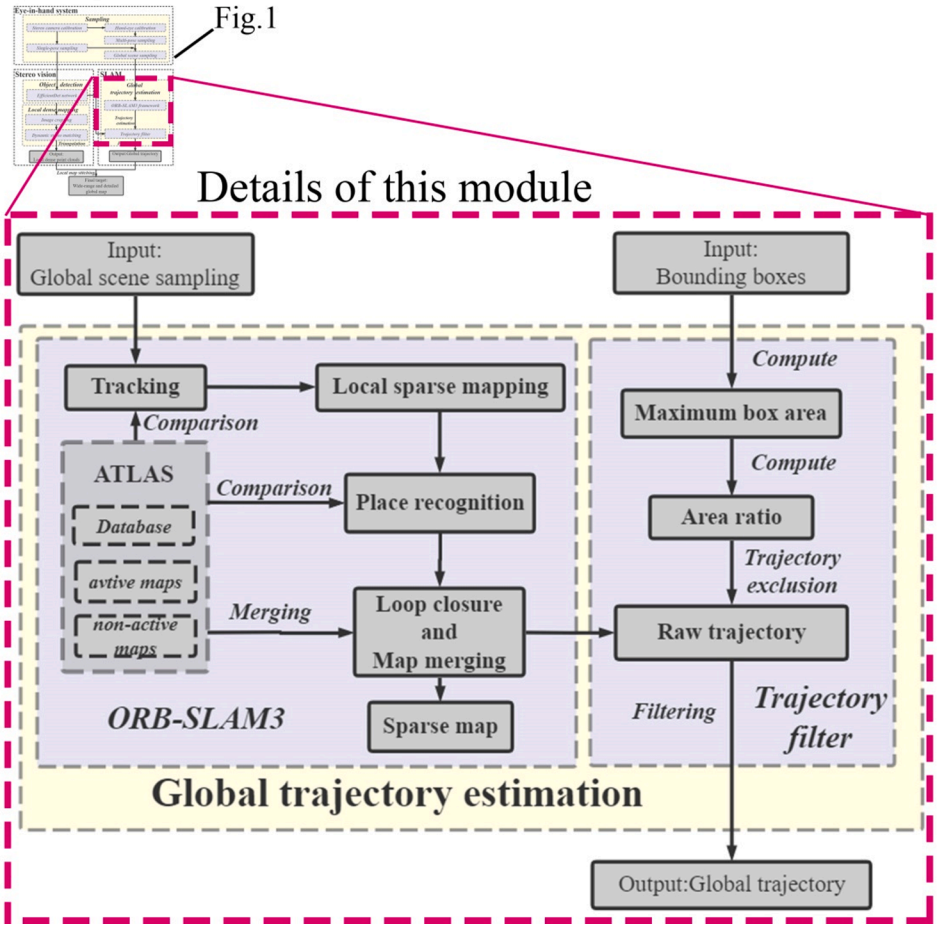


Fig. 7. Global trajectory estimation module.

experiments, detailed and broad global 3D maps of the orchards were obtained, facilitating the validation of the performance of the entire mapping system.

3.1. Local experiments

3.1.1. Experiment for evaluating the accuracy of the eye-in-hand system

The eye-in-hand system is the key to realize sampling. The performance of the eye-in-hand system is affected by the vision algorithm error, hand-eye calibration error, and robotic movement error. The hand-eye calibration error itself is affected by the other two errors, and therefore, the combined error is complicated and coupled. In this study, this error is called "compounding positioning error". The error propagation is shown in Fig. 8.

In this experiment, the compounding positioning error of the checkerboard corners were collected to estimate the accuracy of the eye-in-hand sampling system. The checkerboard featured (11 × 8 =) 88 corners and was placed horizontally, approximately 500 mm in front of the robot, as shown in Fig. 9.

The robot drove the camera to capture the checkerboard from different positions. The 3D coordinates of the corners relative to {R} were calculated as

$$ {}^{R}P_{ij} = {}^{R}_{F}T_i\,{}^{F}_{C}T\,{}^{C}P_{ij} \tag{11} $$

where ${}^{C}P_{ij}$ and ${}^{R}P_{ij}$ are the coordinates of the j-th corner for the i-th pose relative to {C} and {R}, respectively; ${}^{F}_{C}T$ is the output of the hand-eye calibration; and ${}^{R}_{F}T_i$ is the coordinate transformation matrix between {F} and {R} for the i-th pose. For the j-th corner, the average of the 3D coordinates, sampled for different positions, was considered as the optimal estimation, and the average Euclidean distance between each sample and the optimal estimation was considered as the compounding positioning error:

$$ \widehat{d_j} = \frac{1}{N_R}\sum_{i=1}^{N_R}\left\| {}^{R}P_{ij} - \frac{1}{N_R}\sum_{i=1}^{N_R}{}^{R}P_{ij}\right\|_2 \tag{12} $$

where N_R is the number of sampling poses of the robot.

According to the distance between the flange and the checkerboard (noted as "capturing distance"), the experiment was divided into 12 groups. The capturing distance in each group was about z = 380 mm, 430 mm, 480 mm, 530 mm, 580 mm, 630 mm, 680 mm, 730 mm, 1000 mm, 1500 mm, 2000 mm, and 2500 mm. For each distance, the camera captured the checkerboard from 90 different poses ($N_R = 90$), and then calculated the compounding positioning error of each corner according to equation (12). Taking the data at z = 380 mm as an example, its compounding positioning error is visualized in Fig. 10. The abscissa and ordinate represent the column number and row number of the corner point on the 11 × 8 checkerboard, respectively.

The proposed hand-eye calibration method was compared with the most widely used hand-eye calibration methods in the robotics field. There were 5 methods for comparison. For convenience, the author's names are used to indicate the methods, namely TSAI (Tsai and Lenz, 1989), PARK (Park and Martin, 1994), HORAUD (Horaud and Dornaika, 1995), ANDREFF (Andreff et al., 1999), and DANIILIDIS (Daniilidis, 1999). Our method is denoted as "Ours". Based on formula (12), the mean and maximum of the compounding positioning errors of 88 corners were calculated as indicators for comparison, and the result is shown in Fig. 11.

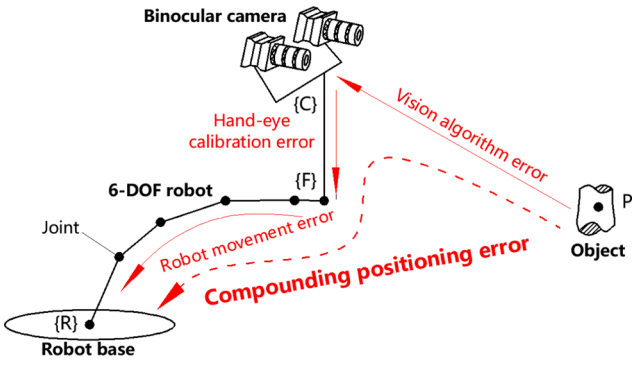


Fig. 8. Error propagation of the eye-in-hand sampling system.

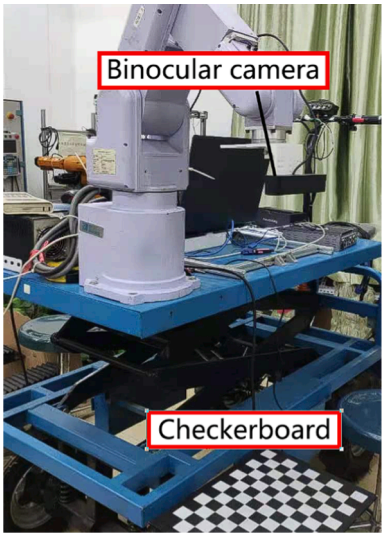


Fig. 9. Physical image of the experimental equipment.

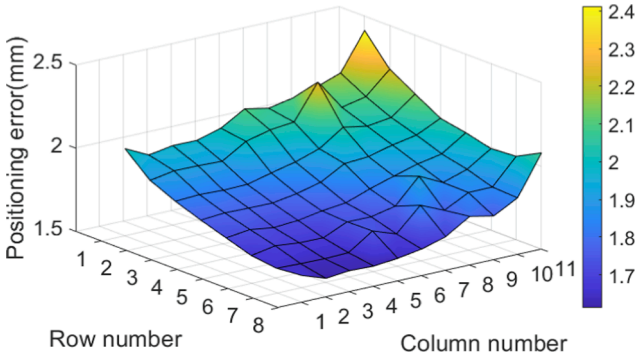


Fig. 10. Visualization of compounding positioning errors.

It can be observed from Fig. 11 that our hand-eye calibration method achieved the best performance in almost all cases. The data of our method is detailed in Table 1.

Table 1 indicates that when z < 1 m, the compounding positioning error of the eye-in-hand system calibrated by the proposed method did not exceed 6 mm; when 1 m < z < 2.5 m, the error did not exceed 15 mm. For large-scale maps, the relative error is small. As what has been presented in section 2, our global orchard map was constructed based on many local maps generated at short capturing distances. In fact, 2.5 m is far enough for local sampling. Therefore, it is reasonable to believe that the calibrated eye-in-hand system reached satisfactory relative error and matched the requirement of our mission.

3.1.2. Experiment for evaluating the performance of the object detection network

A total of 1250 image samples, including banana central stocks, pineapple, papaya, litchi, and camellia oleifera, were collected from unstructured orchards and labeled. The number of each fruit image was

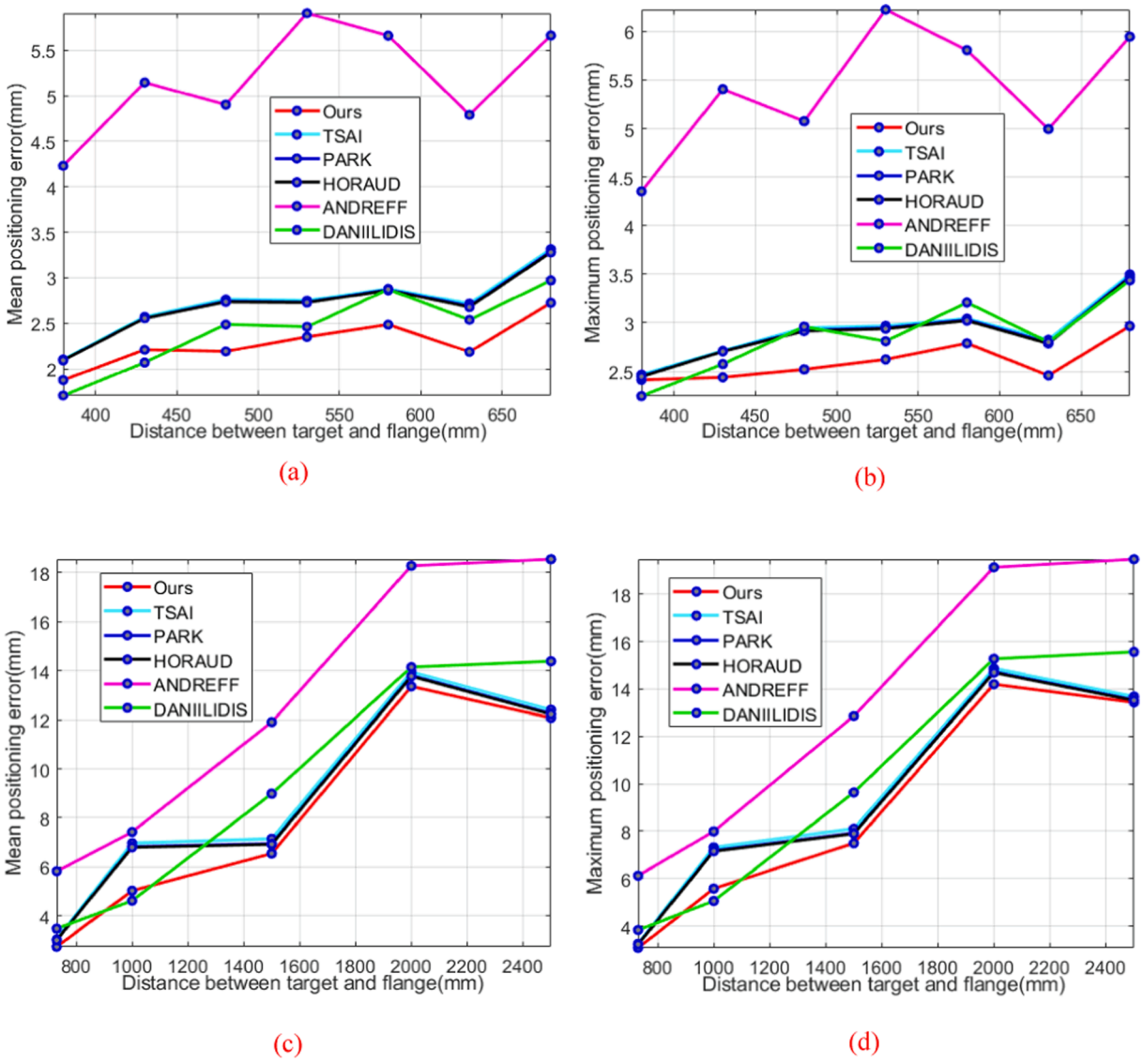


Fig. 11. Error indicators under different calibration methods and capturing distances: (a) 380 mm < z < 680 mm, mean error; (b) 380 mm < z < 680 mm, maximum error; (c) 730 mm < z < 2500 mm, mean error; (d) 730 mm < z < 2500 mm, maximum error.

Table 1
Error indicators corresponding to our method.

Capturing distance (mm)	380	480	580	680	1000	1500	2000	2500
Maximum error (mm)	2.41	2.52	2.79	2.97	5.57	7.49	14.21	13.44
Mean error (mm)	1.88	2.19	2.48	2.72	5.03	6.55	13.36	12.07

250. After the samples were randomly shuffled, they were divided into the training, validation, and test sets, in the ratio of 81:9:10, accounting for 1013, 112, and 125 samples, respectively. The EfficifentDet network was trained using Pytorch. After the training was completed, the forward propagation operation was performed on the test set to obtain the predicted bounding boxes. Thereafter, the precision–recall (P-R) curve

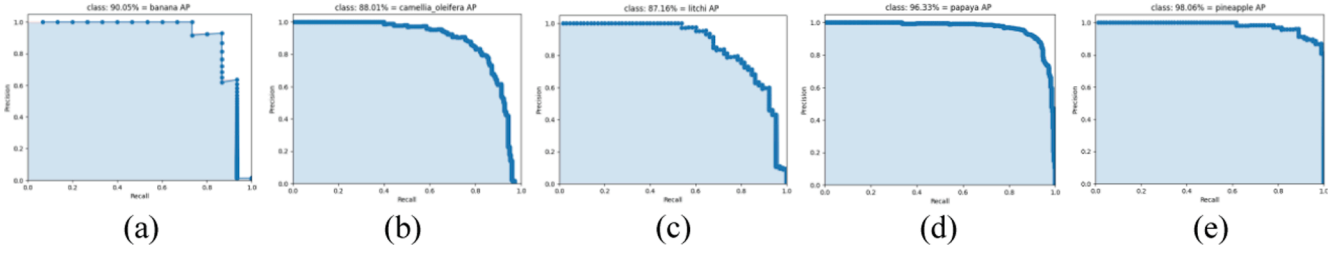


Fig. 12. P-R curve: (a) Pineapple; (b) Camellia oleifera; (c) Litchi; (d) Papaya; (e) Banana central stocks.

corresponding to each class was drawn, as shown in Fig. 12. Finally, the average precision (AP) of each category and the mean average precision (mAP) were calculated:

$$AP = \int_0^1 f_j(r)dr \tag{15}$$

$$mAP = \frac{\sum_{j=1}^{N_k} \int_0^1 f_j(r)dr}{N_k} \tag{16}$$

where $N_k = 5$ is the number of classes, and $f_j(r)$ is the P-R curve corresponding to the j-th class. The AP statistics is shown in Fig. 13. The mAP of the network was 0.9192, indicating that the trained EfficifentDet network is capable of adapting to the scene of unstructured orchards and accurately completing the task of fruit detection.

Meanwhile, the forward propagation time cost for each frame was recorded, as shown in Fig. 14, with most falling within the interval of 0.3–0.4 s; the average is 0.33 s.

3.1.3. Experiment for evaluating the performance of dynamic stereo matching

The binocular images of the banana central stocks and camellia oleifera fruit, whose background had already been removed by EfficientDet, were considered as the object of this experiment.

The traditional stereo matching method and the proposed dynamic stereo matching method were applied to calculate the disparity of the specific positions on the surface of the objects. Both of them were based on the classic semi-global matching method (Hirschmuller, 2005). The former utilized fixed search intervals, while the latter utilized a specially designed dynamic search interval. The disparities were finally compared with the standard disparities manually measured using Photoshop CC.

The experimental objects included three groups of images of bananas and camellia oleifera, denoted as group B_1, B_2, B_3, O_1, O_2, and O_3. To verify the stability of the method, the cameras with different resolutions, focal lengths, and field of views were selected for sampling. Groups B1–B3 were sampled by 2 MV-VD120SC cameras with a resolution of 1280 × 960. Groups O1–O3 were sampled by a ZED2 binocular camera with a resolution of 1920 × 1080. Each group contained 10 pairs of repetitive binocular images, and some of the images are shown in Fig. 15. For each group, the disparity at five specific sampling points on the surface of each object was collected through visual and manual measurements. The average absolute disparity error was subsequently calculated. Finally, the average of 10 repeated samples was obtained as

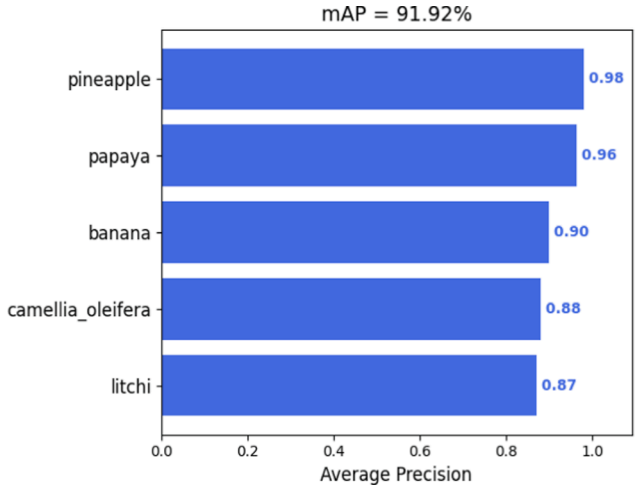


Fig. 13. Statistics of the average precision (mAP: mean average precision).

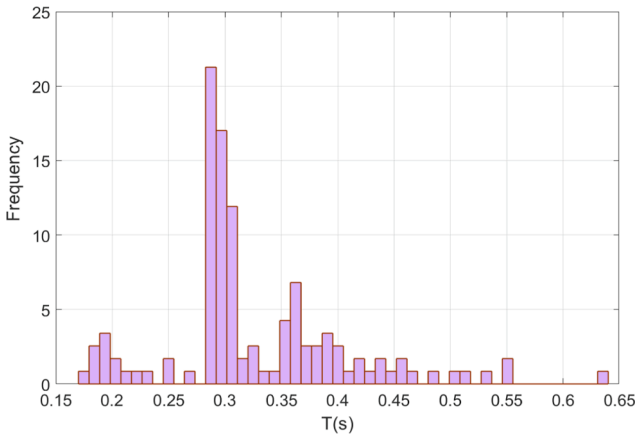


Fig. 14. Distribution of forward propagation time of a single frame.

$$\widehat{e_d} = \frac{1}{N_W} \sum_{i=1}^{N_W} \sum_{j=1}^{N_S} \frac{|d_{ij} - d_j|}{N_S} \tag{17}$$

where $\widehat{e_d}$ is the error of the group, d_{ij} is the disparity of the j-th sampling position of the i-th repeated sample, d_j is the manual measurement of the j-th sampling position, $N_s = 5$ is the number of sampling positions of each object, and $N_w = 10$ is the number of repeated samples. The sampling points on the left subimages are highlighted in red in Fig. 15. The center sampling point is the center of the left bounding box, and the remaining four sampling points are located at the top and bottom of and to the left and right of the center. The distances between the four outer sampling points to the center sampling point are 10% of the height or width of the bounding boxes. For better visualization, the left and right subimages are projected onto the same canvas simultaneously, whereas the transparency of the right sub-image is set to 50%.

We designed 12 different stereo matching searching parameters for the traditional methods to span across different depth ranges. For simplicity, the stereo matching search interval determined by the fixed parameters is called the fixed interval, whereas that of the dynamic stereo matching parameters is called the dynamic interval. The disparity errors are listed in Table 2.

Table 1 reveals that if a target does not fall within the fixed interval, the stereo matching will fail at the beginning of the procedure. In contrast, because the proposed dynamic stereo matching method determines the matching parameters according to the locations of the targets, the targets can be made to always fall within the search interval. Furthermore, it can be observed that for each group, the absolute disparity error corresponding to the dynamic interval was lower than that corresponding to the fixed interval; this implies that the dynamic interval comprehensively covered the targets, thereby avoiding interference due to the complex backgrounds and mismatches.

For the dynamic interval and five fixed intervals that had successfully completed stereo matching, their average matching time was recorded and visualized (Fig. 16), along with the disparity errors shown in Table 2. As shown in Fig. 16(b), the calculation speed corresponding to the dynamic interval was significantly higher than that of the fixed interval. This is because of the flexible length of the interval as well as the reduction in the calculation resulting from the subimage cropping strategy.

3.2. Global experiments

The global experiment was conducted successively in the passion fruit orchard, citrus orchard, guava orchard, and the green jujube orchard. Photographs of the parts of the orchards are shown in Fig. 17. The passion fruit orchard, citrus orchard, and guava orchard were sampled

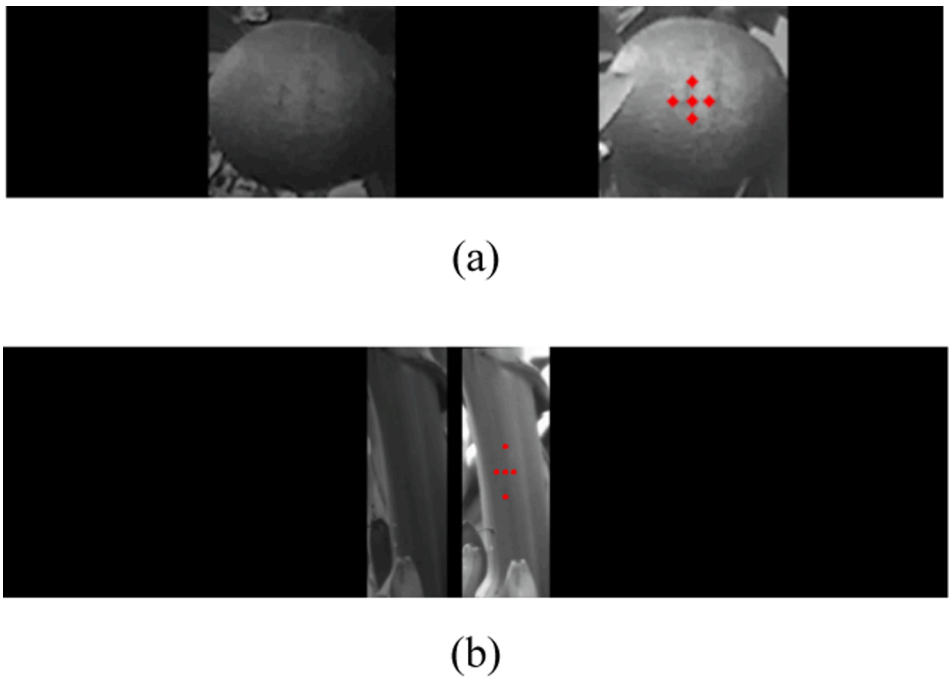


Fig. 15. Subimages and corresponding sampling points: (a) Camellia oleifera; (b) Banana central stocks.

Table 2
Absolute disparity errors.

Searching parameters (pixel)	Group B_1 (pixel)	Group B_2 (pixel)	Group B_3 (pixel)	Group O_1 (pixel)	Group O_2 (pixel)	Group O_3 (pixel)
$d_0 = 0$, s = 64	/	/	/	/	/	/
$d_0 = 0$, s = 256	0.57	4.01	1.71	5.08	1.94	4.97
$d_0 = 0$, s = 512	0.57	4.01	1.71	7.37	4.22	4.97
$d_0 = 192$, s = 64	/	/	/	/	9.58	9.15
$d_0 = 192$, s = 256	/	/	/	/	/	9.03
$d_0 = 192$, s = 512	/	/	/	/	/	13.15
$d_0 = 384$, s = 64	/	/	/	/	/	/
$d_0 = 384$, s = 256	/	/	/	/	/	/
$d_0 = 384$, s = 512	/	/	/	/	/	/
$d_0 = 768$, s = 64	/	/	/	/	/	/
$d_0 = 768$, s = 256	/	/	/	/	/	/
$d_0 = 768$, s = 512	/	/	/	/	/	/
Dynamic d_0, s	0.57	3.56	1.70	2.89	1.32	3.62

"/" denotes that the object falls outside the search interval.

during the day with sufficient lighting, while the green jujube orchard was sampled in the evening with weak lighting. The global maps of various orchards were constructed and the qualities of the maps were evaluated. This experiment involves all the methods described in this study and is a comprehensive investigation of the mapping performance of the proposed mobile robot system.

3.2.1. Equipment and initialization

A mobile robot platform, equipped with an eye-in-hand binocular vision system was built. The platform consisted of a lightweight 6-DOF robotic arm, 4-wheel mobile platform, ZED2 binocular camera, large-capacity lead-acid battery, and a laptop (Intel core i5-9400 CPU, Nvidia GTX 1660Ti GPU, 16 GB ddr4 RAM, Ubuntu 18.04). In addition, a checkerboard (0.3 mm manufacturing precision) for offline calibration and a digital vernier caliper (0.01 mm measurement precision) for measuring standard experimental data were also included.

As shown in Fig. 18, the robotic arm was fixed on the mobile platform. The height of the robot arm from the ground can be changed through the hydraulic system at the bottom. The camera was attached to the end of the robotic arm and the images were continuously sampled. The laptop was used to process visual data and communicate with the robotic arm through the RS232 serial port. The robotic arm, computer, and camera were powered with the removable battery on the mobile platform, whereas the mobile platform itself was powered with its internal battery.

Prior to the experiment, all the angles of the joints of the robotic arm were reset to 0, and the base of the robotic arm was adjusted to approximately 600 mm away from the ground. The mobile platform moved around the orchard at a speed of approximately 1.5 m/s, and the camera continuously sampled the incoming images. The resolution of the camera was 1920 × 1080; the sampling frequency was approximately 5 fps, and the exposure time and gain were set automatically.

In the process of sampling, the EfficientDet network detected the latest sampled binocular images in real time. If the fruits were not detected within 60 consecutive frames, the movement of the mobile platform would be suspended. Thereafter, Joint 1 of the robot would drive the camera to scan a larger field of view within the range, [-90°, 90°], in steps of 30°, until a detectable fruit appears within the field of view. If the fruits were still not detected, the current position was abandoned and the platform continued to move forward. The sampling process of the mobile platform is shown in Fig. 19 .

3.2.2. Experiments and results

A control group was introduced for comparison with our global mapping system (experimental group). The map of the control group was constructed based on a high-performance SLAM pipeline provided the ZED2 commercial product, whereas the map of the experimental group was constructed using our mapping framework. The resolution of

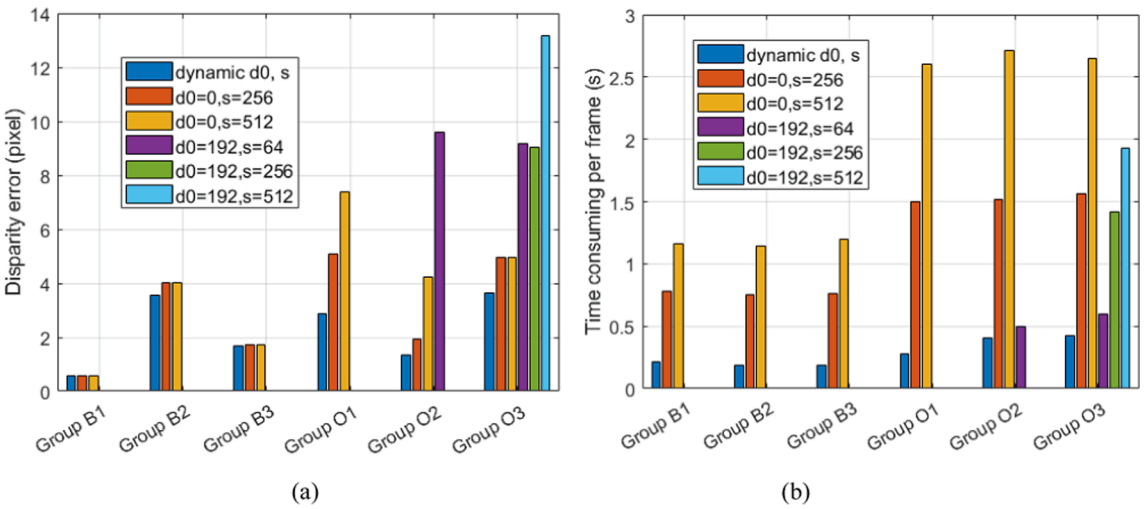


Fig. 16. Statistics of stereo matching performance: (a) Absolute error of disparity; (b) Average time consumption.

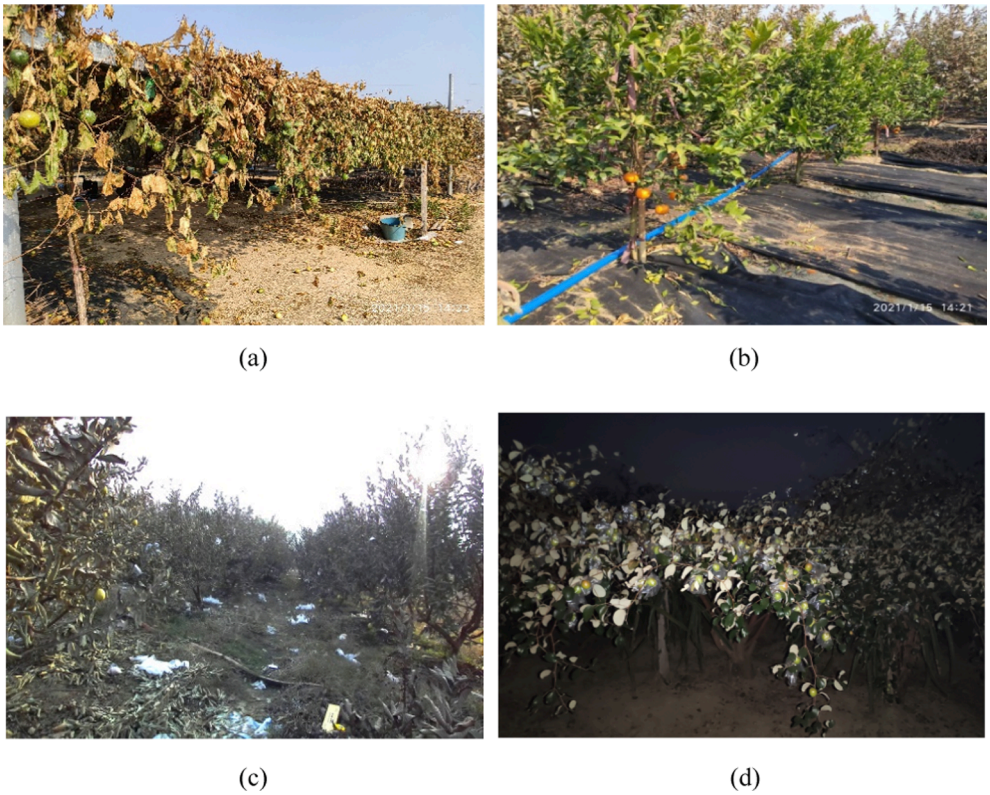


Fig. 17. Photographs of the (a) passion fruit orchard, (b) citrus orchard, (c) guava orchard, and (d) green jujube orchard. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the images collected in both groups was 1920 × 1080. It can be qualitatively determined that the spatial resolution of the control group is significantly lower than that of the experimental group (see Fig. 20).

In addition to the intuitive observation, quantitative experiments were also performed. Unlike indoor scenes with a fixed size, the standard map data for an outdoor orchard were not easy to collect; therefore, the diameter of the fruit and the distance between the fruits were considered to represent the accuracy of the constructed maps. The standard values of the fruit diameters and distances were measured using a digital vernier caliper, whereas the corresponding visual measurement values were directly measured from the global point clouds. Limited to the performance of the mapping system, not all selected fruits can be found in the global map. For example, as shown in Fig. 21, the 3D information of Fruit 1 in the map is lost, therefore, it has been recorded as "lost". In addition, when the point cloud of a fruit is deformed, overlaps with the leaves, or has a large surface undulation, it is also recorded as "lost," and the rest are recorded as "found".

The ratio of the number of fruits marked as "found" to the total number of selected fruits was defined as Recall. It was considered as one of the indicators for evaluating the performance of the orchard mapping system:

$$\eta = \frac{N_F}{N_A} \times 100\% \tag{18}$$

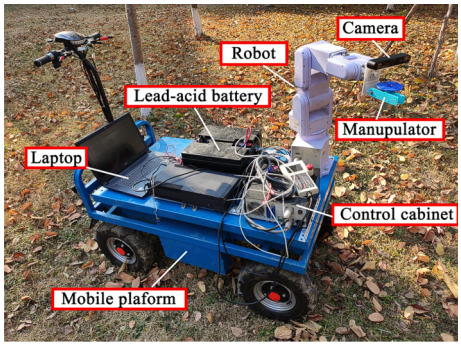


Fig. 18. Photograph of the mobile platform.

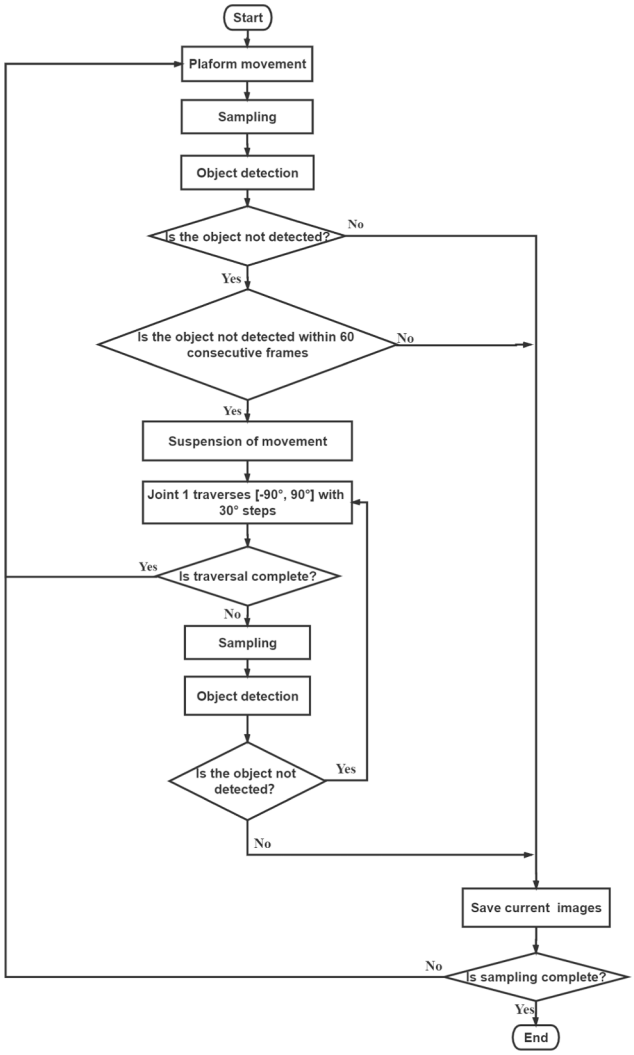


Fig. 19. Sampling process.

where *NF* is the number of fruits marked as "found" and *NA* is the total number of selected fruits. The number of randomly selected fruits in the passion fruit, citrus, guava, and green jujube orchards were 14, 14, 6, and 9, respectively; the corresponding recalls are shown in Fig. 22.

For the fruits marked as "found," the diameters and the distances between the fruits were measured in CloudCompare as visual measurements. Notably, D_i and d_j are the visual measurements of the *i*-th fruit diameter and *j*-th fruit distance on the map, respectively. D_{i0} and d_{j0} are the standard values of the *i*-th fruit diameter and *j*-th fruit distance

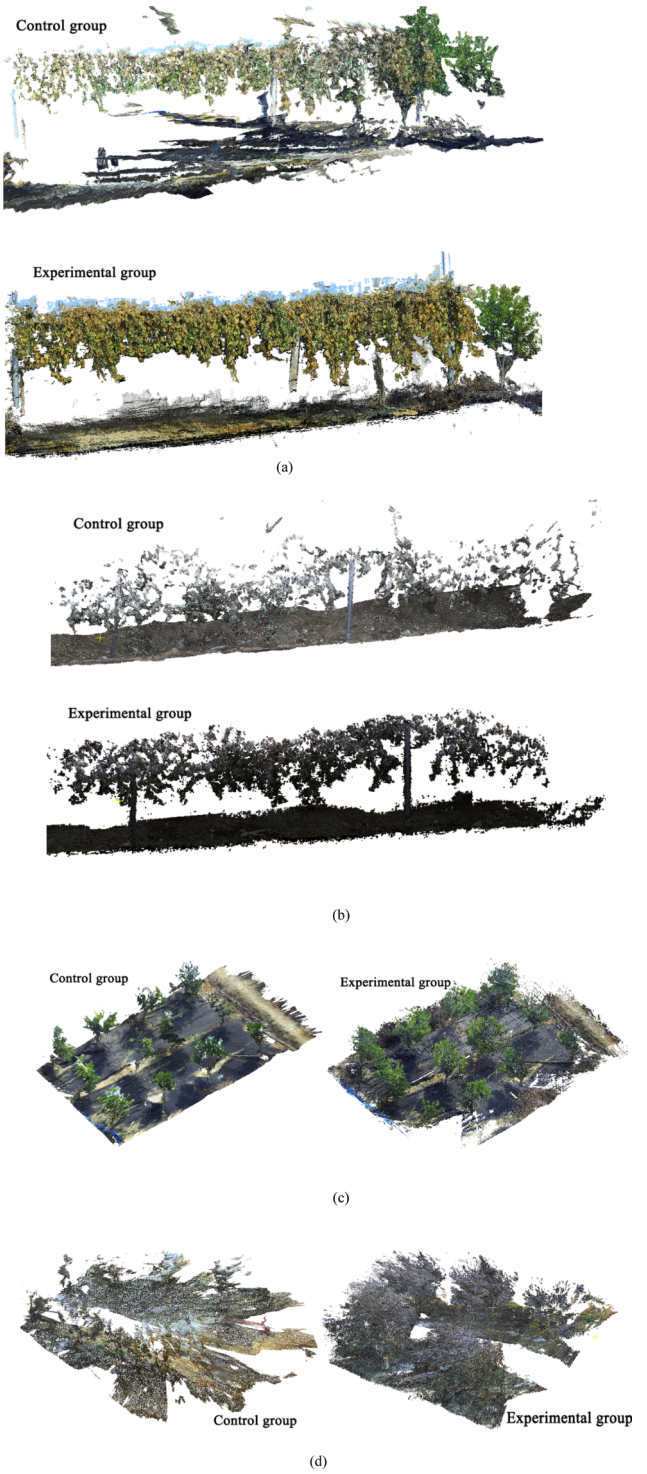


Fig. 20. Control group and experimental group: (a) Passion fruit orchard; (b) Green Jujube orchard; (c) Citrus orchard; (d) Guava orchard. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

obtained using digital vernier calipers, respectively. The average absolute measurement error was calculated as follows:

$$\Delta\overline{D} = \frac{1}{N_F'}\sum_{i=1}^{N_F'}|D_i - D_{i0}| \tag{19}$$

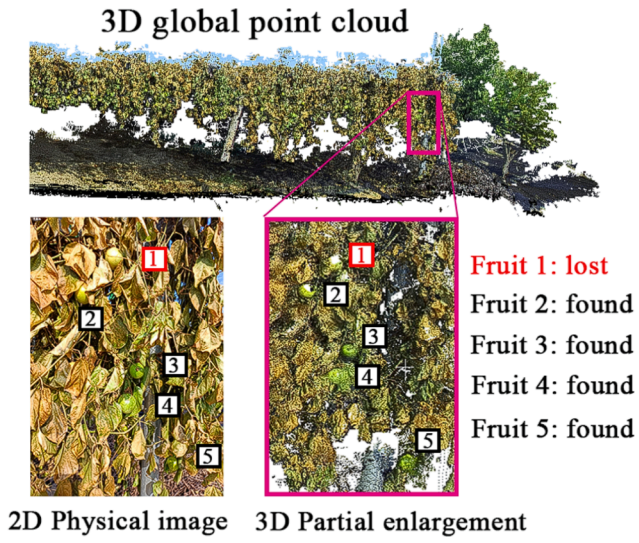


Fig. 21. Illustration of labelling the fruits as "lost" and "found".

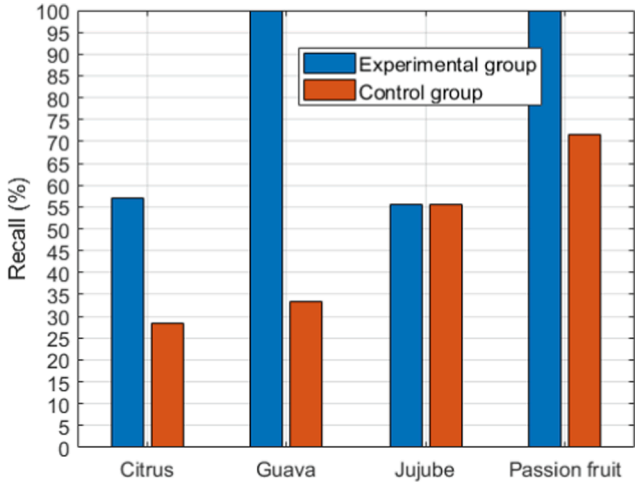


Fig. 22. Recall of the selected fruits.

$$\Delta\bar{d} = \frac{1}{N_d'}\sum_{j=1}^{N_d'}|d_j - d_{j0}| \tag{20}$$

where $\Delta\bar{D}$ and $\Delta\bar{d}$ are the average absolute measurement errors of the fruit diameter and fruit distance, respectively. N_F' is the number of fruits marked as "found" in the global point cloud. N_d' is the number of fruit distances obtained for the fruits marked as "found." For different orchards, $\Delta\bar{D}$ and $\Delta\bar{d}$ were calculated separately, as shown in Fig. 23(a) and Fig. 23(b), respectively.

Fig. 23(a) and Fig. 23(b) indicate that the error of the experimental group is significantly lower than that of the control group. This is highly probable because of the fact that the dynamic stereo matching method adaptively obtained the best search interval for the multi-target stereo matching, thereby reducing the probability of mismatch and the amount of calculation required. For the green jujube orchard with a low-light environment, the improvement in accuracy of the experimental group relative to the control group is not significantly different from that of the other three orchards with sufficient light; this indicates that the proposed method have the potential to be stable and applicable in low-light environments.

Fig. 24 shows the enlargements at the same position in the experimental and control groups. It can be observed that the structural information of the control group is genuinely lost, and the spatial resolution of the fruits is significantly lower than that of the experimental group, indicating that with the enhancement of stereo vision, the experimental group achieved more detailed mapping of the local areas.

4. Conclusions and future work

The basic vision technologies of orchard picking robot, such as image segmentation, 3D positioning and surface reconstruction, have been initially completed with the joint efforts of today's researchers. It is believed that the global perceptions, general frameworks, and practical applications will be the key to the future development of the visual picking robots.

In this study, a mobile platform equipped with an eye-in-hand stereo vision system and a 6-DOF robotic arm was built. A flexible 3D mapping framework for unstructured orchards was established. The advantages of large-scale SLAM and high-accuracy stereo vision system were analyzed and integrated. Local and global experiments completely demonstrated the high stability, adaptability, and accuracy of the mapping system. It does not rely on any artificially designed features,

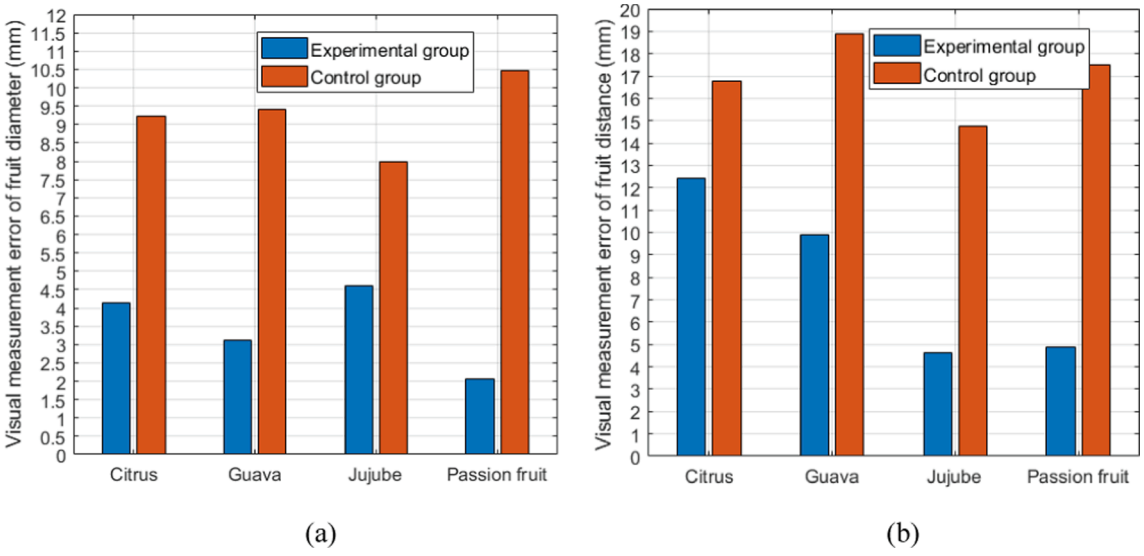
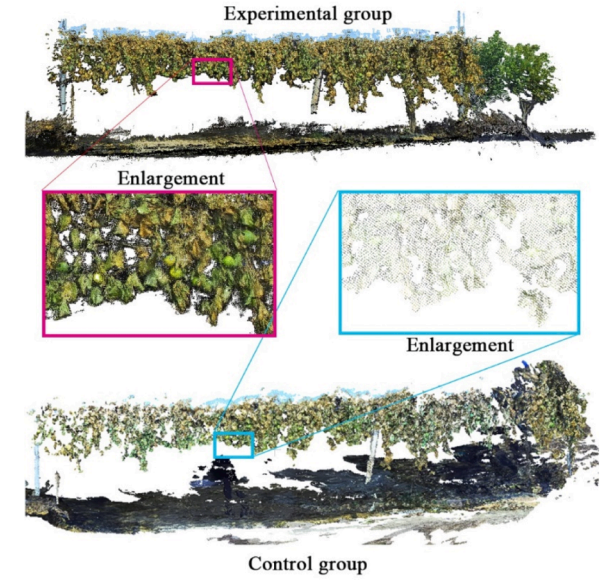
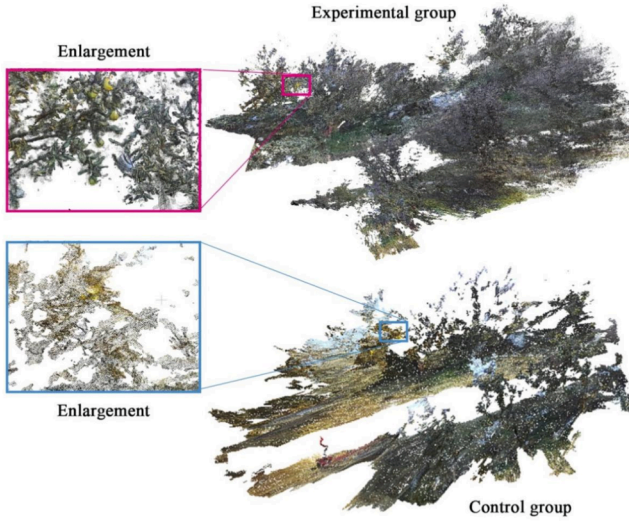


Fig. 23. Average absolute visual measurement errors: (a) Fruit diameter; (b) Distance between fruits.



(a)



(b)

Fig. 24. Enlargements of the global point clouds: (a) Passion fruit orchard; (b) Guava orchard.

therefore, it is expected to be replicable for different types of orchards, contributing stability and practicability to the current agricultural mapping systems.

This study has potential limitations. Some large-scale noises are very close to the main body of the global point cloud, so they cannot be removed in an appropriate manner. This may affect the performances of subsequent navigation or harvesting operations. At the same time, the visual information of the global point cloud seems to be redundant and thereby requires suitable down-sampling algorithms to obtain more compact structures.

Future work will involve navigation and path planning module based on the original mapping system. We will further study and introduced the methods for calculating the yield map from the constructed global map, and finally determine the optimal movement and picking behaviors of the mobile platform. Furthermore, more powerful hardware, such as, the inertial navigation systems, laser rangefinders, and force sensors will be equipped, and more systematic and detailed experiments will be conducted, particularly, at night (in low-light environments).

CRediT authorship contribution statement

Mingyou Chen: Methodology, Software, Writing - review & editing, Writing - original draft. **Yunchao Tang:** Conceptualization, Writing - original draft. **Xiangjun Zou:** Supervision. **Zhaofeng Huang:** Methodology, Data curation. **Hao Zhou:** Investigation, Visualization. **Siyu Chen:** Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Special Funding Project of Guangdong Enterprise Science and Technology Commissioner (GDKTP2020029500), the Key-area Research and Development Program of Guangdong Province (2019B020223003), and the Major scientific research projects of Guangdong Province (2020KZDZX1037).

References

Aguiar, A.S., dos Santos, F.N., Cunha, J.B., Sobreira, H., Sousa, A.J., 2020. Localization and mapping for robots in agriculture and forestry: a survey. Robotics 9, 97.

Andreff, N., Horaud, R., Espiau, B., 1999. On-line hand-eye calibration, in: Second International Conference on 3-D Digital Imaging and Modeling (Cat. No. PR00062). IEEE, pp. 430–436.

Capua, F.R., Sansoni, S., Moreyra, M.L., 2018. Comparative Analysis of Visual-SLAM Algorithms Applied to Fruit Environments, in: 2018 Argentine Conference on Automatic Control, AADECA 2018. IEEE, pp. 1–6. https://doi.org/10.23919/AADECA.2018.8577360.

Chebrolu, N., Lottes, P., Schaefer, A., Winterhalter, W., Burgard, W., Stachniss, C., 2017. Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. Int. J. Rob. Res. 36, 1045–1052. https://doi.org/10.1177/0278364917720510.

Chen, M., Tang, Y., Zou, X., Huang, K., Huang, Z., Zhou, H., Wang, C., Lian, G., 2020. Three-dimensional perception of orchard banana central stock enhanced by adaptive multi-vision technology. Comput. Electron. Agric. 174, 105508 https://doi.org/10.1016/j.compag.2020.105508.

Chen, X., Wang, S., Zhang, B., Luo, L., 2018. Multi-feature fusion tree trunk detection and orchard mobile robot localization using camera/ultrasonic sensors. Comput. Electron. Agric. 147, 91–108. https://doi.org/10.1016/j.compag.2018.02.009.

Daniilidis, K., 1999. Hand-eye calibration using dual quaternions. Int. J. Rob. Res. 18, 286–298.

Dong, W., Roy, P., Isler, V., 2020. Semantic mapping for orchard environments by merging two-sides reconstructions of tree rows. J. F. Robot. 37, 97–121. https://doi.org/10.1002/rob.21876.

Fan, Y., Feng, Z., Mannan, A., Khan, T.U., Shen, C., Saeed, S., 2018. Estimating tree position, diameter at breast height, and tree height in real-time using a mobile phone with RGB-D SLAM. Remote Sens. 10 https://doi.org/10.3390/rs10111845.

Gan, H., Lee, W.S., Alchanatis, V., 2017. A Prototype of an Immature Citrus Fruit Yield Mapping System, in: 2017 ASABE Annual International Meeting. Am. Soc. Agric. Biol. Eng., p. 1. https://doi.org/10.13031/aim.201700164.

Gao, X., Li, J., Fan, L., Zhou, Q., Yin, K., Wang, J., Song, C., Huang, L., Wang, Z., 2018. Review of wheeled mobile robots' navigation problems and application prospects in agriculture. IEEE Access 6, 49248–49268.

Ge, Y., Xiong, Y., From, P.J., 2020. Symmetry-based 3D shape completion for fruit localisation for harvesting robots. Biosyst. Eng. 197, 188–202. https://doi.org/10.1016/j.biosystemseng.2020.07.003.

Ge, Y., Xiong, Y., Tenorio, G.L., From, P.J., 2019. Fruit localization and environment perception for strawberry harvesting robots. IEEE Access 7, 147642–147652. https://doi.org/10.1109/ACCESS.2019.2946369.

Habibie, N., Nugraha, A.M., Anshori, A.Z., Ma'sum, M.A., Jatmiko, W., 2018. Fruit mapping mobile robot on simulated agricultural area in Gazebo simulator using simultaneous localization and mapping(SLAM), in: MHS 2017 - 28th 2017 International Symposium on Micro-NanoMechatronics and Human Science. IEEE, pp. 1–7. https://doi.org/10.1109/MHS.2017.8305235.

Hirschmuller, H., 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. Computer Vision and Pattern Recognition. 807–814. https://doi.org/10.1109/CVPR.2005.56.
Horaud, R., Dornaika, F., 1995. Hand-eye calibration. Int. J. Rob. Res. 14, 195–210.
Ivanov, M., Sergyienko, O., Tyrsa, V., Lindner, L., Flores-Fuentes, W., Rodriguez-Quinonez, J.C., Hernandez, W., Mercorelli, P., 2020. Influence of data clouds fusion from 3D real-Time vision system on robotic group dead reckoning in unknown terrain. IEEE/CAA J. Autom. Sin. 7, 368–385. https://doi.org/10.1109/JAS.2020.1003027.
Jia, Z., Yang, J., Liu, W., Wang, F., Liu, Y., Wang, L., Fan, C., Zhao, K., 2015. Improved camera calibration method based on perpendicularity compensation for binocular stereo vision measurement system. Opt. Express 23, 15205. https://doi.org/10.1364/oe.23.015205.
Katikaridis, D., Moysiadis, V., Kateris, D., Bochtis, D., 2019. Large-Scale Point-Cloud Based Global Mapping for Orchard Operations.
Li, J., Tang, Y., Zou, X., Lin, G., Wang, H., 2020. Detection of fruit-bearing branches and localization of litchi clusters for vision-based harvesting robots. IEEE Access 8, 117746–117758.
Lin, G., Tang, Y., Zou, X., Xiong, J., Li, J., 2019. Guava detection and pose estimation using a low-cost RGB-D sensor in the field. Sensors (Switzerland) 19, 428. https://doi.org/10.3390/s19020428.
Liu, S., Qi, L., Qin, H., Shi, J., Jia, J., 2018. Path aggregation network for instance segmentation, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768.
Nellithimaru, A.K., Kantor, G.A., 2019. ROLS: Robust object-level SLAM for grape counting. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2648–2656. https://doi.org/10.1109/CVPRW.2019.00321.
Park, F.C., Martin, B.J., 1994. Robot sensor calibration: solving AX= XB on the Euclidean group. IEEE Trans. Robot. Autom. 10, 717–721.
Pierzchała, M., Giguère, P., Astrup, R., 2018. Mapping forests using an unmanned ground vehicle with 3D LiDAR and graph-SLAM. Comput. Electron. Agric. 145, 217–225.
Ramírez-Hernández, L.R., Rodríguez-Quiñonez, J.C., Castro-Toscano, M.J., Hernández-Balbuena, D., Flores-Fuentes, W., Rascón-Carmona, R., Lindner, L., Sergiyenko, O., 2020. Improve three-dimensional point localization accuracy in stereo vision systems using a novel camera calibration method. Int. J. Adv. Robot. Syst. 17, 1–15. https://doi.org/10.1177/1729881419896717.
Shalal, N., Low, T., McCarthy, C., Hancock, N., 2015. Orchard mapping and mobile robot localisation using on-board camera and laser scanner data fusion - Part A: Tree detection. Comput. Electron. Agric. 119, 254–266. https://doi.org/10.1016/j.compag.2015.09.025.
Silwal, A., Davidson, J.R., Karkee, M., Mo, C., Zhang, Q., Lewis, K., 2017. Design, integration, and field evaluation of a robotic apple harvester. J. F. Robot. 34, 1140–1159. https://doi.org/10.1002/rob.21715.
Tan, M., Le, Q. V, 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv Prepr. arXiv1905.11946.
Tan, M., Pang, R., Le, Q.V., 2020. In: Efficientdet: Scalable and efficient object detection, in, pp. 10781–10790.
Tang, Y.-C., Wang, C., Luo, L., Zou, X., 2020. Recognition and localization methods for vision-based fruit picking robots: a review. Front. Plant Sci. 11, 510.
Tang, Y., Li, L., Wang, C., Chen, M., Feng, W., Zou, X., Huang, K., 2019. Real-time detection of surface deformation and strain in recycled aggregate concrete-filled steel tubular columns via four-ocular vision. Robot. Comput. Integr. Manuf. 59 https://doi.org/10.1016/j.rcim.2019.03.001.
Tsai, R.Y., Lenz, R.K., 1989. A new technique for fully autonomous and efficient 3 D robotics hand/eye calibration. IEEE Trans. Robot. Autom. 5, 345–358.
Underwood, J.P., Hung, C., Whelan, B., Sukkarieh, S., 2016. Mapping almond orchard canopy volume, flowers, fruit and yield using lidar and vision sensors. Comput. Electron. Agric. 130, 83–96. https://doi.org/10.1016/j.compag.2016.09.014.
Wibowo, T.S., Sulistijono, I.A., Risnumawan, A., 2017. End-to-end coconut harvesting robot. IEEE, pp. 444–449. https://doi.org/10.1109/ELECSYM.2016.7861047.
Williams, H.A.M., Jones, M.H., Nejati, M., Seabright, M.J., Bell, J., Penhall, N.D., Barnett, J.J., Duke, M.D., Scarfe, A.J., Ahn, H.S., Lim, J.Y., MacDonald, B.A., 2019. Robotic kiwifruit harvesting using machine vision, convolutional neural networks, and robotic arms. Biosyst. Eng. 181, 140–156. https://doi.org/10.1016/j.biosystemseng.2019.03.007.
Xiong, J., Lin, R., Liu, Z., He, Z., Tang, L., Yang, Z., Zou, X., 2018. The recognition of litchi clusters and the calculation of picking point in a nocturnal natural environment. Biosyst. Eng. 166, 44. https://doi.org/10.1016/j.biosystemseng.2017.11.005.
Zhang, T., Huang, Z., You, W., Lin, J., Tang, X., Huang, H., 2020. An autonomous fruit and vegetable harvester with a low-cost gripper using a 3D sensor. Sensors (Switzerland) 20. https://doi.org/10.3390/s20010093.
Zhang, Z., 2000. A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach, Intell, p. 22.
Zhao, W., Wang, X., Qi, B., Runge, T., 2020. Ground-level Mapping and Navigating for Agriculture based on IoT and Computer Vision. IEEE Access 8, 221975–221985.