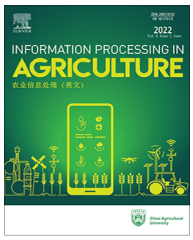
Available at www.sciencedirect.com

INFORMATION PROCESSING IN AGRICULTURE 10 (2023) 276–287

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



An improved binocular localization method for apple based on fruit detection using deep learning

Tengfei Li [a], *Wentai Fang* [a], *Guanao Zhao* [a], *Fangfang Gao* [a], *Zhenchao Wu* [a], *Rui Li* [a], *Longsheng Fu* [a,b,c,*], *Jaspreet Dhupia* [d]

[a] *College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China*
[b] *Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Yangling, Shaanxi 712100, China*
[c] *Shaanxi Key Laboratory of Agricultural Information Perception and Intelligent Service, Yangling, Shaanxi 712100, China*
[d] *Department of Mechanical Engineering, The University of Auckland, Private Bag 92019, Auckland, New Zealand*

ARTICLE INFO

Article history:
Received 22 August 2021
Received in revised form
5 December 2021
Accepted 13 December 2021
Available online 18 December 2021

Keywords:
Deep learning
Object detection
Faster R-CNN
Template matching
Image segmentation
Binocular localization

ABSTRACT

Apple picking robot is now being developed as an alternative to hand picking due to a great demand for labor during apple harvest season. Accurate detection and localization of target fruit is necessary for robotic apple picking. Detection accuracy has a great influence on localization results. Although current researches on detection and localization of apples using traditional image algorithms can obtain good results under laboratory conditions, it is difficult to accurately detect and locate objects in natural field with complex environments. With the rapid development of artificial intelligence, accuracy of apple detection based on deep learning has been significantly improved. Therefore, a deep learning-based method was developed to accurately detect and locate the position of fruit. For different localization methods, binocular localization is a widely used localization method for its bionic principle and lower equipment cost. Hence, this paper proposed an improved binocular localization method for apple based on fruit detection using deep learning. First, apples of binocular images were detected by Faster R-CNN. After that, a segmentation based on chromatic aberration and chromatic aberration ratio was applied to segment apple and background pixels in bounding box of detected fruit. Furthermore, template matching with parallel polar line constraint was used to match apples in left and right images. Finally, two feature points on apples were selected to directly calculate three dimensional coordinates of feature points with the binocular localization principle. In this study, Faster R-CNN achieved an AP of 88.12% with an average detection speed of 0.32 s for an image. Meanwhile, standard deviation and localization precision of depth of two feature points on each apple were calculated to evaluate localization. Results showed that the average standard deviation and the average localization precision of 76 groups of datasets were 0.51 cm and 99.64%, respectively. Results indicated that the proposed improved binocular localization method is promising for fruit localization.

* Corresponding author.
E-mail address: fulsh@nwafu.edu.cn (L. Fu).
Peer review under responsibility of China Agricultural University.
https://doi.org/10.1016/j.inpa.2021.12.003

1. Introduction

Apple is one of the most widely produced fruits and is still picked manually, which is desiring for mechanical harvesting. Apple has rich nutritional value, low fat, high carbohydrates, and vitamins C and E [1]. Meanwhile, good taste of juicy, sweet, and crispy apples are popular among consumers [2]. Besides, apple is feasible to plant in a variety of environments and has high economic value. China is a major apple producer in the world, with the largest planting area and largest fresh apple export [3]. In 2019, China's annual apple production reached 42.43 million tons [4]. There is an urgent need for automatic picking robots because of the great demand for labor during the harvest season [5]. Apple picking robot is now being developed to replace manual picking. Picking based on robots is not only efficient but also low-cost. And the reason for higher cost in hand picking is scarcity of rural labor, which is caused by transfer of rural labor force to cities since reform and opening up in China from 1978. As a result, demand for labor during apple picking season is difficult to meet. A picking robot consists of two main subsystems, a vision system and an end-effector system [6]. The vision system guides the robot end-effector to pick apples from trees by detecting and localizing apples [7].

Localization is the core of apple picking robots. In recent years, researchers have made a deep research on using machine vision to locate objects. According to principle and different hardware used, localization methods can be roughly divided into binocular localization, structured light, and ToF (Time of Flight) method [8]. Structured light is easily disturbed by ambient light, strong reflected light will lead to image saturation which results in phase error, and ultimately causes measurement error [9]. Due to limitation of measurement principle and influence of external environment factors, the ToF method will produce a series of systematic errors and scene errors [10]. These errors will lead image information obtained by the ToF depth camera to be distorted [11]. Binocular localization is more bionic in principle, with wider application and lower equipment cost [12]. Ji et al. used the binocular stereo vision localization method to locate apple branch obstacle, whose error was only 6.20 mm [13]. Ye et al. analyzed error data by using the binocular camera to locate the actual location of target litchi [14]. Considering the influence of complex conditions in natural environment and equipment cost in this experiment, the binocular positioning method was selected. Stereo matching is the key link in binocular stereo vision. Some methods such as graph-cuts, belief-propagation, and SIFT (Scale Invariant Feature Transform), have low efficiency and high complexity [15]. Compared with these methods, template matching was used in this study to obtain feature points on apple, which was simpler and had high localization precision through the experiment. Furthermore, this study adopted binocular localization methods to calculate the real depth of feature points.

The first step to localize apple is to detect fruit from the binocular images, respectively, which can be divided into traditional detection methods and deep learning-based methods. Jiao et al. used LAB color space to process apple images to achieve the purpose of identifying target fruits [16]. Besides, Laplacian operator was also applied for edge detection to extract the contour of target apples [17]. However, results obtained by these methods in complex environment were not ideal and were easily affected by the background of fruit orchard [18]. Compared with these traditional detection methods, deep learning technology shows more promising performance in object detection. It provides higher accuracy and is superior to traditional image processing methods [19]. Gao et al. applied a Faster R-CNN (Region-Convolutional Neural Network) to detect apple and reached 0.879 of average precision (AP) [20]. Xiao et al. used BP neural network to train an apple color identification model, which identified apples on fruit trees effectively [21]. Fu et al. employed ZFNet to detect apples from RGB images segmented by depth threshold, which obtained 0.805 of AP [22]. The result indicated that the application of deep learning technology in object detection and localization of fruit is feasible.

In this study, an improved binocular localization method for apple based on fruit detection by using Faster R-CNN and the traditional segmentation method was proposed. Faster R-CNN was used for detecting apple images. Apple pixels were segmented from detection results by using a traditional segmentation method based on chromatic aberration and chromatic aberration ratio. Besides, this study used a template matching with parallel polar line constraint to match apples. When above work had been completed, three dimensional coordinates of feature points on apple were obtained based on the binocular localization principle. This study tested AP (Average Precision) and the average detection speed of the model. The standard deviation of depth and localization precision were calculated to evaluate the localization of apples.

2. Materials and methods

2.1. *Image acquisition*

In this study, RGB (Red, Green, and Blue) images of apples were taken at Prosser, Washington, USA during 2017 and 2018 harvest season. RGB images were acquired with the binocular camera (BumbleBeeXB3) at a resolution of 1 280 × 960 pixels and saved in PNG format, as shown in Fig. 1. LabelImg was applied to label dataset in this study. All images were manually labeled with rectangular annotations. Meanwhile, corresponding 'XML' annotation files were generated and saved. This study used geometric transformation and image enhancement to annotate images. All images were used as inputs to Faster R-CNN based VGG16 model for training and testing. A total of 800 apple images in the field

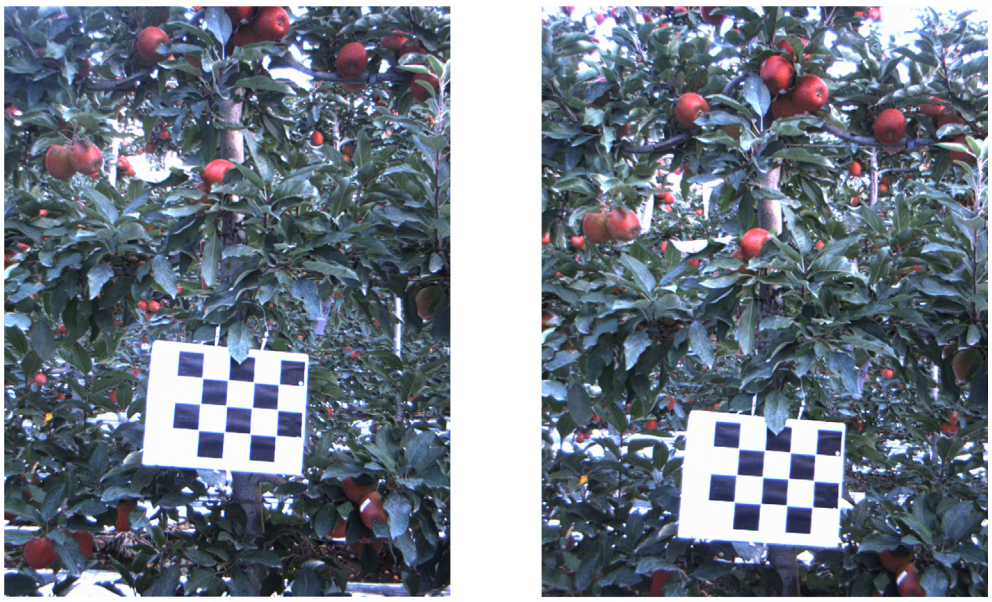


Fig. 1 – Examples of apple images captured by binocular camera BumbleBeeXB3. (a) Image taken by left lens; (b) Image taken by right lens.

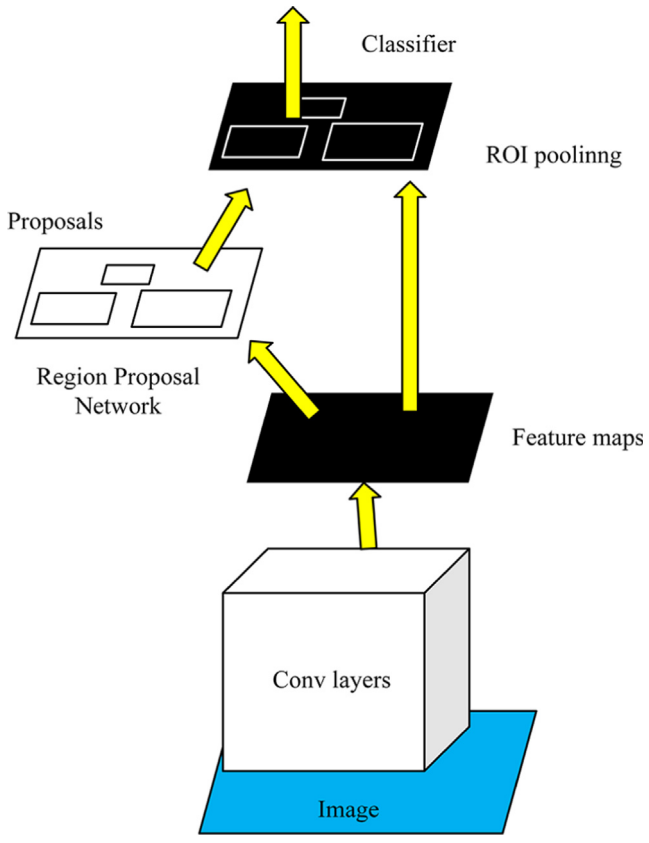


Fig. 2 – The architecture of Faster R-CNN.

were collected and labeled, 560 of which constituted training dataset and 240 composed test dataset.

2.2. *Deep learning model*

Faster R-CNN is an efficient object detection network improved on the basis of Fast R-CNN, which consists of two modules. The first and second modules are a Regional Proposal Network (RPN) and the Fast R-CNN detector using the proposed regions [23]. As shown in Fig. 2, the input of Faster R-CNN is an image of any size. This image is extracted through convolutional layer CNN and other basic networks to obtain the feature map. CNN realizes image detection and classification by combining the forward feature extraction process and the backward parameter adjustment process [24]. The RPN uses a sliding window on the feature map to traverse the whole feature map, and it outputs a series of rectangular object proposals, each of which comes with a target box score. The final detection result is a list of the bounding box (BBox), composed of values containing coordinates of two rectangular corner points. The Four coordinates of BBox were X_{min}, Y_{min}, X_{max}, and Y_{max}, as shown in Fig. 3.

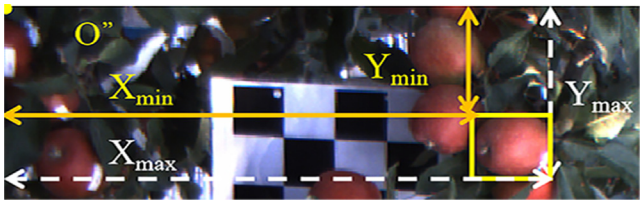


Fig. 3 – The meaning of parameters in BBox. Coordinate origin O" is the boundary point in upper left corner of the image. X_{min} and Y_{min} represent distance from upper left corner of rectangle to left edge and top edge of image, respectively. X_{max} and Y_{max} represent distance from lower right corner of rectangle to left edge and top edge of image, respectively.

Fig. 4 – Apples were shadowed by leaves. Center points of two yellow boxes were overlapped with leaves and could not be used as feature point of apple. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.3. *Image segmentation*

Image segmentation methods were applied to segment apple pixels and background in the detected bounding box by Faster R-CNN. Compared with texture, shape, and other features, selecting color features of apple for segmentation is more simple and effective [25]. As shown in Fig. 4, some apples in yellow boxes were mostly shadowed by leaves. If midpoint of the yellow box was taken as a feature point of the apple, this selected feature point actually belonged to the leaf. Moreover, calculated three dimensional coordinates would be the localization result of leaf rather than an apple. Therefore, image segmentation methods were adopted to remove pixels that did not belong to apple in the image. In this study, three traditional threshold segmentation methods were used, and results were compared with original images. By dividing different values of pixel channels into several types, the effect of segment background and object is realized. This experiment used threshold segmentation for H channel of HSV (Hue, Saturation, and Value) color space, Otsu threshold segmentation for R channel in RGB (Red, Green, and Blue) space and A channel in LAB color space, and segmentation method based on chromatic aberration and chromatic aberration ratio, respectively.

HSV color space uses Hue (H), Saturation (S), and Value (V) to represent color. In OpenCV, value of the H channel ranges from 0 to 179. Besides, S channel represents the degree to which color is close to spectral color, namely, vividness of color. And V channel represents the brightness of color. The value ranges of the S and V channels are both 0–255 in OpenCV. A larger S means the color will be darker, while a larger V shows color will be brighter. Since pixels between [160, 179] or [0, 20] in the H channel are red, this study directly classified pixels between this threshold range as apple pixels, to achieve fixed threshold segmentation of the H channel. HSV distribution diagram of an apple (as shown in Fig. 5(a)) was drawn, as shown in Fig. 5(b). Although there was one peak in the H channel, the segmentation result was greatly affected by light. Hence, this experiment tried to use other methods for segmentation.

Otsu threshold method [26] is a threshold segmentation method with adaptive characteristics. It is one of the most famous methods for automatic image thresholding and is widely used in image binarization applications [27]. The basic principle of the Otsu algorithm is to find an optimal threshold for image segmentation. This threshold maximizes variance between two classes obtained after segmentation. Among them, interclass variance g is defined in Eq. (1).

$$g = \omega_1 * \omega_0 * (u_0 - u_1)^2 \tag{1}$$

where ω_0 and ω_1 represent proportion of segmented objects and proportion of background in a total number of pixels, respectively. Meanwhile, u_0 and u_1 represent an average gray level of foreground and total average gray level, respectively.

Since apples are generally red, the Otsu threshold method was applied to the R channel in RGB color space to segment apple pixels from background. RGB distribution diagram of apple in a red box in Fig. 5(a) was drawn, as shown in Fig. 5 (b). Distribution of B and G channels were concentrated, while the R channel was widely distributed, indicating that the segmentation of the R channel can better separate apple pixel and background. However, the segmentation effect of the R channel was not acceptable, this study tried other channels for segmentation in the Otsu threshold method. LAB color space is a color model based on physiological characteristics, which is composed of luminance (L) and channels A and B. LAB distribution diagram of the apple was also drawn, as shown in Fig. 5(b). Although there was one peak in both A and B channels, colors in the B channel from low to high included bright blue, gray, and yellow, while colors in the A channel from low to high included dark green, gray, and bright red. As A channel contained the color of an apple, this experiment tried A channel in LAB color space for segmentation.

In addition to single-channel segmentation method, other methods based on multi-channel are needed, because of the poor segmentation performance of the single-channel segmentation method. The pixels whose RGB value meets two conditions in Eq. (2) and Eq. (3) are classified as apple pixels by chromatic difference and chromatic difference ratio [28]. Experimental result has shown that the segmentation method based on chromatic aberration and chromatic aberration on ratio was superior to the other two segmentation methods, while the segmented image did not contain leaves. Therefore, the segmentation method based on chromatic aberration and chromatic aberration ratio in threshold segmentation method was adopted to segment apple in the detected bounding box by Faster R-CNN.

$$R - G > 0 \tag{2}$$

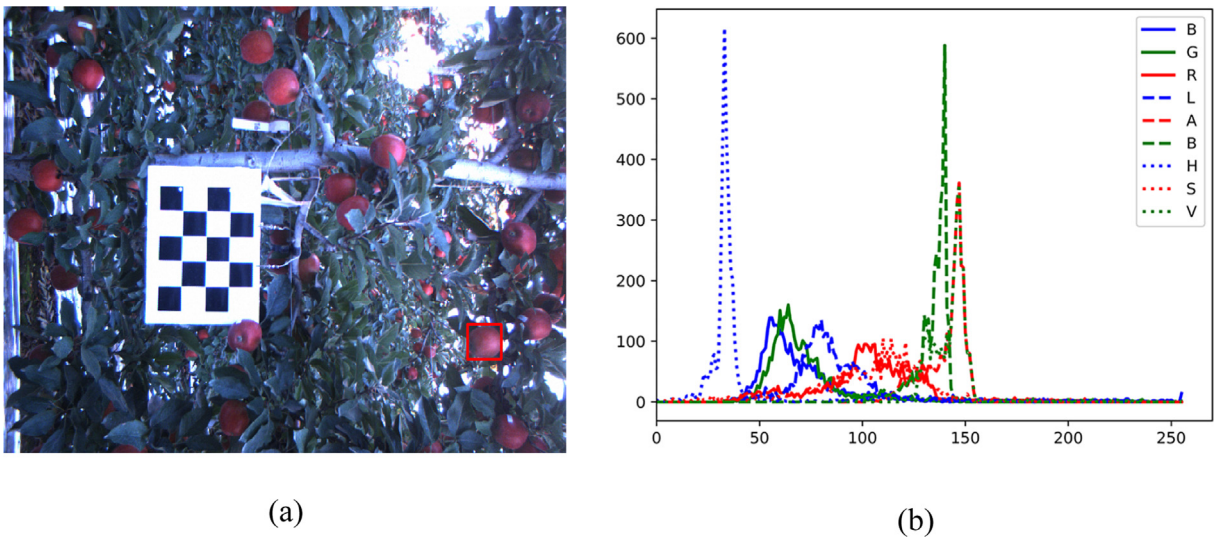


Fig. 5 – Color distribution of related channels of a detected apple by deep learning. (a) Original image of detected apples. Channel distribution map of apple in red box was draw. (b) Channel distribution map of HSV, RGB, and LAB. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$\frac{R-G}{R-B} > 1 \tag{3}$$

2.4. *Principle of binocular localization*

After image segmentation was completed, this study realized the localization of segmented apples through the principle of binocular localization. Binocular localization is based on left and right lenses of BumbleBeeXB3 observing a feature point together in space, as shown in Fig. 6. *O* and *O'* are origins of left and right lenses coordinate systems, respectively. Two lenses simultaneously observe a feature point A, whose three-dimensional coordinates are X_A, Y_A, and Z_A. The projection of feature point A on left lens is A_l (X_l, Y_l), and its projection on right lens is A_r (X_r, Y_r). According to the principle of similar triangles in plane geometry, they can be obtained as shown from Eq. (4) to Eq. (6).

$$X_A = \frac{Baseline}{X_r - X_l} * X_l \tag{4}$$

$$Y_A = \frac{Baseline}{X_r - X_l} * Y_l \tag{5}$$

$$Z_A = \frac{Baseline}{X_r - X_l} * f \tag{6}$$

where *Baseline* of Eq. (4) represents distance between *O* and *O'*, and *f* of Eq. (6) is focal length of the binocular camera. Eq. (4) and Eq. (5) show that the key step in locating a point is to find corresponding coordinates of this point in left and right images. However, photography environment is complex that traditional partial matching algorithm is not applicable to find corresponding relations of feature points of apples, such as SAD (Sum of absolute differences) [29]. To obtain corresponding relationship of feature points on apple, this study first carried out apple stereo matching to determine the

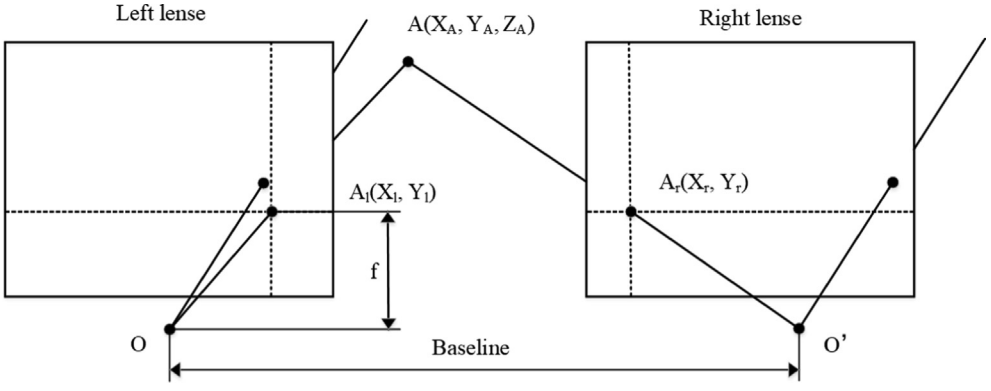


Fig. 6 – Principle of binocular localization. Black intersection of the line between A and origin O and imaging plane of left lens is A_l, while A_r is the intersection of the line between A and origin O' and imaging plane of right lens.

corresponding relationship of the same apple in left and right images. After that, the apple detection results were traversed to get feature points.

2.5. Apple stereo matching

Apple stereo matching focuses on confirming relations between apples of binocular images to further get feature points of apples and calculate parallax to localization. As images captured by the binocular camera were not completely the same, only most of points on right side in left images and most of points on left side in right images can be found corresponding points in the other image. The left side of left images and right side of right images do not appear in another image.

Stereo matching algorithms can be divided into global stereo matching algorithms and local stereo matching algorithms according to the constraint range [15]. Template matching applied in this study belongs to the local stereo matching algorithm. Global stereo matching algorithms, such as graph-cuts and belief-propagation, have low efficiency and high complexity [30]. SIFT algorithm is a local stereo matching algorithm based on image features. Although it has good resistance to distortion, its algorithm efficiency is low and has poor matching accuracy, which is suitable for images with obvious geometric features. Compared with above algorithms, template matching is simpler and more widely applicable.

Template matching used left image and right image as template and target image for matching corresponding apple images, respectively. Besides, it removed apples that do not appear in both views at the same time. Template matching is to slide template on target image pixel by pixel and traverses every position of target image [31]. It extracts several feature vectors from detected image patches and compares them with corresponding feature vectors of the template. At the same time, it calculates the Euclidean distance between image and feature vector of the template. When calculated Euclidean distance is smallest, matching effect is best. Euclidean distance between template and image patches is defined in Eq. (7).

$$D(i,j) = \sum_{m=1}^{M}\sum_{n=1}^{N}[P(m,n) - T(m,n)]^2 \tag{7}$$

where $D(i, j)$ is Euclidean distance between template and image patch. The center of gravity of the image patch is located at (i, j). M and N are width and height of the image template, respectively. $P(m, n)$ is pixel value of the image template, and $T(m, n)$ is pixel value of image with center of gravity located in (m, n).

Normalized square error matching method was one of numerous template matching methods, which was adopted in this study. The normalized square error matching method is calculated based on the square error matching method. Before introducing the normalized square variance matching method, the square error matching method was explained first. The square variance matching method is used to describe the similarity of two pixels. It subtracts the square sum of detected image pixel value from pixel value of template, and its formula is shown in Eq. (8).

$$R(x',y') = \sum \left(T(x',y') - I(x+x',y+y')\right)^2 \tag{8}$$

$$R'(x',y') = \frac{R(x',y')}{\sqrt{\sum T(x',y')^2 * \sum I(x+x',y+y')^2}} \tag{9}$$

where $R(x', y')$ represents matching degree, $T(x', y')$ represents pixel value of template, and $I(x', y')$ represents pixel value of the detected image. Normalized square deviation matching value is denoted as $R'(x', y')$, which satisfies the relationship of Eq. (9). When the matching cost is minimum, the best match of results acquired by the normalized square error matching method is obtained. When the matching cost is closer to 0, the match result is good. While the matching cost is too large, the match result is not ideal, and the matching object cannot be found.

The target of template matching is usually the entire image, but it wastes a lot of time and causes mismatching due to the possibility of fruits with similar colors and contours in the wild. In this study, parallel pole constraint was used to reduce the range of template matching to reduce matching time costs and improve matching accuracy. Since the binocular camera has been rectified for parallelism and distortion, each point should be at the same pixel height as its corresponding point in another view after rectification. Therefore, matching of an apple in left and right images should only be done within the same height range. According to the binocular vision model of parallel optical axis, abscissa of point (X_L, Y_L) in left image should be greater than that of its corresponding point (X_R, Y_R) in right image, namely, $X_L > X_R$. As shown in Fig. 7, coordinate range of the matching template in the original image was $([X_1, X_2], [Y_1, Y_2])$, and range of the target image was ([0, X], [0, Y]). Where X and Y represented width and height of whole image, respectively. Matching of the template with parallel polar constraint only needed to be carried out within range of $([0, X_2], [Y_1, Y_2])$, which greatly reduced matching range and matching time. In addition, meaningless part in the matching range was eliminated, which reduced probability of mismatching of apples and matching time.

2.6. Localization

This study selected feature points on apples to directly calculate three dimensional coordinates of feature points of apples to complete localization work. In this experiment, two feature points on each apple were obtained by a traversal. It traversed segmented apple pixels from top-left and bottom-right corner to center column by column simultaneously, as shown in Fig. 8(b). Traversal stopped when the first point belonging to the apple was found in each direction. After traversal, two feature points on apple were determined, as shown in Fig. 8 (d). Left and right feature points of the apple in the left image corresponded to left and right feature points of the matching apple in the right image, respectively. After that, three-dimensional coordinates of two feature points could be calculated from to Eqs. (4) to (6).

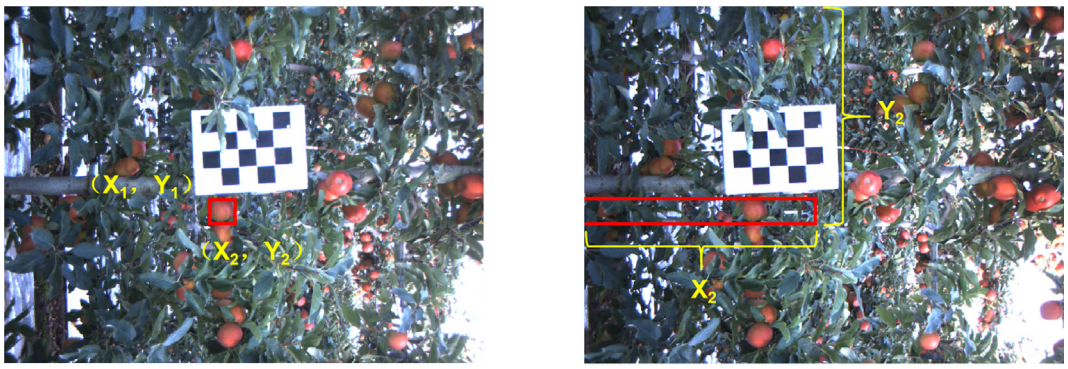


Fig. 7 – Limit of apple matching range. (a) Original images; (b) Range matching limitation based on parallel polar constraint. Two red parallel polar lines limit matching range to ([0, X_2], [Y_1, Y_2]). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

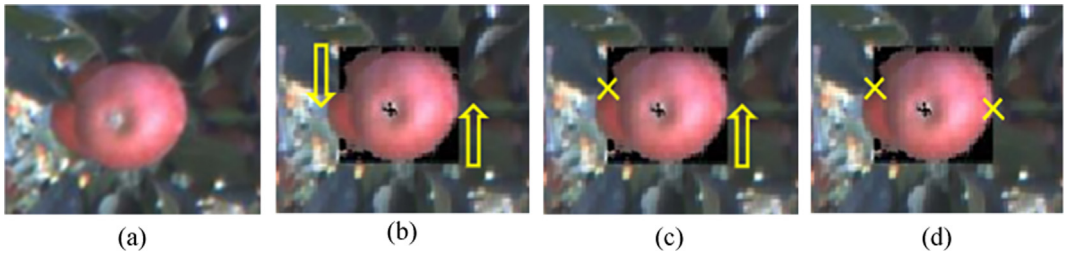


Fig. 8 – Result after traversing two feature points on apple. (a) Original apple image. (b) Traversal began on apple after segmentation. Yellow arrows represented direction of traversing from top-left and bottom-right corner to center column by column simultaneously. (c) Traversal was in progress. The direction from top-left to center firstly encountered feature point on apple and stopped, while another direction was still traversing. (d) Traversal stopped. Two yellow crosses represented two feature points on the apple. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.7. Detection performance

Detection performance in this study was evaluated using average precision (AP) and average detection speed. Among them, AP was calculated by Precision and Recall, which were defined in Eqs. (10) and (11), respectively.

$$Precision = TP/(TP + FP) \tag{10}$$

$$Recall = TP/(TP + FN) \tag{11}$$

where TP (True Positives) meant number of apple images correctly identified as apple positive samples. FP (False Positives) indicated number of images falsely identified as apple false samples. FN (False Negatives) referred to number of apple images falsely identified as images that were not of apples. Precision is percentage of TP in a detected image, and Recall is percentage of all positive samples in the dataset that are correctly identified as positive. AP is defined in Eq. (12), which represents area under precision and recall curve. Higher AP shows classifier is better.

$$AP = \int_0^1 P(R)dR \tag{12}$$

2.8. Evaluation of localization accuracy

In this experiment, localization accuracy was evaluated by calculating the standard deviation (SD) of depths of two feature points and localization precision (LP) on apples. Meanwhile, SD and LP were defined in Eqs. (13) and (14), respectively.

$$SD = \sqrt{\sum_{i=1}^{n} \left(D_i - D'\right)^2 / n} \tag{13}$$

$$LP = 1 - \left(D_{max} - D'\right)/D' \tag{14}$$

where D_i represented distance value obtained by localizing each feature point, D' represented average distance of two feature points on each apple, and n represented the number of feature points on each apple. Besides, D_{max} meant distance

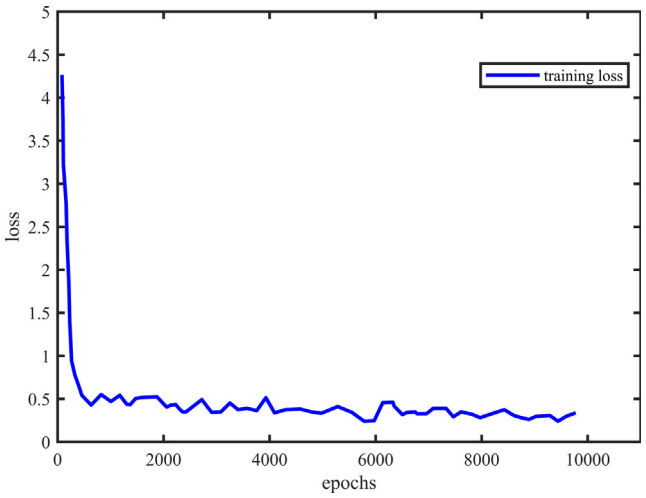


Fig. 9 – Training loss curve of Faster R-CNN.

value of feature point with the largest localization result on apple.

3. Results and discussion

3.1. *Training assessment and performance of the network*

Training performance of deep learning models is heavily affected by iteration times. Training loss curve of Faster R-CNN was converged, as shown in Fig. 9. The abscissa "epochs" refers to the number of times that the training dataset passes through the neural network and is returned during training. Ordinate "loss" represents difference between output value of the model and the real value. The loss function is a way of assessing how well the model predicts the dataset, while smaller loss represents model output value is closer to real value. As the number of iterations increased, the loss value decreased gradually. After approximately 10 000 epochs, Faster R-CNN began to reach a steady loss value of around 0.30. The convergent loss curve showed that Faster R-CNN can effectively identify the target and carry out apple image detection. Faster R-CNN achieved 88.12% of AP and the average detection speed was 0.32 s to detect an image with 1 280 × 960 pixels. There was no case of mistaken background identification as apples for the result of detection, as shown in Fig. 10.

3.2. *Comparison of three threshold segmentation methods*

Three threshold segmentation methods were compared, performance of the segmentation method based on chromatic aberration and chromatic aberration ratio was better than the other two segmentation methods. Their segmentation performances were as shown in Fig. 11. The results of fixed threshold segmentation of the H channel and Otsu segmentation of the R channel and A channel were shown in Fig. 11(b), (c), and (d), respectively. Some apple pixels were not com-

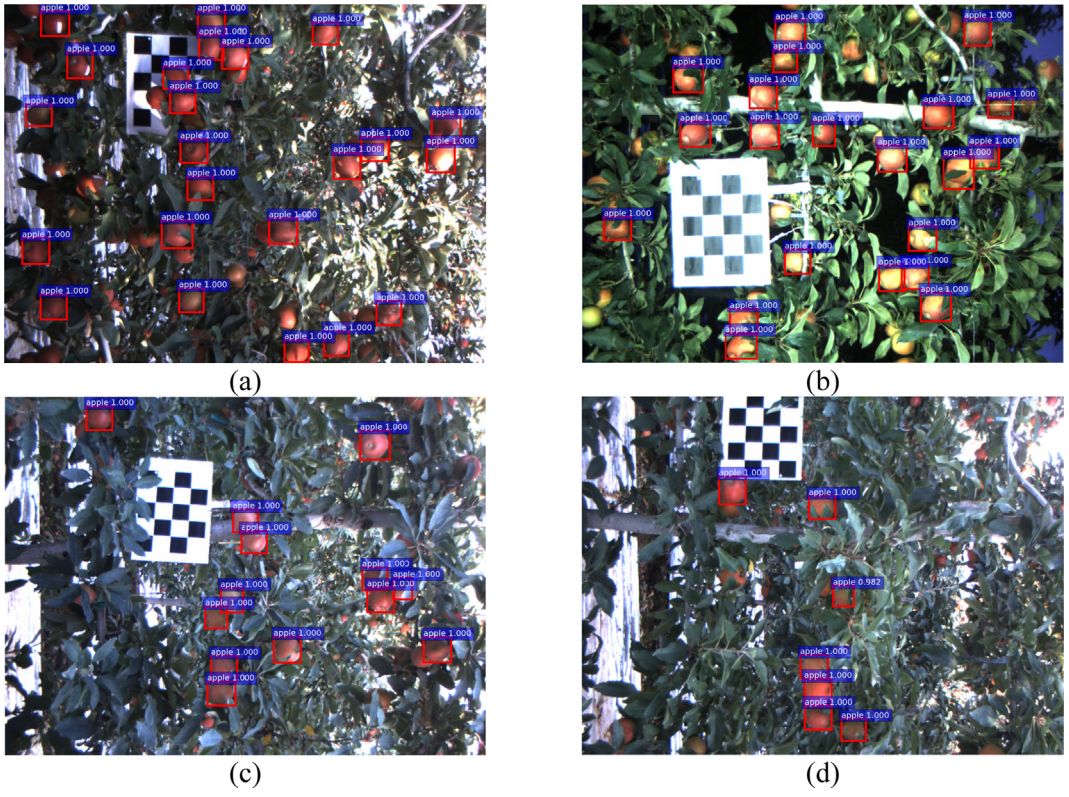


Fig. 10 – Result of apple image detected by Faster R-CNN. (a) Image detected in natural light; (b) Image detected in artificial light; (c) Only one apple was not detected due to coverage of another apple. (e) Every uncovering apple in image was completely identified.

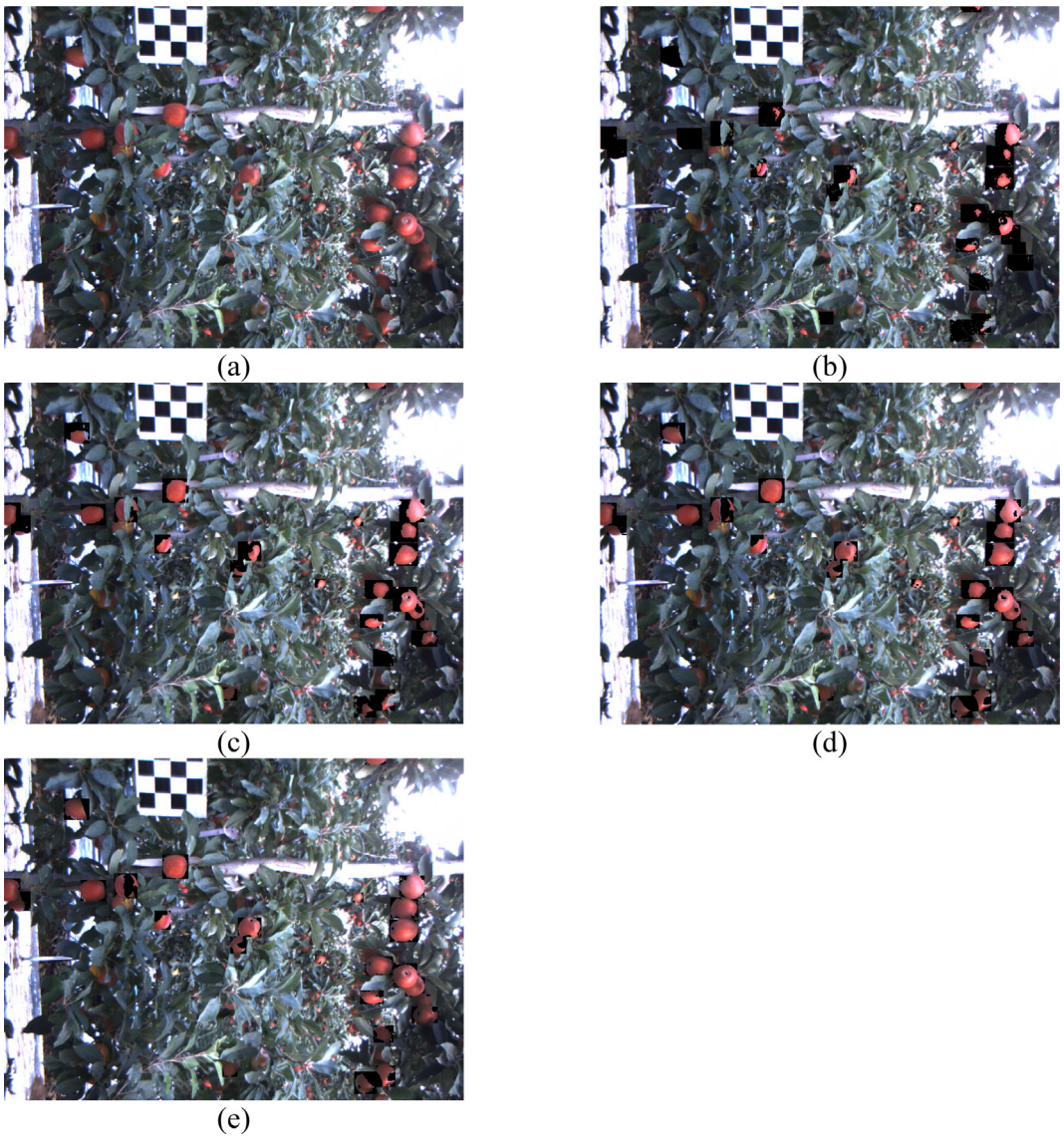


Fig. 11 – Comparison of results of different threshold segmentation methods. (a) Original image; (b) Fixed threshold segmentation of H channel in HSV; (c) Otsu segmentation of R channel in RGB; (d) Otsu segmentation of A channel in LAB; (e) Segmentation method based on chromatic aberration and chromatic aberration ratio.

pletely segmented from the background with these methods. It was because these apples got less light than others and appeared darker than the normal bright red color in the image. The result of the segmentation method based on chromatic aberration and chromatic aberration ratio was as shown in Fig. 11(e). After comparing multiple images manually, this method had the best segmentation effect for apples among these methods. Even if the color of the apple image changed greatly due to the influence of light, this method could also segment apple pixels correctly. Compared with other methods, this segmentation method was more suitable for complex environments and met the localization needs of picking robots. Hence, it was used to segment apple in the detected bounding box by Faster R-CNN.

3.3. *Binocular localization evaluation*

In this study, the depth of feature points on each apple was directly displayed on the image, and the location results were evaluated. After detection, segmentation, stereo matching and localization, one of its localization visualizations was shown in Fig. 12. End points of cyan line segment represented two feature points at the far left and right on this apple. And cyan number near the end points of cyan line segment represented the depth of this point. Since apple is approximately spherical and points on the edge of apple were selected in this study, Z coordinates of two feature points that belong to the same apple should be nearly the same. Besides, this study evaluated the localization accuracy of apples by the standard

Fig. 12 – Visualization of localization results using Faster R-CNN combined with traditional segmentation methods.

deviation of depths of two feature points and localization precision on apples. A higher value of SD and lower value of LP represented localization results of multiple feature points on the same apple differed greatly and localization effect was not good. Meanwhile, lower value of SD and higher value of LP meant that localization results of multiple feature points on the same apple had little difference, and the localization effect was better. This paper counted SD and LP of 76 groups of data in random statistical datasets, as shown in Fig. 13. All the datasets used to calculate SD and LP did not include the training dataset. The average SD and LD of 76 dataset were 0.51 cm and 99.64%, respectively. These localization results basically met the needs of the apple picking robot localization system.

However, due to inevitable errors in the manufacturing process of the binocular camera and different lens locations, left and right images of the left and right apples are not exactly the same, resulting in different segmentation results

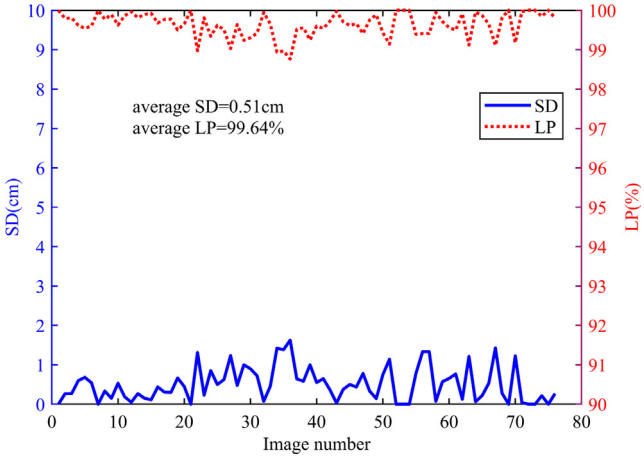


Fig. 13 – SD and LP of 76 groups of data in random statistical dataset.

in two views. And different segmentation results of an apple made mistakes in choosing feature points, which led to a bigger localization deviation, such as the apple in red rectangle of Fig. 12. The depth of two feature points of the apple in red rectangle was 117.10 cm and 122.26 cm, respectively. The average depth of feature points on this apple was 119.68 cm, and depth difference of two feature points on this apple was 5.16 cm. And the SD and LP of this apple were 2.58 cm and 97.84%, respectively. Nonetheless, this result had a large difference from the average SD and LP of 76 datasets, which is not a satisfied localization result for the apple picking robot.

3.4. *Results from other localization studies*

Compared with localization methods based on traditional image algorithm, the binocular localization method based on deep learning had higher detection accuracy, which lead to better performance. Jiao et al. found the maximum of calculated minimum distance from the inner point to the edge. Finally, by finding the minimum distance between center and edge, it obtained the radius to achieve the apple location. Maximum apple's center error in that study reached 23.21 cm [16]. Li and Liu performed segmentation and processing of reducing noise on apple image, using target contour extraction and refinement processing on contour edge [17]. Finally, the study completed localization work by using the method of determining a circle at three points. The average error achieved 6.16%. However, its localization result was influenced by a certain error in edge thinning and invalid point merging of apple images. Meanwhile, there is an error between target area in the binary image and target area in the real scene after processing. Compared with above localization methods, the binocular localization method in this study depended on feature matching of stereo. There was no need to preprocess the image, and there was no area error in the method of edge coincidence. Instead, binocular localization calculated three dimensional coordinates of feature points on apple to determine the localization of the apple. Its principle was simpler and equipment cost was lower.

3.5. *Further works on apple detection*

The limitation of this study was that the traditional segmentation method only separate object from the background, but it is impossible to distinguish pixel from two targets of the same type. In this study, pixels were belonging to two adjacent apples that appeared in a box where only one apple should appear. It resulted in segmentation errors and would have a certain impact on localization results. In the future, Mask R-CNN will be used to detect and segment apple images. Mask R-CNN is a kind of instance segmentation network. It is more straightforward and flexible, which obtains bounding boxes and segmentation masks simultaneously [32]. Instance segmentation obtains the information that the image itself needs to express according to texture, scene, and other features of the image. On this basis, Mask R-CNN classifies different objects of the same category.

4. Conclusions

In this study, Faster R-CNN network was used in detecting binocular images of apples. Segmentation methods of chromatic aberration and chromatic aberration ratio were adopted to segment apple in the detected bounding box by Faster R-CNN. Meanwhile, template matching with parallel polar line constraint was adopted to match example of apples. Finally, three dimensional coordinates of feature points were calculated. Distance between binocular camera lenses and apple in the image was obtained. Meanwhile, the average standard deviation and average localization precision of localization results of 76 groups of datasets were calculated as 0.51 cm and 99.64%, respectively. This experiment showed that results of an improved binocular localization method for apples based on fruit detection using Faster R-CNN could reflect the actual localization of apples in the image from the camera. Therefore, this method will further help the picking robot to locate fruits better. In the future, further research will be adopted in using Mask R-CNN to detect and segment apple images. This study will be applied to the operation of a picking robot, realizing the detection and localization of apples.

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

Funding: This work was supported by the National Natural Science of China (32171897); Youth Science and Technology Nova Program in Shaanxi Province of China (2021KJXX-94); Science and Technology Promotion Program of Northwest A&F University (TGZX2021-29); Recruitment Program of High-End Foreign Experts of the State Administration of Foreign Experts Affairs, Ministry of Science and Technology, China (G20200027075).

REFERENCES

[1] Koutsos A, Tuohy KM, Lovegrove JA. Apples and cardiovascular health—is the gut microbiota a core consideration? Nutrients 2015;7:3959–98. https://doi.org/10.3390/nu7063959.

[2] Blažek J, Paprštein F, Zelený L, Křelinová J. The results of consumer preference testing of popular apple cultivars at the end of the storage season. Hortic Sci 2019;46(No. 3):115–22. https://doi.org/10.17221/146/2017-HORTSCI.

[3] Wang Na, Wolf J, Zhang F-S. Towards sustainable intensification of apple production in China - Yield gaps and nutrient use efficiency in apple farming systems. J Integr Agric 2016;15(4):716–25. https://doi.org/10.1016/S2095-3119(15)61099-1.

[4] UN Food & Agriculture Organization, 2021. Production of Apple (Fruit) by Countries. Retrieved 2021-04-10. link: http://www.fao.org/faostat/en/#data/QCL.

[5] Chu P, Li Z, Lammers K, Lu R, Liu X. Deep learning-based apple detection using a suppression mask R-CNN. Pattern Recognit Lett 2021;147:206–11. https://doi.org/10.1016/j.patrec.2021.04.022.

[6] Zhang Z, Igathinathane C, Li J, Cen H, Lu Y, Flores P. Technology progress in mechanical harvest of fresh market apples. Comput Electron Agric 2020;175:105606. https://doi.org/10.1016/j.compag.2020.105606.

[7] Kang H, Chen C. Fast implementation of real-time fruit detection in apple orchards using deep learning. Comput Electron Agric 2020;168:105108. https://doi.org/10.1016/j.compag.2019.105108.

[8] Robin C, Lacroix S. Multi-robot target detection and tracking: taxonomy and survey. Auton Robots 2016;40(4):729–60. https://doi.org/10.1007/s10514-015-9491-7.

[9] Qi Z, Wang Z, Huang J, Xing C, Gao J. Error of image saturation in the structured-light method. Appl Opt 2018;57(1):A181. https://doi.org/10.1364/AO.57.00A181.

[10] Lu C, Song Y, Wu Y, Yang M. 3D information acquisition and error analysis based on ToF computational imaging. Infrared Laser Eng 2018;47:1–7. https://doi.org/10.3788/IRLA201847.1041004.

[11] Li Z, Yan S, Liu Y, Guo Y, Hong T, Lü S. Error analysis and compensation method for ToF depth-sensing camera. Mod Electron Tech 2021;44:50–5. https://doi.org/10.16652/j.issn.1004373x.2021.07.010.

[12] Jiang Du, Zheng Z, Li G, Sun Y, Kong J, Jiang G, et al. Gesture recognition based on binocular vision. Cluster Comput 2019;22(S6):13261–71. https://doi.org/10.1007/s10586-018-1844-5.

[13] Ji W, Meng X, Qian Z, Xu B, Zhao D. Branch localization method based on the skeleton feature extraction and stereo matching for apple harvesting robot. Int J Adv Robot Syst 2017;14:1–9. https://doi.org/10.1177/1729881417705276.

[14] Ye M, Zou X, Luo L, Liu N, Mo Y, Chen M, et al. Error analysis of dynamic localization tests based on binocular stereo vision on litchi harvesting manipulator. Trans Chinese Soc Agric Eng 2016;32:50–6. https://doi.org/10.11975/j.issn.1002-6819.2016.05.007.

[15] Chen H, Wang L, Liu G. A survey of stereo matching algorithms. Chinese High Technol Lett 2020;30:157–65. https://doi.org/10.3772/j.issn.1002-0470.2020.02.007.

[16] Jiao Y, Luo R, Li Q, Deng X, Yin X, Ruan C, et al. Detection and localization of overlapped fruits application in an apple harvesting robot. Electron 2020;9(6):1023. https://doi.org/10.3390/electronics9061023.

[17] Li H, Liu Q. Study on technology of location of apples. J Agric Mech Res 2016:54–7. https://doi.org/10.13427/j.cnki.njyi.2016.02.013.

[18] Williams HAM, Jones MH, Nejati M, Seabright MJ, Bell J, Penhall ND, et al. Robotic kiwifruit harvesting using machine vision, convolutional neural networks, and robotic arms. Biosyst Eng 2019;181:140–56. https://doi.org/10.1016/j.biosystemseng.2019.03.007.

[19] Kamilaris A, Prenafeta-Boldú FX. Deep learning in agriculture: A survey. Comput Electron Agric 2018;147:70–90. https://doi.org/10.1016/j.compag.2018.02.016.

[20] Gao F, Fu L, Zhang X, Majeed Y, Li R, Karkee M, et al. Multi-class fruit-on-plant detection for apple in SNAP system using Faster R-CNN. Comput Electron Agric 2020;176:105634. https://doi.org/10.1016/j.compag.2020.105634.

[21] Xiao C, Zheng L, Li M, Chen Y, Mai C. Apple detection from apple tree image based on BP neural network and Hough transform. Int J Agric Biol Eng 2015;8:46–53. https://doi.org/10.3965/j.ijabe.20150806.1239.

[22] Fu L, Majeed Y, Zhang X, Karkee M, Zhang Q. Faster R-CNN–based apple detection in dense-foliage fruiting-wall trees using RGB and depth features for robotic harvesting. Biosyst

Eng 2020;197(6):245–56. https://doi.org/10.1016/j.biosystemseng.2020.07.007.
[23] Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017;39(6):1137–49. https://doi.org/10.1109/TPAMI.2016.2577031.
[24] Wang S, Li Y, Yuan J, Song L, Liu X, Liu X. Recognition of cotton growth period for precise spraying based on convolution neural network. Inf Process Agric 2021;8(2):219–31. https://doi.org/10.1016/j.inpa.2020.05.001.
[25] Yu Y, Velastin SA, Yin F. Automatic grading of apples based on multi-features and weighted K-means clustering algorithm. Inf Process Agric 2020;7(4):556–65. https://doi.org/10.1016/j.inpa.2019.11.003.
[26] Otsu A. tlreshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 1979;20:62–6.
[27] Xing J, Yang P, Qingge L. Robust 2D Otsu's algorithm for uneven illumination image segmentation. Comput Intell Neurosci 2020;2020:1–14. https://doi.org/10.1155/2020/5047976.
[28] Si Y, Liu G, Feng J. Location of apples in trees using stereoscopic vision. Comput Electron Agric 2015;112:68–74. https://doi.org/10.1016/j.compag.2015.01.010.
[29] Lv D, Jiao G. Experiment of stereo matching algorithm based on binocular vision. J Phys Conf Ser 2020;1574(1):012173. https://doi.org/10.1088/1742-6596/1574/1/012173.
[30] Yin C, Xiang C, Song J, Qiao S. Fast stereo matching algorithm based on adaptive window and graph cuts. Opt Precis Eng 2008;16:1117–21. https://doi.org/10.1109/ICALIP.2010.5684994.
[31] Shen Y, Xiong W, Huang W, Xu W. Instrument recognition based on template matching and hough circle detection. Comput Technol Dev 2021;31:69–73. https://doi.org/10.3969/j.issn.1673-629X.2021.04.012.
[32] Carvalho OLFd, de Carvalho Júnior OA, Albuquerque AOd, Bem PPd, Silva CR, Ferreira PHG, et al. Instance segmentation for large, multi-channel remote sensing imagery using mask-RCNN and a mosaicking approach. Remote Sens 2021;13(1):39. https://doi.org/10.3390/rs13010039.