Analysis Of Deep Learning-Based Visual Perception Technology for Picking Robots

Xinyue Jiang [1], Hao Zeng [2, *]

[1] College of Big Data and Information Engineering, Guizhou University, Guizhou, 550025, China

[2] College of Mechanical and Energy Engineering, Beijing University of Technology, Beijing, 100124, China

* Corresponding Author Email: zenghaou@emails.bjut.edu.cn

Abstract. With global population growth and rising labor costs, the development of agricultural picking robots is crucial to improving agricultural productivity and reducing costs. A combination of deep learning technology and traditional vision technology has revolutionized the visual perception capability of agricultural picking robots. This has enabled robots to achieve more accurate fruit identification and localization in complex environments. In this paper, we review the visual perception technology of picking robots based on deep learning, which is primarily divided into two core technologies for fruit recognition and 3D reconstruction and localization during fruit picking. Firstly, we analyze the application of deep learning models in fruit recognition. We discuss how to integrate and process the feature information of fruits through deep learning neural networks to improve recognition accuracy and review the development of more typical deep learning models in the past three years. Secondly, we discuss the advantages and disadvantages of several types of traditional stereo vision technology. We synthesize the advantages and disadvantages of several types of current stereo vision technology that are more widely used. We also analyze how deep learning technology has optimized the model and combined it with traditional stereo vision technology in recent years to achieve the three-dimensional reconstruction and precise positioning of fruits. Finally, we summarize the main challenges of current picking robots, mainly the accuracy of the deep learning network, the difficulty of acquiring and calibrating the original image set, and the generality of the models carried by the picking robots. Additionally, we look forward to the future development direction.

Keywords: Deep learning; picking robots; machine vision; current state of research.

1. Introduction

China is the world's largest fruit producer, its production and planting area are among the world's top, according to China's National Bureau of Statistics, fruit production in 2022 for the first time exceeded 300 million tonnes, a variety of fruit production hit a record high. However, the comprehensive mechanization rate of fruit production in China is still low compared with developed countries, only 28.6%, especially in the picking stage mechanization rate is less than 3% [1]. In the process of fruit production operations, fruit picking is an important part and a labor-intensive work. With urbanization and population mobility, the rural population is gradually transferring to the cities, traditional agriculture is facing less and less available labor, and rural labor resources are becoming increasingly tight. At the same time, Chinese agriculture faces challenges such as rising agricultural production costs and unstable farmers' incomes. Therefore, the adoption of mechanical harvesting to replace manual fruit picking is an inevitable trend [2]. Robotic harvesters are designed to enhance the productivity of farming operations, alleviate the workload of agricultural laborers, and address the issue of labor scarcity by employing automated technologies.

The concept of robot-based fruit picking started in the United States in the 1960s, which led to the development of agricultural intelligence and research on intelligent picking robots. In 1983, the first picking robot was introduced [3]. The equipment at this stage was simple, mainly composed of mechanical arms and actuators, with low picking accuracy and efficiency, high damage rate, and could only work in specific environments. In the 1990s, with the development of vision technology, image processing and machine learning, the field of picking robots developed rapidly, and developed

countries represented by Japan, including France, the Netherlands, the United States, Israel and other countries, have achieved great success in the research of fruit picking robots, and developed a variety of picking robots with artificial intelligence [4]. Currently, with the rise of deep learning technology, fruit picking robots have stepped into a new stage of development to improve the recognition accuracy by automatically extracting fruit features and classification. At the same time, combined with machine learning and control theory, the robot has the ability to learn and adapt autonomously, and continuously improves the picking behavior, which is gradually maturing and being applied in real-world scenarios.

Since the 1990s, the field of harvesting robotics has developed rapidly, and many new methods for target recognition have emerged: 1) traditional digital image processing techniques, such as extraction and recognition based on features such as color, texture, shape, etc.; 2) machine learning based image segmentation techniques and classifiers, such as K-means based clustering algorithms, Bayesian classifier based algorithms, KNN based clustering algorithms, Adaboost and Haar-like feature based algorithms, support vector machine based algorithms, etc.; 3) deep learning based neural network algorithms, such as AlexNet, CNN, Faster R-CNN, VGGNet, SSD, YOLO, etc. [5].

Picking robots based on deep learning technology have significant research value in several aspects. Firstly, deep learning improves the recognition of fruits in changing natural environments (e.g., different light conditions during the day, nighttime environments) by learning a large amount of sample data, which can improve the robustness and adaptability of the system and better adapt to complex agricultural environments. Secondly, deep learning-based picking robots can work better under unstable conditions (e.g., sensor imaging blurring due to vibration of the robotic arm picking fruits, fruit shaking due to natural wind, and stroboscopic interference of the camera and sensors), and have stronger generalization ability. This means that the system can rely on the ability of field analysis to maintain high target recognition accuracy, thus improving the efficiency and reliability of the robot in real picking tasks. In addition, this technology can greatly save manpower, freeing manual labor from complex and tedious repetitive actions, allowing the agricultural population to be free from seasonal harvesting, and is expected to improve agricultural management.

Overall, the introduction of deep learning brings more advanced and reliable technical means to the vision system of picking robots, which will be a great help to promote the automation process in the agricultural field in the future. It is of great significance in improving the picking efficiency of agricultural products, reducing the burden of labor, and solving the problems of global food security and sustainable agricultural development.

This paper focuses on a series of deep learning-based techniques for accurate target recognition, 3D reconstruction and localization in the "pre-picking" period, i.e. the vision system that has a significant impact on the picking process.

2. Target Recognition Technology

The current research on target recognition technology for the vision system of picking robots mainly focuses on single-feature vision and multi-feature fusion vision. Although single-feature vision segmentation algorithms with color as the main feature are widely used, they are not effective enough to accurately distinguish target features when performing fruit detection in complex natural environments. As a result, multi-feature techniques are frequently employed to increase the application's constraints' robustness and efficiency. The segmentation of target and background is more reliable when texture differences are used in conjunction with algorithms like image color space and geometric characteristics; nevertheless, multi-feature fusion techniques raise the complexity of the algorithms and the system requirements [6]. In recent years, the rise of deep learning technology provides new ideas for the limitations existing in the result traditional recognition methods. At present, there are mainly CNN, RNN, YOLO, etc., and this part mainly introduces three algorithms, CNN, FAST R-CNN and YOLO, which have a wider range of applications.

2.1. Convolutional Neural Network

Convolutional neural network (CNN) is a widely used neural network model, the core idea of which is local connectivity between neurons and weight sharing in the convolutional layer, based on these two main characteristics makes it perform well in image classification and recognition, target detection and natural language processing.

The basic principle of CNN is to extract and learn image features through different hierarchical structures such as convolutional layer, pooling layer and fully connected layer, whose convolutional and pooling layers are used for image feature extraction and dimensionality reduction, while the fully connected layer is used to map the extracted features to the output layer to achieve the classification and recognition of images.

Li et al. designed and investigated a deep learning-based tomato harvesting robot, proposed an improved classification system that accomplishes the observation of tomato ripeness through a convolutional neural network. The design uses five layers of CNN to extract features, three different sizes of convolutional kernels, ReLU as the activation function of the architecture, and two layers of maximal pooling are inserted to preserve features and reduce unnecessary parameters. The design used a combination of geometric transformations and random noise and added Gaussian noise, Pepper and Salt to each of the three datasets (R, S and R&S datasets) to create nine different datasets for training. Three dataset enhancement methods were used to prevent overfitting in the model training problem. This allowed us to investigate the impact of the various data enhancement methods on the prediction results of this task. Training on the R & S & SN datasets produced the best prediction results, with an accuracy of 91.9% and a prediction time of less than 0.01 seconds, according to trials done on other datasets [7].

Zeeshan et al. proposed a deep learning convolutional neural network model for detecting oranges in complex dynamic environments and built a Keras sequential convolutional neural network model using convolutional layer activation function, maximum pooling and fully connected layers. The performance evaluation measures were substantially enhanced with the use of a diverse dataset of real-time pictures with noise, light variations, and occlusion. The suggested CNN model achieved 93.8% accuracy, 98% precision, 94.8% recall, and 96.5% F1 score [8].

2.2. Faster R-CNN

With the rapid development of deep learning technology, the target recognition algorithm of Faster regional CNN (Faster R-CNN), plays a crucial role in the fruit image recognition process. Before the appearance of Faster R-CNN, R-CNN and Fast R-CNN already existed, but the R-CNN training process will be repeated convolutional computation, which takes too much time; Fast R-CNN also has the same problem and does not achieve the real sense of end-to-end training. Faster R-CNN is a more advanced deep learning algorithm compared to Fast R-CNN, which can be regarded as a combination of Region Proposal Network (RPN) and Fast R-CNN monitoring network, and not only improves the speed of target detection, but also further improves the accuracy of detection by improving on Fast R-CNN.

Zhu et al. used Faster R-CNN technology for image feature recognition of wolfberry flowering stage and fruit ripening stage to achieve automatic observation of developmental stage through the processes of VGG16 network feature extraction, RPN network, ROI pooling, and target classification and regression. The difference between the results of automatic recognition based on Faster R-CNN and the result of expert visual judgement is not large, and the results of expert visual judgement can be used to optimize [9].

Zhu et al. used Faster R-CNN algorithm for detection and recognition analysis of blueberry canopy fruits, firstly, the fruit images were transformed into feature maps through the feature extraction process and shared with region proposal network (RPN) and target region pooling network (ROI Pooling). Secondly, RPN is used to perform a binary classification operation to distinguish the background from the target blueberries through sortmax classifier, and the candidate region locations are initially obtained through Bbox regression. Then pooling and normalization of the coarsely

filtered feature maps are performed using target region pooling network. Finally, after feature integration in the fully connected layer, the softmax classifier is responsible for accurate classification to discriminate the ripeness of blueberries. Meanwhile, Bbox regression is responsible for fine-tuning the position of the prediction box to ensure its accuracy. By statistically analyzing the recognition results of three ripeness levels of blueberry fruits, the average recognition accuracy of Faster R-CNN algorithm for blueberry fruits is 94.05% [10].

Li et al. investigated and improved the Faster R-CNN model for automatic strawberry identification and counting function. By optimizing the model RPN structure and adjusting the area of three kinds of base frames, the detection accuracy of unripe strawberries and ripe strawberries was improved; comparing four kinds of image backbone feature extraction networks (VGG16, ResNet50, VGG19 and ResNet101), it was concluded that ResNet50 was more effective in target detection of strawberries. Under the condition that ResNet50 is used as the backbone extraction network, all kinds of performance evaluation indexes of the improved Faster R-CNN model are better than that of the Faster R-CNN model, and its AP for ripe strawberries is 0.8930, the AP for mature strawberries is 0.8207, and the mAP is 0.8569, which is a big improvement compared to that of the Faster R-CNN model [11].

2.3. YOLO

In order to improve the efficiency of the algorithms, scholars have developed end-to-end single-stage detection algorithms, the most famous of which is the YOLO (You only look once) series of algorithms, of which YOLO v5 and YOLO v8 are the most advanced target detection algorithms currently available, balancing the overall accuracy with the speed, which makes them suitable for fast fruit detection in natural environments.

He et al. proposed a nighttime tomato fruit recognition algorithm based on improved YOLO v5 to achieve normal operation of a picking robot in the nighttime environment of a solar greenhouse as well as fast recognition of tomato fruits. The algorithm model is iterated using the computational function ANCHOR to optimize the detection frame and modifies the target detection loss function in the original data to better suit the night environment. As a result of these improvements, it provides a significant improvement in recognition accuracy. Specifically, the improved YOLO v5 model achieves 96.7% in average accuracy mean, which is 3.3 percentage points higher compared to 93.4% of the original YOLOv5. In addition, the improved YOLO v5 model has significant improvement for tomato fruit recognition in complex situations such as overlapping and occlusion of multiple fruits. In the case of occlusion of multiple fruits, the recognition accuracies of green and red tomato fruits reach 96% and 98%, respectively, with better robustness compared to the YOLO v5 model [12].

Liu et al. proposed an improved YOLO v5 method for orange detection and achieved better experimental results. Firstly, the RepVGG module is used to replace the CSPDarkNet53 backbone feature extraction network in the original YOLO v5 model, which strengthens the backbone network for feature acquisition of difficult-to-extract targets; secondly, the ghost blending convolution is used to replace the standard convolution in the necking network stage, to avoid the loss of feature information; thirdly, the ECA attention module is added to the output of the model, which is able to enhance the network's attention to oranges and locate the target information more accurately; finally, the shortcomings of the CIOU positional regression loss function of the original YOLO v5 network are improved using the EIOU loss function to enhance the regression accuracy of the bounding box. The improved YOLO v5 network achieves an average accuracy of 90.1% in orange detection in natural environments, which is more accurate than the current popular detection networks YOLO v3, YOLO v4 and CenterNet for orange recognition in complex situations such as branch and leaf occlusion and overlapping of neighboring fruits [13].

Yang et al. proposed an improved tomato detection algorithm for YOLO v8 model. It successfully reduces computational cost and increases the model's detection accuracy by combining operations like feature improvement and attention mechanism. Comparison experiments are conducted with a variety of algorithms (Faster R-CNN, YOLOv4, YOLOv5, YOLOv7 and SSD), and the algorithm

based on the improved YOLO v8 performs well under the same experimental environment, with a mAP of 93.4% and a smaller model of 16.1 MB. In addition to satisfying real-time detection needs, it also reduces model size and increases detection accuracy, both of which have greater practical and generalizability [14].

2.4. Transformer

Transformer is a self-attention mechanism-based model that exhibits strong performance in modeling global context and good transferability to tasks downstream after extensive pre-training. In the domains of natural language processing (NLP) and machine translation, this benefit has received extensive validation and use [15].

In target detection tasks, CNNs have been dominant thanks to properties such as translation invariance and local sensitivity of the convolutional kernel. However, it is challenging to gather and combine global picture information because of CNN's narrow receptive field [15]. Consequently, the Transformer model from natural language processing has been applied to computer vision challenges by certain researchers. Compared to CNN, the Transformer model is able to model the global dependencies of an image more efficiently and make fuller use of contextual information. It has been demonstrated to be applicable to a wide range of vision tasks, including image classification, target detection, and image segmentation, and has shown great potential for application. Some transformer-based architectures demonstrate powerful image processing capabilities, such as Vision transformer (ViT), Swin Transformer. The excellent performance of Vision transformer and Swin Transformer in image recognition tasks shows the potential of transformers for applications in vision.

Wang et al. proposed a network model consisting of three parts, Transformer backbone layer, multiscale convolutional layer and classification layer for conducting weed growth cycle identification studies. The modified and adapted VisionTransformer-base16 is used as the backbone network to generate feature information, and the generated information is exchanged for dimensionality as well as reshaping. In the multiscale convolution layer, continuous asymmetric convolution is introduced to replace the conventional convolution to reduce the risk of overfitting in the model. In the classification layer, the feature maps obtained after the convolution operation are input to the Global Average Pooling (GAP) layer for dimensionality reduction, followed by recognition and classification of weed images using the Softmax function. Through the comparison experiments of multiple network models on Leaf-Counting dataset, including ResNet-50, Inception-v3, ViT and the proposed ViT_Inception network, it can be concluded that the accuracy of the proposed model for the recognition of the weed growth cycle is improved by 4.1%, which has the advantage of faster recognition, and outperforms other mainstream models [16].

Zheng et al. proposed a "Swin-MLP" method based on Swin Transformer and multilayer perceptron (MLP) to recognize the appearance quality of strawberries, in order to overcome the disadvantages of convolutional neural networks, which are time-consuming and computationally intensive. Firstly, Swin Transformer is used to extract the strawberry image features, and then it is imported into the MLP neural network to recognize the strawberries. The "Swin-MLP" method's procedure is comparable to transfer learning, and its benefit is that it achieves great efficiency while cutting training time. During the experiment, the feature extraction time of Swin-T is 11.38 seconds and the training time of MLP is only 5.41 seconds, which achieves more than 98% accuracy in recognizing strawberries with different ripeness levels. When compared to alternative combinations of distinct classifiers, this approach can yield higher identification results with greater accuracy and less processing time. Swin Transformer has a wide range of potential applications in agriculture and can be used as a very efficient feature extractor [17].

3. 3D Reconstruction and Localization Techniques

The main challenges in the process of fruit localization include fruit displacement and fruit overlap caused by wind and mechanical vibration during the imaging process. In recent years, many methods

have been proposed by scholars to study 3D reconstruction and localization techniques for machine vision. The most commonly used methods include 3D localization based on monocular color cameras, stereo vision matching, laser range finders, depth cameras, and time-of-flight of light-based 3D cameras.

Early monocular cameras were replaced by binocular or multi-camera systems due to high errors. Although laser rangefinders perform well in complex lighting conditions, they may have limited localization due to occlusion of fruit in natural environments. On the other hand, depth cameras and light-based 3D cameras, which consist of a color camera and a depth camera, complement each other to improve the robustness of localization. These cameras are widely used in the field of target localization for picking robots.

This chapter classifies these methods into two main categories and discusses the application of deep learning in target localization. It also looks at the future directions in the field of agricultural picking.

3.1. Optical Geometry Based Stereo Vision Technology

3.1.1 Monocular stereo vision technology

Three-dimensional positioning using a monocular color camera involves capturing the image of the target object using the camera, and then using the geometric projection relationship between the spatial target feature points and the image feature points to determine the spatial position information of the object. The camera model is also used in the process. Depending on the number of images used for positioning, the monocular camera positioning can be classified into single, two or more images of monocular color camera positioning. The monocular camera positioning system has many advantages, including a simple structure, low cost and easy calibration. However, it has the disadvantage of having a large error in the precise positioning of the target fruit, which can easily cause damage to the end-effector. Monocular recognition and positioning system research is relatively mature and is the longest and most widely used category.

Mehta et al. have conducted a study in which they introduced an approach for determining the three-dimensional coordinates of specific fruits. This approach utilizes a single-lens camera system and perspective transformation techniques, which helps in real-time manipulation of an automated fruit-picking apparatus. The method uses the large field of view of a fixed camera and the high accuracy of a handheld camera (CiH) to gauge the depth of the fruit by computing image-based perspective transformations [18]. The researchers assume that the fruits have ellipsoidal geometrical features and use images captured by a monocular camera to estimate the depth of the fruits based on the known principal axis dimensions of the fruit samples. This approach does not require additional distance sensors, has a lower computational complexity than stereo vision techniques and is suitable for real-time applications. With this approach, the researcher can generate a global map of the fruit location and select the target fruit for picking. In practical applications, this monocular camera-based localization method demonstrated good accuracy and robustness, especially when dealing with partially occluded and clustered fruits.

Feng et al. used a monocular camera for identifying and locating ripe fruit bunches of cherry tomatoes. In the study, the team of investigators applied an R-G color space to boost the differentiation of the desired fruits from their surroundings within the captured image, followed by the identification of potential fruit cluster areas from the modified RGB-G image. Next, individual fruits were identified from the fruit bunches using the CogPMAlignTool tool, which is capable of handling scaling (0.8,1.2) and rotational ($-\pi$, π) variations of the fruits and accepts a scoring threshold of 0.36. With these image processing steps, the camera is able to accurately identify ripe fruit bunches, providing precise target localization for the robot's picking action [19].

3.1.2 Binocular stereo vision technology

Binocular cameras work by mimicking the human eye's visual system, which acquires depth information through the image difference between the two viewpoints. Compared to monocular

cameras, binocular cameras increase the perceptual range of the visual system and improve localization accuracy, but they can be slightly less lightweight and mobile.

Wang et al. proposed a method to compensate for the lack of localization accuracy in the case of occlusion by using two CCD cameras and a geometric centre-based matching method to identify lychee fruits [20]. They used binocular charge-coupled device (CCD) color cameras to acquire lychee images. The technique captures the stereo image of lychee through binocular camera calibration and image acquisition, after which four different supervised classifiers are applied to segment the lychee fruit. A pixel thresholding method is then used to identify predefined lychee aggregation categories, and the identified lychees are accurately matched by a geometric centre matching method. This method performs well in experiments, achieving 94.17% and 92.00% recognition accuracies for lychees under different lighting and occlusion conditions, respectively. Compared with the traditional monocular vision method, this method is more robust in dealing with occlusion and light variations and shows a higher matching success rate in the experiments.

Chen et al. designed a four-camera system based on the binocular vision technique and developed a point cloud correction algorithm to calibrate the camera system [21]. The system ensures the consistency of geometric parameters among the four cameras through accurate global calibration, thus providing a solid foundation for stereo matching and 3D reconstruction. The study adopted an adaptive stereo-matching strategy, which dynamically adjusts the search window to accommodate banana-centred stems of different depths based on the results of semantic segmentation. This adaptive approach significantly improves the efficiency and accuracy of stereo matching, enabling the robot to locate bananas quickly and accurately in an orchard environment. In the experiments, the four-camera system can achieve more than 98% recognition accuracy when dealing with bananas of different ripeness levels.

3.2. Laser-Based Stereo Vision Technology

3.2.1 RGB-D stereo vision imaging technology

Fu et al. conducted a study where they collected 800 sets of images using a Kinect 3D camera in a commercial orchard in Washington State. The images were taken under varying lighting conditions, both during the day and at night. The researchers created two image datasets: original RGB images and foreground RGB images after filtering the background using depth information. These images were then used to train and test two deep-learning models of Faster R-CNN (ZFNet and VGG16) [22].

As shown in Fig.1, when using the VGG16 model, the depth-processed image (blue) is consistently higher than the original image (red) in the average accuracy of the validation set; and when using the ZFNet model, the depth-processed image (violet) is also consistently higher than the original image (green) in the average accuracy of the validation set. The results show that the precision (P), recall (R), and average precision (AP) of training two Faster R-CNN models using foreground RGB images with depth-filtered backgrounds are improved compared to the original RGB images, suggesting that depth information is crucial for accurate fruit detection.

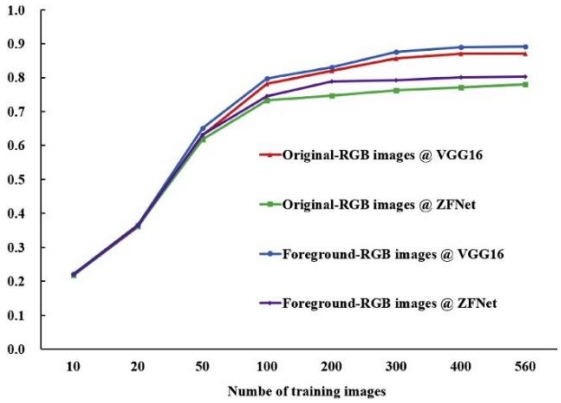

Fig. 1 Plot of average precision (AP) of ZFNet and VGG16 equipped with RGB and RGB-D cameras on the validation set versus the number of training images from the training set [22].

3.2.2 Stereo vision technology based on laser range finder

Laser rangefinders are capable of using scanning sensors to provide range information across a scene. In recent years, researchers have employed laser range finders in agricultural robotics.

Genemola et al. used a mobile terrestrial laser scanner (MTLS) to detect and localize apples. The method first uses MTLS to generate a 3D point cloud of apple trees and then distinguishes apples from leaves and trunks through reflectance analysis. The researchers developed a four-step fruit detection algorithm and achieved 87.5% localization success and 82.4% recognition success in their experiments [23]. Compared to RGB camera-based systems, this method has the advantage of being independent of lighting conditions and provides direct information about the 3D position of the fruit.

3.2.3 Tof-based stereo vision technology

Time-of-flight-based 3D localization is a technique used to measure the depth distance from a camera to an object. This method involves emitting continuous pulses of invisible light to the surface of the object, receiving the light pulses that are reflected past the object's surface, and recording the time between emitting and reflecting to calculate the depth distance. By implementing this technique, the camera can not only immediately create a distance map of the entire scene, but also provide depth information values as well as the 3D coordinates of the object in the field of view.

Zhang et al. proposed a hand-eye calibration method based on a TOF camera for improving the stability and efficiency of a picking robot in a fruit picking scenario. The method first fixes the TOF camera at the end of the robot, and then operates the robot to take pictures of the calibration plate from different poses to obtain clear images of the plate. With the depth information captured by the TOF camera, they were able to accurately locate the centroid points on the calibration plate and calculate the 3D coordinates of these points in the robot's base coordinate system [24]. The method was tested in a simulated peach picking scenario, and by combining deep learning and 3D vision technology, it achieved high-precision recognition and localization of peaches, and the error of the localization accuracy was controlled to be within 5 mm, which meets the accuracy requirements of the picking operation. The method significantly reduces the computation time while improving the recognition accuracy. The TOF camera, as an efficient depth information acquirer, shows great potential for application in the field of agricultural automation, especially in the hand-eye calibration of picking robots.

3.3. Comparative Analysis

Below is a summary of the advantages and limitations of five targeting techniques, grouped into two categories.

Table 1. Analysis of optical geometry and laser-based stereo vision techniques

Classification	Targeting techniques	Advantages	Limitations
Optical geometry based stereo vision technology	Monocular vision system	Streamlined, cost-effective, easy to calibrate and identify	Unable to determine the true size of an object, sensitive to lighting conditions, accuracy affected by changes in lighting
	Binocular stereo vision system	Proximity measurement with high accuracy, low cost and high image resolution	Affected by light variations and target texture, unable to work properly in dark environments, and the stereo matching and calibration process is more complex and takes longer to compute
Laser-based stereo vision technology	Based on depth camera	Proximity measurement with high accuracy, The accuracy remains consistent regardless of variations in light and surface texture.	Susceptible to reflections on the target surface. Low-depth image resolution, low accuracy of object edge detection
	Based on laser rangefinders	Long measuring distance, high accuracy, high speed, simple structure	Out of focus when measuring too long a distance, produces false signals when obscured by branches or overlapping fruits
	Based on light-based ToF camera	Wide measuring range, high immunity to interference, unaffected by light changes and surface texture	Susceptible to multiple reflections from target surfaces. Structured light has a drowned laser dispersion and is not suitable for use in outdoor environments

3.4. Deep Learning in 3D Reconstruction

3D reconstruction techniques based on deep learning have gained widespread use in various industries such as autonomous driving, medical reconstruction, cultural relic restoration, and more, as stated in [25]. While there are numerous deep learning algorithms available for fruit recognition in agriculture, accurately positioning fruit through deep learning-based image processing algorithms remains a challenging task.

In recent years, researchers have shown increasing interest in new visual sensors such as TOF, light field, and chlorophyll fluorescence cameras. These sensors are particularly relevant to the complex natural environments in which picking robots operate. However, limited access to data sets for various growth stages of fruits, individual differences between fruits, and the large computational volume of neural network training can often result in picking inefficiency. To address these challenges, it is important to expand the collected image data sets of fruits and vegetables or extract meaningful characteristics of the fruits through information processing technology. This approach enables accurate target area judgement, decision fusion for lighting conditions, and effective target identification and positioning. Ultimately, such techniques reduce the demand for neural network computation and shorten training time.

The simplified model proposed by Zheng et al. begins with the original image data and expands it to extract the sign features of the fruit from captured images. The model then performs decision fusion to account for occlusion or lighting conditions, before merging to complete the recognition and localization process. This approach effectively reduces the demand for data volume and computation, while shortening network training time [5].

Tang et al. proposed a technique for detecting and localising oleander fruits based on an improved YOLOv4-tiny model and binocular stereo vision. They used a K-means++ clustering algorithm to determine the bounding box to fit the oleifera fruit [26]. The team used an innovative strategy to accurately capture the critical region of interest (ROI) of the fruit using the bounding box produced by the YOLO-Oleifera model and performed stereo vision matching based on the rules generated by the bounding box to effectively estimate the parallax. This approach significantly reduces the computational complexity compared to traditional stereo-matching techniques that rely on binocular camera images. Experimental results show that the detection speed of the YOLO-Oleifera model averages 31 ms, which is sufficient for real-time detection. Compared to other deep learning models, the model achieved the highest detection accuracy of 92.07% in the orchard environment, and the model size was only 29 MB, reducing the hardware requirements for mobile picking robots. This study provides efficient and accurate technical support for the automated picking of oleander fruits, which has the potential to be widely applied in the field of agricultural automation.

4. Challenges and Future Prospects

4.1. Existing Challenges

With the rapid development of science and technology, the application of computer vision technology in the field of agriculture is becoming more and more widespread, especially in the fruit recognition and positioning has made significant progress. However, at the same time, the existing research and prototype still have certain limitations, it is difficult to be applied to the real picking scene to replace manual harvesting, and there is still a considerable gap from the commercial application, mainly facing the following challenges:

Firstly, various fruit detection and recognition techniques are currently facing some challenges, mainly because picking robots need to work in variable natural environments. Traditional digital image processing techniques are more seriously affected by factors such as light and occlusion. Although convolutional neural networks based on deep learning can well solve the problem of target detection and recognition, neural networks have a huge demand for training datasets, and the model training is computationally intensive and time-consuming. Furthermore, despite the relatively high correct rate of target recognition, the deep learning network still exhibits a significant mistake in target placing in the agricultural field because of the environment's complexity and unpredictability. In complex natural environments, affected by many factors such as the environment, the growth state of fruits and vegetables, and changes in the characteristics of the fruits and vegetables themselves, the more factors considered by the deep learning network, the more complex the corresponding network structure is, and the longer the running computation time is, resulting in a low real-time machine vision system and affecting the efficiency of picking.

The second involves the process of acquiring and calibrating the data training set, which is one of the biggest challenges faced when applying deep learning techniques to real-world problems. In order to train an efficient and accurate model, a large amount of high-quality image data is required. Not only do these data need to be acquired under different environmental conditions to ensure that the model can adapt to a variety of real-world scenarios, but they also need to contain a diversity of fruit types, sizes, shapes, and colors so that the model can learn rich features. The process of acquiring these images is a difficult task in itself. Firstly, field photography is required in a variety of different natural environments, which may include different lighting conditions, weather conditions, and seasonal changes. Additionally, in order to ensure diversity in the images, they need to be taken at different points in time to capture the different stages of fruit growth. Once the images have been

acquired, the next step is to calibrate them. This is a meticulous and time-consuming process that requires precise labelling of the fruit in each image, including the location, size and possible defects of the fruit. This usually involves manually drawing bounding boxes using image processing software or using semi-automatic tools to aid in the labelling, and the accuracy of the calibration directly affects the effectiveness of the model training. In addition to manual calibration, cleaning and enhancement of the image data is also required. All these steps are to ensure the quality and diversity of the training dataset, thus improving the generalization ability and robustness of the model. Therefore, the acquisition and calibration of data training sets is a complex and resource-intensive process, which requires professional knowledge, technology, and a large investment of human and material resources.

Third, the generality of the fruit recognition and localization model. Since fruit targets growing in the field have different colors and sizes at different growth stages, the development of a fruit target recognition model with higher generality improves the accuracy of decision-making for picking robots. In addition, increasing model versatility would allow for a wider range of scenarios for robots rather than the current need to match one fruit type with a corresponding robot. This will also reduce the difficulty of commercializing robots and increase the rate of mechanical use in agricultural scenarios.

4.2. Future Prospects

Intelligent agriculture has gradually become a research hotspot in recent years, and autonomous intelligent fruit picking robots are the future development trend, using advanced technology and intelligent equipment to achieve automation, precision and intelligence of agricultural production. Intelligent agriculture includes many aspects, among which the research and development of fruit picking robots is an important aspect. Based on the existing problems, the following aspects are worthy of attention in future research:

(1) Multi-field and multi-disciplinary joint collaboration. Fruit picking robots gather multidisciplinary core technologies and methods such as agronomy, computers, deep learning, intelligent systems, sensors, dynamic control, software design, and crop management, which are technically difficult and costly to develop. Agricultural picking robots have not been commercially promoted on a large scale both domestically and internationally, and the development of real-time, low-cost, and fully automated picking robots still requires further efforts and exploration. In horticulture, a variety of different planting patterns can be developed to reduce shading, improve the visibility of fruits, and reduce the difficulty of recognition by deep learning networks and fruit picking by end-effector. Through the combination of agro-mechanical and agro-technical, the picking efficiency can be improved, thus promoting the development of mechanized fruit picking.

(2) Further research and application of vision system with multi-sensor fusion. Currently, single sensors have certain limitations and are difficult to meet the requirements of vision systems in complex and changing agricultural outdoor scenes. By integrating multiple sensors such as pressure sensors, palm cameras, tactile sensors and other sensors into the end-effector, and integrating the information from multiple sensors, the multi-sensor fusion picking robot can improve the accuracy of fruit recognition and positioning accuracy and enhance its adaptability to complex operating environments.

(3) Optimization and enhancement of recognition and localization algorithms. The performance of the vision system is closely related to the corresponding processing algorithms, and deep learning-based algorithms are still the best choice for fruit recognition and localization in natural environments. In order to further improve the efficiency of fruit picking robots in natural environments, the deep learning algorithms need to be further optimized, including expanding the dataset to improve the generalization ability of the model, optimizing the feature extraction module of the algorithms to capture more subtle image features, introducing the attention mechanism module to strengthen the model's focus on key information, and simplifying the model structure to increase the computing speed, etc., so as to achieve the improvement of the accuracy or speed of fruit recognition. The improvement of fruit recognition accuracy or speed can be achieved.

(4) Integration of 5G communication technology and Internet of Things (IoT) technology to build a big data cloud computing platform. 5G mobile communication technology solves the problem of real-time machine vision, making high-definition real-time transmission, multi-machine collaborative cluster operation, and high-definition processing possible, and providing technical support for the realization of a high degree of automation, intelligence, and even unmanned modern agricultural equipment. With the wide application of IoT technology, the fruit picking robot can complete the identification and positioning of fruit images through the cloud platform; use the cloud platform for network monitoring and collection of control systems, vision systems, robotic arms and sensors of the working parameters, and big data analysis and optimization; at the same time, use the cloud computing platform's powerful computing power to improve the real-time and accuracy of the picking robot.

5. Conclusion

This paper provides a detailed overview of deep learning-based visual perception techniques for picking robots used in agriculture. It can be divided into two main segments. The first segment is fruit recognition using deep learning models. By analyzing the development of typical deep learning models over the last three years, we find that these models demonstrate significant advantages in dealing with fruit detection tasks in complex orchard environments. They gradually improve accuracy and speed while reducing the size of the model. The second segment involves the introduction of deep learning technology for 3D reconstruction and localization of fruits based on traditional classification of the two techniques. Stereo vision techniques applied to picking robots in recent years have still been dominated by traditional optical and laser sensors. Compared with target recognition techniques, deep learning has been less applied to 3D reconstruction and localization techniques in the field of agricultural picking. This paper analyzes the pros and cons of several types of traditional stereo vision technologies that are more widely used. It also discusses cases of combining deep learning technologies with traditional vision technologies to explore future development directions.

In summary, although deep learning can significantly improve the accuracy and robustness of picking robots, the current technology still has limitations in complex natural environments. Future research should focus on exploring more efficient algorithms and lighter hardware support. Developing more versatile deep learning models and mechanical structures to enhance the applicable environment of picking robots will promote a large number of agricultural picking robots to be put into practical applications. Ultimately, it can lead to comprehensive innovation in agricultural automation and free up the agricultural labor force.

Authors contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] Yuanmin Gou, Jianwei Yan, Fugui Zhang, et al. Research Progress on Vision System and Manipulator of Fruit Picking Robot. Computer Engineering and Applications, 2023, 59(9): 13-26.

[2] Qing Chen, Chengkai Yin, Ziliang Guo, et al. Current status and future development of the key technologies for apple picking robots. Transactions of the Chinese Society of Agricultural Engineering, 2023, 39(4): 1-15.

[3] Feng Lan, Zihao Su, Ziming Li, et al. The Actuality and Development Directions of Fruit Harvesting Machine. Journal of Agricultural Mechanization Research, 2010, 32(11): 249-252.

[4] Jian Song, Tiezhong Zhang, Liming Xu, et al. Research Actuality and Prospect of Picking Robot for Fruits and Vegetables. Transactions of the Chinese Society for Agricultural Machinery, 2006(5): 158-162.

[5] Taixiong Zheng, Mingzhe Jiang, Mingchi Feng. Vision based target recognition and location for picking robot:A review. Chinese Journal of Scientific Instrument, 2021, 42(9): 28-51.

[6] Yunchao Tang, Mingyou Chen, Chenglin Wang, et al. Recognition and Localization Methods for Vision-Based Fruit Picking Robots: A Review. Frontiers in Plant Science, 2020, 11: 510.

[7] Zhang Li, Jingdun Jia, Guan Gui, et al. Deep Learning Based Improved Classification System for Designing Tomato Harvesting Robot. IEEE Access, 2018, 6: 67940-67950.

[8] Zeeshan Sadaf, Tauseef Aized, and Fahid Riaz. The Design and Evaluation of an Orange-Fruit Detection Model in a Dynamic Environment Using a Convolutional Neural Network. Sustainability, 2023, 15(5): 4329.

[9] Zhu Yongning, Zhou Wang, Yang Yang, et al. Automatic Identification Technology of Lycium barbarum Flowering Period Based on Faster R-CNN. Chinese Journal of Agrometeorology, 2020, 41(10): 668-677.

[10] Zhu Xu, Ma Hao, Ji Jiangtao, et al. Detecting and identifying blueberry canopy fruits based on Faster R-CNN. Journal of Southern Agriculture, 2020, 51(6): 1493-1501.

[11] Li Jiajun, Zhu Zifeng, Liu Hongxin, et al. Strawberry fruit recognition algorithm based on improved Faster R-CNN model. Hubei Agricultural Sciences, 2023, 62(11): 183-190.

[12] He Bin, Zhang Yibo, Gong Jianlin, et al. Fast Recognition of Tomato Fruit in Greenhouse at Night Based on Improved YOLO v5. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(5): 201-208.

[13] Liu Zhongyi, Wei Dengfeng, Li Meng, et al. Orange fruit recognition method based on improved YOLO v5. Jiangsu Agricultural Sciences, 2023, 51(19): 173-181.

[14] Guoliang Yang, Jixiang Wang, Ziling Nie, et al. A Lightweight YOLOv8 Tomato Detection Algorithm Combining Feature Enhancement and Attention. Agronomy, 2023, 13(7): 1824.

[15] Li Xiang, Zhang Tao, Zhang Zhe, et al. Survey of Transformer Research in Computer Vision. Computer Engineering and Applications, 2023, 59(1): 1-14.

[16] Wang Guishen, Yang Chenglin, Pu Jiajia, et al. Identification of agricultural weed growth cycle based on improved Vision Transformer. Journal of Changchun University of Technology, 2022, 43(6): 712-718.

[17] Hao Zheng, Guohui Wang, and Xuchen Li. Swin-MLP: a strawberry appearance quality identification method by Swin Transformer and multi-layer perceptron. Journal of Food Measurement and Characterization, 2022, 16(4): 2789-2800.

[18] S.S. Mehta, T.F. Burks. Vision-based control of robotic manipulator for citrus harvesting. Computers and Electronics in Agriculture, 2014, 102: 146-158.

[19] Qingchun Feng, Wei Zou, Pengfei Fan, et al. Design and test of robotic harvesting system for cherry tomato. International Journal of Agricultural and Biological Engineering, 2018, 11(1): 96-100.

[20] Chenglin Wang, Yunchao Tang, Xiangjun Zou, et al. Recognition and Matching of Clustered Mature Litchi Fruits Using Binocular Charge-Coupled Device (CCD) Color Cameras. Sensors, 2017, 17(11): 2564.

[21] Mingyou Chen, Yunchao Tang, Xiangjun Zou, et al. Three-dimensional perception of orchard banana central stock enhanced by adaptive multi-vision technology. Computers and Electronics in Agriculture, 2020, 174: 105508.

[22] Longsheng Fu, Yaqoob Majeed, Xin Zhang, et al. Faster R-CNN-based apple detection in dense-foliage fruiting-wall trees using RGB and depth features for robotic harvesting. Biosystems engineering, 2020, 197: 245-256.

[23] Jordi Gené-Mola, Eduard Gregorio, Javier Guevara, et al. Fruit detection in an apple orchard using a mobile terrestrial laser scanner. Biosystems engineering, 2019, 187: 171-184.

[24] Xiangsheng Zhang, Meng Yao, Qi Cheng, et al. A novel hand-eye calibration method of picking robot based on TOF camera. Frontiers in Plant Science, 2023, 13: 1099033.

[25] Long Xiaoxiao, Cheng Xinjing, Zhu Hao, et al. Recent progress in 3D vision. Journal of Image and Graphics, 2021, 26(6): 1389-1428.

[26] Yunchao Tang, Hao Zhou, Hongjun Wang, et al. Fruit detection and positioning technology for a Camellia oleifera C. Abel orchard based on improved YOLOv4-tiny model and binocular stereo vision. Expert Systems with Applications, 2023, 211: 118573.