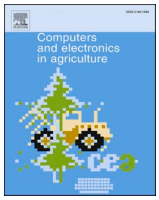
Computers and Electronics in Agriculture 232 (2025) 110077

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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



Design and verification of a litchi combing and cutting end-effector based on visual-tactile fusion

Zhaoshen Yao [a], Juntao Xiong [a,*], Jiayuan Yang [a], Xiao Wang [a], Zexing Li [a], Yuhua Huang [a], Yanan Li [b]

[a] College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China
[b] School of Engineering and Informatics, University of Sussex, Brighton BN1 9RH, UK

ARTICLE INFO

Keywords:
Litchi picking
Visual-tactile fusion applications
End-effector
Segment Anything Model
Picking Robot System

ABSTRACT

According to the characteristics of litchi tree branching and litchi fruit growing in clusters, this paper designs a litchi combing and cutting end-effector based on visual-tactile fusion with reference to human hair-cutting behavior. The litchi information in the image is perceived visually using the zero-sample prediction large model and the Hough circle detection principle, the tactile timing of the litchi fruit stem is captured by the tactile sensors between the clamping fingers, and the fruit stem clamping status is monitored through the tactile timing update mechanism, and finally the fusion at the decision-making level is realized by integrating visual and tactile in information, realizing the perception of the litchi fruit stem for the litchi picking robot. The picking trial results show that the static tactile fruit stem perception rate is only 16.7%, while the use of visual-tactile fusion increases the rate to 86.7%. The average of a single picking achieves 1.08 bunches. The end actuator and clamping and cutting picking method in this paper provide technical basis for efficient and low-loss litchi picking.

1. Introduction

China is the world's largest producer of litchi, with a production area that accounts for about 52.6 % of the global litchi production area (Qi et al., 2019). Fruit picking requires 60–70 % of the labor for the entire planting process, and the aging population has led to a shrinking of human resources in agricultural processing, which affects the efficiency of picking and pushes up the cost of labor, and these changes have become a core challenge for fruit farmers (Zhang et al., 2024).

As agricultural modernization progresses, the mechanization of fruit harvesting has emerged as a crucial component of the production process (Pu et al., 2023). The end-effector, which serves as the interface between the picking robot and its environment, is pivotal in the harvesting process. Research in this area has primarily concentrated on the development of structural designs and control algorithms (Song et al., 2021; Suo et al., 2021; Bac et al., 2014). Li et al. (Li and Liu, 2023) introduced a multi-armed robotic harvesting system, suitable for the automated collection of spherical fruits like apples. Li et al. (Li et al., 2023) analyzed the process of manually grasping pears and designed an adaptive rope-driven end-effector for picking pear. Mu et al. (Mu et al., 2020) designed a biomimetic finger-type kiwifruit end-effector with tactile sensing capability. The use of different cultivation modes affects fruit quality, and further integration of horticultural techniques in the cultivation process can better enhance the convenience of picking operations (Hu et al., 2022; Xiulan et al., 2024). In conclusion, the tailored design of picking objects stands out as a potent strategy for enhancing the efficiency of the picking process.

In recent years, there has been a significant advancement in the field of robotic perception technology. Liang et al. (Liang et al., 2020) successfully achieved the recognition of litchi in natural settings through image processing techniques employing deep learning. For the precise identification of fruit picking points, Qi et al. (Qi et al., 2022) developed an innovative method that integrates YOLOv5 and PSNet. Bosilj et al. (Bosilj et al., 2018) conducted a thorough analysis of plant attributes and proposed a pipeline based on attribute morphology, which is effective for both segmentation and classification tasks. To bolster the generalizability of machine vision, OpenAI has introduced a model grounded in contrastive language-image pretraining (CLIP). This model,

* Corresponding author.
E-mail addresses: 790639993@qq.com (Z. Yao), xiongjt2340@163.com, 309283022@qq.com (J. Xiong), 1360816089@qq.com (J. Yang), 573456679@qq.com (X. Wang), 1064727105@qq.com (Z. Li), 2496040555@qq.com (Y. Huang), yl557@sussex.ac.uk (Y. Li).

https://doi.org/10.1016/j.compag.2025.110077
Received 15 July 2024; Received in revised form 18 November 2024; Accepted 2 February 2025
Available online 12 February 2025

trained on a vast array of datasets, is capable of aligning image and text classification with remarkable accuracy (Radford et al., 2021). Huang et al. (Huang et al., 2023) proposed a specialized neural network that utilizes an enhanced Transformer structure to process the complete tactile time sequence during grasping, which is instrumental for the implementation of tactile sensing in agricultural robotics. Regarding information processing, Liu et al. (Liu et al., 2024) delved into the tactile temporal information during the grasping process by applying the Short-Time Fourier Transform (STFT) and Discrete Wavelet Transform (DWT). They were able to accurately capture the variations in contact force signals within the time–frequency domain during sliding, thereby facilitating the safe grasping of delicate fruits.

To address the current market's dearth of effective litchi picking tools, the limited capacity for real-scene perception, and the suboptimal success rate in picking, this paper introduces an innovative end-effector. This end-effector conducts a comprehensive analysis of the lychee picking environment and the actual operational workflow, integrating visual and tactile data to facilitate the picking process. The end-effector adopts a comb and cut structure to meet the demand for litchi picking and leaf combing, and uses a generalized vision model with zero sample prediction to perceive the pre-picking point scene. Furthermore, it incorporates a tactile time sequence updating mechanism that dynamically processes input from multiple tactile sensors, enabling real-time monitoring of the fruit stem clamping status across each finger-clamping channel. The fusion of visual and tactile information processing realizes accurate perception and operation of litchi picking, and a litchi picking robot is designed to verify its effectiveness in real scenarios.

2. Materials and methods

2.1. End-effector design and operation

2.1.1. Picking scenario analysis

Litchi is widely grown in the south of China's Lingnan region, especially in the rugged mountainous areas. The compact shape of litchi trees in these areas, the complex and variable terrain, and the dense foliage and uneven distribution of fruits create a typical unstructured picking environment, which poses significant challenges to mechanized litchi picking.

Different varieties of litchi have different growth characteristics. In this paper, we take the example of dwarfed "Gui Wei" litchi trees, whose average height is controlled to be less than 5 m. During the summer when the fruits are ripening, the furthest distance from the stump to the sunny side is about 3 m, while the distance to the shady side is about 20 % less than that to the sunny side. At this stage of growth, litchi fruits droop due to gravity and grow interspersed with leaves. The fruiting branch morphology of litchi is quite complex and its structure can be categorized into parent branch, meristem and fruit stem in descending order from the thickest to the thinnest. Fig. 1(a) shows multiple bunches of litchi on a tree, Fig. 1(b) clearly shows the specific construction of a single bunch of litchi, while Fig. 1(c) provides a simplified schematic of the structure of a litchi fruit stem.

Detailed measurements were made on the diameter of the parent branch and the bending angle of the fruiting branch of "Gui Wei" litchi, and the measurement position was set at 5–10 cm above the main node of the parent branch as shown in Fig. 2(a). Collecting 140 sets of data and counting the frequency percentage results as shown in Fig. 2(b)(c), it turned out that the diameter of the parent branch showed a normal distribution, which was mainly concentrated at about 5 mm, with the highest percentage in the range of 5–6 mm. The bending angle of the fruiting peduncle was mainly concentrated in the range of 0° to 60°. During the measurement process, it was observed that the degree of bending of a bunch of litchi showed a positive relationship with its number of fruits. These above values provide parameters for the design of the end-effector cutting mechanism.

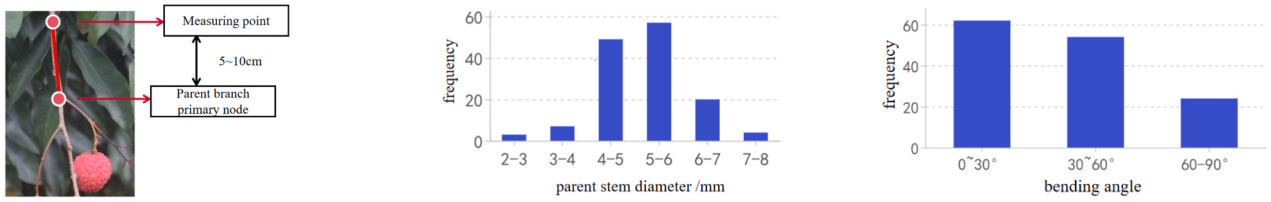
The key steps in the litchi picking process include accurately locating the fruit position, avoiding interference from branches and leaves, and harvesting the fruit efficiently and safely. As litchi branches and trunks are leafy, the end of the picker needs to be equipped with an appropriate tip design to minimize collision with the leaves and reduce localization interference. In addition, litchi fruits often grow in bunches and the distance between neighboring fruits is relatively close, which means that the possibility of harvesting multiple fruits at the same time can be taken into account when designing the picking strategy. Since most parts of litchi fruit trees are rigid, the end-to-end autonomous control system has to cope with possible interferences and ensure the safety and efficiency of the picking process.

2.1.2. Structural design of comb and cut end-effector

In this paper, a litchi combing and cutting end-effector based on visual-tactile fusion is designed, and its specific structure is shown in Fig. 3. The clamping and cutting operation of the motion unit is controlled pneumatically, and each scissor is individually controlled by a cylinder, which adopts a push–pull slot to complete the diagonal cutting of the scissors, and a total of three scissor actuating units are assembled. The clamping arm includes a slide cylinder for uniform

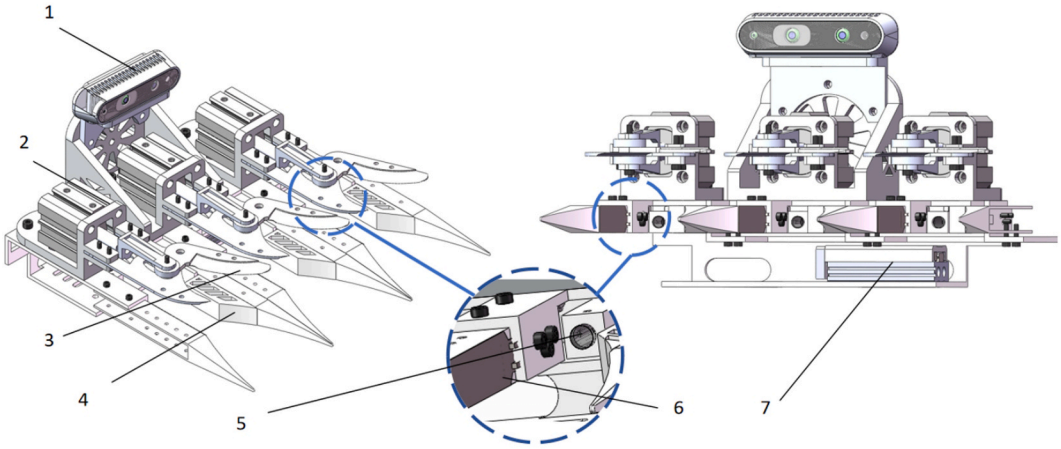
(a) Multiple Bunches of Litchi Figure (b) Single Bunch of Litchi Figure (c) Simplified Diagram of Litchi Fruit Stem

Fig. 1. Litchi diagrams and simplified diagrams.



(a)Measurement position diagram (b) Frequency diagram of diameter distribution (c) Frequency diagram of bending angle distribution

Fig. 2. Measurement and analysis of litchi characteristics.



1. RealSense D435i depth camera 2. SDA25 × 20 cylinder 3. Scissors 4. Soft clamping arm 5. Endoscope 6. RP-L tactile sensor 7. Slide cylinder

Fig. 3. Litchi combing and cutting end-effector.

drive, to prevent the litchi string position backward resulting in miscalculation of the cutting area. One side of the clamping arm is equipped with a fish-fin structure to reduce the possibility of the picking process damage, and the other side is equipped with RP-L tactile sensors for sensing external forces. The robotic arm is equipped with an eye-in-hand RealSense D435i depth camera, which is used to obtain the image information of the surrounding environment and depth information to localize the fruits, and in the middle of each of the two clamping arms is equipped with a borescope, to obtain the local visual information in the picking process.

2.1.3. *Picking point positioning analysis*

In the process of harvesting target fruits by the end-effector, it is necessary to define the picking point first, and there are usually two ways: one defines the center of mass as the picking point to allow the end-effector to harvest single fruits; the other defines the picking point in the main stems of the bunches in order to allow the actuator to harvest bunches of fruits (Xiong et al., 2023; Ye et al., 2021; Tang et al., 2020).

As shown in Fig. 4, the depth camera is used to obtain the position information of the litchi, with the infrared structured light to obtain the depth, the minimum detection distance D as a reference point, while taking into account the litchi bunches of fruit stems relative to the vertical direction of the angle θ. The end of the end-effector is also in the direction of the angle of θ as the entry attitude, to obtain an ideal pre-picking point position, which can be used for the end of the combing to set aside a certain amount of space, or through the endoscope to allow the end-effector to harvest the bunches of fruits. At the same time, the endoscope can be used to identify the fruits in each clamping channel.

The choice of θ angle as the end entry stance is to optimize the effect of scissors cutting the fruiting stems. As shown in Fig. 4(c), we abstract the fruiting parent branch for modeling, keeping the vertical tangent direction. When the end-effector is perpendicular to the fruiting stem, as shown in Fig. 4(d), the equivalent diameter is shortest in the shear direction, and less force can be used to cut the fruiting stem.

The transformation matrix ${}^{W}_{A}R$ of the camera relative to the base is obtained by means of hand-eye calibration (Tsai and Lenz, 1989), and the position of the picking point P_W in the world coordinate system can be calculated according to the 3D rotation matrix. The calculation formula is shown below:

$$P_W = {}^{W}_{A}R \times P_A + T \tag{1}$$

where P_A is the position of the uppermost litchi fruit under the camera coordinate system, T is the displacement matrix, specifically $T = [0, Zsin\theta - D - L, Zcos\theta]$, Z is the counted length of the uppermost litchi fruit from the node of the main stem, and L is the distance from the end clamp finger to the scissors.

2.1.4. *Picking workflow*

The picking flow chart is detailed in Fig. 5, and the implementation process of this paper is as follows: the YOLO series algorithm is used as the target detection algorithm to locate the position of the target fruit (Redmon et al., 2016); combined with the hand-eye calibration and coordinate conversion calculation, the position conversion from the pixel coordinate system to the world coordinate system is realized; in the robot operating system, inverse kinematics calculations are carried out

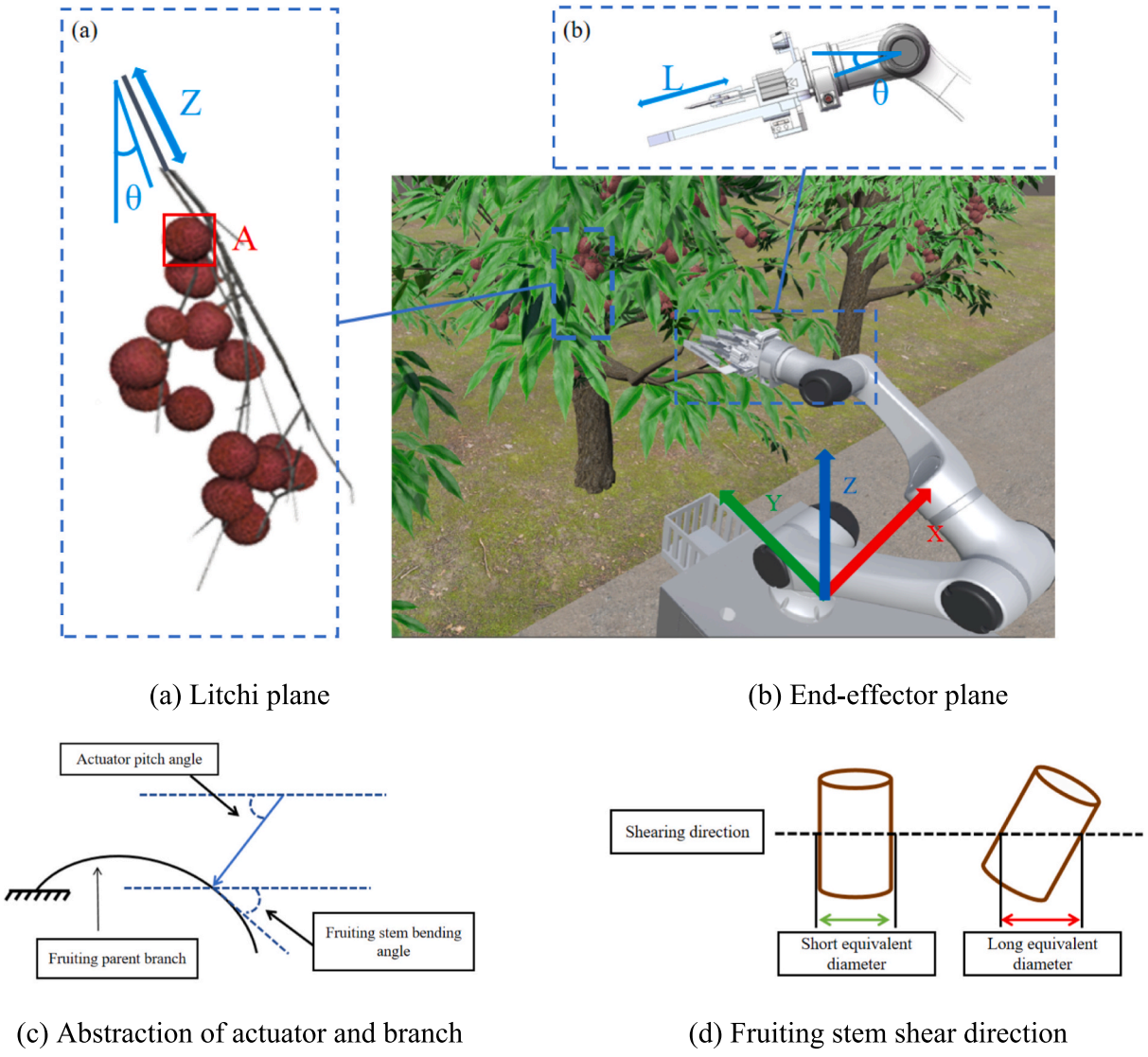


Fig. 4. End-effector pre-picking point pose calculation.

on the given coordinate points to determine the movement of the various joints of the robot. Subsequently, a suitable trajectory is determined through path planning, and the robot arm is controlled to follow the trajectory, so that the end-effector reaches the corresponding position. At the pre-picking point, an endoscope is used to obtain a local visual image, and after visual processing to determine whether litchi fruit exists in this finger-clamping channel, the end-effector penetrates into the picking area to ensure that the litchi bunches reach the scissors cutting area. After necessary combing, the clamping plate is driven by the slide cylinder to clamp the litchi fruit stems, and after tactile processing to determine the clamping condition of the fruit stems, corresponding scissors execute the cutting action, avoiding pulling and tugging action during operation, that may cause damage to the fruit tree and the end-effector structure. Finally, the robot cuts and clamps the target fruit bunch to complete the end-effector picking operation.

2.2. Visual information processing

In order to match the cutting behavior of the end-effector mentioned above, this paper proposes a generalized visual-tactile processing method. The zero-sample prediction of the picking process scene is carried out using the Open Vocabulary Segmentation method (Ren et al., 2024). The Open Vocabulary Segmentation was realized by combining Grounding DINO (Liu et al., 2023)and Segment Anything Model (SAM) (Kirillov et al., 2023). The generated mask image is subjected to an image edge detection algorithm to determine the existence of fruits, and the tactile time sequence updating mechanism with multi-sensors is combined to monitor the clamping state of each finger-clamping channel, which is mainly used for visual-tactile sensing fusion by the decision layer and guides the actions of the cutting mechanism through the concatenation logic.

2.2.1. Mask extraction

For fruit mask extraction in picking scenarios, this paper uses a generalized visual macromodel, SAM, as the base process. Different from traditional segmentation methods, SAM introduces the concept of promptable segmentation task, relying on flexible model architecture and diversified training data. SAM has been trained on 11 million images with more than 1 billion masks, and is capable of handling all kinds of segmentation cues, such as points, box, and mask, and realizes image segmentation without sample learning.

As shown in Fig. 6, SAM consists of three main components: an image encoder, a prompt encoder, and a mask decoder.

Image encoder: SAM employs a vision converter ViT (Dosovitskiy et al., 2020) based on MAE (He et al., 2022) pre-training as an image encoder in order to process high resolution images. This encoder design

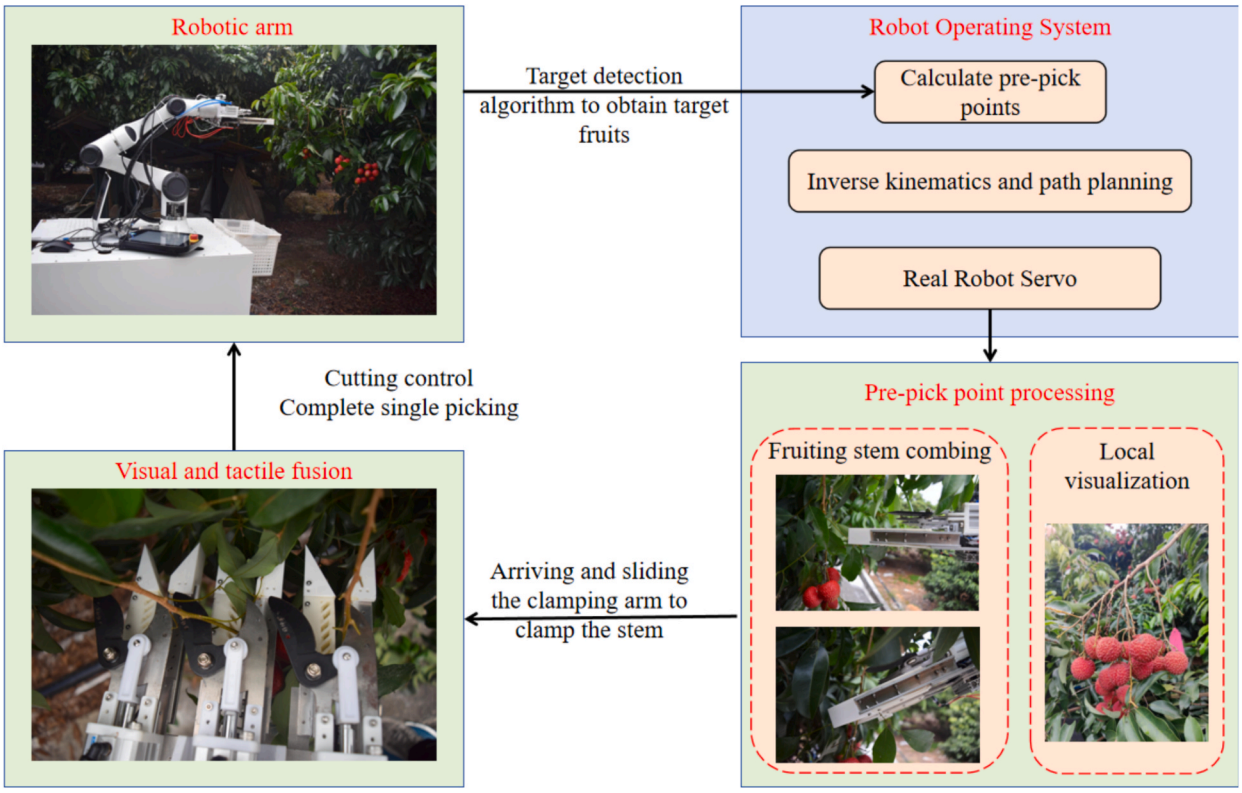


Fig. 5. Picking workflow diagram.

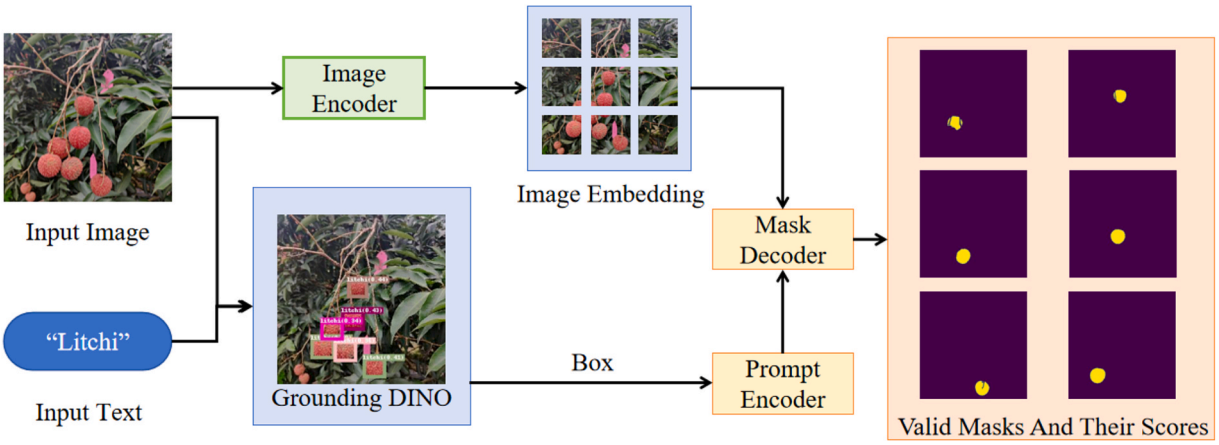


Fig. 6. SAM model structure and mask extraction of litchi images.

takes into account scalability and robust pre-training capabilities, especially for high resolution inputs. To accommodate real-time processing, each image is processed through the encoder only once, instead of repeating the process for each prompt.

Prompt Encoder: SAM considers two types of cues: sparse (e.g., points, box, text) and dense (masks). For sparse prompts, positional coding is combined with embeddings for specific prompt types. In this paper, we use inputs in the form of texts as prompts. However, during processing we introduced the Grounding DINO model. This results in the SAM's prompt encoder receiving what are actually box-based prompts. This approach takes into account both geometric information and textual descriptions, which improves the flexibility and accuracy of the model in handling various input prompts.

Mask Decoder: The Mask Decoder efficiently maps image embeddings, prompt embeddings, and output tokens to masks. A modified version of the Transformer (Vaswani, 2017) decoder is borrowed to incorporate a dynamic mask prediction mechanism, which updates embeddings through self-attention and cross-attention mechanisms, as well as accurately predicts the mask probability at each location through up-sampling and a multilayer perceptron (MLP). In addition, to deal with potential ambiguity of the input prompts, a multi-output strategy is employed to improve the model's adaptability and accuracy to complex scenes.

In this paper, the input image is a litchi image of a pre-picked point, and the input image is first scaled to 1024×1024 resolution by an image encoder and then converted to a refined 64×64 image embedding by the encoder. In addition, the number of channels is uniformly reduced to 256 by 1×1 and 3×3 convolution and layer normalization is applied to maintain data consistency and stability. At the same time, certain prompts are input, and the fused features are fed into a mask decoder,

which optimizes the output embedding by introducing positional coding and careful processing of the input prompts, as well as by transposing the convolutional layers, and through this process, the model is able to efficiently separate the litchi from the complex background.

2.2.2. Fruit circles detection

Based on the base mask extracted from SAM, a Hough transform (Ballard, 1981; Duda and Hart, 1972) based circle detection algorithm in OpenCV is used to determine whether litchi fruits are present in the image or not. The Hough transform is a mathematical parser that maps pixels in an image to a parameter space. For circle detection, the core of the method is to find the possible location and radius of a circle by accumulating the higher part of the circle's count in the parameter space. Since litchi fruits usually have an approximate circular outline, this property allows us to determine the presence of litchi fruits in an image by using the Hough transform.

The general equation of a circle is as follows:

$$(x-a)^2+(y-b)^2=R^2 \tag{2}$$

where (a,b) is the center of the circle and R is the radius. In the Cartesian coordinate system, the point (x,y) on the circle is transformed to a point in polar coordinates, and any of the edge points (x_0,y_0) in the parameter space is mapped to the parameter space with radius R_0. The transformation gives:

$$\begin{cases} a=x_0-R_0cos\theta \\ b=y_0-R_0sin\theta \end{cases} \tag{3}$$

where $\theta\in[0,2\pi]$. From Eq. (2), traversing θ, all the points (x_0,y_0) in the image space are mapped to the parameter space as a circle. From this we can introduce that any edge point in the parameter space corresponding to the parameter space is connected to generate a circle. In the parameter space $\rho-\theta$, we build a cumulative value matrix with the initial value of 0. We traverse all the pixel points to obtain the Hough cumulative matrix. According to the results of the hough accumulation matrix, the location and radius of the specific circle are detected, and the specific circle detection principle is shown in Fig. 7.

2.3. Tactile information processing

2.3.1. Tactile sensor principle analysis

Piezoresistive tactile sensing is one of the most commonly used principles in tactile sensing. Piezoresistive sensors are characterized by high sensitivity and high resolution, and are able to maintain good performance despite the vibration of the machine and bad weather conditions, thus providing robust sensing for picking robots in the field environment.

For the end-effector, a long RP-L thin-film sensor is selected as the source of tactile information, with a length of 10 cm, a range of 20 g ~ 10 kg, and a resistance accuracy controlled within plus or minus three percent. RP-L resistive pressure-sensitive sensors are flexible thin-film sensors whose resistance decreases as the pressure acting on the sensing area increases. As shown in Fig. 8(a), the sensing region of this thin-film sensor consists of two layers: the bottom layer is a highly conductive material and the top layer is a nanoscale pressure-sensitive material. These two layers are isolated from each other by a precise lamination process, but are tightly coupled together to jointly realize pressure sensing and signal transmission. When the sensing area is pressurized, the lines separated from each other in the bottom layer are conducted, and the output resistance of the metal ports changes accordingly with different pressures. In order to amplify the pressure, we add a foam layer under the bottom layer to increase the degree of its deformation.

The output voltage formula is:

$$U_o=(1+R_{AO\cdot RES}\times\frac{1}{R_x})\times 0.1 \tag{4}$$

where U_o is the output voltage value of the sensor module, $R_{AO\bullet RES}$ is the size of the feedback resistance, R_x is the output resistance of the thin-film pressure sensor. Sensor pressure resistance relationship according to the actual calibration is:

$$\frac{1}{R_x}=0.0004\cdot F+0.3749 \tag{5}$$

where F is the magnitude of the force acting on the induction region. According to the product properties, this paper takes the feedback resistance 22KΩ, $R_{AO\bullet RES}$ and $\frac{1}{R_x}$ into the curve calculation and obtains:

$$U_o=0.00088\bullet F+0.9228 \tag{6}$$

2.3.2. Tactile time sequence updating mechanism

In actual picking, when the clamping arm effectively clamps the fruit stems, the microcontroller can successfully capture the change of tactile time sequence from the tactile sensors, which is regarded as a trigger event. In this paper, dynamic pressure sensing is used for collision detection. The response time of the selected tactile sensors is in the range of $<$ 10 ms, and the appropriate frequency is selected to effectively sample each sensor. As shown in Fig. 8(b), the acquired data are averaged as a unit group by taking the average value of every five tactile data to reduce the effect of random errors.

Three sensors are deployed on the end-effector and signal processing is performed through a single microcontroller. Despite the fact that the three sensors are of the same model and mounted in the same way, their sensitivity varies due to errors in the manufacturing and assembly process. In a multi-tactile sensor system, when interacting with the environment, it becomes very difficult to realize the condition of repeated triggering events due to the difficulty of maintaining a consistent distribution of force each time. By implementing the sliding window approach, the tactile time sequence data collected by multiple sensors can be processed efficiently and the independent perception and output of tactile trigger events in each clamping arm can be realized, so as to

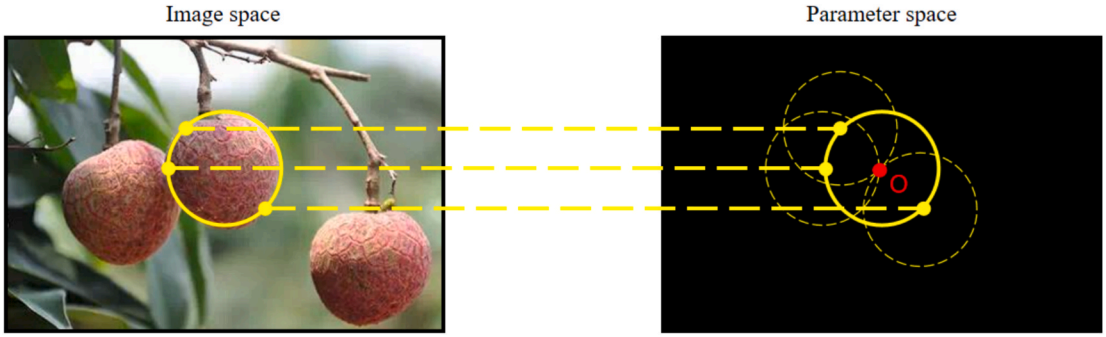


Fig. 7. Hough circle transformation of litchi border.

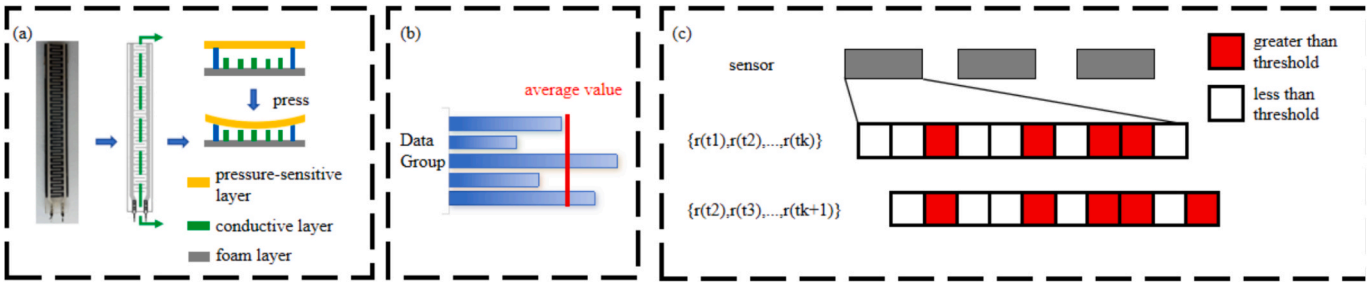


Fig. 8. Tactile information processing.

establish an accurate tactile detection model for fruit stems. As shown in Fig. 8(c), the node only saves the latest k data. The data in the sliding window follows the first-in-first-out rule, assuming that the current window dataset is $\{r(t_1), r(t_2), \cdots, r(t_k)\}$, where $r(t_1), r(t_2), \cdots, r(t_k)$ are the $t_1, t_2, \cdots, t_k$ data collected at the sampling time. When new data $r(t_{k+1})$ arrives, the window slides forward and the dataset in the window changes to $\{r(t_2), r(t_3), \cdots, r(t_{k+1})\}$, and so on. By means of a sliding window, the latest data is recorded, with 10 groups of data, and the number of groups exceeding the set threshold is counted as the basis for clamping judgment, which in turn triggers a change in the state node. The method amplifies the change of actions and is not only limited to the state the object is in.

2.4. Application of visual-tactile fusion in comb and cut end-effector

As shown in Fig. 9, for a certain finger-clamping channel $n(n \in \{1,2,3\})$, based on the processing of visual information and tactile information proposed in Sections 2.2 and 2.3, respectively, the decision-making of multiple scissor end-effector's corresponding channels is realized by aligning the information of the same channel. Specifically, after the endoscope acquires an image at the pre-picking point, it obtains the visual result S_v through visual processing, where S_v represents whether there is fruit on the stem that is ready to be picked under the corresponding channel; the robotic arm executes a clamping operation after arriving at the picking point, and the tactile sensor collects the tactile time sequence and obtains the tactile result S_t through tactile processing, where S_t represents whether the fruit stems are effectively clamped by the end-effector finger-clamping after this end-effector passes through the combing behavior. Under ideal conditions, for the case where both visual and tactile sensations are reliable, the classification cases are subdivided as shown in Table 1.

This decision logic can be clearly demonstrated by the above table. In the decision-making process, we combine the inputs of the visual result S_v and the tactile result S_t. When both the visual result S_v and the tactile result S_t give a 'Yes' result, we make a 'Yes' decision, i.e., the scissors of the corresponding channel perform the cutting action; in any other case, we make a 'No ' decision.

In the actual picking process, due to the machine's weak anti-interference ability, and since this paper uses a large model, the visual result S_v is relatively unreliable. In order to improve the fault-tolerance of this end-effector, the visual-tactile fusion means is used by the decision-making layer.

Table 1
Decision table for visual and tactile outcomes.

Visual results S_v with or without fruits	Tactile results S_t with or without stems	Whether the scissors cut
Yes	Yes	Yes
Yes	No	No
No	Yes	No
No	No	No

By combining different information sources, the decision P_n corresponding to the finger-clamping channel cutting mechanism is obtained by taking the union set processing using the results of the visual-tactile categorization. The formula is as follows:

$$P_n = S_v \cup S_t \tag{7}$$

3. Results and discussion

3.1. Trial platforms and scenarios

The test platform in this paper consists of the Elfin5 robotic arm, computer platform, end-effector, RealSense D435i depth camera, mobile platform, power control sources and fruit collection basket. The end-effector is equipped with an endoscope and RP-L tactile sensors, and is mounted on the Elfin5 robotic arm; the power control source is composed of an STM32 microcontroller together with an air pump and other components to provide motion control for the end-effector; the computer platform adopts an Intel Core i5-8300H CPU and an NVIDIA 1050Ti GPU, equipped with 8 GB of RAM; the operating system selects Ubuntu 18.04 to support ROS to control the robotic arm; the target detection algorithm adopts YOLO v5 model with PyTorch as the framework; the RealSense D435i depth camera is mounted on the end-effector in the way of eye-in-hand, which is in conjunction with the

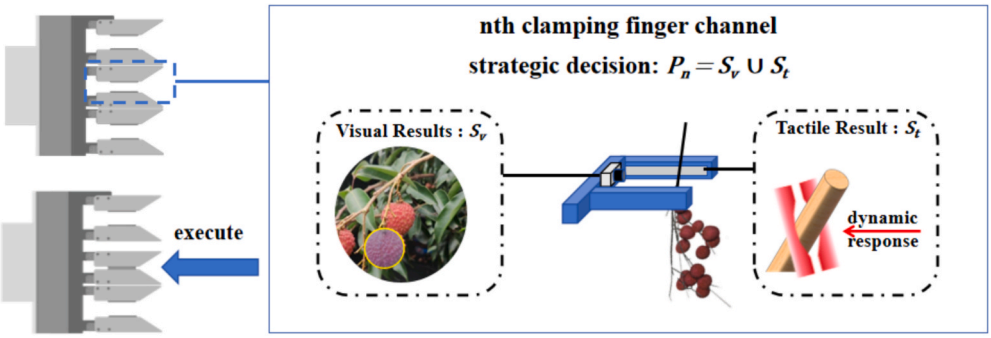


Fig. 9. Schematic diagram of decision making of the finger clamping channel n.

target detection algorithm to complete the localization of the picking point by depth alignment and coordinate conversion. The constructed picking robotic arm platform is shown in Fig. 10.

3.2. Trial platforms and scenarios

In this paper, two trials were designed to evaluate the performance of the proposed method. The first set of trials focused on evaluating the ability of the SAM model to segment litchi in an orchard environment. The second set of trials was conducted in an ideal indoor environment using different perceptual methods to initially assess the effectiveness of each method under ideal conditions. Then, actual picking was carried out in the field to evaluate the efficiency of the whole set of litchi picking equipment. Through these trials, the effectiveness of the proposed methods in practical applications was verified.

Precision (P) and Recall (R) are selected as the evaluation metrics for the picking scene segmentation trial in this paper. Precision rate reflects the ability to correctly predict the precision of positive samples, and recall rate reflects the ability to correctly predict the degree of checking completeness of positive samples, and these metrics pay more attention to the performance of the model in each step, and the requirements of both are slightly different for different steps. The calculation formula is shown below:

$$P = \frac{T_P}{T_P + F_P} \tag{8}$$

$$R = \frac{T_P}{T_P + F_N} \tag{9}$$

where T_P is the number of true positive samples, F_P is the number of false positive samples and F_N is the number of false negative samples. In evaluating the picking trials, success rate and perception rate were used as evaluation metrics. The success rate reflects the effectiveness of the whole picking robot. The perception rate as an extension of the precision rate is calculated in the same formula and refers to the precision of recognizing fruit stems under specific conditions. It measures the consistency between the number of fruit stems recognized by the perception technology and the number of fruit stems actually present. These metrics are more concerned with the ability to validate the significance of the work in this paper on the overall trial.

3.2.1. Picking scene segmentation trial

Different sizes of VIT model weights were tested in order to perform local model inference prediction. VIT_B model performs well in the same image test due to its small size and fast inference speed. Therefore, VIT_B was chosen as the weight checkpoint for the SAM model to ensure efficient real-time modeling. The endoscope in the structure was utilized to collect the local visual map of the pre-picked point. The zero-sample prediction accuracy of the overall sample for picking scenarios with open vocabulary was calculated based on a comparison of the original graphs collected with the resultant graphs derived from the SAM model. 150 images were collected during the picking process for the trial, and the results of each processing step were organized as shown in Table 2.

As can be seen from the data in Table 2, for the open vocabulary, the model recall is 100 %, and the model will always give a corresponding prompt, but the precision of this prompt is only 57.45 %, which means that although the model can perceive the agricultural scene of picking, it is weak in recognizing the litchi category. In particular, in some of the images, the general vision model misjudged and incorrectly labeled litchi on the clamping arm, as shown in Fig. 11(d).

In terms of mask extraction, the precision rate of mask extraction is 100 % and the recall rate is at 80.25 %, which means that the model is able to segment the corresponding mask well if the given prompts are accurate, but if the prompts are inaccurate, there will be incorrectly segmented masks as shown in Fig. 11(b). For the correctly classified litchi masks, the parameters of the Hough transform process are relatively stringent due to the preprocessing step, so the fruit judgment accuracy reaches 100 %, but the recall rate is only 81.5 %, which suggests that as long as the masks are extracted accurately, the fruit judgment step can still be carried out in most cases, despite the fact that there are a few cases of misjudgment due to occlusion of the leaves. The overall accuracy of visual processing was 41.3 %, this value is the ratio obtained by adding the 9 samples correctly categorized in the segmentation of vocabulary phase to the 53 samples correctly categorized in the final detection phase and dividing by the 150 total number of samples. This indicates that the generalized visual model has the potential to be

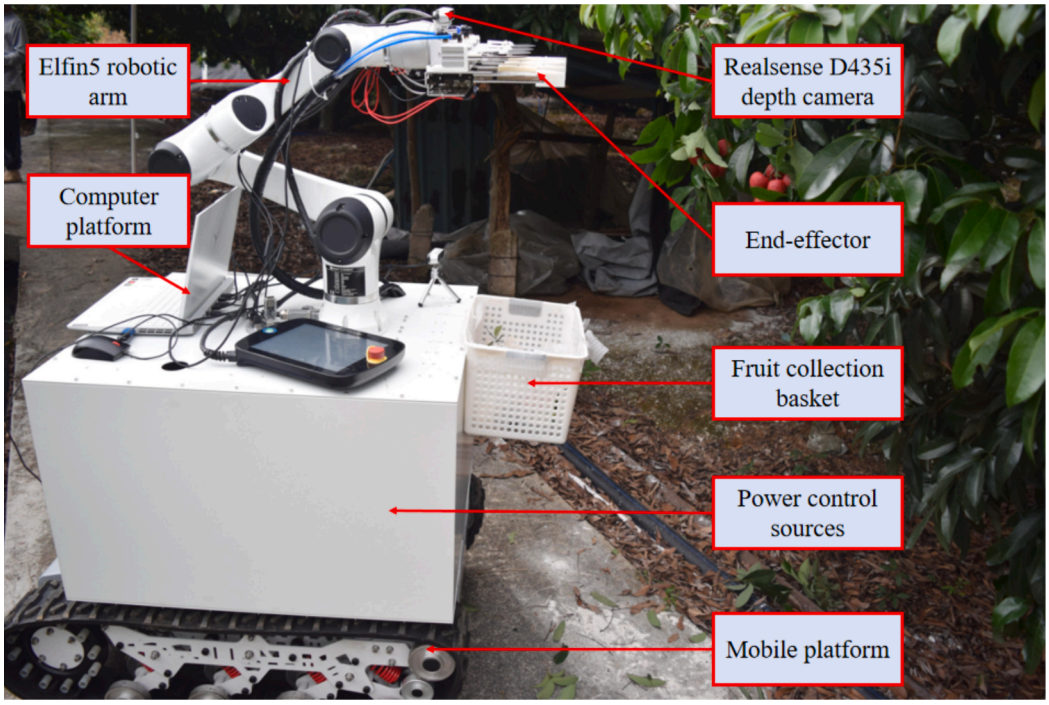


Fig. 10. Field trial robotic picking platforms.

Table 2
Results of open vocabulary segmentation experiment in picking scenarios.

Processing steps	Processing samples	True positive sample	False positive sample	False negative sample	Precision /%	Recall /%
Segmentation of vocabulary	150	81	60	0	57.45	100
Mask extraction	141	65	0	16	100	80.25
Fruit circles detection	65	53	0	12	100	81.54

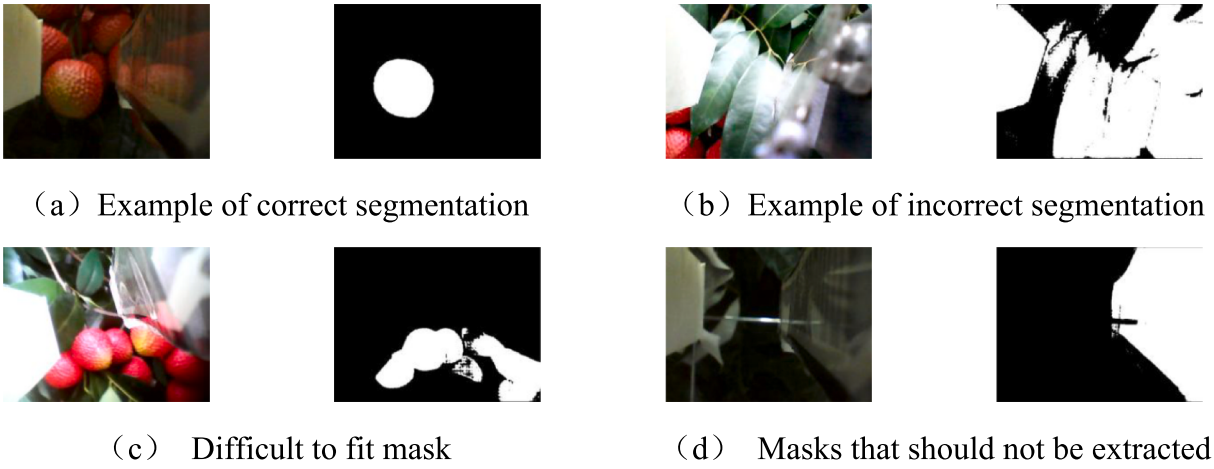


Fig. 11. Four different segmentation results.

applied in the field of agricultural automation without sample learning, but the accuracy is low when relying on only a single modality for judgment.

3.2.2. *Picking trials*

In order to verify the effectiveness of this paper's visual and tactile information processing and its fusion application, this paper conducts laboratory trials in the School of Mathematical and Information Technology, South China Agricultural University. Meanwhile the field harvesting trials were conducted at Litchi Expo Park, South China Agricultural University. A diagram of the trial scenario is shown in Fig. 12.

According to the target detection algorithm to recognize and locate the position of litchi bunches, the robotic arm reaches the pre-picking point, collects the visual information, and then reaches the picking point to carry out the acquisition of tactile information as well as the cutting process. According to the above test steps, 80 groups of tests were conducted and the results were obtained as shown in Table 3.

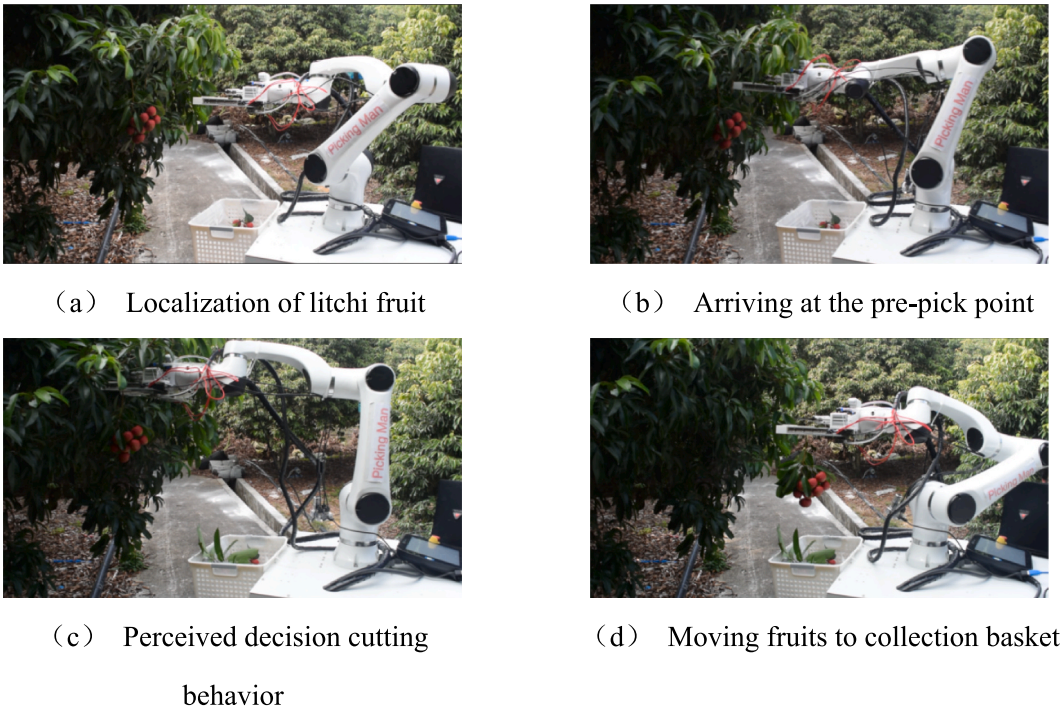


Fig. 12. Robotic arm orchard picking trial.

Table 3
Statistics of picking trial results.

Scene description	Number of trials	Number of true values of fruit stems	Number of fruit stems sensed	Number of successful harvests	Success rate /%	Perception rate /%
Indoor static tactile trials	10	18	14	6	60	77.8
Field static tactile trials	20	36	6	2	10	16.7
Indoor visual-tactile cooperation trials	15	27	23	12	80	85.2
Field vision-tactile cooperation trials	35	60	52	22	62.86	86.7

Comparing the pure tactile sensing method using a fixed threshold with the method in this paper, it can be seen that in the laboratory scenario, there is little change in the thickness of the fruit stems and little environmental interference, and the perception rate is at 77.8 %. However, in a complex field environment, due to structural and other problems, the tactile changes caused by the picking process are extremely small, which makes it difficult for the fixed-threshold method to capture. In addition, the triggered sensing area is relatively random, and the perception rate decreases to 16.7 %, indicating that the fixed threshold setting method is not applicable to field picking. Because this paper adopts visual-tactile sensing fusion in the logic of the union processing, the dynamic tactile modality can be analyzed separately. In the laboratory scenario, the perception rate in the static response of the pure tactile method is comparable. In the field environment, the effective degree of the reduction is much smaller than the static response of the tactile method, the reason being that the tactile sensor's flexible layer in many trials has a small change, and it takes a certain amount of time for recovery. The decision-making mechanism using visual-tactile synergy achieves a perception rate of 86.7 % in the field environment, which indicates that having multimodal information perception will enable the robotic arm to better understand the surrounding environment to help complete the picking in a better way. The field picking trials take Guiwei litchi as an example, with a success rate of single picking at 62.86 %, and due to the parallel picking of multiple bunches, the average number of bunches picked is 1.08 bunches per picking.

The main reason for picking failure in the trial is that the end-effector fruit bunches did not enter the clamping arm. The end-effector touches the leaves causing the fruit stems to move backward, and the positioning generates a deviation, resulting in clamping without cut or bunches ripped off, so it can only stop running. In practice, we have recorded the data thoroughly in the visual-tactile fusion trials conducted in the field. Out of the 35 sets of trials, 18 sets of trials were single fruiting stem only cases, accounting for 51.4 % of the total, and 16 fruiting stems were sensed, with a perception rate of 88.9 %, whereas 9 sets of trials were double fruiting stem cases, with 15 fruiting stems being sensed, and 8 sets of trials were triple fruiting stem cases, with 21 fruiting stems being sensed. From the above, it can be seen that single shearing was more common and had the highest perception rate. Although the efficiency of multi-cut picking is high, there is still a need to optimize the structure to improve the overall success rate of picking. In this paper, we tested the capability of open vocabulary segmentation for agricultural picking, a technology that is still to be developed, and for picking robots that need to be productized, it is recommended to use specially trained visual models. In addition, only the decision layer's visual-tactile sensing fusion was implemented; if the visual model is trained with a dataset, the feature layer's sensing fusion can be done using neural networks.

4. Conclusions

This paper analyzes the key steps of litchi picking process and designs a litchi comb-cutting end-effector based on visual-tactile sensing fusion. By analyzing the litchi fruit tree environment and structured agronomic characteristics, it adopts a comb-cut design with three independently controlled cutting units, which realizes the synchronous picking of multiple bunches of litchi. The visual macromodel is used as a visual pre-processing means, combined with the edge detection algorithm to judge the existence of the picked litchi fruits. Exploiting the application of the visual macromodel in the picking scenario, the trial shows that the model has an accuracy of 41.3 % under zero-sample prediction. The tactile time sequence updating mechanism is used to process multi-tactile sensor information and monitor the fruit stem clamping status of each finger-clamping channel, which reduces the mechanical damage during picking. Indoor simulated picking and orchard field picking tests were designed, and with the application of visual-tactile sensing fusion, the picking success rate in the field test was 62.86 %, and the fruit stem sensing rate was 86.7 %. The average single picking of litchi fruit bunches was 1.08 bunches, realizing the improvement in picking efficiency.

CRediT authorship contribution statement

Zhaoshen Yao: Writing – original draft, Methodology, Formal analysis, Data curation. **Juntao Xiong:** Writing – review & editing, Supervision, Project administration, Investigation, Conceptualization. **Jiayuan Yang:** Writing – original draft, Validation, Software, Data curation. **Xiao Wang:** Methodology, Data curation. **Zexing Li:** Methodology, Conceptualization. **Yuhua Huang:** Writing – review & editing, Visualization, Investigation. **Yanan Li:** Writing – review & editing, Supervision.

Declaration of competing interest

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

Acknowledgements

This work is funded by National Natural Science Foundation of China (Project No. 32071912), The open competition program of top ten critical priorities of Agricultural Science and Technology Innovation for the 14th Five-Year Plan of Guangdong Province (2022SDZG03), Natural Science Foundation of Guangdong Province (2024A1515010226), Guangdong Provincial Department of Science and Technology (2023A0505050130), Key Projects of Guangzhou Science and Technology Program (2024B03J1357), Basic and Applied Basic Research Foundation of Guangdong Province (2022A1515140013). The authors wish to thank the useful comments of the anonymous reviewers to this paper.

Data availability

The data that has been used is confidential.

References

Bac, C.W., Van Henten, E.J., Hemming, J., Edan, Y., 2014. Harvesting robots for high-value crops: state-of-the-art review and challenges ahead. J. Field Rob. 31 (6), 888–911.

Ballard, D.H., 1981. Generalizing the Hough transform to detect arbitrary shapes. Pattern Recogn. 13 (2), 111–122.

Bosilj, P., Duckett, T., Cielniak, G., 2018. Connected attribute morphology for unified vegetation segmentation and classification in precision agriculture. Comput. Ind. 98, 226–240.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Duda, R.O., Hart, P.E., 1972. Use of the Hough transformation to detect lines and curves in pictures. Commun. ACM 15 (1), 11–15.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16000–16009.

Hu, Q., Dian, Y., Gong, Z., Zhang, J., Hu, C., Liu, Y., Zhou, J., 2022. Analyzing fruit quality of Newhall navel oranges with different cultivation patterns. J. Huazhong Agric. Univ. 41 (5), 108–115.

Huang, R., Zheng, W., Zhang, B., Zhou, J., Cui, Z., Zhang, Z., 2023. Deep learning with tactile sequences enables fruit recognition and force prediction for damage-free grasping. Comput. Electron. Agric. 211, 107985.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., 2023. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026.

Li, M., Liu, P., 2023. A bionic adaptive end-effector with rope-driven fingers for pear fruit harvesting. Comput. Electron. Agric. 211, 107952.

Li, T., Xie, F., Zhao, Z., Zhao, H., Guo, X., Feng, Q., 2023. A multi-arm robot system for efficient apple harvesting: Perception, task plan and control. Comput. Electron. Agric. 211, 107979.

Liang, C., Xiong, J., Zheng, Z., Zhong, Z., Li, Z., Chen, S., Yang, Z., 2020. A visual detection method for nighttime litchi fruits and fruiting stems. Comput. Electron. Agric. 169, 105192.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., ... & Zhang, L. (2023). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arxiv preprint arxiv:2303.05499.

Liu, Y., Zhang, J., Lou, Y., Zhang, B., Zhou, J., Chen, J., 2024. Soft bionic gripper with tactile sensing and slip detection for damage-free grasping of fragile fruits and vegetables. Comput. Electron. Agric. 220, 108904.

Mu, L., Cui, G., Liu, Y., Cui, Y., Fu, L., Gejima, Y., 2020. Design and simulation of an integrated end-effector for picking kiwifruit by robot. Inform. Process. Agric. 7 (1), 58–71.

Pu, Y., Wang, S., Yang, F., Ehsani, R., Zhao, L., Li, C., Yang, M., 2023. Recent progress and future prospects for mechanized harvesting of fruit crops with shaking systems. Int. J. Agric. Biol. Eng. 16 (1), 1–13.

Qi, W., Chen, H., Luo, T., Song, F., 2019. Development status, trend and suggestion of litchi industry in mainland China. Guangdong Agric. Sci. 46, 132–139.

Qi, X., Dong, J., Lan, Y., Zhu, H., 2022. Method for identifying litchi picking position based on YOLOv5 and PSPNet. Remote Sens. (Basel) 14 (9), 2004.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., 2021. Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR, pp. 8748–8763.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., ... & Zhang, L. (2024). Grounded sam: Assembling open-world models for diverse visual tasks. arxiv preprint arxiv: 2401.14159.

Song, Z., Zhou, Z., Wang, W., Gao, F., Fu, L., Li, R., Cui, Y., 2021. Canopy segmentation and wire reconstruction for kiwifruit robotic harvesting. Comput. Electron. Agric. 181, 105933.

Suo, R., Gao, F., Zhou, Z., Fu, L., Song, Z., Dhupia, J., Cui, Y., 2021. Improved multi-classes kiwifruit detection in orchard to avoid collisions during robotic picking. Comput. Electron. Agric. 182, 106052.

Tang, Y., Chen, M., Wang, C., Luo, L., Li, J., Lian, G., Zou, X., 2020. Recognition and localization methods for vision-based fruit picking robots: a review. Front. Plant Sci. 11, 510.

Tsai, R.Y., Lenz, R.K., 1989. A new technique for fully autonomous and efficient 3 d robotics hand/eye calibration. IEEE Trans Rob Autom 5 (3), 345–358.

Vaswani, A., 2017. Attention is all you need. Adv. Neural Inf. Proces. Syst.

Xiong, C., Xiong, J., Yang, Z., Hu, W., 2023. Path planning method for citrus picking manipulator based on deep reinforcement learning. J. South China Agric. Univ. 44 (3), 473–483.

Xiulan, B.A.O., Zhitao, M.A., Xiaojie, M.A., Yishu, L.I., Mengtao, R.E.N., Shanju, L.I., 2024. Design and experiment of citrus picking robot in hilly orchard natural environment. Nongye Jixie Xuebao/Trans. Chinese Soc. Agric. Mach. 55 (4).

Ye, L., Duan, J., Yang, Z., Zou, X., Chen, M., Zhang, S., 2021. Collision-free motion planning for the litchi-picking robot. Comput. Electron. Agric. 185, 106151.

Zhang, J., Kang, N., Qu, Q., Zhou, L., Zhang, H., 2024. Automatic fruit picking technology: a comprehensive review of research advances. Artif. Intell. Rev. 57 (3), 1–39.