2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Enhancement on Target-Gripper Alignment: A Tomato Harvesting Robot with Dual-Camera Image-Based Visual Servoing

Lu-Ching Wang, Yen-Cheng Chu, Yennun Huang[1], and Feng-Li Lian[2]

Abstract— **Automation application in crop harvesting has increased in the past decades. Various types of harvesting robots are emerging in both commercial and research areas. One of the main challenges is the precision alignment of the gripper and the target crop. An undesired dislocation can harm both the gripper and the crop, which is mainly caused by uncertainties from the sensors and the manipulator. To solve the problem, the dual-camera setup is designed and implemented on a self-built robot. The perception of the tomato is done by a fixed depth camera and a camera without depth on the gripper. The proposed dual-camera image-based visual servoing (IBVS) controller is designed to deal with the image feedback from both cameras and the proof of asymptotically convergence is provided. Furthermore, the cumulative error compensation reduces the time for the harvesting process. The experiments were conducted in the greenhouse and tested under various conditions. The time cost is formulated as a function and the success picking rate of tomatoes is 68.4%.**

I. Introduction

Automatic harvesting has received more attention in the last few decades. The performance improvement and decreasing price on computational hardware units enable real-time crop perception with high recognition rates. Many companies [1] have invested in the fruit automation market with their harvesting robots. Other kinds of automation harvesting such as sweet pepper [2], apple [3], and blackberry [4] have also been deeply investigated all around the world.

Tomato poses a unique situation compared to other crops. It is one of the major economic crops worldwide [5]. In Taiwan, the annual output value of tomatoes reached over $133 million in 2022 [6], ranking among the top 15% of all crops. In addition, tomatoes pose a special challenge for automatic harvesting because of their delicate physical attributes and growth traits. A tomato plant ripens its fruits from top to bottom. It is best to harvest the tomato with the pedicel for longer storage. However, the pedicel of the unripe tomato is fragile, as is the peduncle. Therefore, it is significant for the robot to aim at the ripe tomato precisely without causing harm to the others.

However, the dislocation problem is one of the main challenges for the robot during automation harvesting. The dislocation is defined as, when the robot finally starts to harvest, the pose of its gripper is actually unable to reach the target crops. There may be distance errors and angle errors. It is a common issue for agricultural robots since the environmental condition is often severe. The sunlight, the dirt, and the moisture may cause inappropriate measurements from the sensors and imprecise motion from the manipulator.

The solution of this paper is to use an image without depth information as the position feedback for the gripper. Images are preferred for their detailed information, and the ability to generalize to different types of fruit is a major advantage. We present a self-built SCARA-style 4-DOF (degree of freedom) autonomous mobile harvesting robot capable of performing tasks within a greenhouse. A fixed depth camera provides the initial position for the gripper, another camera mounted on the gripper then detects the dislocation between itself and the target crop. With the proposed IBVS controller, the error feedback leads the gripper to the desired position, enabling a precise alignment for the target and the gripper. Using Lyapunov stability analysis, the error is proven to approach zero asymptotically with the optimal gain. Moreover, using the error compensation, the time costs for each harvesting process will decrease.

The contribution of this paper is to enhance the target-gripper alignment by addressing the dislocation problem. The self-built robot presented demonstrates that the algorithm can resolve the dislocation problem even in a system with low hardware precision. In Section II, more details about the existing works are discussed. Section III provides an overview for the robot system, and Section IV presents the algorithm for such a system. Experiments are shown in Section V and conclusions are made in Section VI.

II. Related Works

There have been numerous studies on harvesting robots, which can essentially be classified into three categories: crops perception enhancement [7], [8], grippers design [9], [10] and the whole robot system integration [11], [12].

Image recognition technology has developed rapidly in the past decade. Models specify on tomatoes using CNN-based neural network are presented in [13] and [14], which put efforts on fast recognition of the ripeness. Recently, Yolo has gained more popularity due to its lightweight model and fast output performance described in [15] and [16].

The issue of dislocation affects not only tomatoes but also other crops and even leaves. An apple harvesting robot [17] proposed an integrated system with a high harvesting rate. However, the success picking rate was not good due to the dislocation problem, which was caused by uncertainties in the camera and the manipulator. A leaf retrieval robot[18] and a tomato harvesting robot [15] faced a common issue: their process was fast but had a low success rate due to dislocation problems.

[1]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan yennunhuang@citi.sinica.edu.tw
[2]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan fengli@ntu.edu.tw

The problem can be eliminated with further feedback while approaching the target crop. A tactile sensor mentioned in [19] is deployed for apple harvesting, while [20] uses an infrared sensor to determine if the strawberry is in position. The study in this paper adopts image feedback due to its ability to adapt to various crops and the well-established visual servoing technique [21], [22]. The concept from [23] is similar to our works since it enhances the performance by using both position and image features for visual servoing. However, we feed the image error back directly to the system, instead of coupling it with the depth information. Further details will be discussed in the later sections.

III. System Overview

The tomato harvesting robot is designed as an autonomous mobile robot. Fig. 1 shows the details of the robot and the reference point of each frame. The braces indicate the name of the frame. The red, green, and blue axes denote the x, y, and z-axis, respectively. The manipulator is designed to resemble the SCARA robot, which is composed of a 3-DOF robotic arm for revolution motion on the xy-plane and a lift for prismatic motion along the z-axis. The bottom vehicle has differential wheels, which enable the robot to move freely in the greenhouse. For the sensors, a depth camera is fixed on the top of the lift, which is an ETH (eye-to-hand) setup. Another camera without depth information is mounted on the top of the end-effector as an EIH (eye-in-hand) setup.

The bottom of the lift is defined as the base $\{B\}$ of the whole system. The transformation of the position of a detected tomato p from the ETH camera $\{F\}$ to the base frame $\{B\}$ has the following transformation relation:

$$\mathbf{m}^F = \mathbf{M}_F(z_p)\mathbf{p}^F, \tag{1}$$

$$\mathbf{p}^B = {}^B\mathbf{R}_F\mathbf{p}^F + {}^B\mathbf{t}_F, \tag{2}$$

where $\mathbf{m}^F = \begin{bmatrix} u_F & v_F & 1 \end{bmatrix}^\top \in \mathbb{R}^3$ is the homogeneous coordinates of the tomato position in the ETH camera image space. The depth camera intrinsic $\mathbf{M}_F(z_p^F) \in \mathbb{R}^{3\times3}$ projects coordinates from the 3D Euclidean space to the image space according to the pinhole model [23], where z_p^F is the depth (z-axis) of $\mathbf{p}^F$, and therefore, $\mathbf{M}_F(z_p^F)$ is invertible. For conciseness, we denote $\mathbf{M}_F := \mathbf{M}_F(z_p^F)$ and $\mathbf{M}_E := \mathbf{M}_E(z_p^E)$. However, since the EIH camera is a 2D camera, z_p^E is unavailable. $\mathbf{p}^B$ and $\mathbf{p}^F \in \mathbb{R}^3$ are the tomato positions in Euclidean space expressed in $\{B\}$ and $\{F\}$, respectively. The rotation matrix ${}^B\mathbf{R}_F \in \mathsf{SO}(3)$ and translation vector ${}^B\mathbf{t}_F \in \mathbb{R}^3$ transform the target from $\{F\}$ to $\{B\}$. The Denavit-Hartenberg graph is shown in Fig. 2.

Considering the real application and the system simplification, firstly, the y-axis of $\{X\}$ is always parallel with the x-axis of $\{B\}$, meaning that the gripper consistently faces the same direction. We abandon the rotational DOF of the gripper, and thus the inverse kinematics has a closed-form geometry solution. Secondly, the z-axes of ETH and EIH cameras are assumed to be aligned, and the only DOF is the rotation along their z-axis. The rotation from $\{E\}$ to $\{F\}$ is

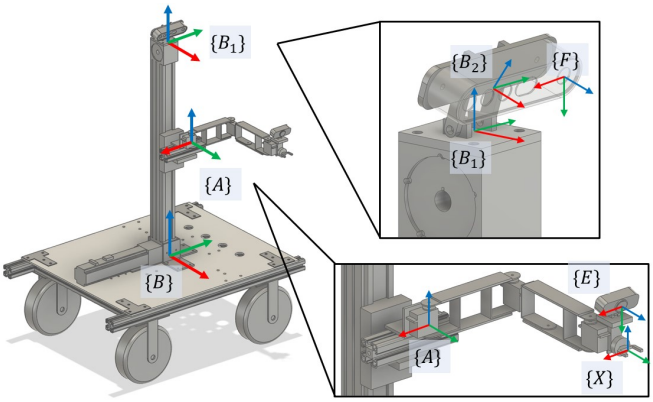


Fig. 1. Tomato harvesting robot with all reference frames.

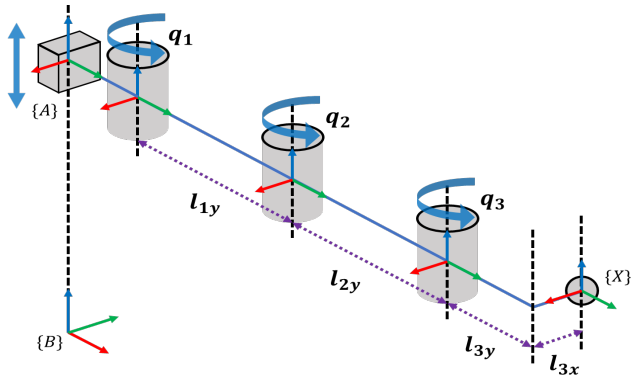


Fig. 2. Denavit-Hartenberg graph of the manipulator.

described as ${}^F\mathbf{R}_E \in \mathbf{R}_z \subset \mathsf{SO}(3)$, where $\mathbf{R}_z$ is the set of basic rotations along the z-axis.

To prevent encountering singularity during harvesting, the ETH camera image is divided into two non-overlapping areas S_i and $\bar{S}_i$. The area S_i is chosen such that if $\mathbf{m}^F \in S_i$, then $\mathbf{p}^B$ is at the reachable workspace. The robot will first detect tomatoes within S_i and select the tomato with the lowest v_F according to its ripeness properties. If no tomato is found in S_i, the robot finds tomatoes in $\bar{S}_i$ and (the vehicle) moves itself according to the position of the detected tomato. In this paper, the vehicle is considered stationary, and the detected tomato has the property that $\mathbf{m}^F \in S_i$, i.e., the tomato is always reachable.

The robot should move the gripper to the desired 3D position $\mathbf{x}_d^B \in \mathbb{R}^3$, such that the gripper can retrieve the target confidently. However, the uncertainties from the environment makes it to an undesired position $\mathbf{x}_u^B$. The dislocation is then represented as $\mathbf{x}_u^B - \mathbf{x}_d^B$. In the next section, the visual servoing controller is designed such that $\mathbf{x}^B$ approaches $\mathbf{x}_d^B$ asymptotically.

IV. IBVS using Dual-Camera

In order to eliminate the dislocation problem, the information from the ETH and the EIH camera is used. When a tomato is detected by the ETH camera, it triggers an event $\tau \in \{1, 2, ..., m\}$, where m is the total amount of tomatoes. The gripper moves closer to the tomato with the

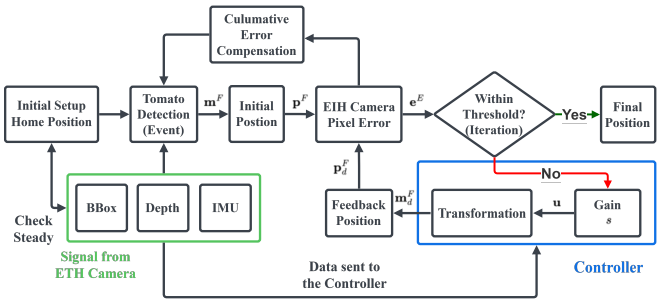


Fig. 3. The flowchart of the system.

3D position estimation from the ETH camera for an event. During each event, the iteration of IBVS control is defined as $k \in \{0, 1, ..., n\}$, where n is the total number of iterations. For iteration k, the dislocation of the robot is corrected by the information from the EIH camera. The IBVS iterations proceed until the image error satisfies a prescribed threshold. After the error converges, the gripper opens based on the size of the tomato, as being too wide or too narrow may result in hitting the stem or the target. Furthermore, all errors during an event are added up and provide compensation for the next event, called *Cumulative Error Compensation*. The flowchart of the system is depicted in Fig. 3.

A. Initial Position from the ETH Camera

Tomato perception, including an HSV filter and image segmentation for image prepossessing, is done by a pre-trained Yolo V5 model. The model classifies if a tomato is mature, and then the coordinate is verified. The output would be the center pixel of a mature tomato, with confidence higher than 80% and with the lowest v_F coordinate, which is selected as target p. The transformation of the tomato from the ETH camera image frame to the 3D space of $\{B\}$ is then derived by (1) and (2). Since collision may occur if the gripper reaches exactly the target position, a distance of 4 centimeters is set between them. Therefore, the gripper moves to the given position $\mathbf{x}^B$ accordingly:

$$\mathbf{x}^B = \mathbf{p}^B - \begin{bmatrix} 0 & 0.04 & 0 \end{bmatrix}^\top. \tag{3}$$

B. Error Feedback from the EIH Camera

After the gripper moves to $\mathbf{x}^B$, dislocations are expected to appear due to model uncertainties. However, it can not be unbounded. An assumption is needed to guarantee the later-mentioned controller design.

Assumption 1. When the process of (3) is finished, the target p must exist in the field of view from the EIH camera.

Assumption 1 indicates that although the dislocation happens, the EIH camera is still able to see the target, enabling the tomato perception model to detect the target. If the assumption holds, the homogeneous position of p observed by the EIH camera is $\mathbf{m}^E = \begin{bmatrix} u_E & v_E & 1 \end{bmatrix}^\top \in \mathbb{R}^3$, where (u_E, v_E) is the centroid of the target bounding box and $\mathbf{m}^E$ denotes the position of target p in the EIH camera image space $\{E\}$. The error is then described by the difference between $\mathbf{m}^E$ and a reference coordinate $\mathbf{m}_r^E$ in image space, which is written as

$$\mathbf{e}^E = \mathbf{m}^E - \mathbf{m}_r^E, \tag{4}$$

where $\mathbf{e}^E \in \mathbb{R}^3$ and $\mathbf{m}_r^E = \begin{bmatrix} 320 & 240 & 1 \end{bmatrix}^\top$, which is the image centroid of $\{E\}$. The pose of the EIH camera has been calibrated such that, if $\mathbf{x}^B \to \mathbf{x}_d^B$, then $\mathbf{m}^E \to \mathbf{m}_r^E$, and $\mathbf{e}^E \to 0$.

C. Dual-Camera IBVS Controller

To find the desired position $\mathbf{x}_d^B$, the desired image position $\mathbf{m}_d^F(k+1)$ will be designed first. The dual-camera visual servoing control in the discrete form is

$$\mathbf{m}_d^F(k+1) - \mathbf{m}_d^F(k) = \mathbf{u}, \tag{5}$$

$$\mathbf{u} = s\mathbf{J}\,{}^F\mathbf{R}_E\mathbf{e}^E(k), \quad \mathbf{J} = \begin{bmatrix} \mathbf{I}_{2\times 2} & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix} \tag{6}$$

where $\mathbf{u}$ is the control input. Since ${}^F\mathbf{R}_E$ is a pure rotation along the z-axis and the last element of $\mathbf{e}^E(k)$ must be zero, $\mathbf{u}$ is equivalent to a scaled error observed in image space of $\{E\}$ expressed in image space of $\{F\}$. The gain $s \in \mathbb{R}$ indicates how strongly the error affects the next position. The initial condition $\mathbf{m}_d^F(0)$ is set to the initial position $\mathbf{m}^F$. Unlike the conventional IBVS control law from [21] and [22], the errors from the EIH camera are fed back to the initial position of the ETH camera. If both sides of (5) are left multiplied by $\mathbf{M}_F^{-1}$, it would have a similar form with $\mathbf{v} = -\lambda\mathbf{L}^\dagger\mathbf{e}$, which is the common expression of IBVS control. In Section IV-E, the range and the numerical value of the gain will be derived based on Lyapunov analysis. The experiment in Section V-B also shows the results of different gains.

After $\mathbf{m}_d^F(k+1)$ is designed, $\mathbf{p}_d^F(k+1)$ can be derived by the inverse mapping of (1). The gripper position is then set to the position at the next iteration, which is

$$\mathbf{x}^F(k+1) = \mathbf{p}_d^F(k+1). \tag{7}$$

The 3D position of the gripper $\mathbf{x}^B$ is then calculated again with similar procedures as (2) and (3). The processes (4), (6), and (7) iterate until the error $\mathbf{e}^E$ converges within a threshold, i.e., $\mathbf{e}^E \in \mathcal{E}_T^E$ where

$$\mathcal{E}_T^E = \left\{ \begin{bmatrix} u_E & v_E & 0 \end{bmatrix}^\top \in \mathbb{R}^3 \,\middle|\, \begin{matrix} |u_E| \le 20, \\ |v_E| \le 20 \end{matrix} \right\}. \tag{8}$$

D. Cumulative Error Compensation

After an event ends, the dislocation can be formulated as the difference between $\mathbf{m}_d^E(n)$ and $\mathbf{m}_d^E(1)$, which is

$$\mathbf{d}(\tau) = \sum_{k=1}^{n} s\mathbf{J}\,{}^F\mathbf{R}_E\mathbf{e}^E(k), \tag{9}$$

where $\mathbf{d}(\tau) \in \mathbb{R}^2$ is the dislocation of position measurements. The dislocation $\mathbf{d}(\tau)$ would be added as a compensation of $\mathbf{m}^F$ for the next event. The goal is to reduce the total iteration n of a single event, thereby reducing the time cost. The dislocation would be verified in every event and added to the compensation. After the first event, i.e.,

$\tau > 1$, the detected tomato at the ETH camera image space is represented as

$$\hat{\mathbf{m}}^E(\tau+1) = \mathbf{m}^E(\tau+1) + \sum_{\tau=1}^{m} \mathbf{d}(\tau) \tag{10}$$

where $\hat{\mathbf{m}}^E$ indicates the cumulative measurement and the error compensation.

The time cost t within an event is then formulated as

$$t = t_0 + \sum_{k=1}^{n} t_k = t_0 + n\overline{T}_k, \tag{11}$$

where t_0 is the time needed for the zeroth iteration and t_k is the time needed for each iteration. The average time of an iteration $\overline{T}_k$ is used to assess the overall performance. It is more practical to evaluate the required iterations n rather than directly focusing on the time cost of an event. In section V-C, we will delve into the effects of cumulative error compensation.

E. Proof of Error Convergence

The Lyapunov function is established to prove the control law from (6) has the property of error convergence. Let the position of the EIH camera, i.e., the origin of $\{E\}$ depicted in Fig. 1, be named as x^E. The time derivative of (4) is

$$\dot{\mathbf{e}}^E = \dot{\mathbf{m}}^E. \tag{12}$$

Since there is no angular velocity for the gripper, $\dot{\mathbf{x}}^E$ can be expressed as

$$\dot{\mathbf{x}}^E = {}^F\mathbf{R}_E^\top \dot{\mathbf{x}}^F. \tag{13}$$

The discrete forms of (12) and (13) are

$$\Delta\mathbf{e}^E(k) = \Delta\mathbf{m}^E(k) = -\mathbf{M}_E \Delta\mathbf{x}^E(k), \tag{14}$$

$$\Delta\mathbf{x}^E(k) = {}^F\mathbf{R}_E^\top \Delta\mathbf{x}^F(k). \tag{15}$$

Because the EIH camera and gripper are mounted fixed, $\Delta\mathbf{x}^F(k)$ can be expressed as $\Delta\mathbf{p}_d^F(k)$ by (7). Also from the control law (6), $\Delta\mathbf{x}^F(k)$ is written as

$$\Delta\mathbf{x}^F(k) = \Delta\mathbf{p}_d^F(k) = \mathbf{M}_F^{-1}\mathbf{u}. \tag{16}$$

Substituting (6), (16) and (15) into (14),

$$\Delta\mathbf{e}^E(k) = -s\mathbf{M}_E\, {}^F\mathbf{R}_E^\top \mathbf{M}_F^{-1} \mathbf{J}\, {}^F\mathbf{R}_E \mathbf{e}^E(k).$$

Since $\mathbf{M}_F^{-1}\mathbf{J}$ is diagonal, and with the fact that $\mathbf{J}\mathbf{e}^E(k) = \mathbf{e}^E(k)$, it is commutative. Let $\mathbf{M} := \mathbf{M}_E\mathbf{M}_F^{-1}$, the error at $(k+1)$-th iteration is written as

$$\begin{aligned}\mathbf{e}^E(k+1) &= \mathbf{e}^E(k) + \Delta\mathbf{e}^E(k) \\ &= (\mathbf{I}_{3\times3} - s\mathbf{M})\mathbf{e}^E(k).\end{aligned} \tag{17}$$

In order to show the convergence of the error, the Lyapunov function candidate is selected as

$$V(k) = \|\mathbf{e}^E(k)\|^2 = (\mathbf{e}^E(k))^\top \mathbf{e}^E(k). \tag{18}$$

According to (17) and (18), and we let $\mathbf{e} := \mathbf{e}^E(k)$, the difference of the Lyapunov function at k-th and $(k+1)$-th iteration is

$$\begin{aligned}\Delta V &= \|\mathbf{e}^E(k+1)\|^2 - \|\mathbf{e}^E(k)\|^2 \\ &= \mathbf{e}^\top s(s\mathbf{M}^\top\mathbf{M} - \mathbf{M} - \mathbf{M}^\top)\mathbf{e}.\end{aligned} \tag{19}$$

Let $\mathbf{Q} := s(s\mathbf{M}^\top\mathbf{M} - \mathbf{M} - \mathbf{M}^\top)$, the error converges only when $\mathbf{Q}$ is negative definite. The property holds when the gain satisfies

$$0 < s < \frac{\max(\mathrm{eig}(\mathbf{M} + \mathbf{M}^\top))}{\max(\mathrm{eig}(\mathbf{M}^\top\mathbf{M}))}, \tag{20}$$

where $\mathrm{eig}(\cdot)$ indicates the eigenvalues of the given matrices. From (19) and (20), the optimal solution and the range of the gain for the system can be derived. However, $\mathbf{M}$ is time-varying during an event since $\mathbf{M}_E$ depends on z_p^E. To inspect the effect of z_p^E on $\mathbf{M}$, the eigenvalues of $\mathbf{Q}$ are examined by the lower and upper bound of z_p^E and z_p^F, which are $z_p^E \in [0.16, 0.18]$ and $z_p^F \in [0.56, 0.68]$ (in meters), respectively. Since the error $\mathbf{e} \in \mathbb{R}^3$ is expressed in the homogeneous form, only the eigenvalues from the upper-left 2×2 submatrix of $\mathbf{Q}$ are considered. Fig. 4 shows the two cases in which extreme values appear and the optimal gain lies within $[0.1835, 0.2507]$.

V. Experiments

The system is running under the architecture of ROS 1 with Intel i5-10300H CPU and NVDIA RTX 2060 GPU. For MCU, an Arduino Uno is implemented to control a step-motor for vertical prismatic motion and five servos for the plane revolution motion. The ETH camera Intel Realsense D455i has an IMU sensor to detect the pose of itself. The EIH camera is a webcam without depth information. Both cameras have the same resolution 640×480.

The greenhouse environment and the robot setup are shown in Fig. 5. The rack was secured to the concrete floor, with tomato plants tied to it. The robot is placed at the center of the road and facing the tomatoes.

A. Performance of the Optimal Gain

Based on the conclusion in Section IV-E, the gain is chosen as 0.2 after several experiments. A step-by-step process for the gripper reaching the target is shown in Table I. The process is classified into three stages, A) error convergence, B) gripping, and C) retrieving. The side view is an extra perspective, and it is not used by the controller. The yellow dots at the EIH camera row mark the center position of the

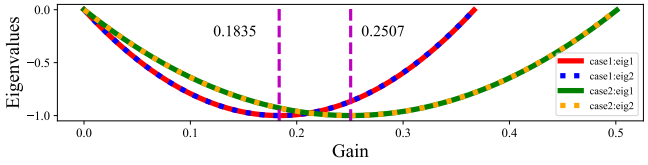


Fig. 4. The range and the optimal solution of the gain in different cases. Case 1, $z_p^F = 0.68$ and $z_p^E = 0.16$. Case 2, $z_p^F = 0.18$ and $z_p^E = 0.56$.

TABLE I

TOMATO HARVESTING PROCESS DURING AN EVENT

Stage	A				B	C
Iter. k	0	1	2	3	–	–
Side View						
EIH Camera						
x_x^B	-0.368	-0.144	-0.110	-0.102	-0.102	-0.368
x_y^B	0.141	0.269	0.292	0.297	0.347	0.141
x_z^B	0.360	0.505	0.551	0.561	0.561	0.561
e_u^E	–	107	26	16	–	–
e_v^E	–	-160	-36	-9	–	–

(a) (b)

Fig. 5. The experiment setup in the greenhouse: (a) the tomato harvesting robot inside the greenhouse; (b) the gripper is facing the tomato.

target. The rows x_x^B, x_y^B, and x_z^B are the 3D positions of the gripper expressed in $\{B\}$ in units of meter. The error (4) is represented in e_u^E and e_v^E.

At the 0-th iteration of stage A, the system is set to an initial state. After the ETH camera detects the target, the gripper moves to the position listed in iteration 1. Due to the dislocation problem, the gripper is still far from the desired position $\mathbf{x}_d^B$. Iteration 1 is where the feedback signal starts. The errors are detected ($\mathbf{e}^E = [82 \quad -182 \quad 0]^\top$) and fed back using (6). The process is repeated until the third iteration since the errors remain within the threshold specified in (8). Observing the EIH camera view, the yellow dot is nearly centered in the image, indicating the desired position $\mathbf{x}_d^B$ has reached. At Stage B, the gripper reaches the target. According to (3), the gripper is positioned 4 centimeters from the target. Here, 5 centimeters is added to ensure the gripper encompasses the entire target. At stage C, the target is retrieved and the manipulator moves back to the initial position. The height does not need to return to the initial position as it will simply be the starting position for another event at stage A.

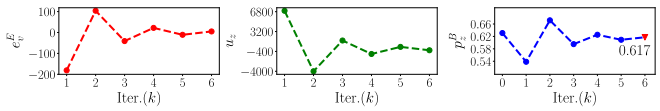
B. Performances of Different Gains

In Section IV-C, it is mentioned that the gain selection is based on the Lyapunov analysis. To verify the impact, the same experiment setup in Section V-A with different gains is tested and the results are shown in Fig. 6, including the error e_v^E, the input of the prismatic motion u_Z, and the gripper position p_z^B. For gain equals to 0.5, 0.2, and 0.1, the iteration counts to 6, 4, and 10, respectively. Also, the system performs underdamping, critically damping, and overdamping, respectively.

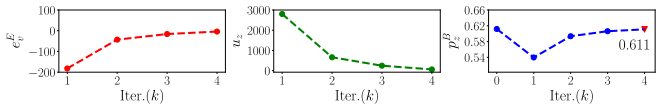
A high gain may lead to an overshoot, which can damage the tomato. As for a low gain, the input may be too small to drive the servos due to their insufficient resolution. Therefore, it is important to find the optimal gain setting that balances precision and stability in controlling the servos for the best performance of the harvesting robot.

C. Evaluation of the Cumulative Error Compensation

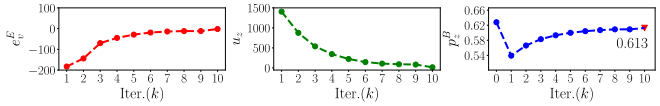
To verify the performance, the experiment in this section uses real tomatoes with artificial leaves. Fig. 7 shows the 3D positions of the gripper within seven continuous events. The events τ are labeled in black dashed lines, and the data between two events are the IBVS iterations k. The target remains almost stationary from the first to the fifth event. Except for the first event, a decrease of one or two iterations is observed in subsequent events due to the cumulative error compensation benefit. At the sixth event, the target shifts upward vertically by 6 centimeters, which can be observed from the x_z^B plot. Even if the position changes, only an additional iteration is required to complete the task. At the seventh event, the gripper reaches the desired position at the first iteration.



(a) Gain $s = 0.5$. Iteration counts to 6.



(b) Gain $s = 0.2$. Iteration counts to 4.



(c) Gain $s = 0.1$. Iteration counts to 10.

Fig. 6. Performances of different gains.

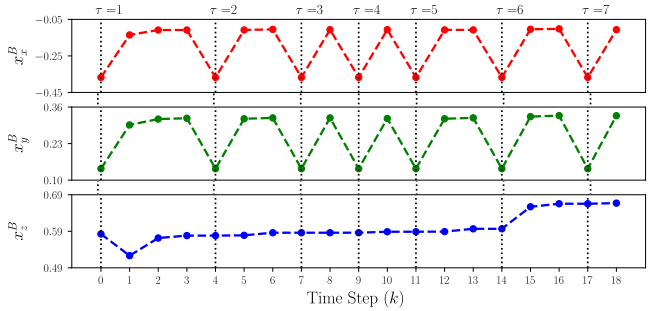
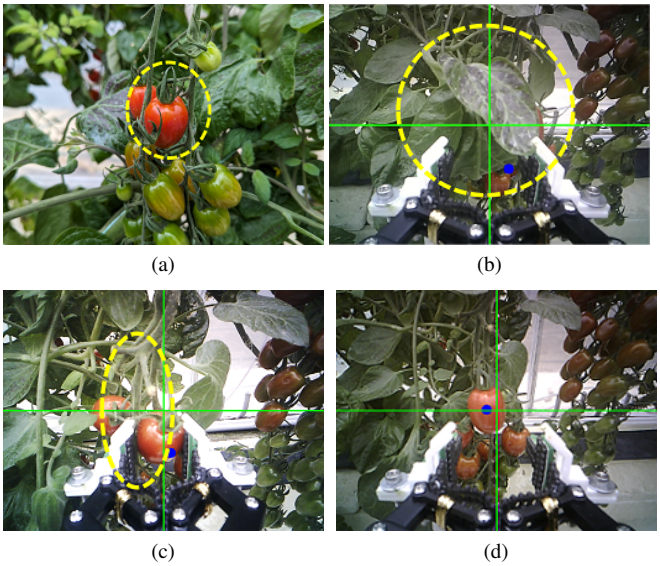


Fig. 7. The effect of the cumulative error compensation with seven events in sequence. Even if the position changes ($\tau = 6$), the cumulative error compensation reduces the iteration.

D. Performance Evaluation

From (11), the average value of t_0 and $\overline{T}_k$ are 6.26 and 4.82 seconds, respectively. The value of t_k depends highly on the stability of the tomato. If the tomato swings due to wind or gripper collisions, the controller must wait for stable detection from the EIH camera, resulting in longer times for an iteration. The average number of iterations of the first event is 3.10, and the average time cost for the initial event is 21.20 seconds. The result is not exceptional, however, subsequent events can be completed faster since cumulative error compensation reduces the number of iterations. Namely, the robot may initially be slow at picking the first tomato, but it will become faster in subsequent.

The success rate for picking tomatoes is 68.4%. The reasons for failure can be cataloged as four types: A) Pedicel Occlusion, B) Gripper's Motion, C) Gripping Foreign Objects, and D) Hardware. The situations of failure are depicted in Fig. 8. A) Pedicel Occlusion: It usually happens at the first iteration, i.e., the tomato in Fig. 8a in the yellow dashed circle observed by the EIH camera, which is occluded by the pedicel. The situation causes the Yolo model to fail to recognize the tomato confidently, resulting in altering output between the actual target and the tomatoes in the background. B) Gripper's Motion: The gripper collides with the leaves or stems while moving, causing the tomato to be obstructed.



(a) (b) (c) (d)

Fig. 8. The reasons of failures: (a) the tomato is occluded by the pedicel; (b) the leaf covers the tomato as the gripper collides with it; (c) the gripper holds the stem along with the tomato in stage B; (d) the error is too small and the servos lack sufficient resolution.

Fig. 8b illustrates an issue where a leaf obstructs the tomato due to the gripper's movement. This obstruction can lead to a failed harvesting process. Moreover, the collision could cause the tomato to sway, prolonging the time needed for the robot to locate its coordinates. C) Gripping Foreign Objects: When the gripper grasps the tomato, stems or leaves may be included. This occurs when the tomato is very close to the stem. The tomato may slip away or even be damaged during the retrieval stage. D) Hardware: Fig. 8d depicts the situation that the errors do not converge because the feedback signal to the servos is too small. The main cause is the low resolution of the servos, causing the gripper to remain in the same position until the input signals exceed a certain threshold.

VI. Conclusion

This paper presents a self-built tomato harvesting robot. The dual-camera IBVS algorithm has been proven and experimented to solve the dislocation problem. This results in a precise alignment of the target and the gripper. In the greenhouse experiment, the harvesting time is formulated, and it can be reduced by the cumulative error compensation. The success rate for picking tomatoes is 68.4% and the failure situations are discussed. Future work will focus on optimizing the robot's navigation capabilities in the greenhouse environment, as well as fine-tuning the harvesting mechanism to deal with more complex scenarios.

VII. Acknowledgments

This work is supported by Academia Sinica in "Design and Research of Smart Digital Farmer Platform" program (AS-TP-110-M07), and by National Science and Technology Council, Taiwan, under MOST 111-2221-E-002-191 and NSTC 112-2221-E-002-192.

REFERENCES

[1] *Introducing AI-equipped Tomato Harvesting Robots to Farms May Help to Create Jobs*, Panasonic Newsroom Global, May 23, 2018. [Online]. Available: https://news.panasonic.com/global/stories/814 (visited on 07/25/2023).

[2] C. Lehnert, A. English, C. McCool, A. W. Tow, and T. Perez, "Autonomous Sweet Pepper Harvesting for Protected Cropping Systems," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 872–879, Apr. 2017.

[3] A. Silwal, J. R. Davidson, M. Karkee, C. Mo, Q. Zhang, and K. Lewis, "Design, integration, and field evaluation of a robotic apple harvester," *J. Field Robot.*, vol. 34, no. 6, pp. 1140–1159, 2017.

[4] A. Qiu, C. Young, A. L. Gunderman, M. Azizkhani, Y. Chen, and A.-P. Hu, "Tendon-Driven Soft Robotic Gripper with Integrated Ripeness Sensing for Blackberry Harvesting," in *2023 IEEE Int. Conf. Robot. Autom. ICRA*, May 2023, pp. 11 831–11 837.

[5] C. W. Bac, E. J. van Henten, J. Hemming, and Y. Edan, "Harvesting Robots for High-value Crops: State-of-the-art Review and Challenges Ahead," *J. Field Robot.*, vol. 31, no. 6, pp. 888–911, 2014.

[6] A. a. F. A. Taiwan, *Production Value of Agricultural Products in Taiwan*, Agriculture and Food Agency, Council of Agriculture, Executive Yuan, R.O.C(TAIWAN), May 2022. [Online]. Available: https://www.afa.gov.tw/eng/index.php? (visited on 07/25/2023).

[7] A. Tafuro, A. Adewumi, S. Parsa, G. E. Amir, and B. Debnath, "Strawberry picking point localization ripeness and weight estimation," in *2022 Int. Conf. Robot. Autom. ICRA*, May 2022, pp. 2295–2302.

[8] I. Sa, C. Lehnert, A. English, *et al.*, "Peduncle Detection of Sweet Pepper for Autonomous Crop Harvesting—Combined Color and 3-D Information," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 765–772, Apr. 2017.

[9] A. L. Gunderman, J. A. Collins, A. L. Myers, R. T. Threlfall, and Y. Chen, "Tendon-Driven Soft Robotic Gripper for Blackberry Harvesting," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2652–2659, Apr. 2022.

[10] Y. Xiong, Y. Ge, and P. J. From, "Push and Drag: An Active Obstacle Separation Method for Fruit Harvesting Robots," in *2020 IEEE Int. Conf. Robot. Autom. ICRA*, May 2020, pp. 4957–4962.

[11] W. Lili, Z. Bo, F. Jinwei, *et al.*, "Development of a tomato harvesting robot used in greenhouse," *Int. J. Agric. Biol. Eng.*, vol. 10, no. 4, pp. 140–149, 4 Jul. 31, 2017.

[12] H. Yaguchi, K. Nagahama, T. Hasegawa, and M. Inaba, "Development of an autonomous tomato harvesting robot with rotational plucking gripper," in *2016 IEEERSJ Int. Conf. Intell. Robots Syst. IROS*, Oct. 2016, pp. 652–657.

[13] C. Hu, X. Liu, Z. Pan, and P. Li, "Automatic Detection of Single Ripe Tomato on Plant Combining Faster R-CNN and Intuitionistic Fuzzy Set," *IEEE Access*, vol. 7, pp. 154 683–154 696, 2019.

[14] L. Zhang, J. Jia, G. Gui, X. Hao, W. Gao, and M. Wang, "Deep Learning Based Improved Classification System for Designing Tomato Harvesting Robot," *IEEE Access*, vol. 6, pp. 67 940–67 950, 2018.

[15] J. Jun, J. Kim, J. Seol, J. Kim, and H. I. Son, "Towards an Efficient Tomato Harvesting Robot: 3D Perception, Manipulation, and End-Effector," *IEEE Access*, vol. 9, pp. 17 631–17 640, 2021.

[16] C. Song, K. Wang, C. Wang, *et al.*, "TDPPL-Net: A Lightweight Real-Time Tomato Detection and Picking Point Localization Model for Harvesting Robots," *IEEE Access*, vol. 11, pp. 37 650–37 664, 2023.

[17] K. Zhang, K. Lammers, P. Chu, N. Dickinson, Z. Li, and R. Lu, "Algorithm Design and Integration for a Robotic Apple Harvesting System," in *2022 IEEERSJ Int. Conf. Intell. Robots Syst. IROS*, Oct. 2022, pp. 9217–9224.

[18] M. Campbell, A. Dechemi, and K. Karydis, "An Integrated Actuation-Perception Framework for Robotic Leaf Retrieval: Detection, Localization, and Cutting," in *2022 IEEERSJ Int. Conf. Intell. Robots Syst. IROS*, Oct. 2022, pp. 9210–9216.

[19] L. M. Dischinger, M. Cravetz, J. Dawes, *et al.*, "Towards Intelligent Fruit Picking with In-hand Sensing," in *2021 IEEERSJ Int. Conf. Intell. Robots Syst. IROS*, Prague, Czech Republic: IEEE, Sep. 27, 2021, pp. 3285–3291.

[20] Y. Xiong, P. J. From, and V. Isler, "Design and Evaluation of a Novel Cable-Driven Gripper with Perception Capabilities for Strawberry Picking Robots," in *2018 IEEE Int. Conf. Robot. Autom. ICRA*, May 2018, pp. 7384–7391.

[21] S. Hutchinson, G. Hager, and P. Corke, "A tutorial on visual servo control," *IEEE Trans. Robot. Autom.*, vol. 12, no. 5, pp. 651–670, Oct. 1996.

[22] F. Chaumette and S. Hutchinson, "Visual servo control. I. Basic approaches," *IEEE Robot. Automat. Mag.*, vol. 13, no. 4, pp. 82–90, Dec. 2006.

[23] E. Malis, F. Chaumette, and S. Boudet, "2 1/2 D visual servoing," *IEEE Trans. Robot. Autom.*, vol. 15, no. 2, pp. 238–250, Apr. 1999.