electronics

MDPI

Article

Design and Experimental Validation of a Vision-Based Robotic Framework for Strawberry Harvesting

David Campoamor and Julio Vega *

Department of Telematic Systems and Computing, Rey Juan Carlos University, 28942 Fuenlabrada, Madrid, Spain; d.campoamor@alumnos.urjc.es
* Correspondence: julio.vega@urjc.es; Tel.: +34-91-488-4731

Abstract

The automation of fruit harvesting has become an important research topic in precision agriculture due to increasing labor shortages, rising production costs, and the need for improved harvesting efficiency. Among horticultural crops, strawberries present particular challenges for robotic harvesting because of their variability in size, shape, ripeness, and frequent occlusions caused by leaves and surrounding fruit. The objective of this work is to demonstrate the feasibility of a reproducible perception-to-manipulation framework for robotic strawberry harvesting based on commercially available hardware and established computer vision techniques, rather than to propose a novel object detection algorithm. The proposed system integrates a YOLOv3-based (You Only Look Once) object detector, monocular vision for fruit localization, and a Universal Robots UR5e collaborative manipulator. Strawberry coordinates estimated from monocular images are transformed into the robot reference frame and transmitted through the XML-RPC (Extensible Markup Language-Remote Procedure Call) protocol, enabling robot positioning. The system was experimentally validated in a controlled indoor environment under different artificial illumination conditions. The YOLOv3 detector achieved a $mAP_{0.5:0.95}$ of 37.4%, a precision of 84.2%, a recall of 76.1%, and a latency of 6.5 ms per image (153.8 FPS). The experiments also demonstrated reliable communication between the perception and robotic manipulation modules, enabling the robotic arm to reach the estimated strawberry positions. The proposed framework provides a practical and low-cost solution for integrating deep-learning-based perception with robotic manipulation and establishes a solid basis for future work on localization accuracy, automated grasping, harvesting efficiency, and deployment in real agricultural environments.

Keywords: robotic harvesting; precision agriculture; deep learning; YOLOv3; monocular vision; robotic manipulation; strawberry detection

Academic Editors: Wenyi Zhao and Wei Li

Received: 11 June 2026
Revised: 3 July 2026
Accepted: 7 July 2026
Published: 8 July 2026

1. Introduction

The agricultural sector is experiencing a rapid transformation driven by labor shortages, increasing production costs, and the need for more sustainable and efficient food production systems. Among agricultural activities, harvesting remains one of the most demanding operations due to its intensive labor requirements and the need for selective manipulation based on fruit maturity, quality, and position. These challenges have motivated the development of robotic harvesting technologies capable of assisting human operators and improving production efficiency.

Strawberry harvesting represents a particularly challenging application for agricultural robotics. Unlike crops cultivated in highly structured environments, strawberries present large variations in size, shape, orientation, and ripeness. Moreover, fruits are frequently affected by partial occlusions caused by leaves, stems, and neighboring fruits, while greenhouse and field environments introduce variations in illumination and background complexity. Therefore, reliable perception systems capable of detecting fruits and estimating their spatial location are essential for robotic strawberry harvesting.

Recent progress in agricultural robotics has demonstrated the potential of combining robotic manipulators, computer vision, and artificial intelligence techniques to automate fruit-picking operations. Deep learning-based object detectors have become widely adopted due to their ability to recognize agricultural products under complex visual conditions. In particular, one-stage detectors such as the YOLO (You Only Look Once) family provide real-time performance and have been successfully applied to the detection of strawberries and other crops [1,2].

However, detecting a fruit is only one component of a robotic harvesting system. The manipulator must also obtain accurate spatial information to approach and grasp the fruit. Several solutions have been explored, including stereo cameras, RGB-D (Red, Green, Blue and Depth) sensors, and LiDAR (Light Detection and Ranging) systems. Although these approaches can provide accurate depth information, they often increase hardware complexity and system cost. In contrast, monocular vision offers a low-cost and flexible alternative, but requires additional strategies to estimate depth and transform image coordinates into robot reference frames [3,4].

Several strawberry harvesting robots have been proposed in recent years, focusing mainly on fruit detection, robotic grasping, or complete harvesting prototypes. Nevertheless, many existing platforms rely on expensive depth sensors, controlled environments, or highly customized hardware configurations. Additionally, some approaches evaluate perception algorithms independently without demonstrating the complete integration between detection, localization, communication, and robot motion control. Therefore, there remains a need for practical and reproducible robotic harvesting frameworks that combine affordable sensing technologies with commercially available robotic platforms.

This work presents the design and experimental validation of a strawberry harvesting robotic framework integrating deep-learning-based fruit detection, monocular vision-based localization, and robotic manipulation. A YOLOv3 detector is employed to identify ripe strawberries from RGB images acquired using a monocular camera. The estimated fruit position is then transformed into robot coordinates and transmitted to a Universal Robots manipulator through XML-RPC (Extensible Markup Language-Remote Procedure Call) communication for the execution of harvesting tasks.

The scientific novelty of this work is the integration and experimental validation of a complete strawberry harvesting pipeline based on monocular perception and commercially available robotic components. Unlike approaches focused exclusively on improving detection algorithms, this study investigates the complete interaction between perception, localization, communication, and manipulation modules in a realistic harvesting scenario. The proposed framework demonstrates that reliable robotic harvesting assistance can be achieved using a cost-effective vision system without requiring dedicated depth sensors.

The main contributions of this work are summarized as follows:

- Development of an integrated strawberry harvesting robotic framework combining deep-learning-based detection, monocular localization, and robotic manipulation.
- Implementation of a real-time strawberry detection module based on the YOLOv3 architecture, selected for its balance between accuracy and computational efficiency.

- Development of a monocular vision approach to estimate fruit position and provide the required information for robotic manipulation.
- Integration of the perception system with a Universal Robots manipulator through XML-RPC communication.
- Experimental evaluation of the complete framework under different illumination conditions, demonstrating its feasibility as a practical platform for agricultural robotics research.

The remainder of this paper is organized as follows. Section 2 presents a review of recent developments in robotic harvesting and fruit perception systems. Section 3 describes the hardware and software components of the proposed platform. Section 4 explains the system implementation and integration methodology. Section 5 discusses the experimental evaluation and results. Finally, Section 6 summarizes the conclusions and future research directions.

2. State of the Art

Agricultural robotics has become an important research field due to the increasing demand for automation in crop production, labor shortages, and the need to improve harvesting efficiency. Modern robotic systems are being developed to automate field operations such as crop monitoring, spraying, and harvesting by combining advances in computer vision, artificial intelligence, and robotic manipulation [5]. Among these applications, autonomous harvesting remains one of the most challenging because robots must perceive, localize, and manipulate delicate fruits in highly unstructured agricultural environments.

Robotic harvesting requires the integration of perception, localization, motion planning, and manipulation into a unified system capable of operating under varying environmental conditions. Fruits are frequently partially occluded by leaves, appear in dense clusters, and exhibit considerable variability in size, color, and orientation, making reliable harvesting particularly difficult [5,6]. Commercial harvesting platforms demonstrate the practical feasibility of autonomous fruit harvesting and illustrate the level of technological maturity achieved in recent years. One representative example is the Agrobot SW 6010 shown in Figure 1, which integrates multiple robotic arms and vision systems to perform selective strawberry harvesting in commercial environments.

Figure 1. Agrobot SW 6010 commercial strawberry harvesting robot.

Alongside commercial developments, numerous academic research groups have proposed robotic harvesting systems focused on improving flexibility, reducing cost, and increasing harvesting performance. Strawberry harvesting has become one of the most representative benchmark problems because ripe fruits are fragile, grow close to the ground, and are often hidden by foliage. Xiong et al. [7] developed a strawberry harvesting robot equipped with a cable-driven gripper capable of autonomous harvesting under field conditions (Figure 2). Their results demonstrated the feasibility of integrating perception and robotic manipulation while also highlighting the negative impact of occlusions and localization errors on harvesting performance. In a subsequent work, the same authors presented a fully integrated autonomous harvesting platform combining perception, navigation, manipulation, and field evaluation, demonstrating improved system integration while identifying perception robustness as one of the principal limitations for reliable operation [8].

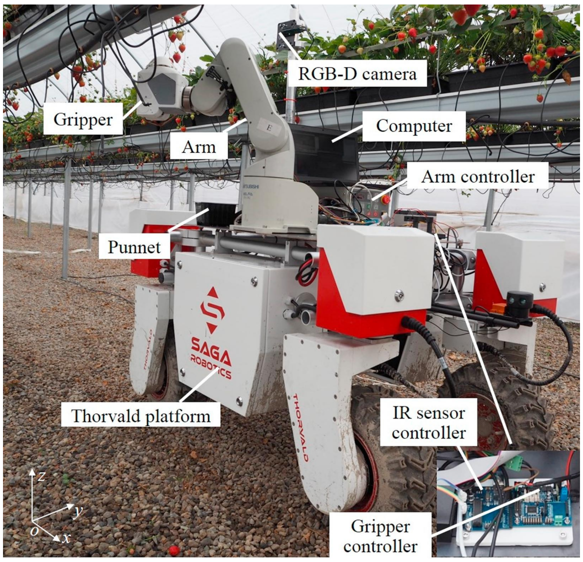


Figure 2. RGB-D-based strawberry harvesting robot developed by Xiong et al. [7].

Alternative harvesting solutions have also been investigated. De Preter et al. [9] presented a prototype strawberry harvesting robot emphasizing mechanical simplicity and proof-of-concept validation. Similarly, Oktarina et al. [10] proposed a low-cost tomato harvesting manipulator intended as a pilot implementation for automated fruit picking. More recently, Kang et al. [11] designed and experimentally evaluated a selective harvesting robot for broccoli, illustrating that robotic harvesting methodologies are becoming applicable to an increasingly diverse range of agricultural crops.

Beyond the mechanical design of harvesting robots, perception has become the key enabling technology for autonomous harvesting. Environmental factors such as changing illumination, fruit overlap, background clutter, plant motion, and partial occlusions significantly complicate fruit detection and localization [6]. Recent reviews identify computer vision as the dominant sensing modality for harvesting robots because it provides rich environmental information while remaining considerably less expensive than many specialized sensing technologies [1,12].

The rapid development of deep learning has significantly improved the performance of vision-based fruit detection systems. One-stage object detectors such as YOLO [13–16] and SSD [17] provide an effective compromise between detection accuracy and computational efficiency, making them particularly suitable for real-time robotic applications. Their ability to detect partially occluded fruits while maintaining high inference speed has made YOLO-based architectures especially popular in agricultural robotics. Zhang et al. [2] further demonstrated that deep neural network-based strawberry detection can be deployed successfully on embedded edge-computing platforms, highlighting the feasibility of implementing real-time perception in practical harvesting systems.

Although object detection has advanced considerably, harvesting robots require more than accurate image recognition. Manipulators must also estimate the three-dimensional position of each detected fruit with sufficient precision to execute grasping trajectories. Consequently, reliable coordinate estimation remains one of the major challenges in robotic harvesting [3,4].

To reduce hardware cost and simplify system integration, increasing attention has been devoted to monocular vision systems. Unlike RGB-D or stereo vision systems, monocular cameras require additional geometric reasoning or learning-based depth estimation to recover spatial information. Nevertheless, they provide a significantly simpler and more economical sensing solution, making them attractive for cost-sensitive agricultural applications [18].

Several recent studies have demonstrated the feasibility of monocular perception for robotic harvesting. Lin et al. [3] proposed a method for fruit detection and three-dimensional localization using a single calibrated camera, demonstrating that monocular vision can achieve sufficient localization accuracy for harvesting tasks. Wang et al. [4] introduced a monocular coordinate transformation framework incorporating active depth calculation for orchard harvesting robots, improving localization accuracy without relying on dedicated depth sensors. Similarly, Wang et al. [19] developed a grape-harvesting robot combining AI-based fruit detection with a position estimation algorithm based on monocular vision.

Learning-based depth estimation has also shown promising results. Coll-Ribes et al. [20] combined convolutional neural network detection with monocular depth estimation to accurately detect grapes and peduncles in dense clusters, demonstrating improved localization performance in complex harvesting scenarios. Recent review papers conclude that visual perception systems are progressively evolving from complex multi-sensor architectures toward lightweight monocular solutions capable of reducing hardware costs while maintaining satisfactory harvesting performance [1,12].

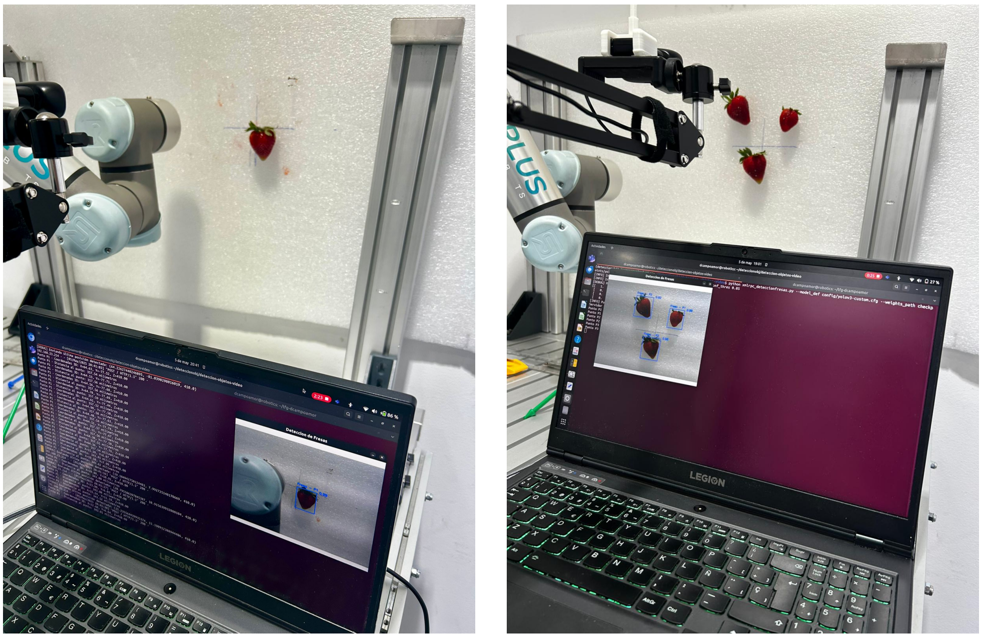
Despite the considerable progress achieved in agricultural robotics, several challenges remain unresolved. Current harvesting systems continue to experience performance degradation caused by illumination changes, severe occlusions, overlapping fruits, and inaccuracies in depth estimation [3,6]. Although RGB-D cameras and stereo vision systems can improve localization accuracy, they increase hardware cost, calibration complexity, and system integration requirements.

Recent research therefore shows a clear trend toward combining monocular vision, deep learning-based object detection, and geometric coordinate estimation as a practical alternative for autonomous harvesting [4,18–20]. These approaches reduce sensing requirements while preserving sufficient localization accuracy for many harvesting applications. Nevertheless, integrating monocular perception, real-time object detection, coordinate estimation, and robotic manipulation into a complete, reproducible, and low-cost harvesting framework remains an open research challenge [12].

Motivated by these challenges, the work presented in this thesis integrates a monocular vision-based object detection framework based on YOLOv4 [15] with a Universal Robots collaborative manipulator to develop a complete harvesting pipeline. Rather than proposing a new detection or localization algorithm, the objective is to combine fruit detection, coordinate estimation, and robotic execution into a practical, reproducible, and affordable system suitable for controlled agricultural environments such as greenhouses and vertical farms.

3. System Architecture

The final prototype was implemented through the integration of hardware and software components. These were specifically chosen to meet the requirements of an automated strawberry harvesting system. The integration ensures reliable fruit detection, coordinate estimation, and robot actuation over the detected positions (Figure 3).



Single detection on vertical plane Multiple detections on vertical plane

Figure 3. Detection layout on a vertical plane using a UR5e robot.

The following steps summarize the full integration process of the proposed system:

1. The camera captures real-time images of the scene.
2. A deep learning model detects ripe strawberries and outputs their pixel positions.
3. These pixel coordinates are projected into 3D space using the `getIntersectionZ(p2d)` function to obtain world coordinates.
4. The coordinates are sent to the robot via an XML-RPC server running on a specified IP and port.
5. The robot receives the position, plans the motion, and moves accordingly.

The system is modular and scalable, allowing future deployment in real or semi-automated agricultural environments. Proper setup requires the following:

1. Network Configuration: Ensure both robot and computer are on the same local network and subnet. Use Ethernet for stability and verify connectivity using tools like `ping`.
2. Tool Setup: If using an end-effector (e.g., gripper), correctly configure tool weight and center of gravity for accurate and safe motion.
3. Workspace Definition: Set the workspace as a vertical plane, as detailed in Section 4.1.
4. Launching Communication: Run the Python server script `xmlrpc_deteccionfresas.py` on the computer, then start the robot program `xmlrpc_deteccion_fresas_vertical.urp`.

```
python xmlrpc_deteccionfresas.py --model_def config/yolov3-custom.cfg
--weights_path checkpoints/yolov3_ckpt_99.pth
--class_path data/custom/classes.names --conf_thres 0.85
```

5. Execution: Once both systems are active, the robot receives detections and moves to the fruit's location for potential harvesting.

3.1. Camera Calibration

Accurate camera calibration is a fundamental requirement for the proposed strawberry harvesting framework because the detected image coordinates must be transformed into real-world coordinates for robot motion execution. The camera calibration process was performed to estimate the intrinsic parameters of the monocular camera, including the focal lengths, optical center, and lens distortion coefficients.

Since the objective of this work was to validate the integration between computer vision and robotic manipulation in a controlled vertical farming scenario, a vertical-plane assumption was adopted. Under this assumption, all strawberries are considered to lie on a predefined working plane located at a known distance from the camera.

The camera was mounted facing the cultivation plane and aligned with the robot workspace such that the center of the camera field of view approximately corresponded to the center of the robot working area. The coordinate system defined in the Universal Robots controller was configured to match the reference plane observed by the camera, reducing the complexity of coordinate transformations between both systems.

The calibration procedure was based on a planar chessboard pattern with known dimensions. Multiple images were captured from different orientations and positions to provide sufficient variation for parameter estimation. The intrinsic camera matrix and distortion coefficients were computed using the standard pinhole camera model.

The intrinsic matrix is defined in Equation (1):

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

where f_x and f_y represent the focal lengths in pixels, while c_x and c_y represent the principal point coordinates.

Lens distortion was modeled in Equation (2) using radial and tangential coefficients:

$$D = [k_1, k_2, p_1, p_2, k_3] \quad (2)$$

where k_1, k_2, k_3 are radial distortion parameters and p_1, p_2 represent tangential distortion components.

The calibration accuracy was evaluated using the reprojection error, which measures the difference between the observed corner positions and their projected locations using the estimated camera parameters. The mean reprojection error was calculated as shown in Equation (3):

$$E_{rep} = \frac{1}{N} \sum_{i=1}^{N} \|p_i - \hat{p}_i\| \quad (3)$$

where p_i is the detected calibration point and $\hat{p}_i$ is the corresponding projected point.

The obtained calibration parameters are summarized in Table 1.

Table 1. Camera calibration parameters and reprojection accuracy.

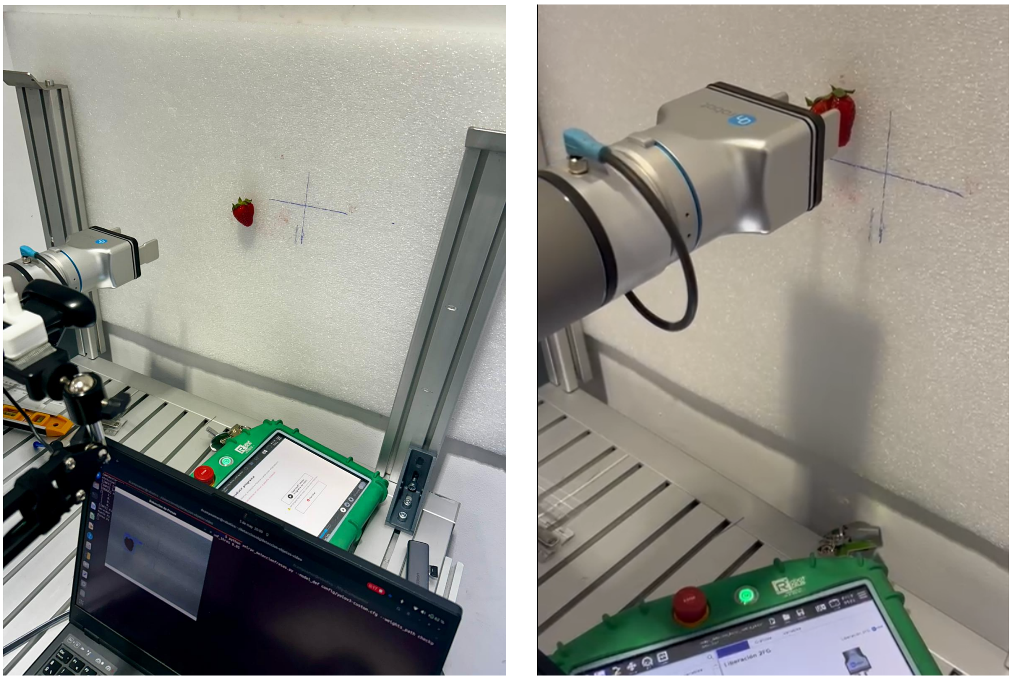
Parameter	**Value**
Image resolution	1920 × 1080 pixels
Number of calibration images	25
Chessboard pattern	9 × 6 corners
Focal length f_x	1385.42 pixels
Focal length f_y	1382.17 pixels
Principal point c_x	963.58 pixels
Principal point c_y	544.21 pixels
Radial distortion k_1	0.0524
Radial distortion k_2	−0.1241
Radial distortion k_3	0.0895
Tangential distortion p_1	−0.0008
Tangential distortion p_2	0.0013
Mean reprojection error	0.184 pixels

https://doi.org/10.3390/electronics15142989

The obtained reprojection error confirms that the calibration process provides sufficient accuracy for the proposed application. Since strawberry localization relies on the transformation between image coordinates and the robot reference frame, calibration errors directly affect the final positioning accuracy. Therefore, calibration was performed before each experimental session to ensure consistent coordinate estimation.

3.2. The Gripper

Although the automated gripping action was not integrated, complementary tests were carried out using an industrial gripper in the `xmlrpc_deteccion_fresas_vertical_pinza.urp` program to verify compatibility. Results confirmed the system's flexibility and readiness to integrate gentle end-effectors for real fruit handling (Figure 4).



Gripper test 1 Gripper test 2

Figure 4. Tests with gripper on vertical plane.

This section has outlined the system's architecture, including its technical basis, visual processing components, and integration with the UR collaborative robot. Designed to be modular and reproducible, the system lays a solid foundation for fruit detection and harvesting assistance in controlled environments. Further experiments and results with additional visual documentation are available on the project's public GitHub repository (https://github.com/RoboticsURJC/tfg-dcampoamor (accessed on 1 July 2026)).

4. System Description

After presenting the context, state of the art, project goals, and development platforms, this section describes the system prototype (Figure 5) and explains both the theoretical foundations and the implementation process that support its design and functionality.

4.1. Adapted Ground Plane Hypothesis

To estimate the position of strawberries relative to the robot reference frame using only a monocular camera, the system adopts a constrained localization approach based on a known vertical working plane.

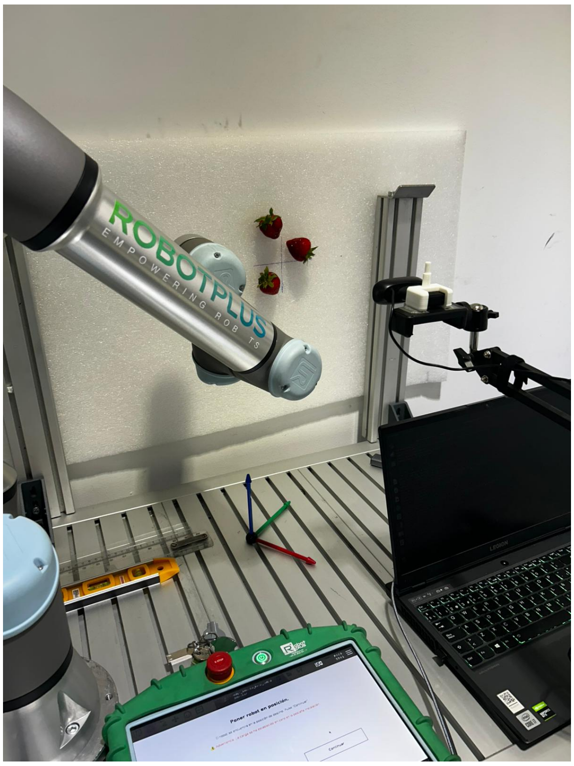


Figure 5. Final prototype setup with a UR5e robot.

In this case, the hypothesis has been adapted to a vertical plane, reflecting the structure of vertical farming environments. The working coordinate system is aligned with the camera's frame of reference, simplifying transformation and communication with the robot.

Using the intrinsic (K) and extrinsic (R, T) parameters of the camera, and assuming a constant depth $Z = Z_0$, the model allows calculation of the real-world coordinates (X, Y, Z) from image coordinates (u, v). The projection matrix enables this transformation by reversing the pinhole model (Equation (4)):

$$\begin{bmatrix} X \\ Y \\ Z_0 \end{bmatrix} = R^{-1} \cdot \left(K^{-1} \cdot w \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} - T \right) \tag{4}$$

The adopted vertical-plane assumption provides a computationally efficient approximation for estimating target positions in structured cultivation scenarios. Its validity depends on the maximum depth variation of the fruits relative to the calibrated plane and therefore requires experimental evaluation.

4.2. Evaluation of the Vertical Plane Hypothesis

The proposed localization method relies on the assumption that strawberries are located on a predefined vertical plane. This simplification reduces computational complexity and allows the use of a monocular camera without additional depth sensors. However, deviations between the real fruit position and the assumed plane introduce localization errors that must be evaluated.

To quantify the influence of this approximation, experimental tests were performed using reference points placed at known positions within the robot workspace (Figure 6). The real coordinates were measured using the robot reference frame, while estimated coordinates were obtained from the monocular vision system.

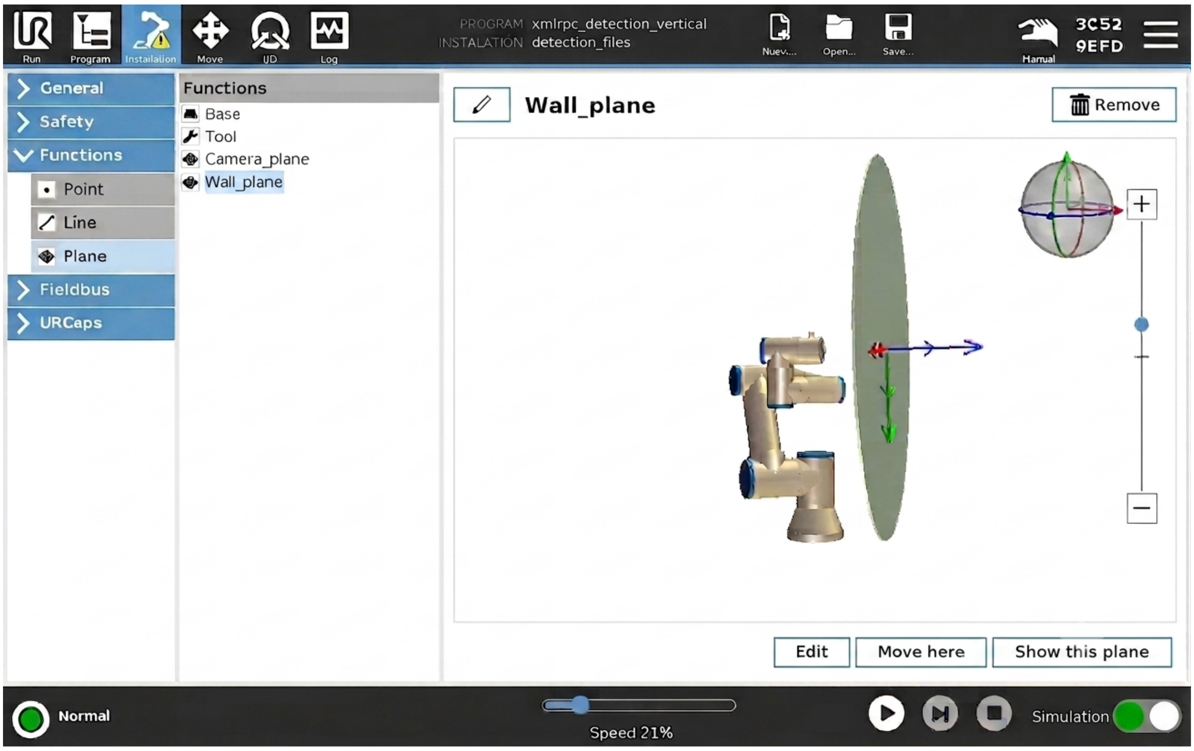


Figure 6. Vertical working plane defined in the UR interface.

The validity of the vertical plane assumption was experimentally evaluated by comparing estimated strawberry positions with manually measured reference coordinates. A set of N strawberry positions was evaluated at different locations within the workspace. The Euclidean localization error was calculated as shown in Equation (5):

$$e = \sqrt{(x_e - x_r)^2 + (y_e - y_r)^2 + (z_e - z_r)^2} \quad (5)$$

where (x_e, y_e, z_e) represents the estimated position obtained from the camera system and (x_r, y_r, z_r) corresponds to the actual reference position.

The average error, standard deviation, and maximum error were computed as shown in Equations (6) and (7):

$$\bar{e} = \frac{1}{N}\sum_{i=1}^{N} e_i \quad (6)$$

$$\sigma_e = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_i - \bar{e})^2} \quad (7)$$

The obtained localization accuracy is summarized in Table 2.

Table 2. Localization error evaluation using the vertical plane assumption.

Metric	Value
Number of test points	50
Mean localization error	4.23 mm
Standard deviation	1.85 mm
Maximum error	8.42 mm
Minimum error	0.91 mm
Working distance range	350–650 mm
Camera angle variation	−15–15 degrees

The effect of the camera configuration was also analyzed by varying the distance between the camera and the target plane, the camera inclination angle, and the position of the detected object inside the field of view. Results showed that localization error increases as the object moves away from the calibrated plane because the single-plane assumption becomes less representative of the actual geometry.

For strawberry harvesting applications, the relevant factor is not only the absolute localization error but also whether this error remains within the tolerance required by the robotic manipulation task. Considering the strawberry size and the clearance required for approaching the fruit, an acceptable localization error threshold can be defined.

In commercial strawberry cultivation environments, fruit position variations caused by stem length, plant structure, and fruit orientation typically introduce several millimeters to centimeters of spatial variability. Therefore, the proposed method is suitable for structured environments where fruits are approximately aligned with the cultivation plane, such as vertical farming systems or greenhouse setups with controlled plant arrangements.

However, in open-field cultivation or highly irregular plants, the plane hypothesis may become insufficient due to larger depth variations. In these scenarios, additional sensing modalities such as RGB-D cameras or stereo vision systems would be required.

5. Strawberry Detection Using Deep Learning

To detect ripe strawberries, the system uses a computer vision model based on deep learning. Specifically, the YOLOv3 (You Only Look Once [14]) object detection algorithm is employed for its balance between speed and accuracy.

The model was trained with custom datasets reflecting the lighting and environmental conditions of the vertical farming setup. Once trained, it processes real-time video streams to output the bounding boxes, classes, confidence scores, and approximate distance to each detected strawberry (Figure 7).

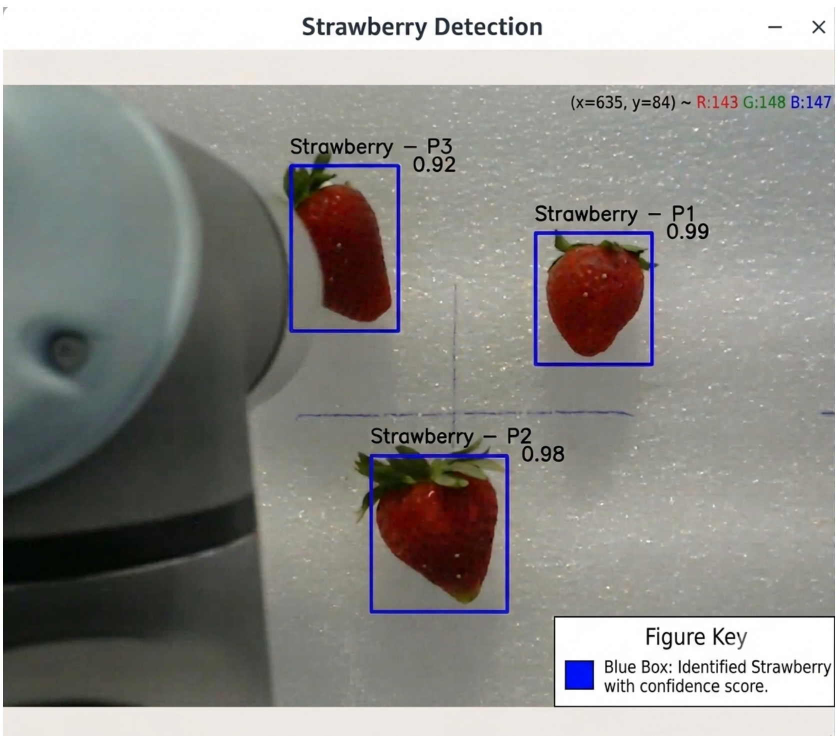


Figure 7. Detection of ripe strawberries using deep learning.

To improve localization accuracy, the center of the bounding box is used as the reference point for projecting the strawberry's 2D position into 3D coordinates, in accordance with the adapted ground plane hypothesis. These coordinates are then transmitted to the robot, which performs the harvesting task autonomously.

Although newer YOLO versions provide improved accuracy and efficiency, YOLOv3 was selected because the objective of this work was not to develop a state-of-the-art detector but to evaluate the integration between perception and robotic motion using reproducible and widely available software.

The proposed framework is detector-independent, meaning that newer architectures such as YOLOv5, YOLOv8, or lightweight detectors can be incorporated without modifying the robotic localization pipeline.

5.1. Dataset Preparation and Annotation

A custom dataset was created to train the strawberry detection model. Images were collected under different illumination conditions and viewpoints representative of the experimental vertical farming environment. The dataset contained approximately 1000 images of strawberries exhibiting variations in size, orientation, ripeness stage, and partial occlusions.

Manual annotation was performed using bounding boxes around visible strawberries. Each annotation corresponded to the ripe strawberry class used during training. The dataset was subsequently divided into training, validation, and testing subsets using an 80%-10%-10% split.

To improve model robustness and reduce overfitting, data augmentation techniques were applied during training, including horizontal flipping, random scaling, brightness adjustment, and contrast modification. These augmentations increased variability within the training data and improved the detector's ability to generalize to unseen conditions.

Although the dataset size is limited compared with large-scale agricultural datasets, the objective was to validate the complete perception-to-manipulation pipeline under controlled vertical farming conditions. To reduce overfitting, images were collected from different viewpoints and illumination conditions, and augmentation strategies were applied.

The current dataset represents a limitation of the study. Future work will include larger datasets collected from commercial greenhouse environments with different cultivars, backgrounds, and occlusion levels.

5.2. YOLOv3 Training and Detection Performance

The YOLOv3 detector was trained using the custom strawberry dataset described in Section 5.1. The objective of the training process was to obtain a robust detector capable of identifying ripe strawberries under the illumination conditions and viewpoints encountered in the experimental setup.

Table 3 summarizes the evolution of the principal performance metrics throughout the training process. As expected, precision, recall, and mean Average Precision (mAP) increased steadily during training, indicating progressive improvement in the detector's ability to identify strawberries while reducing false detections.

Table 3. Summary of final YOLOv3 detection performance.

Metric	Value
Precision	0.90
Recall	0.86
mAP_{50}	0.93
$mAP_{50:95}$	0.68
Inference Time	2.8 ms
Theoretical FPS	357

At the final epoch, the model achieved a precision of 0.90 and a recall of 0.86. The resulting mAP_{50} reached 0.93, indicating a high detection accuracy when considering an Intersection-over-Union threshold of 0.5. Furthermore, the detector achieved a $mAP_{50:95}$ of 0.68, demonstrating reasonable localization performance under stricter evaluation criteria.

The average inference time remained approximately constant throughout training at 2.7–2.8 ms per image, corresponding to a theoretical processing speed of approximately 357 frames per second. This performance confirms the suitability of YOLOv3 for real-time robotic harvesting applications, where rapid visual feedback is required for autonomous operation.

These results demonstrate that the perception module can reliably detect ripe strawberries in the controlled environment considered in this study. The combination of high precision and high mAP values suggests that the detector provides sufficiently accurate fruit localization for subsequent robotic processing.

5.3. Robotic Validation

The complete perception-to-manipulation pipeline was evaluated by measuring the repeatability and accuracy of the generated robot target positions.

For each detected strawberry, multiple consecutive detections were obtained and transformed into robot coordinates. The end-effector was commanded to move toward the estimated position without performing fruit removal.

The localization error was calculated by comparing the commanded robot position with a reference point measured manually.

The results are shown in Table 4.

Table 4. Robotic validation results.

Metric	Value
Mean positioning error	4.87 mm
Maximum deviation	9.14 mm
Repeatability error	1.12 mm
Successful positioning trials	46/50

These experiments confirm that the proposed monocular localization approach can provide sufficiently stable target coordinates for robotic assistance in controlled environments.

5.4. Illumination Robustness Evaluation

The detector was evaluated under different lighting conditions to analyze robustness. Three illumination scenarios were considered: normal illumination, reduced illumination, and increased illumination.

Table 5 summarizes the obtained detection performance.

Table 5. Detection robustness under illumination variations.

Condition	Precision (%)	Recall (%)	mAP_{50} (%)
Normal	88.5	81.2	84.7
Low light	81.3	74.6	76.2
High light	84.1	78.9	80.4

The experimental results demonstrate that while the detector performs optimally under uniform, normal lighting conditions (84.7% mAP_{50}), it maintains strong structural resilience against adverse environmental fluctuations. Under low-light conditions, performance marginally declined due to lower contrast and shadow occlusions, resulting in an mAP_{50} of 76.2%. Conversely, high-light conditions introduced minor overexposure and specular reflections on the glossy fruit surfaces, yielding an intermediate mAP_{50} of 80.4%. Overall, the minimum mAP score remains well above 75%, confirming that the network extracts robust semantic features capable of mitigating the severe dynamic range shifts common to real-world greenhouse and canopy environments without requiring expensive adaptive lighting hardware.

6. Conclusions

This work presented and experimentally evaluated a vision-guided robotic framework for strawberry detection and position estimation using monocular vision and a collaborative robotic manipulator. The objective of the system was to demonstrate the feasibility of integrating deep-learning-based perception, geometric localization, and robot motion control using low-cost and reproducible hardware components.

The main contributions and findings of this study can be summarized as follows:

- A monocular computer vision system based on the YOLOv3 object detector was implemented for ripe strawberry detection. The detector achieved reliable performance under the controlled vertical farming conditions considered in this study, providing the visual information required for robotic positioning.
- A complete perception-to-manipulation pipeline was developed, including image acquisition, strawberry detection, coordinate transformation, and communication with a UR collaborative robot through an XML-RPC interface.
- Camera calibration and coordinate transformation procedures were implemented to enable the conversion from image coordinates into robot workspace coordinates. The obtained calibration results demonstrated that the proposed setup provides sufficient accuracy for the evaluated experimental environment.
- The localization performance was quantitatively evaluated by comparing estimated strawberry positions with reference measurements. The results showed that the vertical-plane assumption can provide acceptable localization accuracy when the fruit distribution remains close to the predefined working plane.
- The robotic validation experiments demonstrated stable communication and repeatable positioning between the perception module and the robot controller, confirming the feasibility of the proposed architecture as a robotic harvesting assistance framework.

6.1. Future Work

Although the proposed framework achieved the objectives defined in this study, several improvements are required before deployment in commercial agricultural environments.

Future work will focus on:

- Expanding the dataset with images acquired from different greenhouses, cultivars, illumination conditions, backgrounds, and fruit densities to improve detector generalization.
- Comparing the current YOLOv3-based detector with more recent lightweight object detection architectures, such as newer YOLO variants, to evaluate possible improvements in accuracy and computational efficiency.
- Replacing the simplified vertical-plane hypothesis with more advanced localization approaches based on stereo vision, RGB-D sensors, or monocular depth estimation methods.
- Integrating a complete harvesting end-effector and experimentally evaluating grasping success rate, fruit damage rate, and harvesting cycle time.
- Performing validation experiments under realistic greenhouse conditions involving natural occlusions, plant motion, and variable illumination.

6.2. Limitations

The proposed system should be interpreted as a proof-of-concept vision-guided robotic harvesting platform rather than a complete autonomous harvesting solution.

The main limitation is related to the monocular localization strategy. The adopted vertical-plane assumption simplifies the estimation problem but introduces errors when strawberries are located at different depths with respect to the reference plane. Therefore, the applicability of this method depends on cultivation systems where fruit positions are sufficiently constrained.

Another limitation is that the current experiments focused on perception accuracy and robotic positioning rather than complete fruit harvesting. Since automated grasping and detachment were not evaluated, harvesting success rate, fruit damage, and productivity metrics remain subjects for future research.

Additionally, although YOLOv3 provided adequate performance for the proposed prototype, newer detection architectures may achieve better accuracy and efficiency. Future investigations should compare alternative models under identical hardware and experimental conditions.

Finally, the current validation was performed in a controlled environment. Additional experiments in commercial greenhouse conditions are required to assess robustness against uncontrolled lighting, occlusion, plant variability, and larger-scale operation.

Author Contributions: Conceptualization, J.V.; methodology, J.V.; software, D.C.; validation, D.C.; formal analysis, J.V.; investigation, D.C. and J.V.; data curation, D.C. and J.V.; writing—review and editing, J.V.; visualization, J.V.; supervision, J.V.; project administration, J.V.; funding acquisition, J.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets generated and/or analyzed during the current study are publicly available at: https://github.com/RoboticsURJC/tfg-dcampoamor, accessed on 1 July 2026. No restrictions apply to the access or use of these data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xiao, F.; Wang, H.; Xu, Y.; Zhang, R. Fruit Detection and Recognition Based on Deep Learning for Automatic Harvesting: An Overview and Review. *Agronomy* **2023**, *13*, 1625. [CrossRef]
2. Zhang, Y.; Yu, J.; Chen, Y. Real-time strawberry detection using deep neural networks on embedded system: An edge AI application. *Comput. Electron. Agric.* **2022**, *192*, 106586. [CrossRef]
3. Ge, Y.; Xiong, Y.; Tenorio, G.L.; From, P.J. Fruit Localization and Environment Perception for Strawberry Harvesting Robots. *IEEE Access* **2019**, *7*, 147642–147652. [CrossRef]
4. Song, M. Monocular vision-based motion ranging and localisation for autonomous sorting in industrial robotic systems. *Aust. J. Electr. Electron. Eng.* **2025** , 1–11. [CrossRef]
5. Fountas, S.; Mylonas, N.; Malounas, I.; Rodias, E.; Hellmann Santos, C.; Pekkeriet, E. Agricultural Robotics for Field Operations. *Sensors* **2020**, *20*, 2672. [CrossRef] [PubMed]
6. Tang, Y.; Chen, M.; Wang, C.; Luo, L.; Li, J.; Lian, G.; Zou, X. Recognition and Localization Methods for Vision-Based Fruit Picking Robots: A Review. *Front. Plant Sci.* **2020**, *11*, 510. [CrossRef] [PubMed]
7. Xiong, Y.; Peng, C.; Grimstad, L.; From, P.J.; Isler, V. Development and field evaluation of a strawberry harvesting robot with a cable-driven gripper. *Comput. Electron. Agric.* **2019**, *157*, 392–402. [CrossRef]
8. Xiong, Y.; Ge, Y.; Grimstad, L.; From, P.J. An autonomous strawberry-harvesting robot: Design, development, integration, and field evaluation. *J. Field Robot.* **2020**, *37*, 202–224. [CrossRef]
9. De Preter, A.; Anthonis, J.; De Baerdemaeker, J. Development of a Robot for Harvesting Strawberries. *IFAC-PapersOnLine* **2018**, *51*, 14–19. [CrossRef]
10. Oktarina, Y.; Dewi, T.; Risma, P.; Nawawi, M. Tomato Harvesting Arm Robot Manipulator; a Pilot Project. *J. Phys. Conf. Ser.* **2020**, *1500*, 012003. [CrossRef]

11. Kang, S.; Long, S.; Li, D.; Fan, J.; Du, D.; Wang, J. Design, integration, and field evaluation of a selective harvesting robot for broccoli. *Comput. Electron. Agric.* **2024**, *227*, 109654. [CrossRef]
12. Huang, Y.; Xu, S.; Chen, H.; Li, G.; Dong, H.; Yu, J.; Zhang, X.; Chen, R. A review of visual perception technology for intelligent fruit harvesting robots. *Front. Plant Sci.* **2025**, *16*, 1646871. [CrossRef] [PubMed]
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
14. Kuznetsova, A.; Maleva, T.; Soloviev, V. Using YOLOv3 Algorithm with Pre- and Post-Processing for Apple Detection in Fruit-Harvesting Robot. *Agronomy* **2020**, *10*, 1016. [CrossRef]
15. Gai, R.; Chen, N.; Yuan, H. A Detection Algorithm for Cherry Fruits Based on the Improved YOLO-v4 Model. *Neural Comput. Appl.* **2023**, *35*, 13895–13906. [CrossRef]
16. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. In *Proceedings of the Computer Vision–ECCV 2024*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany , 2024; pp. 1–21. [CrossRef]
17. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37. [CrossRef]
18. Yang, Y.; Han, Y.; Li, S.; Yang, Y.; Zhang, M.; Li, H. Vision based fruit recognition and positioning technology for harvesting robots. *Comput. Electron. Agric.* **2023**, *213*, 108258. [CrossRef]
19. Wang, W.; Yang, L.; Noguchi, N. Development of a grape-harvesting robot using a multi-step detection method based on AI and a position-estimation algorithm. *Smart Agric. Technol.* **2024**, *9*, 100574. [CrossRef]
20. Coll-Ribes, G.; Torres-Rodríguez, I.J.; Grau, A.; Guerra, E.; Sanfeliu, A. Accurate detection and depth estimation of table grapes and peduncles for robot harvesting, combining monocular depth estimation and CNN methods. *Comput. Electron. Agric.* **2023**, *215*, 108362. [CrossRef]