Recognition and Position Estimation of Pears in Complex Orchards Using Stereo Camera and Deep Learning Algorithm

Siyu Pan[1], Ayanori Yorozu[2], Akihisa Ohya[2] and Tofeal Ahamed[3]

[1]*Graduate School of Science and Technology, University of Tsukuba, 1-1-1 Tennodai, Tsukuba 305-8573, Japan*

[2]*Institute of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan*

[3]*Faculty of Life and Environmental Science, University of Tsukuba, Tsukuba, Japan*

Keywords: Pear Recognition, Position Estimation, Stereo Camera.

Abstract: Complex orchards present difficulties for fruit-picking robots due to shadows, overlapping fruits, and obstructing branches, resulting in errors during grasping. To improve the robustness of fruit-picking robots in the complex environment, this study compared the performance of different types of deep learning algorithms (Mask R-CNN, Faster R-CNN, and YOLACT) for pear recognition under different conditions (high and low light). Additionally, the ZED2 stereo camera with the algorithm of the highest precision for estimating the position of separating and aggregating pears. For pear recognition, the mAPs of Mask R-CNN were 95.22% and 99.45%, Faster R-CNN were 87.90% and 87.52%, YOLACT were 87.07% and 97.89% in the validation and test set. For position estimation, the mean error of separating pears was 0.017m, the standard deviation was 0.015m and the goodness of fit reached 0.896; The mean error of aggregating pears were 0.018m and the standard deviation was 0.021m and the goodness of fit reached 0.832. A pear recognition and positioning system was developed by ZED2 stereo camera with deep learning algorithm. It aimed to generate precise bounding boxes and recognize pears in a complex orchard within the range of 0.1 to 0.5m. The mean error of separating pears and less than 0.27m for aggregating pears. This demonstrated the system's capability to accurately position and differentiate between individual pears and clusters in challenging orchard environments.

1 INTRODUCTION

Modern fruit harvesting is still predominantly conducted manually throughout the different regions of the world. Among common fruits, the Japanese pear (such as Pyrus pyrifolia Nakai) is one of the most widely grown fruit in Japan(Saito, 2016).because of the shortage of labor in harvesting season, the cost of pear picking has gradually increased. With the development of computer science in agriculture, modern agricultural technology has gradually evolved from manual planting and picking to full automation and intelligence.The recognition and positioning of pears in complex orchards becomes a prerequisite for the development of fruit picking robots. Overtime, most countries in the world have developed intelligent picking robots through different methods and techniques to load and unload agricultural products and detect fruit and positioning issues (Bechar and Vigneault, 2016).The recognition and positioning of a certain number of pears in a complex orchard becomes the focus of this research.

However, due to the complexity of the orchard environment, the precise recognition and localization of each pear by the robot in order to improve the robustness of the robot to the environment has become a major research challenge. Deep learning with convolutional neural networks was widely used for image processing tasks, allowed detection of objects wherever positioned in an image and extracted complex visual concepts(Koirala et al., 2019). Especially in agricultural field, the detection and classification of different fruits were applied based on CNN (Zhang et al., 2019). In the realm of traditional image processing algorithms in deep learning algorithms, pears within orchards can be effectively recognized. These algorithms leverage multiple vision sensors, enabling the detection of fruits with high accuracy (Sa et al., 2016). This paper chose three typical CNN-based deep learning algorithms to recognize the pears in complex orchard.

Due to the different focuses of the three algorithms, Faster R-CNN focuses on an increase in speed to recognize the objects, while Mask R-CNN favors

Pan, S., Yorozu, A., Ohya, A. and Ahamed, T.
Recognition and Position Estimation of Pears in Complex Orchards Using Stereo Camera and Deep Learning Algorithm.
DOI: 10.5220/0012179200003543
In *Proceedings of the 20th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2023) - Volume 1*, pages 632-639
ISBN: 978-989-758-670-5; ISSN: 2184-2809

the separation of individual pears. And YOLACT (Bolya et al., 2019) provides an increase in speed with the separation of individual pears. By employing the Faster R-CNN (Girshick, 2015), which is a two-stage object detection method and only generates the bounding boxes in recognition, cameras precisely recognize individual pears even when the pears are densely clustered together. This aids in facilitating the subsequent picking of the recognized pears. YOLACT is a real-time instance segmentation algorithm which was employed for the recognition of pears by robots. And YOLACT is functioned as a one-stage method, swiftly generates bounding boxes and masks for the rapid recognition of pears. Mask R-CNN (He et al., 2017), as another two-stage instance segmentation method uses intra-station segmentation, it over-detects different individuals of the same species, so that overlapping parts of fruits are detected accurately and shape variations are adjusted to improve recognition accuracy.

Furthermore, the position estimation of the orchard pears is indispensable, the distance from the pears to the camera provides the reference coordinates for fruit picking robot to grab. ZED2 stereo camera provides a platform can be for the position estimation of the recognized pears. However, the irregular aggregation of pears adds difficulty to the position estimation. The camera needs to acquire the precise bounding box coordinates of the recognized pears in the complex orchard, This allows for the calculation of the coordinates of the centroid, which in turn helps determine the distance between this point and the left lens of the ZED2 stereo camera.

Therefore, to enhance the robustness in unstable environments of pear recognition and position estimation in complex orchards, a more accurate method for pear recognition and position is developed for fruit-picking robots. This helps avoid misgrasping of pears by robots and reduce the reliance on labor for agricultural operations. In this paper, three different deep learning algorithms are compared to assess the accuracy of pear recognition, aim to select one algorithm with the lowest recognition error and the highest accuracy of generated bounding box and mask for pears in complex orchards. This chosen algorithm can be combined with the ZED2 stereo camera for accurate position estimation.

2 METHODOLOGY

2.1 System Overview

An overview of the proposed framework was shown in (Figure.1). This paper chose the moving robot named SCOUT MINI developed by AGILEX ROBOTICS to be equipped with ZED2 stereo camera. and the angle of mechanical grip was simulated to recognize pears in complex orchards to measure the distances of pears. This paper was divided into two parts, the first part tested the recognition performance of different deep learning algorithms in separating pears and aggregating pears in complex orchards, and selected the deep learning algorithm with the highest mean average precision for pear recognition. The second part evaluated the distance error of the ZED2 camera for separating and aggregating pears already identified by the first part under different light intensities.

2.2 Data Preparation

A stereo camera named ZED2 (Stereolabs Inc. San Francisco, CA, USA) was utilized to capture 3018 original images from the T-PIRC ($36°07'04''N$,$140°05'45''E$), and measured distance from the pear was less than 0.5m. Considering the influences of different light conditions effected the results, the data were collected at 9:00-10:00 am and 6:00–7:00 pm at Tsukuba-Plant Innovation Research Center (T-PIRC). Among these datasets, there were 1818 images used for training, 900 images for validation, and 300 images for testing with the proportion of 6:3:1.(Table.1)

Table 1: Dataset collection times and light conditions.

Date	Time	Light Condition
24 August 2021	9:00-10:00	High Light
24 August 2021	18:00-19:00	Low Light

Pears and leaves exhibited similar shapes under low light conditions, the dataset underwent augmentation through the inclusion of inverted and rotated images. This manipulation resulted in pears appearing spherical from various angles, in contrast to the distinct shapes exhibited by leaves. Consequently, the dataset was expanded to comprise a training set of 5454 images, a validation set of 2700 images, and a test set of 900 images.(Table.2)

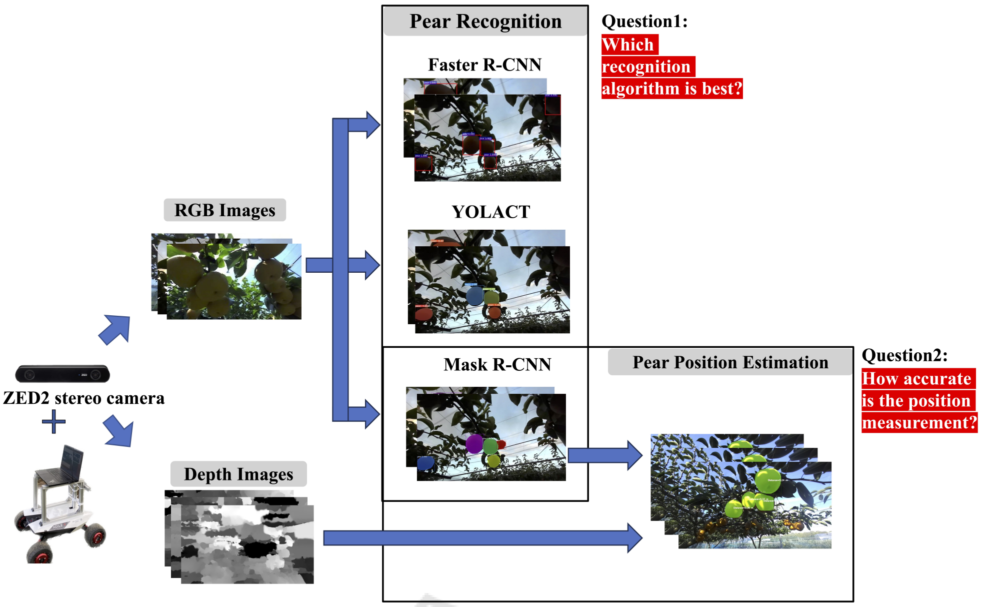


Figure 1: The overview of pear recognition and position estimation in complex orchard.

Table 2: The Amount of dataset in different set.

Images	training	validation	testing
Original	1818	900	300
Augmentation	3636	1800	600
Total	5454	2700	900

2.3 Pear Recognition with Deep Learning Algorithms

Faster R-CNN is composed of two modules used by VGG-16 backbone network: a Region Proposal Network(RPN) (Girshick, 2015), used for detection of RoIs in the images followed by 2) a classification module, which classifies the individual regions and regresses a bounding box around the objects. (Bargoti and Underwood, 2017). YOLACT as a real-time instance segmentation model, which not only performs target detection but also recognizes individual targets under each identified category, is mainly implemented through two parallel networks for strength segmentation (Bolya et al., 2019). And Mask R-CNN is an instance segmentation method, which is efficiently detected objects in images while generating high-quality segmentation masks for each instance (He et al., 2017). Mask R-CNN extended the Faster R-CNN by mask branch at the end of the model (Girshick et al., 2015). And ROI-Align is different from ROI-Pooling (Girshick, 2015) in Fasetr R-CNN, which cancels the quantization and used bilinear interpolation (Kirkland and Kirkland, 2010) to obtain the image values on pixel points with floating-point coordinates (Figure.2). We compared the mean average presion (mAP) of Mask R-CNN, Faster R-CNN and Yolact for pear recognition, and we chose Mask R-CNN as the subsequent recognition method used for pear position estimation.

2.4 Pear Position Estimation

The ZED2 stereo camera has been applied to target reconstruction, position acquisition, and other fields (Tran et al., 2020), it simulated and emulated the imaging principle of the human eyes, which perceives differences (depth) between images formed from the right and left eyes (Ortiz et al., 2018). To generate the depth image, stereo camera utilizes two RGB cameras to capture images of the same scene from different positions. The 3D position is calculated through triangulation based on corresponding points found on both images (Condotta et al., 2020).

Mask R-CNN was used to generate precise bounding boxes and masks, by adjusting the median around the bounding box to achieve a relatively precise coordinate value, the spatial location of the pears was identified.The left images were acquired by the left lens, showed the detection and measurement information in the real situation. And the right depth images were acquired by the parallax between the left lens and the right lens, showed the depth information. Typically, darker colors (black) in the depth images indicated more distant objects, while brighter colors (white) indicated closer objects (Figure.3).

The spatial correspondence between the pixel

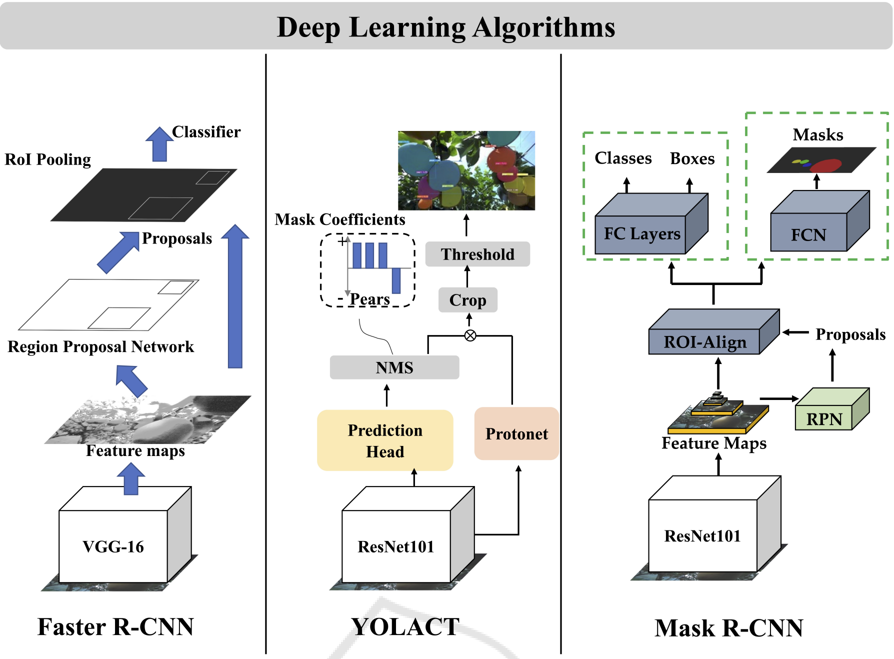


Figure 2: Pear Recognition using Different Deep Learning Algorithms.

plane of the pear and the camera is shown in (Figure.4). $P(x,y,z)$ was the spatial coordinates of the pear centroid. By using the more accurate bounding boxes generated by Mask R-CNN, the pixel coordinates $P_m(x_m,y_m)$ of the centroid of the recognised pear were calculated using the coordinates of the upper left and lower right corners $P_0(x,y)$ of the bounding boxes, with the left lens of the ZED2 camera as the origin $O(0,0,0)$, and the centroid pixel coordinates of the RGB image were matched with the depth image obtained from the parallax. In the depth image, a 2×2 pixel block (ROI) was extracted with the centroid as the center, the median value of the depth value in the pixel block was calculated to z_m, and finally the depth value was converted to the 3D coordinates $P(x,y,z)$ of the center point of the pear.

3 RESULTS AND DISCUSSION

3.1 Training Details

The loss function was used to measure the gap between the model predictions and the actual labels.

$$\begin{aligned} L &= L_{RPN} + L_{MASK} \\ L_{RPN} &= L_{CLS} + L_{BOX} \end{aligned} \tag{1}$$

The L defined as training loss, and it included two parts, which were defined as the loss of RPN networks L_{RPN} and the mask branches L_{MASK}, and define L_{MASK} as the average binary cross-entropy loss (He et al., 2017). The L_{CLS} and L_{BOX} were defined as the classification loss and bounding box loss in RPN (Girshick, 2015). From the performance of the training results in different learning rates, when the learning rate was set to 0.001, the training loss (L) of models dropped to 0.3099 and the validation set loss dropped to 0.4637 in Mask R-CNN. We also compared the different loss trends of the three deep learning algorithms for detecting pears in training set and validation set. The overall loss of Faster R-CNN and YOLACT both fell below 0.2 after 40,000 training steps. The loss curves demonstrated the applicability of the three models to actual-world situations (Figure.5).

3.2 Evaluation of Model Metrics

In this paper, the Precision (P), Recall (R), Average Precision (AP), and mean Average Precision (mAP) were employed as the primary parameters to evaluate the performance of different models. The weights obtained from the training set after 80 epochs were used to test and compared the performance on both the test set and validation set, with an Intersection over Union (IoU) threshold of 50%(Table.3).

With the IoU threshold of 50%, we compared the overlap between the predicted bounding box and segmentation mask with ground truth of bounding box and mask. A prediction was classified as a true positive (TP) if the overlap exceeded 0.5. Conversely, a

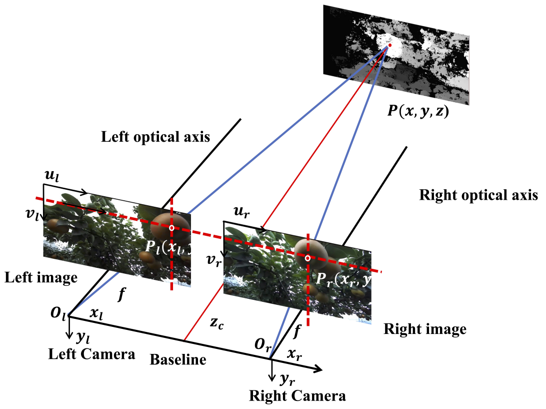


Figure 3: Depth images acquisition of ZED2 stereo camera.

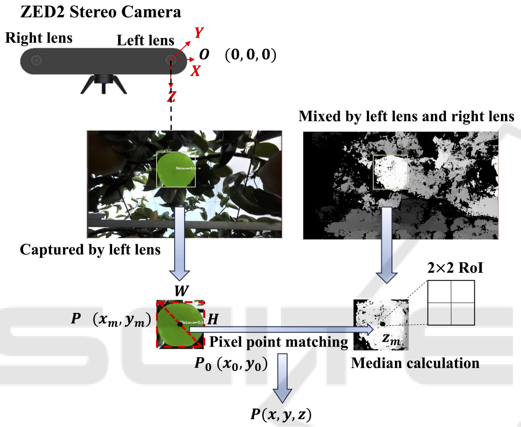


Figure 4: Depth matching with deep learning algorithms.

Table 3: mAP(IoU=50%) results from 3D camera datasets using Mask R-CNN, Faster R-CNN and YOLACT in the testing set and validation set.

Model	Validation Set	Testing set
Faster R-CNN	87.90%	87.52%
YOLACT	87.07%	97.89%
Mask R-CNN	95.22%	99.45%

false positive (FP) was assigned when the predicted category diverged from the actual category. Furthermore, a true negative (TN) was assigned when a category not identified as a pear by the model. A false negative (FN) was applied when the actual pears went undetected (Missing box and mask) (Figure6).

Precision over-affected the proportion of correct classification in the number of positive samples classified by the model. Recall was the ratio of the number of correct samples to the number of positive samples. Its expression was

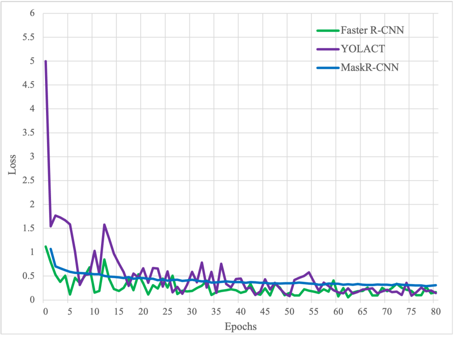


Figure 5: Total losses of three deep learning models.

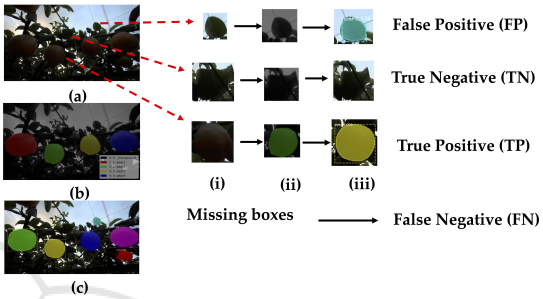


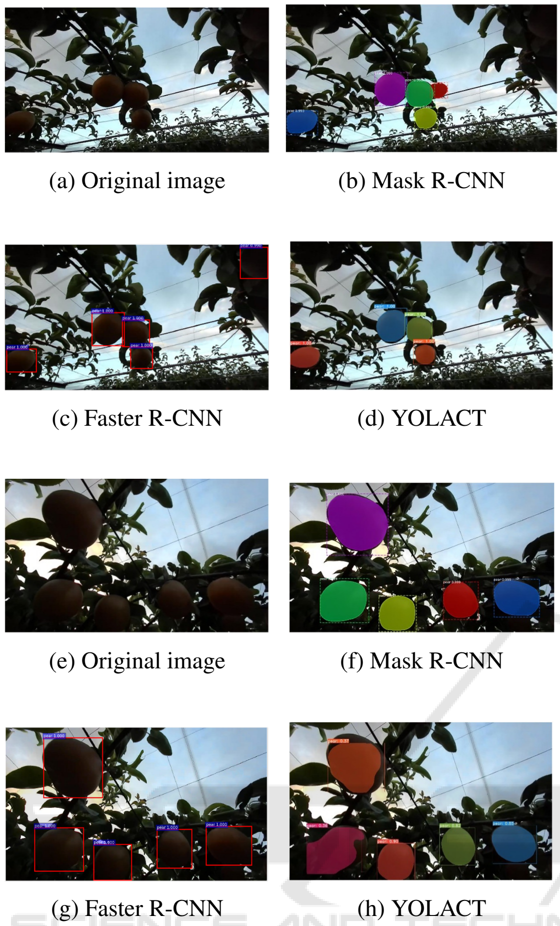
Figure 6: FP, TN, TP, and FN in pear recognition.

$$
\begin{aligned}
P &= \frac{TP}{TP+FP} \\
R &= \frac{TP}{TP+FN}
\end{aligned}
\tag{2}
$$

3.3 Evaluation of Model Effectiveness

By creating a dataset and deep learning models using a 3D stereo camera, we found the best weights by comparing the fitting effect of Mask R-CNN, Faster R-CNN and YOLACT at the same learning rate of lr=0.001. Since the tested orchard was a semi-enclosed structure, there was an indoor-like area structure in the orchard and outdoor-like structure that has direct exposure to sunlight. This paper compared the different effects of separating pears and aggregating pears under different lighting (strong and low light) by testing 900 images of the test set taken at different periods(Figure.7) (Figure.8).

This paper undertook a comparative analysis of the Mean Average Precision (mAP) scores achieved by the three deep learning algorithms on both the validation and test datasets. Mask R-CNN attained an impressive mAP of 95.22% on the validation set and further exceled with a remarkable score of 99.45% on the test set. In comparison, Faster R-CNN, another two-stage algorithm akin to Mask R-CNN, marginally trailed behind with a validation set mAP of 87.90%



(a) Original image (b) Mask R-CNN

(c) Faster R-CNN (d) YOLACT

(e) Original image (f) Mask R-CNN

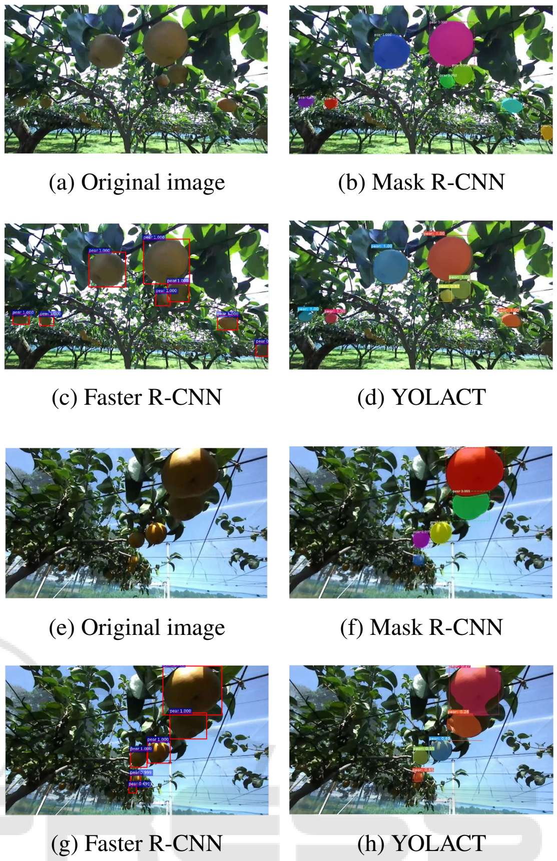
(g) Faster R-CNN (h) YOLACT

Figure 7: Results in low light situation.(a-d):Aggregating pears,(e-h)Separating pears.

and a corresponding test set mAP of 87.52% . Meanwhile, YOLACT, despite being an instance segmentation algorithm akin to Mask R-CNN, achieved a validation set mAP of 87.07% and a commendable test set mAP of 97.89%.

This study categorized conditions into two distinct factors: light intensity and pear aggregation. Both Mask R-CNN and YOLACT generated masks and bounding boxes, whereas Faster R-CNN exclusively generated bounding boxes.

Regarding light intensity, it was divided into high and low light. Mask R-CNN outperformed Faster R-CNN and YOLACT in generating bounding boxes under various light conditions. For mask generation, although both Mask R-CNN and YOLACT were instance segmentation algorithms, Mask R-CNN, which was two-stage method was notably superior to YOLACT, and YOLACT encountered situations where the mask area exceeds the predicted bounding box or the predicted bounding box area is smaller than that of the pears.



(a) Original image (b) Mask R-CNN

(c) Faster R-CNN (d) YOLACT

(e) Original image (f) Mask R-CNN

(g) Faster R-CNN (h) YOLACT

Figure 8: Results in high light situation.(a-d):Aggregating pears,(e-h)Separating pears.

In terms of aggregation of pears, it was categorized into separating pears and aggregating pears. In separating pears, Mask R-CNN accurately recognized pears in different light intensity, while Faster R-CNN generated the missing bounding boxes, and YOLACT exhibited errors in mask area and boxes generation. In aggregating pears, Mask R-CNN outperformed accurately in generating bounding boxes and masks. However, Faster R-CNN misidentified some leaves as pears under low light conditions, while YOLACT not only misidentified but also generated unstable masks.

3.4 Estimation of Pear Positioning Using Mask R-CNN

This paper contrasted the performance of the same dataset using three deep learning algorithms. Mask R-CNN excelled in producing bounding boxes with higher accuracy. Therefore, we intended to calculate the distance from the already recognized pears to the left lens of camera using mask R-CNN and compared errors of the distance measurement in different cases.

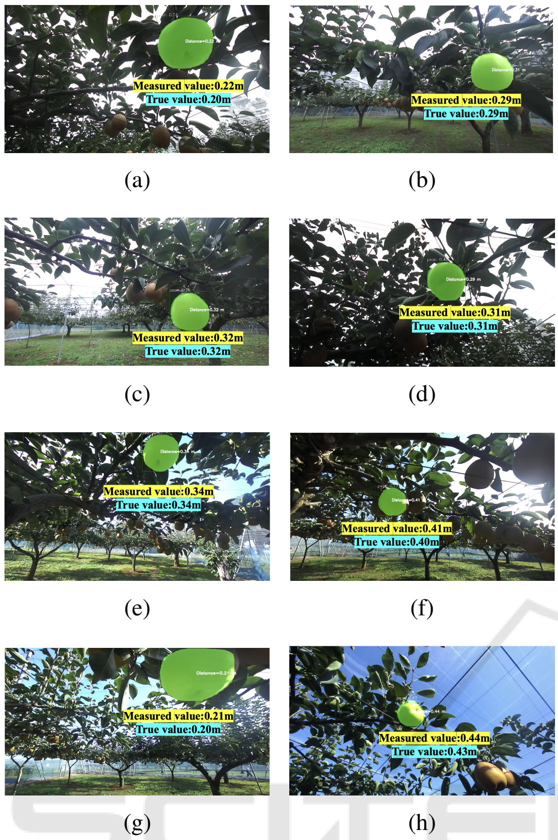


Figure 9: Results of separating pears distance measurement in low and high light.

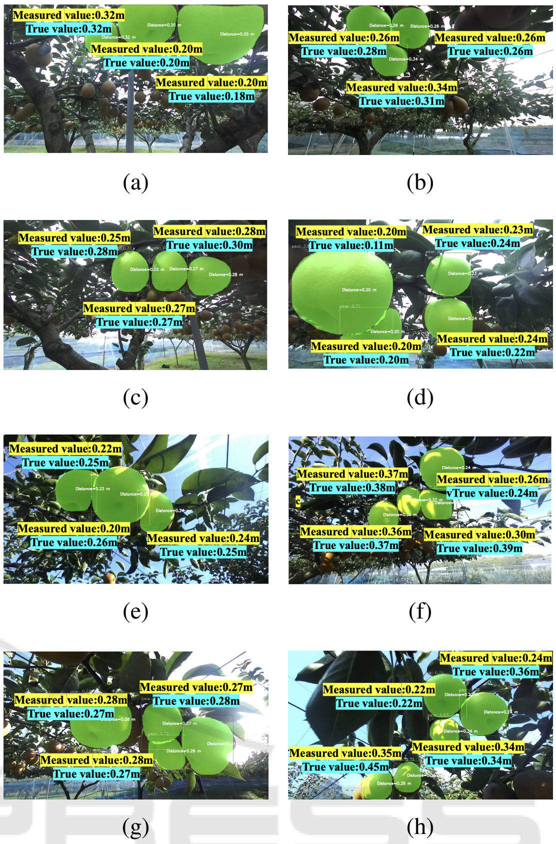


Figure 10: Results of aggregating pears distance measurement in low and high light.

The fact that ZED2 cameras only identified and distance measure information between the two lens when they were matched.This paper found that within 0.1m-0.5m,due to the random of pear growth, different conditions of pears showed different errors in distance measurement. In this paper, we estimated the distance error from the left lens to the pears with different situations (separating and aggregating).(Figure.9) (Figure.10). Moreover, we compared the error means of measured distance and true distance $\bar{x}_n$, standard deviations σ_n and the goodness of fit R^2 between measured values and true values of separating and aggregating pears under low and high light. These metrics were used to evaluate the accuracy of the distance measurements by ZED2 stereo camera (Table.4).

In the range of 0.1-0.5m, the incomplete data of bounding boxes and masks resulting from recognition errors were rounded off. The total $\bar{x}_n$ for separating pears was 0.017m, while for aggregating pears, which involved multiple identified targets, it was slightly less accurate with $\bar{x}_n$ of 0.018m in generating bounding boxes and masks. The standard deviation was used to estimate the degree of dispersion of the measurement errors. For separated pears, the total σ_n was 0.015m, indicating a relatively stable error in measuring them. The camera measured aggregating pears significantly higher, with a value of 0.021m, suggesting that the degree of pear aggregation affected the measurement errors. The R^2 was used to evaluate how closely the measured distances of pears aligned with the true values under different conditions. For separating pears, the R^2 reached 0.896, indicating a tendency for the measured and true values to be similar. However, for aggregating pears, after discarding some larger errors associated with bounding boxes and mask generation, the R^2 reached to 0.832, suggesting a lower level of agreement with the true values.

The results also demonstrated that pears along the edge of the camera exhibited significant errors; This study discarded samples which have significant errors in measuring when calculating the R^2. This was due to the inherent distortion of the ZED2 stereo camera, the incomplete bounding boxes, and the resulting masks of Mask R-CNN, where pears immediately adjacent to the camera showed completely incorrect measurements when measured by the ZED2 camera.

Table 4: The distance errors estimation of recognised pears in different situations.

Pear condition	Separating pears			Aggregating pears		
Light condition	Low	High	Overall	Low	High	Overall
$\bar{x}_n$ (m)	0.013	0.020	0.017	0.012	0.023	0.018
σ_n (m)	0.016	0.013	0.015	0.011	0.028	0.021
R^2	0.834	0.884	0.896	0.848	0.812	0.832

4 CONCLUSIONS

In this paper, we proposed a method to achieve accurate recognition and position estimation in complex orchard environments to reduce the grasping errors caused by problems such as branch occlusion and pear aggregation, which improved the robustness of robots working in the complex orchard. Also, we compared the performance of different deep learning algorithms for the recognition of separating and aggregating pears under different light intensities. The results showed that Mask R-CNN outperforms Faster R-CNN and YOLACT in terms of recognition accuracy for separating and aggregating pears under both high and low light conditions. In further experiments, we chose Mask R-CNN as the recognition algorithm for pear position estimation and compared the error mean $\bar{x}_n$, standard deviation σ_n, and goodness-of-fit R^2 of separating and aggregating pears at a distance of 0.1-0.5 m. The results showed that $\bar{x}$ and σ_n were significantly higher for aggregated pears than for separated pears in the same cases, and R^2 reached more than 0.8 in different cases. Therefore, this paper exhibited commendable efficacy in the precise recognition and position of pears within the range of 0.1-0.5 meters. This outcome substantially bolsters the precise recognition and position estimation of pears by agricultural fruit-picking robots.

ACKNOWLEDGEMENTS

The authors would like to thank the Tsukuba Plant Innovation Research Center (T-PIRC), University of Tsukuba, for providing facilities for conducting this research in its orchards.Also,This work was supported by JST SPRING, Grant Number JPMJS2124.

REFERENCES

Bargoti, S. and Underwood, J. (2017). Deep fruit detection in orchards. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3626–3633. IEEE.

Bechar, A. and Vigneault, C. (2016). Agricultural robots for field operations: Concepts and components. *Biosystems Engineering*, 149:94–111.

Bolya, D., Zhou, C., Xiao, F., and Lee, Y. J. (2019). Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166.

Condotta, I. C., Brown-Brandl, T. M., Pitla, S. K., Stinn, J. P., and Silva-Miranda, K. O. (2020). Evaluation of low-cost depth cameras for agricultural applications. *Computers and Electronics in Agriculture*, 173:105394.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Kirkland, E. J. and Kirkland, E. J. (2010). Bilinear interpolation. *Advanced Computing in Electron Microscopy*, pages 261–263.

Koirala, A., Walsh, K. B., Wang, Z., and McCarthy, C. (2019). Deep learning–method overview and review of use for fruit detection and yield estimation. *Computers and electronics in agriculture*, 162:219–234.

Ortiz, L. E., Cabrera, E. V., and Gonçalves, L. M. (2018). Depth data error modeling of the zed 3d vision sensor from stereolabs. *ELCVIA: electronic letters on computer vision and image analysis*, 17(1):0001–15.

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., and McCool, C. (2016). Deepfruits: A fruit detection system using deep neural networks. *sensors*, 16(8):1222.

Saito, T. (2016). Advances in japanese pear breeding in japan. *Breeding Science*, 66(1):46–59.

Tran, T. M., Ta, K. D., Hoang, M., Nguyen, T. V., Nguyen, N. D., and Pham, G. N. (2020). A study on determination of simple objects volume using zed stereo camera based on 3d-points and segmentation images. *International Journal*, 8(5).

Zhang, Y.-D., Dong, Z., Chen, X., Jia, W., Du, S., Muhammad, K., and Wang, S.-H. (2019). Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation. *Multimedia Tools and Applications*, 78:3613–3632.