Sensors and Materials, Vol. 35, No. 10 (2023) 3777–3790
MYU Tokyo

S & M 3450

Locating Open-field Broccoli Plants with Unmanned Aerial Vehicle Photogrammetry and Object Detection Algorithm: A Practical Prediction Approach

Hiroki Hayashi,[1] Hiroto Shimazaki,[1,4*] Ryoji Korei,[2] and Kazuo Oki[3,4]

[1]National Institute of Technology, Kisarazu College,
2-11-1 Kiyomidai-higashi, Kisarazu-shi, Chiba 299-0041, Japan
[2]AIR WATER Co., Ltd., 2-12-8 Minamisenba, Chuo-ku, Osaka-shi, Osaka 542-0081, Japan
[3]Kyoto University of Advanced Science, 18 Yamanouchi-gotanda-cho, Ukyo-ku, Kyoto-shi, Kyoto 615-8577, Japan
[4]Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan

(Received February 23, 2023; accepted August 16, 2023)

Keywords: unmanned aerial vehicle, object detection, YOLOv5, broccoli, precision farming

We developed a practical approach to locate individual open-field broccoli plants with a position error of less than 5 cm, using the georeferenced high-resolution orthomosaic imagery generated through the unmanned aerial vehicle-based photogrammetry and the YOLOv5 object detection model. The feasibility of our method was evaluated on the basis of two angles: the cost of preparing training data and the accuracy of object detection. The orthomosaic imagery was generated for two plots: Plot A, which experienced large variations in plant growth due to drought-induced mortality and replanting, and Plot B, which showed small variations under normal growing conditions. On the basis of the result of analysis under our recommended settings for the training data, we found that (1) the detection accuracy improved with an increase in the amount of training data in both Plots A and B; (2) in Plot A, 95% of a total of 21277 plants were detected using training data for approximately 630 plants selected to represent the individual differences in growth; and (3) in Plot B, 98% of all 7836 plants were detected using training data for approximately 126 plants selected randomly. Our findings can guide the optimal balance between the cost of training data preparation and the desired accuracy level of object detection in precision crop management, particularly for broccoli production.

1. Introduction

The attention towards smart agriculture technology has increased in recent years as a means of improving productivity and efficiency in crop management.[1,2] The main goal of smart agriculture is to optimize the use of resources such as water, fertilizer, and pesticides and to predict the best time and yield for harvest by using data on weather conditions, field environment, and plant growth.[2] This data-driven approach leads to the effective management of crops grown in open fields.[2] Although attempts have been made to optimize and predict crop

*Corresponding author: e-mail: shimazaki@c.kisarazu.ac.jp
https://doi.org/10.18494/SAM4364

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

management, they have primarily been focused on the field or plot level within a field, rather than the individual plant level.(3,4) To achieve more precise crop management, it is necessary to focus on optimization and prediction at the individual plant level, utilizing accurate data on the location of each plant.

Determining the locations of individual plants in an open field can be accomplished by using the georeferenced high-resolution orthomosaic imagery generated by photogrammetry with unmanned aerial vehicles (UAVs).(5) This imagery allows for the estimation of individual plant locations through both human image recognition and more efficient computer-based image processing techniques. However, the visual inspection is not a practical method for completing the task of precisely locating all plants in a field, because it is labor-intensive and time-consuming. Conventional image processing techniques, such as template matching,(6) Hough transform,(7) and image segmentation,(8) have been used to recognize individual plants on the basis of their morphological characteristics, but their versatility is limited and their effectiveness is largely impacted by the shape and size of the target.

Machine learning and deep learning have also been utilized to efficiently detect plant locations from imagery. Machine learning algorithms, such as Random Forest and Support Vector Machine, have been utilized to distinguish between plants and background soil on the basis of spectral and geometric features.(9–12) These algorithms have high classification performance, but local radiometric distortions in orthomosaic imagery can hinder their success. Deep learning algorithms designed for general object detection, such as YOLO, Faster regional convolutional neural network (R-CNN), and Mask R-CNN, have gained popularity in identifying individual plants as they are more robust to image quality variations.(13,14) However, training these algorithms requires a large amount of high-quality training data, creating a trade-off between the cost of training data preparation and the performance of object detection.

We estimated the locations of individual broccoli plants (*Brassica oleracea var. italica*) grown in open fields using the georeferenced high-resolution orthomosaic imagery generated through UAV-based photogrammetry and the YOLOv5 object detection algorithm. The practicality of our method was evaluated from two angles: the cost of generating training data and the detection accuracy. Our results will offer guidelines for finding the optimal balance between the cost of training data generation and the desired accuracy level of object detection in precision crop management, particularly for broccoli production.

2. Materials and Methods

2.1 Orthomosaic imagery for broccoli test plots

Georeferenced high-resolution orthomosaic imagery was generated for two broccoli test plots, Plots A and B, to examine the impact of the difference in plant growth level on object detection performance. Plot A was grown under unfavorable conditions, causing several plants to die in two weeks after planting owing to drought and requiring replanting. This resulted in significant variations in growth among individual plants. In contrast, Plot B was grown under normal conditions, resulting in relatively small differences in growth among individual plants. The shapes and dimensions of the two plots are depicted in Fig. 1.

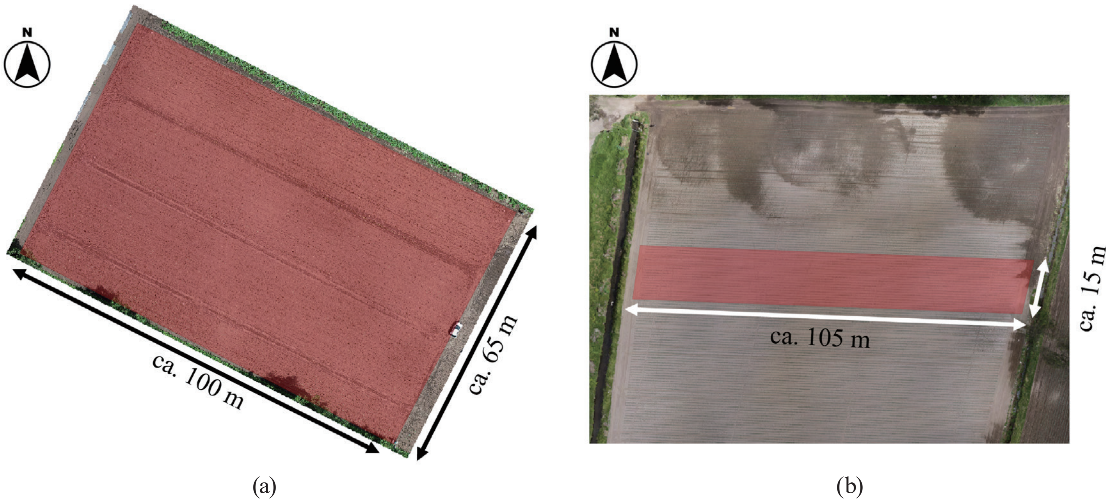

Fig. 1. (Color online) Shapes and dimensions of broccoli test plots: (a) Plots A and (b) B. Red areas, overlaid on the georeferenced high-resolution orthomosaic imagery, denote the extent of Plots A and B. The numbers of individual plants identified were 21277 in Plot A and 7836 in Plot B.

The orthomosaic imagery for two broccoli test plots was generated through UAV-based photogrammetry with a DJI Phantom 4 RTK (P4RTK) and Agisoft Metashape software version 1.8.4. Aerial photography was conducted 2 and 4 weeks after planting broccoli in Plots A and B, respectively. The P4RTK was flown at a height of 25 m above the ground, capturing multiple aerial images with an RGB camera mounted on it. The camera's shutter speed was set to 1/1000 s, and the shooting direction was set at different off-nadir angles of 0 and 30° to reduce the systematic distortion known as the "doming distortion,"[15] which can occur in the Structure-from-Motion and Multi-View-Stereo (SfM-MVS) process. The aerial images had a spatial resolution of approximately 7 mm and an end-lap and side-lap rate of 80%. These multiple aerial images were merged into a georeferenced orthomosaic imagery for each test plot using the SfM-MVS function in Metashape. The resulting orthomosaic imagery had a spatial resolution of approximately 7 mm for each plot.

As described in the following subsections, the georeferenced high-resolution orthomosaic imagery, generated using the aforementioned method, was used to estimate the locations of broccoli. It is important to note that the decision to not utilize individual aerial images captured without overlaps for estimating broccoli locations was based on two reasons. The primary reason was to accurately estimate the geographic coordinates of the aerial image center and the shooting range, while the second reason was to reduce the cost of preparing training data for the object detection algorithm. The details of these reasons are described below.

In general, four types of information are necessary to determine the geographic coordinates of the aerial image center and the shooting range: (1) the position of the lens center during aerial photography; (2) the camera tilt angles during aerial photography; (3) optical parameters within the camera, including focal length, principal point displacement, and lens distortion; and (4) the surface height of the terrain and objects in the target area.

The P4RTK employs the real time kinematic (RTK) method to measure the 3D position of the lens center during aerial photography. The resulting data is recorded in the metadata of each aerial image. However, the P4RTK lacks the capability to measure or maintain the camera tilt angles during aerial photography. As a result, the camera tilt during image capture, affected by variations in aircraft flight speed and direction, as well as changes in wind speed and direction, remains unknown. Moreover, information regarding the camera's internal optical parameters and the surface height of the terrain and objects in the target area is also unavailable. To overcome these limitations, we estimated the aforementioned factors (1) to (4) by utilizing multiple aerial images captured with a high overlap ratio and employing the SfM-MVS method. It is important to note that the lens-center position measured by the RTK method at the time of aerial photography was used as an initial value for accurately estimating factors (1) to (4).

While it is possible to determine the geographic coordinates of the center position of the aerial image and the shooting range by estimating factors (1) to (4), each aerial image utilizes frame central projection. As a result, the terrain and objects in the image may appear to lean outward from the nadir point based on their height. This distortion caused by central projection can introduce variations in the appearance of broccoli, even when the variety and growing conditions are the same. Moreover, the central projection can affect the background appearance of broccoli depending on the relative positions of the camera and the target object, and it can also lead to the overlapping of adjacent broccoli, reducing visibility. These changes in appearance, background, and visibility significantly impact the performance of object detection algorithms.

On the other hand, the orthomosaic imagery is a composite of each aerial image that has been converted from the center projection to an orthographic projection, resulting in every point being represented as if viewed from directly above. Consequently, when utilizing orthomosaic imagery to estimate the location of individual broccoli, there is no longer a need to consider the impact of changes in appearance, background, and visibility. This, in turn, is expected to reduce the cost of preparing training data for object detection algorithms. Moreover, the use of georeferenced orthomosaic imagery simplifies the conversion of detected positions of individual broccoli in the image into geographic coordinates, thereby streamlining subsequent data processing for crop management.

2.2 Background soil and weeds

Aerial images of the open field captured not only the broccoli plants to be detected but also backgrounds such as soil and weeds. When utilizing an object detection algorithm to identify broccoli in an orthorectified mosaic imagery generated from a collection of these aerial images, variations in background conditions can impact the detection performance. The following two reasons outline why background conditions affect the detection performance:

(1) When there is a low contrast between the broccoli plants and the background, or when the color or texture of broccoli is similar to the background, the performance of the object detection model is expected to decrease. This problem is particularly likely to occur when the leaves and stems of broccoli being detected have a similar color to the background soil or when the background contains numerous weeds with a similar appearance to broccoli.

(2) If the background includes distinctive colors, textures, or patterns, these factors can affect the performance of the object detection model. It is important to ensure that the object detection algorithm does not incorrectly learn distinctive colors and textures caused by wet or coarse

soil clods, as well as identifiable patterns such as footprints or dirt ruts, as crucial features for detecting broccoli.

The soil color in both Plots A and B was observed to be either light or dark brown (Fig. 1), creating a distinct contrast with the green color of healthy broccoli plants. Furthermore, the aerial images of Plots and Plot B were captured 2 and 4 weeks, respectively, after planting broccoli in a carefully maintained field with rows, where weeds were effectively controlled. As a result, there were relatively fewer weeds present in these test plots, aside from the targeted broccoli plants. Consequently, concerns regarding potential detection errors associated with the factors mentioned in (1) above were considered to be minimal.

On the other hand, careful consideration was required for the background features and patterns mentioned in (2) above. In the planting areas of Plots A and B, the footprints and vehicle ruts left by farmers created distinct patterns in the orthomosaic imagery (Figs. 2 and 3). If these features and patterns were included in the annotations used to train the object detection algorithm, the model might learn not only the features of the broccoli plants but also those of the background soil. To address this issue, we investigated the optimal size of the annotations.

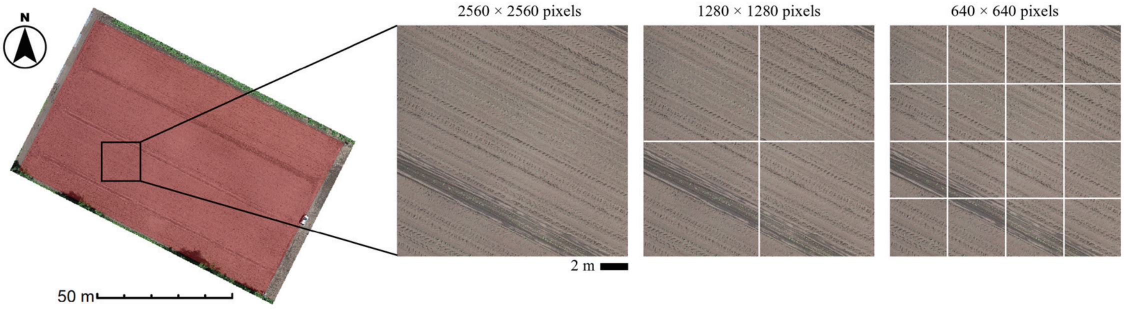

Fig. 2. (Color online) Image chips cropped from the georeferenced high-resolution orthomosaic imagery.

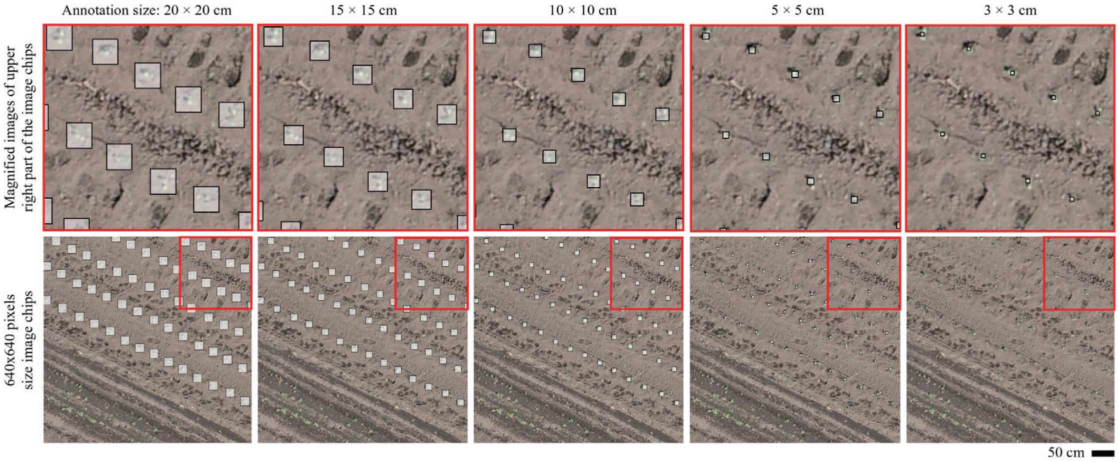
Annotation size: 20 × 20 cm
15 × 15 cm
10 × 10 cm
5 × 5 cm
3 × 3 cm
Magnified images of upper right part of the image chips
640x640 pixels size image chips
50 cm

Fig. 3. (Color online) An example of annotation bounding boxes with different sizes (black outline), overlayed on the image chip with a size of 640 × 640 pixels, which was cropped from the orthomosaic imagery of Plot A.

2.3 Ground reference data

Ground reference data for the locations of individual broccoli plants in Plots A and B were generated through the visual inspection of the georeferenced high-resolution orthomosaic imagery. Specifically, the center of each plant body was precisely identified, and its geographic coordinates were determined from the orthomosaic imagery. In total, we identified 21277 individual plants in Plot A and 7836 in Plot B by this method.

2.4 Object detection algorithm

The YOLOv5 object detection algorithm was utilized to estimate the locations of individual broccoli plants in the georeferenced orthomosaic imagery. YOLO stands for “You Only Look Once” and is a state-of-the-art, real-time object detection algorithm.(16) Unlike R-CNN and the single-shot multibox detector (SSD), which repurpose classifiers for detection, YOLO frames object detection as a regression problem that predicts spatially separated bounding boxes and associated class probabilities. With a single neural network, it predicts these values directly from full images in one evaluation. This end-to-end single network architecture allows for optimization directly on detection performance, leading to YOLO’s reported superior speed and accuracy in comparison with other object detection algorithms such as R-CNN and SSD.(17)

YOLOv5 that was released in June 2020 is one of the versions of the YOLO series. It has four training models: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, which vary in terms of detection accuracy and computational speed.(18) However, there has been evidence suggesting that there is no significant difference in detection performance for a single class of objects among the four models.(17,19) In light of this, in this study, which only focused on detecting broccoli, we used the fastest YOLOv5s model to train the model and perform object detection.

2.5 Training data preparation

The performance of object detection models and the cost of training them are closely related. In general, the higher the performance of a model, the higher the cost of training the model. To determine the optimal balance between the cost of generating training data and the desired accuracy level of object detection, we assessed the impact of four factors on object detection accuracy: (1) the size of image chips used for annotation when cropped from the orthomosaic imagery; (2) the method employed to crop the image chips from the orthomosaic imagery; (3) the size of the annotations surrounding individual target plants captured in an image chip; and (4) the total number of annotations. Initially, we explored the optimal settings for factors (1) to (3). Subsequently, while maintaining the optimal settings for factors (1) to (3), we examined the impact of factor (4) on object detection accuracy. The specific settings for each factor are described below.

2.5.1 Size of image chips for training

The orthomosaic imagery was too large in terms of file size to serve as an image chip in the YOLOv5 training process. Additionally, during the training, the image chips are automatically resized to 640 × 640 pixels, which could potentially impact the detection performance of the trained model by compromising important features for object detection. To assess the effect of resizing the image chip on detection performance, image chips were extracted from the orthomosaic imagery at different sizes: 2560 × 2560, 1280 × 1280, and 640 × 640 pixels. The image chip with a size of 2560 × 2560 pixels was first cropped from the orthomosaic imagery, then divided into four parts to create image chips with a size of 1280 × 1280 pixels, and finally divided into 16 parts to create image chips with a size of 640 × 640 pixels (Fig. 2).

2.5.2 Method of cropping image chips for training

Since substantial variations in plant growth were observed in Plot A, the detection performance of the trained model may vary depending on whether the image chips were cropped to reflect these variations or cropped randomly, even if the size of the image chips was the same. To address this issue, we compared two cropping methods in Plot A: “stratified cropping”, which considered the growth variations by selecting appropriate cropping positions, and “random cropping”, which selected cropping positions randomly. In Plot B, where only small variations in plant growth were observed, only random cropping was used to extract image chips.

2.5.3 Annotation size

Annotation data was prepared for each image chip to be used in the YOLOv5 training process. The annotation data consisted of three elements: (1) the object class, represented by an arbitrary integer that distinguished the object to be detected from others; (2) the object coordinates, represented by the pixel coordinates of the center of the bounding box that surrounded each individual object to be detected in the image chip; and (3) the height and width of the bounding box, which indicated the size of the annotation.

The object class was fixed at 1 since the focus was solely on detecting broccoli. For each individual broccoli plant captured in an image chip, the object coordinates were calculated using the pre-prepared ground reference data. To assess the effect of annotation size on detection performance, we varied the height and width of the bounding box, setting them at different square sizes: 20 × 20, 15 × 15, 10 × 10, 5 × 5, and 3 × 3 cm for Plot A (Fig. 3), and 30 × 30, 25 × 25, 20 × 20, 15 × 15, and 10 × 10 cm for Plot B. The differences in the sizes of the bounding boxes between Plots A and B were due to variations in the growth levels of the plants.

2.5.4 Total number of annotations

To investigate the impact of different annotation quantities on detection performance, we varied the total number of annotations for the YOLOv5 training process by adjusting the number of image chips. As previously mentioned, image chips were extracted from orthomosaic imagery using stratified or random cropping in Plot A and only random cropping in Plot B. The number of broccoli plants captured in an image chip can vary depending on the image chip size and cropping position from which the image chip is extracted, regardless of the cropping method used. Therefore, we increased the number of annotations by gradually adding image chips with the same size while changing their combinations.

2.6 Model training and prediction

To determine the optimal balance between the cost of generating training data and the desired level of accuracy, we trained object detection models using the training data prepared under different settings. Subsequently, we used the trained models to detect individual plants in the orthomosaic imagery and evaluated their detection accuracy on the basis of the metrics explained in the next subsection. The experiment settings used for preparing the training data, which consisted of a set of image chips and their corresponding annotations, are summarized in Table 1.

Since it is impractical to train the model using all possible combinations of training data settings, we investigated the impact of different settings in three steps. In the first step, we assessed the effects of different image chip and annotation sizes on detection accuracy. We used the random cropping method for extracting image chips and approximately 1000 annotations, based on the orthomosaic imagery of Plot A. An annotation quantity of approximately 1000 corresponds to the number of individual plants captured in an image chip with a size of 2560 × 2560 pixels, a set of four image chips with a size of 1280 × 1280 pixels, and a set of 16 image chips with a size of 640 × 640 pixels.

In the second step, we assessed the effect of annotation size on detection accuracy under the optimal setting for image chip size. We used the orthomosaic imagery of Plots A and B, keeping the cropping method of image chips as random cropping with an annotation quantity of approximately 1000 plants. In the final step, we assessed the impact of different image chip cropping methods and annotation quantities on detection accuracy, while keeping the optimal settings for image chip size and annotation size. We used the orthomosaic imagery of Plots A and B for this evaluation.

Table 1
Experiment settings used for preparing training data.

Test plot	Image chip		Annotation	
	Size (pixels)	Cropping method	Size (cm^2)	Quantity
Plot A	2560 × 2560 1280 × 1280 640 × 640	Random cropping Stratified cropping	20 × 20 15 × 15 10 × 10 5 × 5 3 × 3	Ca. 60–1800 plants
Plot B	640 × 640	Random cropping	30 × 30 25 × 25 20 × 20 15 × 15 10 × 10	Ca. 60–1000 plants

Note that the trained model was used to detect individual plants within the georeferenced high-resolution orthomosaic imagery. However, the size of the orthomosaic imagery made it unsuitable for use in the YOLOv5 prediction process. Therefore, we created a set of image chips with the same size as the training image chips to use for prediction. The prediction image chips were obtained by cropping the orthomosaic imagery with 50% overlaps between adjacent image chips, ensuring that all individual plants were fully captured in a set of image chips.

To account for potential variations in detection accuracy caused by differences in the cropping position of the training image chips extracted from the orthomosaic imagery, we conducted the model training and prediction process 20 times. In each repetition, we utilized different sets of image chips cropped from various positions in the orthomosaic imagery, along with their corresponding annotations, while maintaining the same training data settings.

2.7 Model evaluation

The center coordinates of individual plants in the ground reference data were used as the correct positions, and those of the bounding boxes predicted with the trained model were used as the predicted positions. Cases where the distance between the correct and predicted positions was less than 5 cm were classified as true positive (TP), whereas those where the distance was greater than 5 cm were classified as false positive (FP). Cases where the object detection method failed to generate a predicted position within 5 cm of the correct position were classified as false negative (FN). As there can be an infinite number of true negative (TN) cases, where the trained model did not generate a predicted location outside of a 5 cm radius from the correct location, we did not classify them.

The performance of the trained model was evaluated on the basis of commonly used metrics in object detection, such as precision, recall, and F1 score.(20) Precision measures the proportion of correct detections [Eq. (1)]. A high precision indicates that the model is making few FP detections. Recall measures the proportion of actual objects that are correctly detected by the model [Eq. (2)]. A high recall indicates that the model is making few missed detections. The F1 score is the harmonic mean of precision and recall [Eq. (3)], providing a single metric that balances both precision and recall. A high F1 score indicates good overall performance of the model.

As explained in the previous subsection, we repeated the model training and prediction process 20 times using different sets of image chips and corresponding annotations, all of which were obtained by cropping from various positions of the orthomosaic imagery under the same training data settings. We considered models with median scores from the 20 repetitions exceeding 0.95 for all three metrics to have practical detection performance.

$$Precision = \frac{Total\,number\,of\;objects\,correctly\,detected}{Total\,number\,of\;objects\,detected} = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{Total\,number\,of\;objects\,correctly\,detected}{Total\,number\,of\;actual\,objects} = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

3. Results and Discussion

The locations of individual broccoli plants (*Brassica oleracea var. italica*) in open fields were predicted with an error of less than 5 cm, using the georeferenced high-resolution orthomosaic imagery generated through UAV-based photogrammetry and the YOLOv5 object detection algorithm. The bounding boxes detected in the image chips sizes of 2560 × 2560, 1280 × 1280, and 640 × 640 pixels are shown in Fig. 4.

Detection accuracy in terms of precision, recall, and F1 score varied among the models trained under the different training data settings. On the basis of the results of the analysis using the orthomosaic imagery of Plot A, it was found that the most stable detection performance was achieved when using the image chip with a size of 640 × 640 pixels [Fig. 5 (a)]. When the image chip size was 2560 × 2560 pixels and the annotation sizes were 3, 5, and 10 cm, the precision, recall, and F1 scores were all below 0.3. The same was true when the image chip size was 1280 × 1280 pixels and the annotation size was 3 cm. This could be due to the resizing effect of the image chips, where larger chip sizes, such as 2560 × 2560 or 1280 × 1280 pixels, compromise important features for object detection surrounded by smaller annotations. Therefore, the optimal image chip size was determined to be 640 × 640 pixels.

Even with image chips optimized at 640 × 640 pixels, the detection accuracy varied depending on the annotation size in Plots A and B [Figs. 5(a) and 5(b)]. Practical detection performance was achieved with an annotation size of 5 cm for Plot A and annotation sizes of 10 and 15 cm for Plot B. The diameters of the plants in Plot A ranged from a maximum of 10 cm to a minimum of 3 cm, with an average of 5 cm. In contrast, the diameter of the plants in Plot B was generally 10 cm. On the basis of these findings, we recommend matching the annotation size to the size of the individual plants to be detected to achieve a higher detection performance.

If the annotation size is set larger than the plant body, the object detection model will learn not only the features of the plant body but also the background soil pattern. This has both advantages and disadvantages: FPs may be detected if the soil patterns are similar even in areas

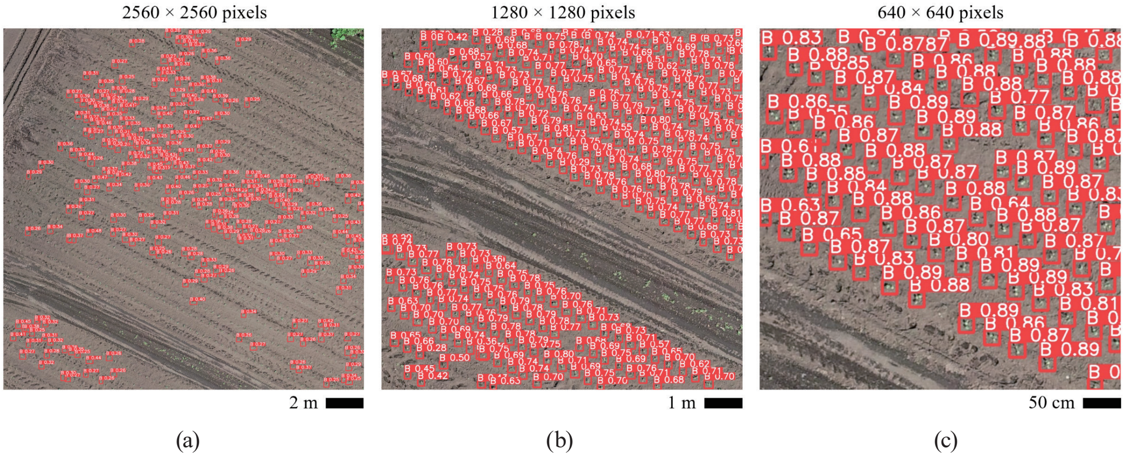

Fig. 4. (Color online) The bounding boxes predicted with the trained object detection model were overlaid on the image chips cropped from the orthomosaic imagery in Plot A, with sizes of (a) 2560 × 2560, (b) 1280 × 1280, and (c) 640 × 640 pixels.

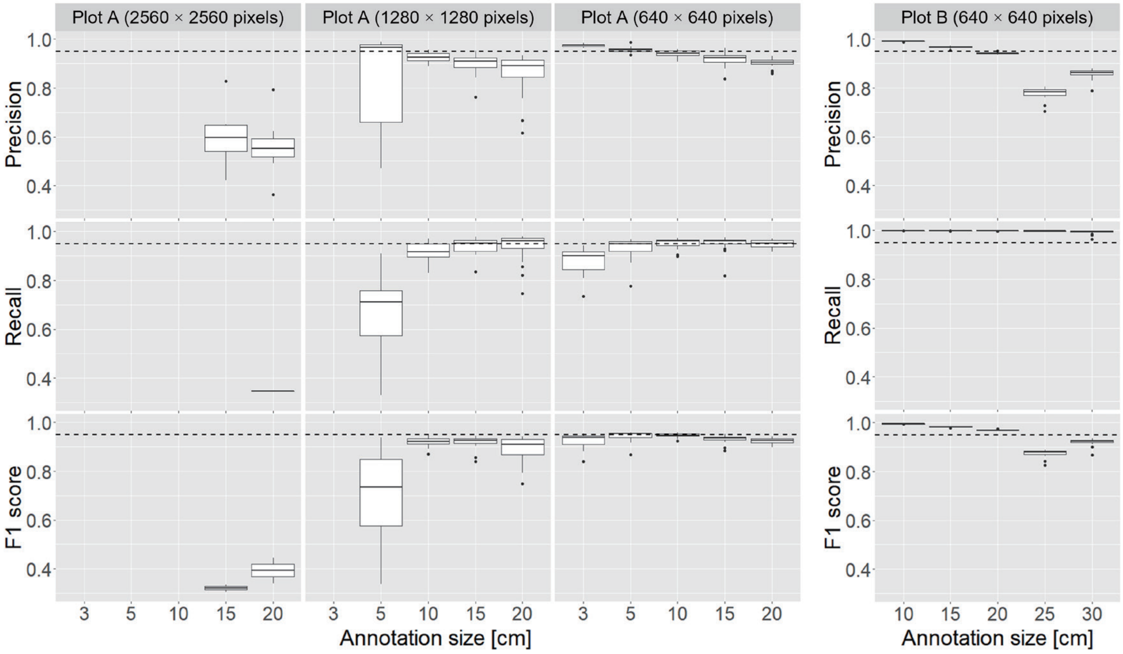

Fig. 5. Effects of different image chip and annotation sizes on detection accuracy. Image chips were extracted using the random cropping method from the orthomosaic imagery in both Plots A and B, and each set of image chips was annotated with approximately 1000 plants. The dashed lines indicate the practical detection performance level of 0.95, as defined in this study.

where individual plants are not present, while TPs may be based on the soil pattern even when individual plants are not clearly visible in the image. Conversely, if the annotation size is set smaller than the plant body, it is likely to detect local features of individual plants, resulting in duplicated detections of the same individual or making it difficult to distinguish them from other non-target plants, which can increase the FP rate.

As outlined in Sect. 2.2, the performance of the object detection algorithm can be affected by backgrounds such as soil and weeds, in addition to the appearance of the target being detected. For instance, when there is a low contrast between broccoli and background, or when the color or texture of broccoli closely resembles the background, it is expected that the performance of the object detection model will decrease. However, this issue is unlikely to arise in a typical open-field broccoli growing environment. In both Plots A and B, there was a clear contrast between broccoli and the background, and there were minimal similarities in color and texture between broccoli and the background. Moreover, by aligning the annotation size with the size of the detection target, as explained previously, the impact of distinct colors, textures, or patterns present in the backgrounds, caused by wet or coarse soil clods, footprints, or dirt ruts, on the overall detection performance could be effectively reduced. Therefore, our method remains applicable even when there are distinct colors, textures, or patterns in the background.

The impact of different image chip cropping methods and annotation quantities on detection accuracy, while maintaining the optimal settings for image chip size and annotation size, is shown in Fig. 6. Increasing the number of training image chips improved the detection accuracy in both Plots A and B. In Plot A, practical detection performance was achieved with more than 10 image chips extracted by stratified cropping or more than 11 image chips extracted by random cropping. Stratified cropping, which accounts for growth differences, achieved practical performance with slightly less training data. In Plot B, where there were few differences in plant growth, the model trained with two or more image chips extracted by random cropping achieved a detection performance exceeding 0.98 for all three metrics.

Considering that a single 640 × 640 size image chip contained approximately 63 individual plants, our findings mentioned in the previous paragraph can be summarized as follows: (1) detection accuracy improved with an increase in the amount of training data in both Plots A and B; (2) in Plot A, 95% of a total of 21277 plants were detected using training data from approximately 630 plants selected to represent the individual differences in growth; and (3) in

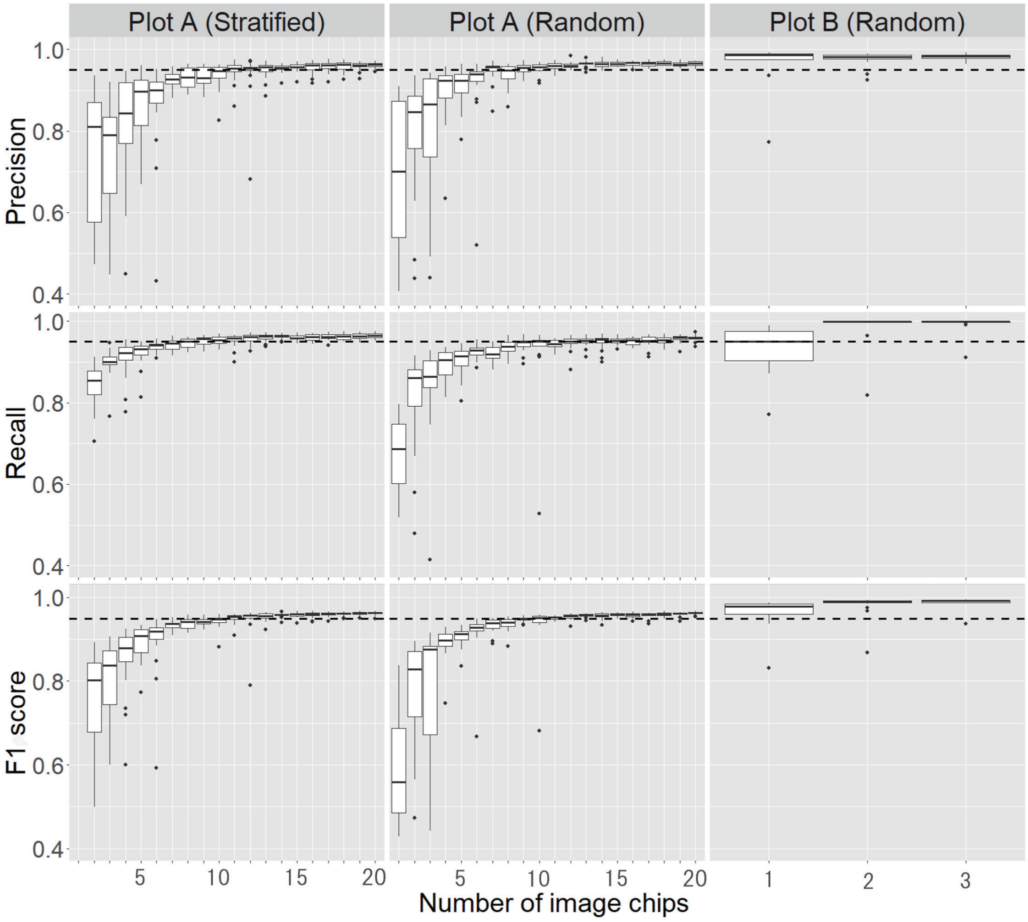

Fig. 6. Impact of different image chip cropping methods and annotation quantities on detection accuracy. Image chip size was fixed at an optimum of 640 × 640 pixels in both Plots A and B. Annotation size was set optimally at 5 cm for Plot A and 10 cm for Plot B. The dashed lines indicate the practical detection performance level of 0.95, as defined in this study.

Plot B, 98% of all 7836 plants were detected using training data from approximately 126 plants selected randomly. These results can help find the optimal balance between the cost of creating training data and the desired accuracy level of object detection in precision crop management, especially for broccoli production.

4. Conclusions

We developed a practical approach to locate individual open-field broccoli plants with a position error of less than 5 cm, using the georeferenced high-resolution orthomosaic imagery generated through UAV-based photogrammetry and the YOLOv5 object detection model. The feasibility of our method was evaluated on the basis of two angles: the cost of preparing training data and the accuracy of object detection. The orthomosaic imagery with a spatial resolution of approximately 7 mm was generated for two broccoli test plots: Plot A, which experienced large variations in plant growth due to drought-induced mortality and replanting, and Plot B, which showed small variations under normal growing conditions.

To determine the optimal balance between the cost of generating training data and the desired accuracy level of object detection, we assessed the impact of four factors on object detection accuracy: (1) the size of the image chips used for annotation when cropped from the orthomosaic imagery; (2) the method used to crop the image chips from the orthomosaic imagery; (3) the size of the annotations surrounding individual plants captured in an image chip; and (4) the total number of annotations. The recommended settings for each factor are as follows: (1) the size of the image chips used for annotation should be 640 × 640 pixels; (2) stratified cropping should be used for the plot with large growth variation, while random cropping should be used for the plot with small growth variation; and (3) the size of the annotations should match the size of the individual plants to be detected.

On the basis of the results of analysis under the recommended settings for the first three factors, we found the following regarding the fourth factor: (1) detection accuracy improved with an increase in the amount of training data in both Plots A and B; (2) in Plot A, training data for approximately 630 plants selected to represent the individual differences in growth led to the detection of 95% of the total 21277 plants; and (3) in Plot B, training data for approximately 126 plants selected randomly led to the detection of 98% of all 7836 plants. These results suggest that our approach for locating individual open-field broccoli plants is practical in terms of the cost of preparing training data and detection accuracy.

Acknowledgments

This work was financially supported by the collaborative research framework of "Technology for IoT Sensing and Analysis" established between The University of Tokyo and Air Water Inc.

References

1 Y. Inoue: Soil Sci. **66** (2020) 798. https://doi.org/10.1080/00380768.2020.1738899

2 Institute of Vegetable and Floriculture Science, NARO: https://www.naro.go.jp/english/laboratory/nivfs/divopenfieldprodres/fieldprodmansystemgroup/index.html (accessed December 2022).

3 D. Kim, H. S. Yun, S. J. Jeong, Y. S. Kwon, S. G. Kim, W. S. Lee, and H. J. Kim: Remote Sens. **10** (2018) 563. https://doi.org/10.3390/rs10040563

4 R. Barzin, R. Pathak, H. Lotfi, J. Varco, and G. C. Bora: Remote Sens. **12** (2020) 2392. https://doi.org/10.3390/rs12152392

5 Y. Chen, J. Ribera, C. Boomsma, and E. Delp: Proc. 2017 IEEE Int. Conf. Computer Vision Workshops (IEEE ICCVW, 2017) 2030. https://doi.org/10.1109/ICCVW.2017.238

6 A. Aeberli, K. Johansen, A. Robson, D. W. Lamb, and S. Phinn: Remote Sens. **13** (2021) 2123. https://doi.org/10.3390/rs13112123

7 D. Flores, I. G. Hernandez, R. Lozano, J. M. Nicolas, and J. L. Hernandez: Drones **5** (2021) 4. https://doi.org/10.3390/drones5010004

8 J. Aval, J. Demuync, E. Zenou, S. Fabre, D. Sheeren, M. Fauvel, K. Adeline, and X. Briottet: ISPRS J. Photogramm. Remote Sens. **146** (2018) 97. https://doi.org/10.1016/j.isprsjprs.2018.09.016

9 K. Johansen, Q. Duan, Y. H. Tu, C. Searle, D. Wu, S. Phinn, A. Robson, and M. F. Mccade: ISPRS J. Photogramm. Remote Sens. **165** (2020) 28. https://doi.org/10.1016/j.isprsjprs.2020.04.017

10 J. Campbell, P. E. Dennison, J. W. Tune, S. A. Kannenberg, K. L. Kerr, B. F. Codding, and W. R. Anderegg: Remote Sens. Environ. **245** (2020) 111853. https://doi.org/10.1016/j.rse.2020.111853

11 A. Castro, J. T. Sanchez, J. M. Pena, F. M. Brenes, O. Csillik, and F. L. Granados: Remote Sens. **10** (2018) 285. https://doi.org/10.3390/rs10020285

12 M. Dalponte, H. O. Orka, L. T. Ene, T. Gobakken, and E. Nasset: Remote Sens. Environ. **140** (2014) 306. https://doi.org/10.1016/j.rse.2013.09.006

13 K. Li, G. Wan, L. Meng, and J. Han: ISPRS J. Photogramm. Remote Sens. **159** (2020) 296. https://doi.org/10.1016/j.isprsjprs.2019.11.023

14 J. Zheng, H. Fu, W. Li, W. Wu, L. Yu, S. Yuan, W. Y. Tao, T. K. Pang, and K. D. Kanniah: ISPRS J. Photogramm. Remote Sens. **173** (2021) 95. https://doi.org/10.1016/j.isprsjprs.2021.01.008

15 M. R. James and S. Robson: Earth Surf. Processes Landforms **39** (2014) 1413. https://doi.org/10.1002/esp.3609

16 R. Joseph, S. Divvala, R. Girshick, and A. Farhadi: Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition (IEEE CVPR, 2016) 779. https://doi.org/10.1109/CVPR.2016.91

17 T. Jintasuttisak, E. Edirisinghe, and A. Elbattay: Comput. Electron. Agric. **192** (2022) 106560. https://doi.org/10.1016/j.compag.2021.106560

18 G. Jocher: yolov5 (2020). https://github.com/ultralytics/yolov5

19 H. Tian, X. Fang, Y. Lan, C. Ma, H. Huang, X. Lu, D. Zhao, H. Liu, and Y. Zhang: Remote Sens. **14** (2022) 4208. https://doi.org/10.3390/rs14174208

20 A. A. Taha and A. Hanbury: BMC Med. Imaging. **15** (2015) 1. https://doi.org/10.1186/s12880-015-0068-x