Scale-Adaptive Small Object Detection Algorithm for Complex Agricultural Environments: A Case Study on Bees (Postprint)

Authors: Guo Xiuming, Zhu Yeping, Li Shijuan, Zhang Jie, Lü Chunyang, Liu Shengping

Date: 2023-02-17T00:00:00+00:00

Abstract

Objects targeted for recognition in agricultural production environments are often characterized by dense distribution, small size, and high density. Coupled with variable illumination and complex backgrounds in farmland environments, existing object detection models fail to achieve satisfactory performance. This study aims to improve the recognition performance of small targets and proposes a scale-adaptive small target recognition algorithm for complex agricultural environments, using bee recognition as a case study. The algorithm overcomes the influence of complex and variable background environments and the difficulty in feature extraction caused by small target sizes, achieving scale-independent small target recognition. First, the original image is split into smaller-sized sub-images to increase the target scale, and the annotated targets are assigned to the split sub-images to form a new dataset. Then, transfer learning is employed to retrain and generate a new target recognition model. During model usage, to ensure proper restoration of sub-image recognition results, the split sub-images must have a certain overlap ratio. All sub-image target recognition results are collected, and Non-Maximum Suppression (NMS) is used to remove redundant bounding boxes generated by the model itself. An Intersection over Small NMS (IOS-NMS) is proposed to further eliminate redundant boxes in the overlapping regions of sub-images. Verification experiments were conducted with sub-image pixel sizes of 300$\times$300, 500$\times$500, *and*700$\times$$700, and sub-image overlap ratios of 0.2 and 0.05. The results show that using SSD (Single Shot MultiBox Detector) as the object detection model in the framework, the proposed scale-adaptive algorithm achieved higher recall and precision than the SSD model, with maximum improvements of 3.8% and 2.6%, respectively, and also showed certain improvements over the original-scale YOLOv3 model. To further verify the superiority of the algorithm in small target recognition against complex backgrounds, bee images in complex farmland environments with different scales and

scenes were crawled from the Internet, and comparative tests were conducted using the proposed algorithm and the SSD model. The results demonstrate that the algorithm can improve target recognition performance and possesses strong scale adaptability and generalization capability. Since the algorithm requires multiple forward inferences for a single image, its timeliness is not high, making it unsuitable for edge computing.

Full Text

Scale Adaptive Small Objects Detection Method in Complex Agricultural Environment: Taking Bees as Research Object

GUO Xiuming, ZHU Yeping, LI Shijuan, ZHANG Jie, LYU Chunyang, LIU Shengping*

(Agricultural Information Institute, Chinese Academy of Agricultural Sciences/Key Laboratory of Agri-information Service Technology, Ministry of Agriculture and Rural Affairs, Beijing 100081, China)

Abstract: Objects in agricultural production environments often exhibit characteristics of small size, high density, and clustered distribution. Combined with variable lighting and complex backgrounds in farmland settings, existing object detection models fail to achieve satisfactory performance. This study aims to improve small target recognition performance, using bee detection as a case study, and proposes a scale-adaptive small object recognition algorithm for complex agricultural environments. The algorithm overcomes the influence of complex and variable background environments and addresses the difficulty of feature extraction caused by small target size, enabling scale-independent small object recognition. First, the original image is divided into smaller sub-images to increase target scale, and annotated objects are assigned to these sub-images to form a new dataset. Transfer learning is then employed to retrain and generate a new object recognition model. During model deployment, adjacent sub-images must overlap to ensure proper restoration of detection results. After collecting results from all sub-images, standard non-maximum suppression (NMS) removes redundant boxes generated by the model itself, while a novel Intersection over Small NMS (IOS-NMS) is proposed to further eliminate redundant boxes in overlapping regions between sub-images. Validation experiments were conducted with sub-image pixel dimensions of 300$\times$300, 500$\times$500, *and*700$\times$$700, and overlap rates of 0.2 and 0.05. Results show that when using Single Shot MultiBox Detector (SSD) as the detection framework, the proposed scale-adaptive algorithm achieves higher recall and precision than the original SSD model, with maximum improvements of 3.8% and 2.6%, respectively, and also outperforms the original-scale model. To further verify the algorithm' s superiority for small target recognition in complex backgrounds, additional bee images from farmland environments with varying scales and scenes were crawled from the internet for comparative testing against the YOLOv3 model. Results demonstrate that the

proposed algorithm enhances recognition performance with strong scale adaptability and generalization capability. However, due to the requirement for multiple forward inferences per image, the algorithm suffers from low time efficiency and is unsuitable for edge computing applications.

Keywords: object detection; machine vision; small object; agricultural environment; bee; Single Shot MultiBox Detector; YOLOv3

1 Introduction

With the development of convolutional neural networks and deep learning technologies [1], machine vision-based object detection has attracted widespread attention and achieved breakthrough progress [2,3]. Numerous scenarios in agriculture involve target recognition and counting, where machine vision technology can enhance agricultural intelligence and modernization levels. Agricultural production environments are typically outdoor settings with variable lighting and complex backgrounds, where target objects are generally small in volume and high in density. Recognizing and detecting small targets in such complex environments represents a common application scenario, including small objects in agricultural remote sensing images, fruits on trees, and bees in hives. Addressing specific application requirements in agriculture and developing specialized algorithms to achieve superior performance on particular metrics represents a research trend in agricultural intelligent recognition for the coming years.

Small target detection has long been a challenge in object detection due to limited effective pixels and small scale, resulting in poor feature representation capability. Many researchers have designed and optimized detection models from various perspectives to improve small target detection performance. Some have optimized and improved backbone network structures [5-11] to extract richer features, some have optimized anchor boxes [12-17] to improve target localization accuracy, and others have optimized loss functions [18-20] to enhance training efficiency and model performance. While these improvements can enhance small target recognition to some extent, the fundamental cause of poor performance remains the limited number of pixels and small scale. Increasing the effective pixel count and scale of small targets constitutes the primary approach for improving recognition performance. Most existing methods focus on model optimization rather than addressing target scale (the ratio of target pixels to total image pixels) directly. This study targets the essential cause of poor small target recognition—insufficient effective pixels and small scale—and employs image splitting to effectively increase target scale and improve recognition performance.

Bees are small in size and appear at small scales in images, often clustering together, making them typical small targets for agricultural recognition and counting. Taking bees at hive entrances as an example, this study proposes a small target recognition algorithm based on image splitting that is independent

of input image size and target scale. The original input image is first divided into multiple sub-images with overlapping regions between adjacent sub-images. These sub-images serve as model inputs, and their outputs are aggregated. A two-stage non-maximum suppression (NMS) method then removes redundant boxes generated by both the model itself and sub-image overlaps. To evaluate algorithm performance, validation experiments were conducted using this algorithm alongside SSD and YOLOv3 (You Only Look Once) models. Additionally, bee images of various scales and backgrounds were crawled from the internet for comparative testing against the SSD model to assess the algorithm' s scale adaptability and generalization capability.

2.1 Algorithm Framework Overview

Deep learning-based object detection algorithms consist of three main components: preprocessing, feature extraction, and post-processing (Figure 1 [Figure 1: see original paper]). Traditional algorithms use the entire image as network input. To enhance recognition performance for difficult small targets, the proposed algorithm splits the input image into multiple sub-images to increase target scale and pixel count. Post-processing primarily employs NMS to remove redundant candidate boxes output by the convolutional neural network and identify optimal target locations, thereby improving detection accuracy. NMS is a crucial step in deep learning-based object detection. The original NMS algorithm [23] sorts all candidate boxes by confidence score, selects the highest-scoring box, and removes all boxes with overlap exceeding a predefined threshold. Overlap is measured using Intersection over Union (IOU)—the ratio of intersection area to union area between two boxes. Various improved NMS algorithms such as Soft-NMS [24] and A-NMS [25] have been proposed for different application scenarios.

The proposed algorithm faces redundant detections from both the deep learning network model and overlapping image regions. To address the latter, an Intersection over Small NMS (IOS-NMS) method is introduced for more accurate target localization. Figure 1 compares the framework of the proposed algorithm with traditional deep learning-based object detection methods.

2.2 New Dataset Generation Method

Data were collected in June 2020 at the Agricultural Information Institute of the Chinese Academy of Agricultural Sciences during peak bee activity hours. The hive entrance—the boundary between the hive and external environment—offers unobstructed lighting and active bee movement. A camera with $1280\$\times720 pixel resolution was positioned directly above the hive entrance, capturing images from 8:00 AM to 6:00 PM at 45-second intervals. The dataset covers multiple time periods (morning, noon, evening) and u

pixels—3.57 times larger than the average bee pixel scale—resulting in inaccurate localization and degraded recognition performance.

To increase effective pixels and scale for small targets, a grid-based splitting method divides the original image into sub-images. The number of sub-images and their dimensions, along with the overlap rate between adjacent sub-images, are key parameters. The newly generated sub-image collection forms a new dataset for model training (Figure 3 [Figure 3: see original paper]). Sub-image dimensions relate to the model's normalized input size, target scale, and original image resolution. To avoid imbalanced positive-negative sample ratios and improve training efficiency, sub-images without targets are removed, while those containing targets are added to the new dataset. Since bee annotations were made on original images, annotation information must be recalculated for sub-images. The algorithm proceeds as follows:

1. Let the original dataset be A, with any original image a A.
2. Let the width and height of a be w and h, respectively, and let O be the set of objects in a containing location and category information. Define sub-image width as zw and height as zh.
3. Divide the original image horizontally at zw intervals and vertically at zh intervals, creating w/zw × h/zh sub-images, with edge regions padded using solid color.
4. For object set O, perform reallocation and coordinate recalculation, where o is an element of O.
5. Extract sub-images containing targets and add them to new dataset B.

During sub-image splitting, original targets must be reallocated and their coordinates adjusted within sub-images. Figure 4 [Figure 4: see original paper] illustrates the target reallocation process. If a target lies entirely within one sub-image, it is assigned to that sub-image. If a target spans two adjacent sub-images (bees A and B in Figure 4), the proportion of the smaller portion is calculated. If this proportion is below a set threshold, the smaller portion is discarded, retaining only the larger portion (bee B). If above the threshold, both portions are retained and assigned to their respective sub-images (bee A) with recalculated coordinates. For targets divided into four parts, the same proportion-based criterion determines retention, and new coordinates are computed. The algorithm flowchart for target reallocation and coordinate recalculation is shown in Figure 5 [Figure 5: see original paper].

2.3 Model Training and Deployment

Since the new dataset differs from the original only in pixel scale while target features and backgrounds remain unchanged, a model trained on the original dataset has already learned many target features that remain highly similar after scale adjustment. Therefore, transfer learning is employed to continue

training the original-scale model on the new dataset, accelerating convergence and reducing training time.

The complete deployment workflow is illustrated in Figure 6 [Figure 6: see original paper]. Since the new model targets images with larger object scales, the original image must also be split into multiple sub-images during inference. To ensure accurate recognition of targets at sub-image boundaries, adjacent sub-images overlap at a rate related to target pixel scale—similar to target scale is sufficient, as excessively large overlap rates would create too many sub-images and reduce algorithm efficiency.

Each sub-image is fed into the new model to obtain its target set. Target coordinates are then restored to the original image based on their sub-image positions. After collecting results from all sub-images, standard NMS removes redundant boxes generated by the model itself (Figure 7(a)). However, overlapping sub-images may cause duplicate detections of the same target with nested inner and outer boxes (marked as A in Figure 7(a)). This occurs because standard NMS uses Intersection over Union (IOU) (Figure 8 [Figure 8: see original paper]) as the localization metric (Equation (1)). When two bounding boxes differ significantly in size and their intersection occupies a large proportion of the smaller box, the IOU value falls below the threshold, preventing standard NMS from removing such redundant boxes. To eliminate these incomplete target detection redundancies at nested locations, Intersection over Small (IOS) (Equation (2)) is proposed as a similarity metric, and IOS-NMS is developed to remove internal redundant boxes. Figure 7(b) shows detection results after IOS-NMS processing.

3.1 Experimental Design

To validate algorithm performance, bee recognition in apiaries was used as a test case. The hardware environment consisted of an Intel Core i7-6700k CPU with a GeForce GTX Titan X GPU, running Ubuntu OS, with the PyTorch deep learning framework.

Manual annotation of 2,613 collected images created the original bee image dataset, with mean bee scale (ratio of bee pixels to total image pixels) of 0.0037. Using the splitting method from Section 2.2, a new dataset was created with sub-image dimensions of 360$×$320 pixels, yielding 6,269 images with mean bee scale of approximately 0.028. SSD and YOLOv3 deep learning models were selected as detection frameworks. Original-scale models were first trained on the original dataset, then fine-tuned via transfer learning on the new dataset to obtain new-scale SSD and new-scale YOLOv3 models—the scale-adaptive models. The same validation dataset was used for comparative analysis, with identical confidence thresholds for detection and original-scale models.

3.2 Performance Validation and Analysis

To analyze the impact of sub-image size and overlap rate, validation experiments were conducted using three sub-image sizes (300$×300, 500×500, *and*700×$700 pixels) and two overlap rates (0.2 and 0.05). Model performance was evaluated using precision, recall, and average single-image processing time. Results are presented in Table 1 .

Table 1 Comparison results for the three detection models

Model	Recall (%)	Precision (%)	Average Time per Image (s)
Original-scale SSD	94.6	87.3	0.035
Original-scale YOLOv3	95.1	88.5	0.042
Scale-adaptive (300$×300, $os =$ 0.2)|98.4|89.9|0.235||$Scale-adaptive$(300×300, $os =$ 0.05)|98.0|88.3|0.198||$Scale-adaptive$(500×500, $os =$ 0.2)|98.4|89.9|0.412||$Scale-adaptive$(500×500, $os =$ 0.05)|97.8|88.9|0.356||$Scale-adaptive$(700×700, $os =$ 0.2)|97.2|88.7|0.689||$Scale-adaptive$(700×$700, os=0.05)	96.5	88.1	0.584

Note: zr and zc represent sub-image height and width; os represents overlap ratio between sub-images.

Results show that compared to the original-scale SSD model, the scale-adaptive algorithm generally improves recall. The highest recall of 98.4% was achieved at sub-image sizes of 300$×300*and*500×500*pixelswith*0.2*overlaprate*, *representinga*3.8×300*to*700×700*pixels*, *wit* pixel size at 0.05 overlap rate showing lower recall than at 0.2 overlap rate.

In terms of precision, the scale-adaptive algorithm also generally outperforms the original SSD model, reaching maximum precision of 89.9% at 300$×300*and*500×500*pixelsizeswith*0.2*overlaprate*—*a*2.6×$300 pixel size with 0.05 overlap rate.

The scale-adaptive algorithm also surpasses the original YOLOv3 model in both recall and average precision. While YOLOv3 demonstrates superior small object

recognition performance through its residual network architecture and multi-layer feature fusion, achieving 95.1% recall at its default 416$×$416 pixel input scale, it still underperforms compared to the proposed method. However, when small targets have insufficient effective pixels, feature expression remains inadequate, leading to unsatisfactory detection of difficult small targets.

Regarding speed, the scale-adaptive algorithm is significantly slower than original models. Processing time increases substantially with sub-image size, roughly doubling from $300\$\times 300 to 500\times 500 pixels and again to 700\times\$700$ pixels. At the same scale, 0.2 overlap rate requires approximately 20% more time than 0.05 overlap rate.

Figure 9 [Figure 9: see original paper] compares partial detection results. The scale-adaptive algorithm can recognize bees with incomplete or indistinct features—such as partially visible bees (marked 1), blurred bees due to lighting or motion (marked 2), and bees with even smaller pixel scales due to posture or position (marked 3). This improvement stems from image splitting, which increases target scale and enriches feature representation.

3.3 Algorithm Performance Testing in Complex Environments

To evaluate scale adaptability and generalization, three field bee images of different scales and backgrounds were crawled from the internet for comparative testing against the SSD model. Image details are provided in Table 2 .

Table 2 Information for test images of scale adaptive small objects detection method

Image	Resolution	Mean Bee Scale	Scene
pic1	$670\$\times 420 \mid 4.975e-3 \mid Sideviewoffieldhive \mid\mid pic2 \mid 1440\times 1080 \mid 0.900e-3 \mid Partialviewoffieldhive \mid\mid pic3 \mid 1920\times\$1280$	0.559e-3	Full view of field hive

The scale-adaptive algorithm was tested with two configurations: $300\$\times 300 and 500\times\$500$ pixel sub-images, both with 0.2 overlap rate. Detection results are shown in Figure 10 [Figure 10: see original paper].

SSD recognized only 1-2 bees in the small pic1, 3-8 bees in the medium-sized pic2, and failed to detect any bees in the largest pic3. In contrast, the scale-adaptive algorithm adjusts bee scale through sub-image splitting, better adapting to different input scales. Particularly at 300$×$300 pixel sub-image size, recognition performance did not significantly degrade with increasing original image size.

Since the training dataset lacked images of hive backgrounds and complete beehives, the model did not learn relevant background features. Combined with

high bee density in test images, the scale-adaptive algorithm' s accuracy and recall were not entirely satisfactory. However, comparative results with SSD clearly demonstrate the algorithm' s strong scale adaptability and generalization capability.

4.1 Discussion

The scale-adaptive algorithm improves recognition performance by splitting original images into sub-images for deep network input, thereby increasing target scale and enabling richer feature extraction. This approach proves particularly effective when target absolute pixel count is sufficient but relative scale is small, preventing excessive reduction of effective pixels during input normalization that would otherwise cause feature extraction difficulties. The algorithm can effectively recognize difficult targets with indistinct features.

Deep network inference constitutes the most time-consuming portion of object recognition. Splitting original images for multiple inferences significantly reduces algorithm time efficiency, causing single-image processing time to increase substantially. Time efficiency decreases as sub-image size decreases, and higher overlap rates increase sub-image count, further reducing speed. Selecting appropriate sub-image size and overlap rate based on target pixel count and model input dimensions can improve both precision and recall while enhancing model time efficiency.

Diverse image acquisition devices in agricultural production yield variable resolution and quality. If images of all sizes are fed directly into models, performance inevitably degrades due to excessively small and inconsistent target scales. The proposed algorithm first performs splitting based on target scale, enabling automatic processing of all-scale images and improving model scale adaptability and generalization.

4.2 Conclusion

This study addresses the challenges of small target recognition and scale variation in farmland environments by developing a method to increase effective pixel count and target scale to improve model performance. The original image is split into multiple sub-images fed into the detection model, followed by a two-stage NMS process for final target computation. Experimental results demonstrate that the method can effectively recognize difficult targets with indistinct features. The scale-adaptive algorithm generally achieves higher recall and precision than original algorithms, with maximum improvements of 3.8% in recall and 2.6% in precision, and also outperforms YOLOv3. However, due to poor time efficiency, the algorithm is suitable for non-real-time applications with high recall and precision requirements rather than edge computing scenarios.

References

[1] NAUATA N, HU H, ZHOU G T, et al. Structured label inference for visual understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(5): 1257-1271.

[2] HUANG K, REN W, TAN T. A review on image object classification and detection[J]. Chinese Journal of Computers, 2014, 36(12):1-18.

[3] ZOU Z, SHI Z, GUO Y, et al. Object detection in 20 years: A survey[J/OL]. arXiv: 1905.05055v2 [cs. CV], 2019.

[4] LI K, WANG X, LIN H, et al. Survey of one stage small object detection methods in deep learning[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(1): 41-58.

[5] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[J/OL]. arXiv: 2004.10934, 2020.

[6] REDMON J, FARHADI A. YOLOv3: An incremental improvement[J/OL]. arXiv:1804.02767 [cs.CV], 2018.

[7] MAHTO P, GARG P, SETH P, et al. Refining YOLOv4 for vehicle detection[J]. International Journal of Advanced Research in Engineering and Technology, 2020, 11(5): 409-419.

[8] ZHAI S, SHANG D, WANG S, et al. DF-SSD: An improved SSD object detection algorithm based on image denseNet and feature fusion[J]. IEEE Access, 2020.

[9] XI Q, ZHANG Z, PENG L. Small object detection algorithm based on improved dense network and quadratic regression[J]. Computer Engineering, 2021, 47(4): 241-247, 255.

[10] SHENZ Q, LIU Z, LI J G, et al. DSOD: Learning deeply supervised object detectors from scratch[C]// The 2017 IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Computer Society, 2017: 1937-1945.

[11] LI H, ZHU M. A small object detection algorithm based on deep convolutional neural network[J]. Computer Engineering & Science, 2020, 42(4): 649-657.

[12] ZHOU H, YAN F, ZHU N, et al. An approach to improve the detection model for small object in complex scenes[J/OL]. Computer Engineering and Applications: 1-8. [2021-10-04]. http://kns.cnki.net/kcms/detail/11.2127.TP.20210419.1404.049.html.

[13] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// The Second

International Conference on Knowledge Discovery and Data Mining. Portland, Oregon, USA: AAAI, 1996: 226-231.

[14] LI Y, ZHANG X, LI C, et al. Improved YOLOv3 target detection algorithm combined with DBSCAN[J/OL]. Computer Engineering and Applications: 1-12. [2021-10-04]. http://kns.cnki.net/kcms/detail/11.2127.TP.20210327.1437.002.html.

[15] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2019: 658-666.

[16] YANG Y, LIAO Y, CHENG L, et al. Remote sensing aircraft target detection based on GIoU-YOLOv3[C]// 2021 6th International Conference on Intelligent Computing and Signal Processing. Piscataway, New York, USA: IEEE, 2021: 474-478.

[17] ZHENG Z, ZHAO H, LIU W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]// The 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference, the 10th AAAI Symposium on Educational Advances in Artificial Intelligence. Piscataway, New York, USA: AAAI, 2020.

[18] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, (99).

[19] ZHANG B, QIN H, JIANG S, et al. A method of vehicle detection at night based on RetinaNet and optimized loss functions[J]. Automotive Engineering, 2021, 43(8): 1195-1202.

[20] ZHENG Q, WANG L, WANG F. Small object detection in traffic scene based on improved convolutional neural network[J]. Computer Engineering, 2020, 46(6).

[21] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[22] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]// In European Conference on Computer Vision. Cham, Switzerland: Springer: 2016.

[23] Neubeck A, Gool L J V. Efficient non-maximum suppression[C]// International Conference on Pattern Recognition. Piscataway, New York, USA: IEEE Computer Society, 2006: 848-855.

[24] LI J, JIANG J, DOU Y, et al. A redundacy-reduced candidate box accelerator based on soft-non-maximum suppression[J]. Computer Engineering & Science, 2021, 43(4): 586-593.

[25] ZHANG C, ZHANG C, WANG H, et al. Research on non-maximum suppression based on attention mechanism in object detection[J]. Electronic Measurement Technology, 2021, 44(19): 82-88.

Note: Figure translations are in progress. See original paper for figures.