ACADEMIC
World Academic Union

**ISSN** 1746-7659, England, UK
**Journal of Information and Computing Science**
Vol. 14, No. 4, 2019, *pp*.291-299

# Insect real-time detection in complex environments based on improved YOLOV3

Juan Wang [1]

[1] *College of Information and Science Technology, Jinan University, Guangzhou, 51000, China*

*(Received November 02 2019, accepted December 28 2019)*

**Abstract.** The combination of advanced computer vision technology and insect image recognition technology can be effectively applied to environmental monitoring, pest diagnosis, epidemiology and other fields. However, accurate location and classification of relatively small insects in complex scenes has always been a difficult problem for this technology. Although YOLOv3 combines deep features with shallow features to facilitate the detection of small objects, experiments have found that YOLOv3 has more undetected cases for small target insects in complex backgrounds. In order to solve this problem, this paper optimizes YOLOv3 algorithm. Firstly, the SE blocks are embedded into YOLOv3 network to learn global features and enhance the expression ability of feature maps, so that the network can detect more objects. Because YOLOv3 itself has a complex network structure and the amount of parameters and calculations are increased after embedding the SE block, so this paper also uses depthwise separable convolutions, which greatly reduces the amount of parameters and computation under the condition of little loss of accuracy, thus improving the detection speed. Training and testing on the insect dataset made in this paper, the original YOLOv3 runs at 33 f / s, and the mean Average Precision (mAP) is only 86.8%, While the improved YOLOv3 runs at 38 f / s, the mAP reaches 90.6%. The improved algorithm can detect more targets, reduce the omission factor and improve the detection speed.

**Keywords:** target detection, real-time detection, convolution neural network, YOLOv3.

## 1. Introduction

Accurate and timely identification of insects is the basis for monitoring agricultural pest information and crop pest control. Through remote image acquisition, photo detection and other means, combined with machine vision and image recognition of pest identification technology, will help to improve accuracy and efficiency, reduce losses caused by pests, thus promoting the implementation of precision agriculture, while also saving labor and time costs.

Classical insect image recognition technology usually requires three steps: image preprocessing, feature extraction, and classifier classification. For example, Kandalkar et al. [1] proposed a wavelet transform algorithm to extract features and use BP neural network to classify and identify agricultural pests. . Han Ruizhen [2] used gray level co-occurrence matrix, geometric invariant moment and other methods to extract features, and used support vector machine for classification and recognition. The recognition accuracy on the test set was 89.5%, and it took 1.5 seconds to identify an image. The process of image preprocessing is complicated and requires steps such as image graying, image denoising, image segmentation, etc. The result of target detection is directly related to the feature extraction algorithm used, while the manually designed feature extraction algorithm is not very robust to the diversity of target features. In recent years, researchers have begun to apply the relevant techniques of deep learning to insect image recognition. For example, Liang Wanjie et al. [3] applied convolution neural network to the identification of rice pests, reaching 89.14% identification accuracy. Cheng Xi et al. [4] used deep convolution neural network to classify and identify seven stored grain pests, and the test accuracy on Alexnet[5] and GoogLeNet[6] reached 97.61%. Liu et al. [7] introduced the convolution neural network, and the accuracy rate of identifying 12 kinds of rice field pests reached 93.2%, and the average time to identify an image was about 4 milliseconds.

YOLOv3 [8] is a target detection algorithm with relatively balanced speed and precision, which combines deep features with shallow features, retains fine-grained features and obtains more meaningful semantic information, so that small objects can be detected in real time. However, experiments have found that YOLOv3

[1] Corresponding author. E-mail address: satakiolo@163.com.

has more missed detection for small target insects in complex background. In order to solve this problem, based on the YOLOv3 network, this article optimizes the YOLOv3 algorithm for the self-made insect dataset. The SE blocks [9] are embedded in the YOLOv3 network to reduce the interference of light, background complexity, and high similarity between insects and the environment, and improve the mAP of target detection. On this basis, depthwise separable convolutions [10] are used to improve mAP while reducing the amount of computation and parameters, so as to achieve the purpose of accurate and real-time insect detection.

The main work of this article are as follows:

- To solve the problem of missed detection of target insects in complex background by YOLOv3 network, SE blocks are embedded into YOLOv3 network to obtain SE-resistant structures, which will introduce the original information into the deep layers to inhibit the degradation of information, then pool and expand the receptive field, fuse the shallow layers information and the deep layers information from multiple angles, so that the combined output contains multi-level information, can learn the global features, and enhance the expression ability of the feature map.
- To solve the problem that YOLOv3 itself has a complex network structure and increases the amount of parameters and computation after embedding SE block, the use of depthwise separable convolutions can greatly reduce the amount of parameters and computation under the condition of little loss of accuracy.

## 2. YOLOv3

YOLO algorithm is a regression-based target detection method proposed by Redmon et al. [11] in 2016. YOLOv3 has been developed in 2018. It can detect a variety of objects by only one forward operation, so Yolov3 series of algorithms have a fast detection speed. YOLOv3 borrows the ideas of ResNet [12], introduces a plurality of residual network modules and uses a multi-scale prediction method to improve the defects of YOLOv2 in small target recognition, so YOLOv3 still maintains the fast detection speed of YOLOv2[13], and the recognition accuracy rate is greatly improved, especially in the detection and recognition of small targets, the accuracy rate is greatly improved.

YOLOv3 designed the basic model of classification network Darknet53. Darknet53 include convolution layers and Residual layers. The convolution layers are obtained by integrating convolution layers with better performance from various mainstream network structures, and each convolution is followed by Batch Normalization and linkyRelu activation operations. The Residual layers uses ResNet's Residual structure for reference and are mainly composed of $1 \times 1$convolutions and $3 \times 3$convolutions. Using this structure can make the network structure deeper and avoid gradient disappearance and gradient explosion while extracting deeper features [12]. YOLOv3 also introduced the idea of anchor boxes[14]. For COCO dataset and VOC dataset, it uses three scales for prediction. These three different scales come from the output of convolution layers at different levels, and each scale has three anchor boxes. This method borrows from the idea of FPN [15]: compared with prediction using shallow feature maps directly and up-sampling prediction with deep feature maps, the latter helps retain fine-grained features and obtain more meaningful semantic information, and is more conducive to detecting small objects.

Fig. 1 shows the network structure of YOLOv3. YOLOv3 first converts the input image into a size of $416 \times 416$, extracts image features through Darknet53, and then uses a feature pyramid structure similar to FPN network to output a feature map of three scales. Among the output feature maps, the $13 \times 13$ size feature map is responsible for detecting large objects, the $26 \times 26$ size feature map is responsible for detecting medium objects, and the $52 \times 52$ size feature map is responsible for detecting smaller objects.