Ear Detection Method Based on YOLOv4 and Adaptive Anchor Box Adjustment (Postprint)

Authors: Wangli Hao, Peiyan Wei, Hao Fei, Han Meng, Han Jiwan, Sun Weirong, Li Fuzhong

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

Abstract

Detection and counting of millet ears are crucial for predicting millet yield and breeding. However, traditional millet ear counting is mainly based on manual statistics, which is both time-consuming and labor-intensive. To address the aforementioned issues, this study first established a millet ear detection dataset containing 784 images and 10,000 millet ear samples. A millet ear detection method based on YOLOv4 and adaptive anchor box adjustment is proposed, which can quickly and accurately detect millet ears in specific boxes. By adaptively adjusting the anchor boxes, candidate boxes that conform to millet ear targets can be generated, thereby improving detection accuracy. To validate the effectiveness of the proposed method, multiple metrics were adopted for evaluation, including mean Average Precision (mAP), F1-Score, Precision, and Recall. Furthermore, comparative experiments were designed to verify the effectiveness of the proposed method, including comparing with other models (YOLOv2, YOLOv3, and Faster-RCNN) to evaluate model performance, evaluating model performance under different Intersection over Union (IOU) thresholds, evaluating millet ear detection performance under adaptive anchor box adjustment, evaluating the reasons causing changes in model evaluation metrics, and evaluating model performance under different original input image sizes. Experimental results demonstrate that YOLOv4 achieved favorable millet ear detection performance. YOLOv4 achieved an mAP of 78.99%, an F1-score of 83.00%, a Precision of 87%, and a Recall of 79.00%, exceeding other comparative models by 8% across all evaluation metrics. Experimental results indicate that the proposed method possesses good accuracy and efficiency.

Full Text

Preamble

Foxtail Millet Ear Detection Approach Based on YOLOv4 and Adaptive Anchor Box Adjustment

HAO Wangli[1], YU Peiyan[1], HAO Fei[2], HAN Meng[1], HAN Jiwan[1], SUN Weirong[1], LI Fuzhong[1]*

[1] School of Software, Shanxi Agricultural University, Jinzhong, Shanxi 030801, China
[2] School of Computer Science, Shaanxi Normal University, Xi' an 710119, China

Abstract: Foxtail millet ear detection and counting are essential for estimating foxtail millet production and supporting breeding programs. However, traditional manual counting approaches are time-consuming and labor-intensive. To address this challenge, we established a foxtail millet ear detection dataset comprising 784 images with 10,000 ear samples and proposed a novel detection method based on YOLOv4 (You Only Look Once) with adaptive anchor box adjustment for rapid and accurate detection of ears within specific regions. By adaptively adjusting anchor boxes to generate candidate boxes that better match foxtail millet ear characteristics, detection accuracy was significantly improved. Multiple evaluation metrics including mean Average Precision (mAP), F1-score, Precision, and Recall were employed to validate the proposed method. Comprehensive ablation studies were conducted to verify effectiveness: (1) performance comparison with other models (Faster-RCNN, YOLOv2, YOLOv3); (2) evaluation across different Intersection over Union (IOU) thresholds to identify optimal values; (3) assessment of detection performance with and without adaptive anchor box adjustment; (4) analysis of factors influencing model performance metrics; and (5) evaluation of different input image sizes. Experimental results demonstrated that YOLOv4 achieved superior detection performance with mAP reaching 78.99% and F1-score reaching 83.00%. Precision attained 87% and Recall reached 79.00%, representing approximately 8% improvement over YOLOv2, YOLOv3, and Faster-RCNN across all metrics. These results confirm that the proposed method delivers both high accuracy and efficient performance.

Key words: foxtail millet ear detection; YOLOv4; deep neural network; dataset; adaptive anchor box adjustment

CLC number: S24
Documents code: A
Article ID: 202102-SA066

Citation: HAO Wangli, YU Peiyan, HAO Fei, HAN Meng, HAN Jiwan, SUN Weirong, LI Fuzhong. Foxtail millet ear detection approach based on YOLOv4 and adaptive anchor box adjustment[J]. Smart Agriculture, 2021, 3(1): 63-74. (in English with Chinese abstract)

Received date: 2021-02-25
Revised date: 2021-03-26

Foundation items: Shanxi Province Higher Education Innovation Project of China (2020L0154)

Biography: HAO Wangli (1988-), female, Ph.D., lecturer, research interests: artificial intelligence and smart agriculture. E-mail: hanmeng10@126.com

Corresponding author: LI Fuzhong (1969-), male, Ph.D., professor, research interest: smart agriculture. Tel: 0354-6287093. E-mail: hualimengyu@163.com

1 Introduction

Effective foxtail millet breeding increases food production and ensures food security, making production estimation a critical research priority. Foxtail millet yield is primarily determined by three factors: ear number, grains per ear, and grain quality, with their relative contributions following the order: grains per ear > ear number > grain quality. Consequently, accurate ear number estimation is essential for predicting foxtail millet production.

Traditional manual estimation methods are subjective and inefficient. Deep neural networks offer a promising alternative for efficient and accurate foxtail millet ear detection, with detected bounding boxes further facilitating production estimation. Recent advances in deep learning and improved hardware performance have drawn significant attention to neural networks for target detection, semantic segmentation, and instance segmentation tasks. For wheat ear detection specifically, various approaches have been developed. Lu proposed a wheat ear recognition method based on back propagation (BP) neural networks. Shi extracted color, shape, and texture parameters of wheat grains for BP neural network classification, employing mean error square (MES) and mean impact value (MIV) optimization to improve recognition rates by 11.45% compared to unoptimized models. Zhang et al. designed a winter wheat ear detection and counting system using convolutional neural networks. Gao applied YOLOv3 and Mask R-CNN for field wheat ear detection, achieving 87.12% mAP. Alkhudaydi et al. developed a fully convolutional model to estimate wheat ears from high-resolution RGB images. Xie et al. proposed a Feature Cascade SVM (FCS R-CNN) method achieving 81.22% mAP.

While deep learning-based detection methods have demonstrated success in wheat ear detection, few approaches have been developed specifically for foxtail millet ear detection. This research addresses this gap by exploring foxtail millet ear detection and proposing an effective method. Leveraging YOLOv4' s promising detection capabilities, we employed it for foxtail millet ear detection and counting. To adapt YOLOv4 for this specific task, anchor box sizes were adjusted using the K-means algorithm based on our foxtail millet ear detection

dataset, significantly enhancing detection performance. The dataset was collected from farmland, containing approximately 784 images with 10,000 foxtail millet ear samples, with 588 images used for training and the remainder for testing.

---

2 Dataset

Data were collected from the foxtail millet experimental field at Shanxi Agricultural University in Taigu County, Jinzhong City, Shanxi Province, covering varieties including Male sterile line GBS, Datong 27, and Dragon Claw.

2.1 Data Collection

The collection period spanned one month from August 10th to September 10th, 2020. To ensure sample diversity, data were collected every other day at 10 a.m., encompassing three foxtail millet varieties (Male sterile line GBS, Datong27, and Dragon Claw) under varying light conditions and weather. A white PVC pipe frame (0.5 m width × 0.6 m length × 0.5 m height) was placed 0.5 m above ground to delineate the sampling area. Images were captured using a Canon EOS 70D camera with 35 mm focal length positioned 1.5–2 m from the frame, producing high-resolution *.jpg images at 4864$×$3648 px resolution. Figure 1 shows representative foxtail millet ear samples.

2.2 Data Cleaning and Annotation

To develop an effective training model, images with blurry ears or excessive weeds were eliminated to minimize background interference and image degradation effects on detection accuracy. Accurate data annotation is crucial for model performance. We employed labelImg to annotate each foxtail millet ear within the white box using rectangular bounding boxes defined by four vertex coordinates. After annotating all ears in an image, an XML file was generated containing image dimensions, label frame names, and target frame locations. These XML files were subsequently converted to text format for network input. The final dataset comprised 784 images with 10,000 foxtail millet ear samples, partitioned into 588 training images (80%) and 196 test images (20%).

---

3 Methods

3.1 YOLO Models

YOLO is an excellent object detection model that effectively balances speed and accuracy. As a one-stage detector, YOLO directly detects objects without generating candidate proposals. The detection process involves: (1) extracting features from the input image through a feature extraction network to obtain an

N×N pixel feature map; (2) dividing the input image into N×N grid cells, where each cell containing an object's ground-truth center coordinate is responsible for predicting that object; (3) each grid cell predicts M bounding boxes of different sizes, selecting the box with maximum Intersection over Union (IOU) for final prediction. Each bounding box contains five prediction values: tx, ty, tw, th, and confidence, where tx, ty, tw, th represent predicted center coordinates, width, and height, and confidence indicates prediction reliability. The predicted box' s center coordinates (cx, cy), width (cw), and height (ch) are calculated using:

$$c_x = \sigma(t_x) + b_x c_y = \sigma(t_y) + b_y c_w = p_w e^{t_w} c_h = p_h e^{t_h}$$

where σ(x) is the logistic function; tx, ty, tw, th are model predictions; pw and ph are prior box dimensions relative to the feature map; bx and by are grid cell coordinates.

3.2 YOLOv4 Architecture

YOLOv4' s architecture (Figure 2) consists of four modules: Input, Backbone, Neck, and Head. The Input employs Mosaic data augmentation. The Backbone is CSPDarknet-53, which integrates five CSP modules into Darknet-53. CSPDarknet-53 includes 29 convolutional layers $(3\$\times 3 kernels) with a 725\times\$725$ receptive field and 27.6M parameters. By leveraging CSPNet' s advantages in reducing computational costs while maintaining accuracy and reducing memory consumption, YOLOv4 adds CSP to each large residual block of Darknet-53, splitting feature mappings into two parts and merging them through cross-stage hierarchical structures.

The Neck comprises Spatial Pyramid Pooling (SPP) and Path Aggregation Network (PANet). SPP is added over the CSPDarknet53 backbone with max pooling sizes of $5\$\times 5, 9\times 9, and 13\times\$13$, significantly increasing receptive field and extracting crucial contextual features without reducing network speed. PANet aggregates parameters from different backbone levels.

To detect foxtail millet ears of varying sizes, anchor boxes provide initial width and height estimates, preventing the model from blindly learning target positions and scales during training. Since foxtail millet ears are smaller than typical objects, we employed the K-means algorithm to adaptively generate nine anchor boxes based on our dataset: (5, 7), (6, 12), (9, 8), (7, 18), (10, 13), (13, 10), (10, 21), (14, 15), and (17, 25), where coordinates represent width and height. The first three detect small ears, the middle three detect medium-sized ears, and the last three detect larger ears.

3.3 Loss Function

YOLOv4' s loss function comprises three components: localization loss, confidence loss, and classification loss.

Localization Loss uses Complete Intersection Over Union (CIOU) Loss:

$$L_{ciou} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{obj} \left[1 - IOU + \frac{d^2(c, c^{gt})}{l^2} + \frac{4}{\pi^2} \left(\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2 \right]$$

where IOU is calculated as:

$$IOU = \frac{C \cap D}{C \cup D}$$

C and D represent ground-truth and predicted bounding boxes; IOU measures prediction accuracy; d($\cdot$) is Euclidean distance; l is diagonal distance between predicted and ground-truth boxes; c, w, h and c^{gt}, w^{gt}, h^{gt} are center coordinates, width, and height of predicted and ground-truth boxes, respectively.

Confidence Loss is formulated as:

$$L_{conf} = -\sum_{i=0}^{S^2} \sum_{j=0}^{B} \left[\mathbb{I}_{ij}^{obj} \log(\hat{C}_j^i) + \lambda_{noobj} \mathbb{I}_{ij}^{noobj} \log(1 - \hat{C}_j^i) \right]$$

where S^2 and B indicate feature map scale and number of prior boxes; λ_{noobj} is a hyperparameter balancing the terms; $\hat{C}_j^i$ represents confidence scores of annotated and predicted boxes; $\mathbb{I}_{ij}^{obj}$ and $\mathbb{I}_{ij}^{noobj}$ are indicator functions (1 and 0 if target exists at grid i, box j; 0 and 1 otherwise).

Classification Loss is:

$$L_{cls} = -\sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{obj} \left[\hat{P}_j^i \log(P_j^i) + (1 - \hat{P}_j^i) \log(1 - P_j^i) \right]$$

where L_{cls} is classification loss; $\hat{P}_j^i$ and P_j^i denote class probabilities of annotated and predicted boxes.

The total loss is: $L = L_{ciou} + L_{conf} + L_{cls}$

3.4 Experimental Setup

Experiments were conducted on a GTX TITANXP 12G GPU with an I7 7800X processor. Software configurations included CUDA 10.1, CUDNN 7.6.4, Python 3.6.9, and PyTorch. Training parameters were set as: learning rate = 0.001, iterations = 12,000, momentum = 0.949.

3.5 Evaluation Criteria

Model performance was evaluated using Precision, Recall, F1-score, and mean Average Precision (mAP). Precision measures prediction accuracy; Recall indicates target detection completeness; F1-score is the harmonic mean of Precision and Recall (ranging from 0 to 1); mAP represents average detection accuracy. mAP calculation follows PASCAL VOC2007 definitions, where detection is considered correct when IOU between detection and ground-truth boxes exceeds a threshold and category confidence score surpasses a specified value.

$$Precision = \frac{TP}{TP + FP} Recall = \frac{TP}{TP + FN} F1\text{-}score = \frac{2 \times P \times R}{P + R} mAP = \int_0^1 P(R)dR$$

where TP, FP, FN indicate True Positive, False Positive, and False Negative samples.

---

4 Experiments and Results

4.1 Model Comparison

To validate YOLOv4' s effectiveness for foxtail millet ear detection, we compared it against Faster-RCNN, YOLOv2, and YOLOv3 using identical training parameters. During testing, confidence and IOU thresholds were set to 0.35 and 0.5, respectively, meaning predictions with confidence > 0.35 and IOU > 0.5 were considered correct.

Table 1 shows that YOLOv4 significantly outperformed all comparison models across all metrics. Specifically, YOLOv4' s Precision exceeded Faster-RCNN, YOLOv2, and YOLOv3 by 1.9%, 13%, and 1%; Recall surpassed them by 4.4%, 8.2%, and 2.6%; F1-score improved by 2.6%, 10.6%, and 2.4%; and mAP increased by 3.9%, 10.4%, and 2.6%, respectively.

Table 1: Comparison results of different models (confidence = 0.35, IOU = 0.5)

Models	Precision/%	Recall/%	F1-score/%	mAP/%
Faster-RCNN				
YOLOv2				
YOLOv3				
YOLOv4				

Figure 3 illustrates performance curves across training iterations, showing YOLOv4' s mAP, Precision, Recall, and F1-score consistently above YOLOv2 and YOLOv3 throughout training, confirming YOLOv4' s superior performance.

Qualitative results (Figure 4) demonstrate YOLOv4 produces more accurate predictions than YOLOv2 and YOLOv3, particularly under occlusion and challenging conditions.

YOLOv4' s superior performance stems from several factors: (1) Mosaic data augmentation merges four training images, allowing detection of objects outside their normal context and enabling batch normalization statistics from four distinct images per layer, reducing large mini-batch requirements; (2) CSPDarknet53 backbone with Mish activation and DropBlock strategies improves model performance—CSP structure enhances capability, Mish activation' s smoothness facilitates better information flow for improved accuracy and generalization, and DropBlock gradually increases dropout units during training for better robustness; (3) CIOU loss and DIOU-NMS enhance convergence speed and regression accuracy.

4.2 IOU Threshold Evaluation

Unlike general object detection, foxtail millet ears in our dataset are small and densely distributed. To determine the optimal IOU threshold, we evaluated YOLOv4 performance at IOU values of 0.2, 0.35, 0.5, and 0.65 (confidence fixed at 0.35). Table 2 shows all metrics decreased as IOU increased. While IOU = 0.2 and 0.35 yielded higher scores, the small overlap between predicted and ground-truth boxes made predictions less reliable. At IOU = 0.65, performance dropped sharply by 20–30%, indicating many ears were missed. Balancing IOU and detection performance, we selected IOU = 0.5 as optimal.

Table 2: Impact of IOU values on model performance

Model	IOU	Precision/%	Recall/%	F1-score/%	mAP/%
YOLOv2	0.2				
	0.35				
	0.5				
	0.65				
YOLOv3	0.2				
	0.35				
	0.5				
	0.65				
YOLOv4	0.2				
	0.35				
	0.5				
	0.65				

4.3 Anchor Box Adjustment Evaluation

We compared models with and without adaptive anchor box adjustment: YOLOv3 vs. YOLOv3_{adj} and YOLOv4 vs. YOLOv4_{adj}. Adjusted

anchors for YOLOv3 were (3,5), (4,8), (6,5), (5,12), (7,9), (9,7), (7,14), (10,10), (11,17); for YOLOv4 they were (5,7), (6,12), (9,8), (7,18), (10,13), (13,10), (10,21), (14,15), (17,25).

Table 3 shows YOLOv3_{adj} and YOLOv4_{adj} outperformed their non-adjusted counterparts, confirming the effectiveness of adaptive anchor boxes. Adjusted anchors provide higher relative offset for foxtail millet ears, enabling more accurate width and height predictions.

Table 3: Comparison results with/without anchor box adjustment

Model	Precision/%	Recall/%	F1-score/%	mAP/%
YOLOv3				
YOLOv3_{adj}				
YOLOv4				
YOLOv4_{adj}				

4.4 Analysis of Performance Metric Variations

Equations (7)-(10) show that TP and FP values directly relate to model performance. To understand performance differences among models, we analyzed TP and FP values on the test set. TP and FP calculation involved: (1) removing predictions below a confidence threshold (0.5); (2) sorting remaining predictions by confidence; (3) calculating IOU between the highest-confidence prediction and ground-truth; (4) if IOU exceeded the threshold (0.35), the prediction was counted as TP and the ear marked as tested, with all subsequent predictions for that ear counted as FP.

Table 4 presents TP and FP statistics. Higher TP values indicate better performance. YOLOv4 outperformed YOLOv3 with 63 more TP (2.92% increase), 23 fewer FP (6.53% decrease), and 2.63% mAP improvement. Compared to YOLOv2, YOLOv4 achieved 168 more TP (8.19% increase), 294 fewer FP (47.19% decrease), and 10.44% mAP improvement. YOLOv3 also outperformed YOLOv2 with 105 more TP (5.12% increase), 271 fewer FP (43.50% decrease), and 7.60% mAP improvement. The mAP change ratio closely matched TP changes, confirming that better detection requires higher TP values.

Table 4: TP and FP values for ear target prediction

Model	TP increase	FP increase	mAP increase
YOLOv2			
YOLOv3			
YOLOv4			

4.5 Input Image Size Evaluation

Original images (4864$\times$3648px)$wereresizedto$608$\times$608$pxforYOLOv4input,creatingalargeresizeratiothatma$ px) would improve detection. Table 5 shows cropped images yielded better results than uncropped originals because reduced resize ratios enhanced detection performance.

Table 5: Impact of original image size on detection performance

Original image size/px	YOLOv4 input size/px	Precision/%	Recall/%	F1-score/%	mAP/%
4864$\times$3648	608$\times$608				
2000$\times$1500	608$\times$\$608				

5 Conclusions

This research proposed an adaptive anchor adjustment approach for foxtail millet ear detection based on YOLOv4, achieving promising results. We established a novel large-scale dataset of 784 images containing 10,000 ear samples collected from farmland. The YOLOv4 model with adaptive anchor adjustment was applied to foxtail millet ear detection. Extensive experiments validated both the dataset and YOLOv4' s effectiveness, demonstrating superior performance compared to YOLOv2, YOLOv3, and Faster-RCNN across all evaluation criteria.

Future work should explore detection of other millet ear categories worldwide and expand the dataset scale. More effective detection approaches for foxtail millet ears will be investigated.

References

[1] LI Y. Millet breeding[M]. Beijing: China Agriculture Press, 1997: 22-23.
[2] CHEN G. Analysis of the total contribution of millet yield components[J]. Miscellaneous Crops, 2000, 20(3): 25-26.
[3] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New York, USA: IEEE, 2017: 7263-7271.
[4] TANG C, HU H, WEI P, et al. An improved Yolov3 algorithm to detect molting in swimming crabs against a complex background[J]. Aquacultural Engineering, 2020, 91(3): 102-115.
[5] CHENG Z, ZHANG F. Flower end-to-end detection based on YOLOv4 using a mobile device[J]. Wireless Communications and Mobile Computing, 2020(2): 1-9.
[6] GONG B, ERGU D, CAI Y, et al. A method for wheat head detection based on Yolov4[EB/OL]. DOI: 10.21203/rs.3.rs-86158/v1.

[7] LU X. Research on wheat ear recognition based on image processing technology[D]. Shijiazhuang: Hebei Agricultural University, 2012.
[8] SHI X. Detection and grading of wheat appearance quality based on image processing[D]. Zhengzhou: Henan University of Technology, 2013.
[9] ZHANG Q, CHEN Y, LI Y, et al. Winter wheat ear detection and counting system based on convolutional neural network[J]. Transactions of the Chinese Society of Agricultural Machinery, 2019, 50(3): 144-150.
[10] GAO Y. Research on field wheat ear detection method based on deep neural network[D]. Beijing: Beijing Forestry University, 2019.
[11] ALKHUDAYDI T, ZHOU J, IGLESIAS B D L. SpikeletFCN: Counting Spikelets from infield wheat crop images using fully convolutional networks[M]. Boca Raton: Artificial Intelligence and Soft Computing, 2019.
[12] XIE Y, HE C, YU Z, et al. Optimization method of wheat ear detection cascade network in complex field scene[J]. Transactions of the Chinese Society of Agricultural Machinery, 2020, 51(12): 212-219.
[13] TZUTA. labelImg[EB/OL]. [2021-2-10] https://github.com/tzutalin/labelImg.
[14] KHARCHENKO V, CHYRKA I. Detection of airplanes on the ground using YOLO neural network[C]// 2018 IEEE 17th International Conference on Mathematical Methods in Electromagnetic Theory (MMET). Piscataway, New York, USA: IEEE, 2018: 294-297.
[15] WANG C, LIAO H M, YEH I, et al. CSPNet: A New Backbone that can Enhance Learning Capability of CNN[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, New York, USA: IEEE, 2020.
[16] TAN X, WANG Z. Ping-pong table recognition based on YOLOv4 improved algorithm[J]. Science and Technology Innovation and Application, 2020(27): 74-76.
[17] WANG J, WANG J, SONG J, et al. Optimized cartesian k-means[J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(1): 180-192.
[18] Microsoft. PASCAL-VOC2012[EB/OL]. (2012-02-20) [2019-08-02]. http://host.robots.ox.ac.uk/pascal/VOC/voc2012.
[19] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New York, USA: IEEE, 2019: 658-666.

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