Artificial Intelligence Research and Development
K. Trejo et al. (Eds.)

doi:10.3233/FAIA250601

An Improved YOLO11 Model for Accurate Cotton Growth-Stage Detection

Guillem GONZÁLEZ [a,1], Sergi FOIX [a] and Guillem ALENYÀ [a]

[a] *Institut de Robòtica i Informàtica Industrial, CSIC-UPC*

ORCiD ID: Guillem González https://orcid.org/0009-0009-7627-0887, Sergi Foix https://orcid.org/0000-0001-9249-6696, Guillem Alenyà https://orcid.org/0000-0002-6018-154X

Abstract. Accurate crop detection in orchard environments is one of the primary challenges for agricultural systems today, as vision plays a crucial role in modern solutions for plant monitoring, crop harvesting, early disease detection, and analysing water and nutritional status, among other applications. In our case, we want to detect cotton bolls and accurately label them in different stages of growth. To address this challenge, we present a custom YOLO11 model with a selection of state-of-the-art vision modules to improve cotton detection and reduce possible mismatches with a custom cotton dataset. Our approach achieves better detection accuracy with around 7M parameters and 16 GFLOPS, less than the official YOLO versions' small to medium variants (8, 10, 11, 12). The final model will be used to count and monitor cotton plants in a greenhouse.

Keywords. Object detection, Deep learning, Custom YOLO, Cotton detection, Cotton dataset

1. Introduction

Since the emergence of the first deep learning vision models, accompanied by the rapid development of graphics cards with high graphics memory, object detection in complex environments has been a significant area of study, ranking as one of the most important fields in computer vision. Object detection has been pushing boundaries in detecting a wide range of complex objects in challenging, dynamic scenery, while minimising computational costs. Convolutional neural networks (CNNs) are today's most accurate, less latent solution to approach this task. Current generic models aim to increase detection accuracy irrespective of the dataset, while maintaining or decreasing computational overhead, but dataset-specific models can be used or be inspirational, for this generalisation task too. Task-specific models propose modifications to state-of-the-art models with custom vision blocks that are used in the next generation of object detectors.

[1] Corresponding Author: Guillem González, ggonzalez@iri.upc.edu

2. State Of The Art

Object detection models for agricultural purposes are specially designed to work in difficult environments, presenting constant changes in brightness, cluttered environments, and noisy data. Currently, agricultural computer vision is focused on the early detection of crops and diseases. Most of these models are custom architectures of strong YOLO object detectors that improve the original version to meet the needs of a specific custom dataset [1,2,3]. These works tend to use attention mechanisms [4,5] and image augmentation techniques to cushion the lack of representativeness in the set of images, as well as to adapt to their complex environments. Li et al. [1] also incorporates CARAFE, an alternative to traditional upsampling that enables learning the upsampling operation with gradients, resulting in better refinement of feature maps upon training [6]. These algorithms aim to achieve better accuracy while maintaining or reducing latency, as these models are to be deployed in mobile robotics, usually with fewer computational resources than conventional robotic solutions.

3. Dataset

In our previous work, we proposed a custom dataset to discern the ripeness of cotton bolls [7]. This time, we can not use it for this task, as it has few images to represent the complexity of the boll growth-stage classification. In consequence, we are labelling more images with five classes. In collaboration with CEBAS-CSIC[2], we created a dataset called CONDIS++, which states 5 different classes for differentiated growth stages: ripe boll, open boll, square, fertilised flower and early flower. The dataset consists of 1052 up-close images of cotton bolls in different growing stages and 3303 labels. Without data augmentation, there are 1191 labels of ripe cotton, 304 open bolls, 1494 squares, 213 fertilised flowers, and 101 early flowers.

4. Objectives and Proposed Detection Model

Following the same approach as the mentioned papers in the state-of-the-art, this research focuses on modifying a YOLO model to improve the accuracy of detecting the different stages of cotton growth, while keeping or reducing prediction latency, as well as parameter size. The developing detection model COT-YOLO11 starts with a YOLO11 base, which consists of a CSP backbone, a PAN+FPN neck, and the original YOLO11 head. The custom YOLO11 model has been modified to add attention as the main feature and contribution to the work. A graphic view of the model can be found in Fig. 1.

Instead of the traditional upsampling operator used in YOLO11, with nearest neighbours as a criterion for new pixels, we are using the Content Aware ReAssembly of Features block (CARAFE) [6] after a convolution block. This operation allows for the same dimension transformation from $H \times W \times C$ to $2H \times 2W \times C/2$ with only the parameter addition of the convolutional block, as CARAFE computational requirements are almost negligible ($\approx 65K$ per block, $130K$ in COT-YOLO11) compared to total parameter count. CARAFE enlarges the field of view by aggregating information from a large receptive

[2]Centro de Edafología y Biología Aplicada del Segura, Consejo Superior de Investigaciones Científicas

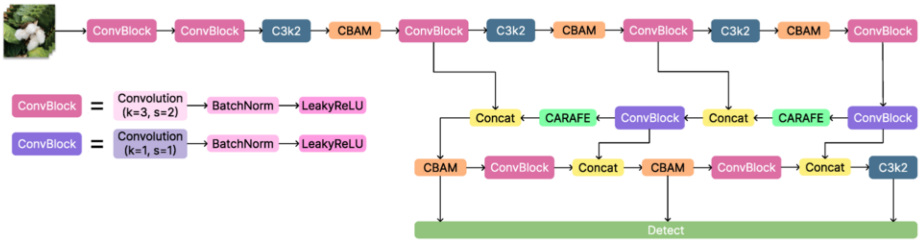

Figure 1. Proposed model's current architecture for improved cotton detection

field, translating to better refinement of feature maps in upsampling and less information loss.

CBAM [5] is incorporated into COT-YOLO11 as our main attention mechanism for its expressivity, lightness, versatility and dimension enhancement. This block boosts the representation power in CNNS, as well as highlighting channel and space dimensions. In other words, CBAM potentiates what is meaningful in an image, as well as the location of relevant information, respectively.

5. Early Results and Conclusions

The first results show a slight improvement of the standard models of the YOLO family, for small to medium (n,s,m) variants, with a reduction of the number of parameters and floating point operations by 20% (Table 1), compared by parameter size resemblance to YOLO11s. The proposed model incorporates attention as its primary feature, with an added module that emphasises the relevant areas in the image, thereby highlighting important regions of the feature maps. As a result, COT-YOLO11 achieves a good balance between accuracy and parameter size, as well as a suitable precision/recall balance, being eligible for mobile systems with limited computational power without losing prediction accuracy. This methodology allows us to test state-of-the-art vision modules to obtain better, faster models.

Model	Params (M)	GFLOPS	Prec.	Rec.	mAP50	mAP50-95
YOLOv8n	3.1	8.9	0.87	0.61	0.766	0.511
YOLOv10n	**2.3**	6.7	0.879	0.642	0.753	0.512
YOLO11n	2.6	**6.6**	0.88	0.68	0.802	0.525
YOLOv12n	2.6	6.7	0.802	0.701	0.774	0.527
YOLOv8s	11.1	28.8	0.835	0.696	0.757	0.52
YOLOv10s	7.2	21.6	0.819	0.623	0.732	0.51
YOLO11s	9.4	21.7	0.814	0.718	0.808	0.543
YOLOv12s	9.2	21.7	**0.891**	0.682	0.791	0.544
YOLOv8m	25.9	79.3	0.785	0.713	0.77	0.542
YOLOv10m	15.4	59.1	0.781	0.711	0.777	0.536
YOLO11m	20.1	68.5	0.805	**0.761**	0.808	**0.555**
YOLOv12m	20.1	67.8	0.799	0.751	0.801	0.55
COT-YOLO11	7.5	16.3	0.88	0.713	**0.823**	0.548

Table 1. Early results from YOLOv8, YOLOv10, YOLO11 and YOLOv12 (nano, small and medium) compared to COT-YOLO11. Training was performed on an Intel(R) Xeon(R) Gold 6326 CPU @ 2.90GHz, 2x Nvidia A10 GPUs. Ubuntu 20.04.6 LTS operating system with CUDA 10.1.

Although the dataset serves as a potentially powerful representation of cotton fruits in greenhouse environments, CONDIS++ lacks images for the accurate representation of cotton morphology, especially the most complex classes, from the square to the ripe fruit. More images from different points of view and different seasons would be required for a stronger phenological learning. Finally, COT-YOLO11 used the following hyperparameters for the training phase: SGD optimiser with $lr_0 = lr_f = 0.01$. Complete Intersection over Union (CIoU), Binary Cross Entropy with logits (BCE) and Distributed Focal Loss (DFL) are used as loss functions, with (7.5/0.5/1.5) values on weight applied respectively. The default YOLO training configuration file defines other hyperparameters.

6. Acknowledgments

We want to thank the personnel at CEBAS-CSIC for providing us with the information and pictures of cotton, with which we could develop CONDIS++. In addition, this work has been partially funded by the DEMETER 5.0 project PLEC2022-009289 funded by MCIU/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR".

7. Future Work

Our future objectives are to upgrade the current model to achieve better accuracy, precision, and recall values in all classes, as well as to update the cotton dataset to enhance class representation. Cotton is a difficult crop to detect, as the fruits can develop in diverse morphologies. The movement of the plant can tear the cotton, forcing the model to learn and adapt to multiple possible shapes, so flexible, robust models are needed. More images will be added to the dataset to strengthen the model.

References

[1] Li J, Li C, Zeng S, Luo X, Chen CLP, Yang C. A lightweight pineapple detection network based on YOLOv7-tiny for agricultural robot system. Computers and Electronics in Agriculture. 2025;231:109944. Available from: https://www.sciencedirect.com/science/article/pii/S016816992500050X.

[2] Shi H, Li L, Zhu S, Wu J, Hu G. FeYOLO: Improved YOLOv7-tiny model using feature enhancement modules for the detection of individual silkworms in high-density and compact conditions. Computers and Electronics in Agriculture. 2025;231:109966. Available from: https://www.sciencedirect.com/science/article/pii/S0168169925000729.

[3] Yang H, Yang L, Wu T, Yuan Y, Li J, Li P. MFD-YOLO: A fast and lightweight model for strawberry growth state detection. Computers and Electronics in Agriculture. 2025;234:110177. Available from: https://www.sciencedirect.com/science/article/pii/S0168169925002832.

[4] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018. .

[5] Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional Block Attention Module. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. .

[6] Wang J, Chen K, Xu R, Liu Z, Loy CC, Lin D. CARAFE: Content-Aware ReAssembly of FEatures. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2019. .

[7] González G, Martínez A, Martínez V, Foix S, Alenyà G. Fast Ready-To-Harvest Cotton Detection and Classification with YOLOv8 in Greenhouse Crops. In: 2024 7th Iberian Robotics Conference (ROBOT); 2024. p. 1-6.