A post-processing pipeline on foundation model for zero-shot learning segmentation of maize ear

Nacir Boutra
Université d'Angers, Phymea Systems, France
Timothé Leroux
Phymea Systems, France
timothe.leroux@phymea-systems.com

Vincent Oury
Phymea Systems, France
David Rousseau
Université d'Angers, France
david.rousseau@univ-angers.fr

Abstract

In this communication we investigate the recently introduced segment anything model (SAM) for a plant science computer vision problem. We consider the segmentation of maize ear in a framework of variety testing protocol. We propose a zero-shot learning procedure followed by post-processing to solve the problem from images acquired with an affordable RGB and IR device. This illustrates the potential huge impact of foundation models to reduce cost of image annotation.

1. Introduction

To study the stress response of a wide range of maize varieties, a large number of cobs were captured using the Earbox imaging system [3]. This imaging system and associated segmentation algorithm (based on the standard MaskRCNN [1]) provided useful phenotypic information on cob's dimension and grain organization. While the system showcased its potential in maize ear analysis, it also brought to light certain limitations including high computational requirements, limited scalability, challenges with small objects, complexities in handling occlusions as well as time-consuming and labor-intensive annotation. In this study, we introduce a simple yet effective approach to address some of these limitations and to improve the results obtained in [3] by harnessing the recently released segment anything model (SAM) [2] to generate maize ear segmentation.

2. Method

A subset of 127 maize ears from different varieties acquired with the Earbox described in [3] were selected and manually annotated with grain count per ear ranging from around 0 to about 500. In short, this device is a box with imaging and lighting system arranged to produce automated acquisition in controlled conditions of samples from top view, with motorized rollers to modify the side of ears facing the camera. The produced pairs of registered RGB and IR images were provided to the pipeline of Fig. 1. As they show a higher contrast in IR than in RGB, ears are segmented from the background with a first pass of IR images into the SAM algorithm. Only large objects are kept to produce the mask of ears. The threshold in size was empirically set based on prior knowledge of the typical size of a maize ear. The semantic mask of each ear is then applied on the RGB images. The bounding box around each RGB ear is cropped and sent to a second pass of SAM. Again a threshold is applied on size and only small objects are kept. This threshold was here again set based on prior knowledge of the typical size of a maize grain. The segmentation was evaluated based on Dice and compared with the results obtained in [3]. After this two pass application of SAM, we used a convolutional neural network architecture (Residual Network composed of 50) to classify small regions as either grains or non-grains. In order to construct the classifier model, we labeled manually 24881 small object RGB images (100x100 pixels), we used 19704 (80%) images for training, and 4927 (20%), for testing.

3. Results

The proposed pipeline produces segmentation results with Dice scores comparable with the results obtained in [3]. SAM holds its promises and enable a zero-shot learning approach even without fine tuning in our case. Some qualitative illustration of the segmentation is provided in Fig. 2. The results of the grain classification is provided in the confusion matrix of the Fig. 2 and compete with the state-of-the-art recently provided in [3].

Grain count per ear determined using the proposed pipeline was significantly correlated with the manual counting method as provided in Fig. 3. These results

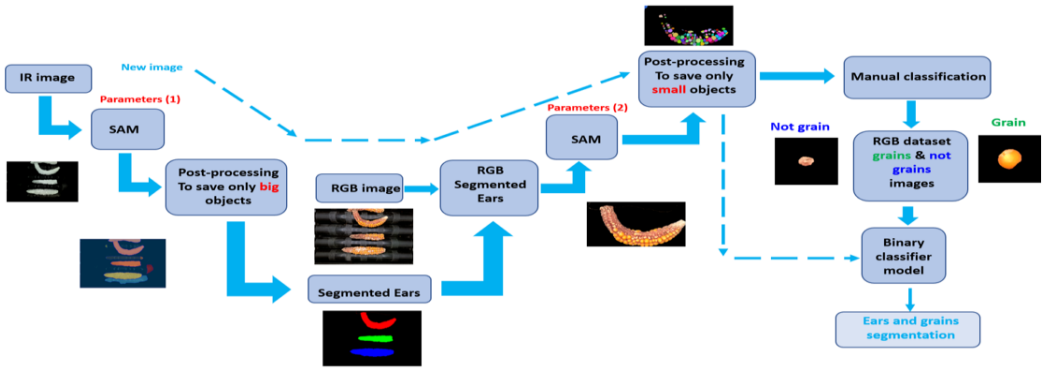

Figure 1. Pipeline proposed for non supervised Grain segmentation and then shallow learning Grain classification.

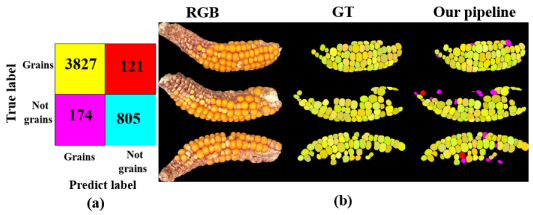

Figure 2. Result of the segmentation on three maize ear and visual comparison with the ground truth. Colors in the GT and result subfigures are the one of the confusion matrix on the left.

validate the potential of our pipeline for accurately segmenting grain instances. In the evaluation of our post-processing pipeline, we observed certain segmentation errors. Specifically, misclassification of non-grain objects as grains in images with low grain count, and difficulties separate crowded objects. To address these limitations, future work could explore specialized grain detection models built on top of the ResNet50 used in the present work, fine-tuned to recognize individual grains accurately in images with low grain counts. We could also integrate contextual information using graph-based approaches, to establish relationships between segmented regions and aid in precise grain identification within crowded regions. While not implemented in the current study, further investigation and refinement of these strategies are warranted to advance the accuracy and efficiency of the proposed pipeline.

4. Conclusion

One of the key advantages of the proposed pipeline lies in the use of the foundation model, which allows us to bypass the need for labor-intensive and precise manual annotation of maize grains for segmentation. By leveraging SAM's pre-trained knowledge on a wide range of visual

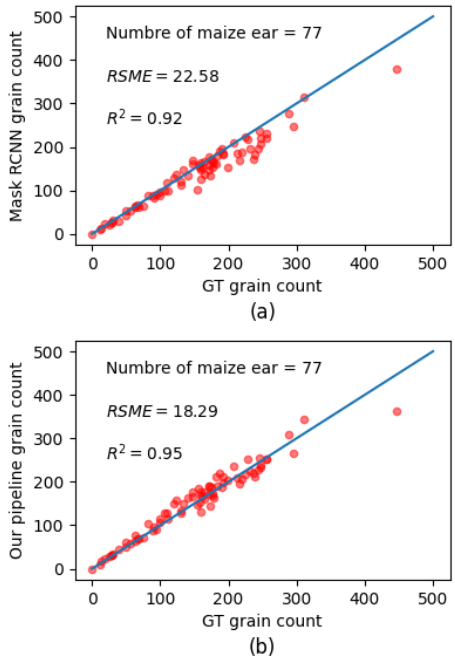

Figure 3. Comparison of the result for grain count (GT) baseline of [3] and (b) the proposed solution based on SAM.

concepts, we adopt a zero-shot learning approach. This means that we can generalize the model to segment maize ears it has never seen before, without requiring explicit annotation of each grain instance during training. Instead, we rely on simpler annotations of the resulting segments, such as image-level labels indicating if the segment is a grain or not. This significant reduction in annotation complexity not only saves time and effort but also opens up the potential for applying the pipeline to new tasks such as identifying diseases and pest damage by adding class labels on already existing segments.

References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[3] V Oury, T Leroux, Olivier Turc, Romain Chapuis, Carine Palaffre, Francois Tardieu, S Alvarez Prado, Claude Welcker, and Sébastien Lacube. Earbox, an open tool for high-throughput measurement of the spatial organization of maize ears and inference of novel traits. *Plant Methods*, 18(1):96, 2022.