Grape counting in RGB videos – comparing two instance segmentation models

Precision agriculture ’23

Ariza-Sentís, M.; Vélez, S.; Baja, H.; Valente, J.

https://doi.org/10.3920/978-90-8686-947-3_1

1. Grape counting in RGB videos – comparing two instance segmentation models

M. Ariza-Sentís[1*], S. Vélez[1], H. Baja[2] and J. Valente[1]
[1]*Information Technology Group, Wageningen University & Research, Wageningen, the Netherlands; mar.arizasentis@wur.nl*
[2]*Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, Wageningen, the Netherlands*

Abstract

The number of grape berries per bunch between pea-size and bunch closure stages provides useful information to the farmer in planning and decision-making since it is an early indicator of the final yield to be harvested. The aim of this study is to count the number of grapes per cluster by comparing two different instance segmentation models (YOLACT and Spatial Embeddings) trained on RGB videos acquired with a UAV. YOLACT tends to undercount the number of grapes, with count estimations ranging from 0% to overcounts of 148%. Nevertheless, the lowest estimation achieved by Spatial Embeddings is 30% and the highest is 116%. In general, Spatial Embeddings segments and detects berries more accurately than YOLACT.

Keywords: grape counting, instance segmentation, YOLACT, spatial embeddings

Introduction

Viticulture plays an important role in the European socioeconomic sector (Fraga *et al.*, 2012), representing 45% of the worldwide land devoted to vine cultivation (International Organisation of Vine and Wine, 2021). Farmers have to improve and adapt their farms to make their business more competitive, and precision agriculture is a growing field that is gaining much traction due to rapid research and development in the sector. Many agricultural tasks involve a lot of labour, especially in agricultural settings. In the last years, there has been an increase in interest and quite rapid developments in using artificial intelligence (AI) tools to assist agricultural tasks (Dharmaraj and Vijayanand, 2018). The usage of unmanned aerial vehicles (UAVs) has also risen in agriculture due to their ease of use for combining many tasks and integrating smart farming into a day-to-day use case on the farm. Some of the tasks that UAVs can handle are seed sowing, fertiliser spraying, monitoring (growth assessment), disease detection and mapping, and early phenotyping, among others (Ariza-Sentís *et al.*, 2022; Kim *et al.*, 2019; Vélez *et al.*, 2023). For the case of berry counting, the usage of UAVs offers a solution with the addition of cameras and other related sensors on the vehicle, which could potentially reduce labour and time.
Regarding grape berry counts, Nuske *et al.* (2011) explored the computer vision field with the Radial Symmetry Transform (Loy and Zelinsky, 2003), which employed the transform to find berry candidates in images. This is further filtered with a K-nearest neighbour classifier, a machine learning technique, which then finally performed linear regression on the detected grape berries. In a further study, Nuske *et al.* (2014) relayed the difficulty of grape berry cluster association due to touching clusters from adjacent grape clusters. Hence, a deep learning method that first detects clusters and subsequently detects berries from that cluster could potentially solve this problem.
This study aims to compare two state-of-the-art instance segmentation methods to count individual grape berries on RGB videos recorded by UAVs on a commercial vineyard.

DOI 10.3920/978-90-8686-947-3_1, © Wageningen Academic Publishers 2023

Materials and methods

The workflow followed during this research is presented in Figure 1. The procedure started with the planning of the UAV flights and their execution to collect data. Afterwards, the processing part took place by cleaning and annotating two datasets, one with grape clusters and the second one with berries. The datasets were used first to train the grape cluster detection model, using the PointTrack (Xu *et al.*, 2020) algorithm, and secondly for berry detection using two models: YOLACT (Bolya *et al.*, 2019) and Spatial Embeddings (SE) (Neven *et al.*, 2019). Finally, a qualitative assessment based on the number of berries detected versus annotated per cluster was carried out.

Data collection

The study acquired the data on a vineyard in Tomiño, Pontevedra, Spain, on the 24th of June 2021. The vineyard is located at the coordinates X: 516992.1, Y: 4644818.2 (ETRS89 / UTM zone 29N). The weather conditions were sunny with quite harsh illuminations. The grape variety grown in the vineyard is Loureiro (*Vitis vinifera*). The UAV platform used to acquire the images from the vineyard rows was the DJI Matrice 210. The platform was equipped with a DJI Zenmuse X5S camera. It was also equipped with an infrared sensor at the bottom of the UAV, which was flown slowly between the rows to capture the video in a 'drive-by' style. The dataset used in this study is made available (Ariza-Sentís and Valente, 2021).

Data setup

From the data acquisition, four rows were recorded, with each row having five to twelve videos. In total, there were 41 videos acquired from all the rows of the vineyard. Out of these 41 videos, some were annotated in the MOTS style (Voigtlaender *et al.*, 2019), following the same procedure as de Jong *et al.* (2022), and some were annotated in COCO style (Lin *et al.*, 2014). All the annotations used the CVAT software2, an annotation tool that was developed by Intel. The videos in the MOTS style contain the grape clusters annotated, meanwhile, the videos in the COCO style have the grape berries annotated. Regarding the grapes annotation, each visible berry was annotated in each cluster, so occluded berries were ignored in the annotation process.

Algorithms

To count the number of grapes per cluster, first, the grape clusters were detected using PointTrack. Afterwards, two different instance segmentation models, YOLACT and SE, were trained to identify and count the number of grapes per bunch.

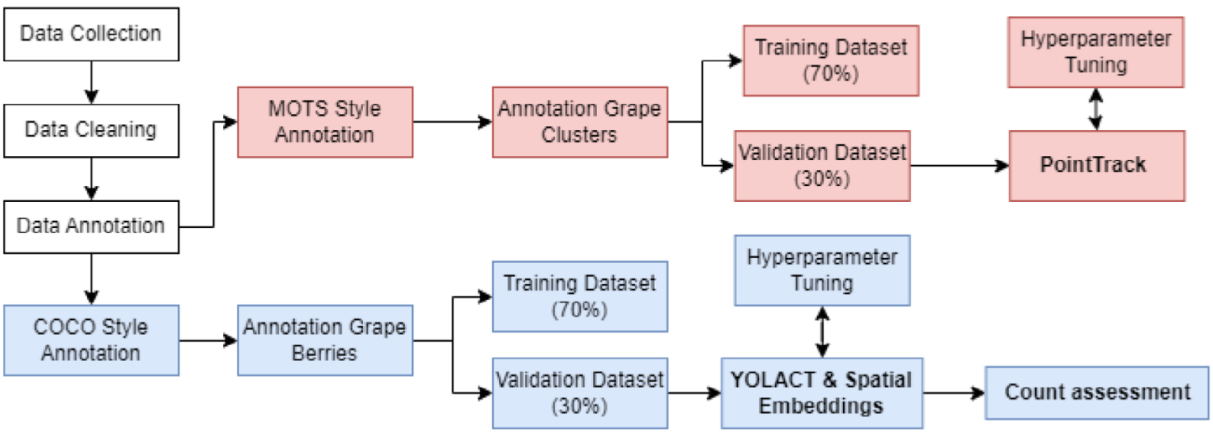

Figure 1. Flowchart of the methodology of this study. There is a common start point which consists of data collection and cleaning and afterwards, the flowchart is divided into two main branches, the red one for cluster detection and the blue one for berry detection.

YOLACT is a state-of-the-art instance segmentation algorithm that could segment objects in real-time. YOLACT achieves this by running two simpler tasks in parallel, as opposed to doing a two-step feature localization as Mask R-CNN (He *et al.*, 2017) does. The parallel tasks that YOLACT does are (1) to produce prototype masks and (2) to predict mask coefficient vectors.
Spatial Embedding is an instance segmentation method that is proposal-free, which means it does instance segmentation without bounding boxes (object detection). Embedding performs instance segmentation in only one step, compared to Mask R-CNN, which performs it in two steps. Hence, SE is much faster in segmenting images. Proposal-free methods have performed worse than their proposal-based counterparts (Hsu *et al.*, 2018; Liang *et al.*, 2018) because these methods sacrifice the accuracy of segmentations for the speed that comes with a one-stage segmentation.

Count assessment
A count assessment between YOLACT and Spatial Embeddings with grape berries was carried out. The performance of YOLACT and SE was compared to assess whether deep learning algorithms were able to accurately segment small objects such as grape berries. The assessment of grape berry detection was done to evaluate how accurately could YOLACT or SE count grape berries compared with the ground truth number of berries per cluster.

Results

The grape berry annotated dataset contained 4,905 manually annotated grape berry masks over 33 images, with a roughly 70/30 training/testing split. 27 images were used for training (containing 4408 berry masks) and 6 images for testing (including 497 berry masks), which is roughly a 70/30 training and testing split.

Ground truth count
The ground truth count of the berries was done manually. The visible berries in the obtained clusters were counted manually and then noted as the ground truth count. In some cases, it was quite easy to obtain the count of the clusters, as shown in Figure 2 (a), (b) and (c). The detected clusters are nice and clear. It has some shadows going across the cluster, but the contours and shape of each grape berry are still very clear-cut and obvious. On the other hand, Figure 2 (d)-(h) has very limited visibility of the grape berries. It is very difficult for a human to determine each grape berry. The images are very dark, and the low resolution of the clusters makes it very hard to differentiate between a grape berry or a simple curl in a leaf. Hence, the manually counted berries in these kinds of clusters were potentially erroneous. Notwithstanding, it gave insight into how the models count estimation compared to a human's ability in counting.

Berry count estimation
A visual example of how the algorithm identified the berries inside each detected cluster using YOLACT and SE is shown in Figure 3. Furthermore, Table 1 presents the count assessment of the same grape clusters as Table 1 for the two algorithms compared to the ground truth count. It can be observed that SE estimates better the berry count compared to YOLACT. Cluster numbers 3, 4b, 5, 8, and 9 are almost estimated perfectly by SE. YOLACT accurately estimates cluster number 8. YOLACT tends to underestimate the count estimates, with values ranging from 0 to 148%, whereas SE's range from 30 to 116%.

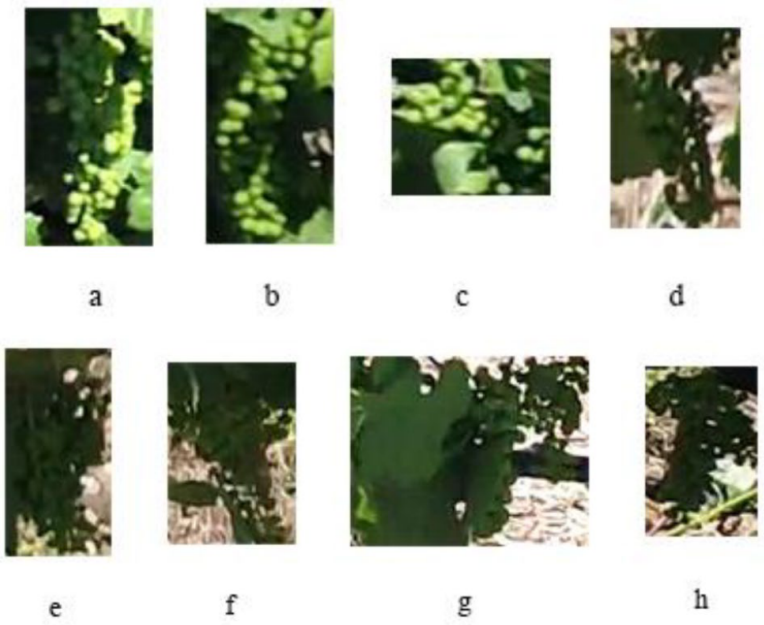

Figure 2. Grape clusters for which the number of individual berries is counted. (a) to (c) show clear clusters for which counting berries is an easy task. (d) to (h) present difficulties with berry counting due to limited visibility.

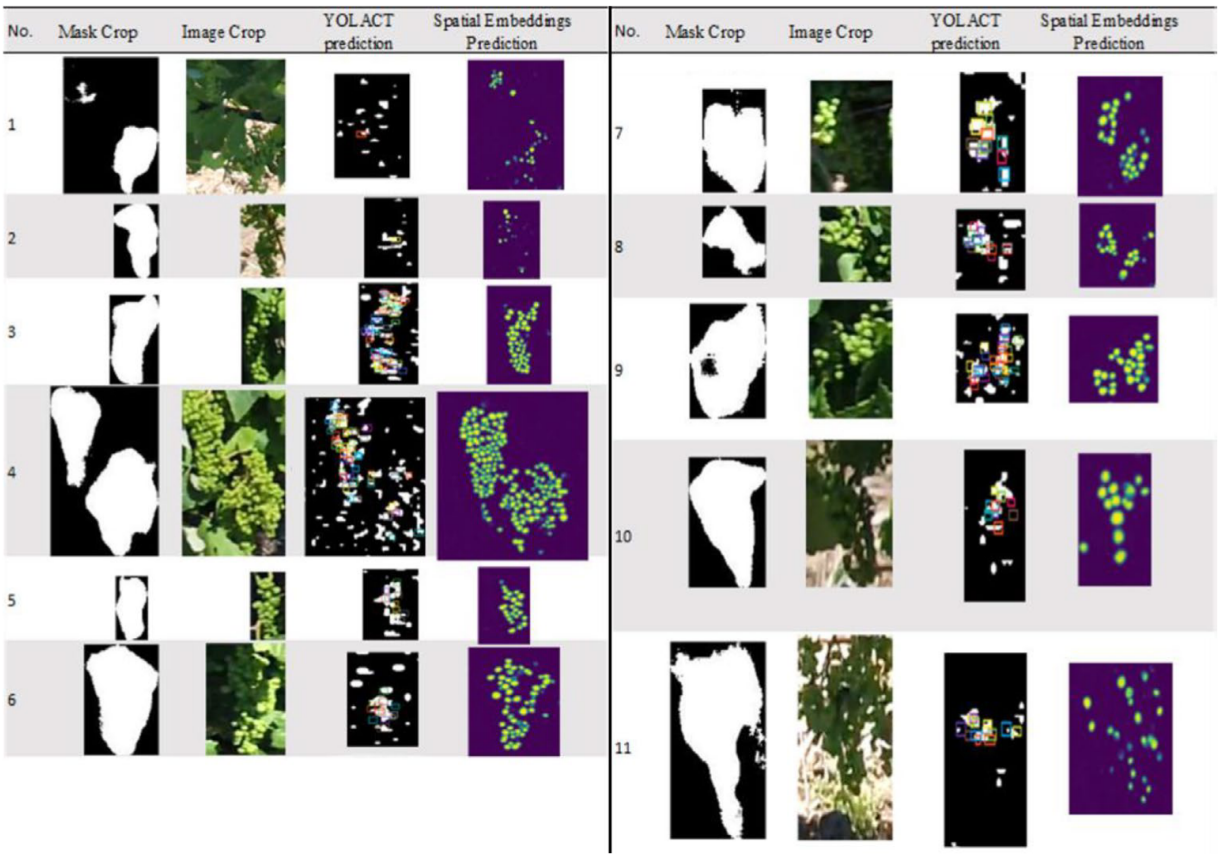

Figure 3. Grape berry count per grape cluster detected using the two algorithms: YOLACT and Spatial Embeddings, along with the ground truth count of the berries per cluster.

Table 1. Count assessment comparing the ground truth berry counts with YOLACT and Spatial Embeddings detections.

No.	Ground truth berries	YOLACT prediction	YOLACT estimated amount	Spatial embeddings prediction	SE estimated amount
1a	19	0	0%	13	68%
1b	41	1	2%	17	41%
2	33	2	6%	10	30%
3	43	35	81%	47	109%
4a	68	39	57%	79	116%
4b	56	7	13%	58	104%
5	22	10	45%	23	105%
6	45	13	29%	39	87%
7	31	11	35%	21	68%
8	15	14	93%	16	107%
9	21	31	148%	22	105%
10	21	9	43%	11	52%
11	46	13	28%	19	41%

Discussion

The grape berries were counted manually. This means, the author inspected each image belonging to each cluster and subsequently isolated the detections from that cluster to obtain the number of counts. Hence, with this method, the false positives from the images are not considered. A method that automatically obtains a berry count from the detected cluster is needed to avoid the necessity of manually counting the berries in the cluster.

YOLACT count estimations are quite chaotic, with one estimation having 0%, and on the other end, a different cluster overestimated 148% (Table 2). Meanwhile, the lowest count achieved by SE is 30%, and the highest overestimate is 116%. Overall, SE tends to overestimate the counts compared to YOLACT. However, SE yields more accurate count results than YOLACT.

The potential reason why cluster number 8 is more accurate is that the grape cluster has well-defined berries. In general, spatial embeddings could segment and detect the grape berries quite well, except in cases where the grape berries were very hard to see under the shade. It can be argued that the detected clusters that do not have visible grape berries to count are not valid, since it is also difficult for a person to count them by looking at the image. Nevertheless, those berries were included in this study to make the model robust and check how well the algorithm detected even with harsh illumination conditions.

It is important to point out that the counts of the grape berry only represent one side of the grape cluster that is visible. Hence, if the model has a 100% estimation amount, it is still an underestimation of the real grape berry counts. Nuske *et al.* (2011) addressed this issue of grape berry occlusion by explaining that occlusion is not a problem if there are few false positives, saying that the portion of visible grape berries could be used to represent the total number of berries from a cluster. Notwithstanding, their further research (Nuske *et al.*, 2014) stated that their method gave difficulty in associating berries with clusters, due to many grapes that have close adjacent clusters.

Conclusions

This study succeeded in counting the number of grape berries per cluster on UAV RGB videos by comparing two instance segmentation models, YOLACT and Spatial Embeddings. It was a challenging task due to the homogeneous environment and harsh illumination in the video sequences. YOLACT tended to underestimate the number of berries with count estimations between 0 and 148%. However, the lowest estimation for Spatial Embeddings was 30%, and the highest 116%, showing more accurate results and potential for berry counting. For future studies, RGB videos with less challenging environment conditions will be used to confirm the expected increase in count estimation of the trained algorithms.

Acknowledgements

This work has been carried out in the scope of the H2020 FlexiGroBots project, which the European Commission has funded in its H2020 programme (contract number 101017111, https://flexigrobots-h2020.eu/). The authors acknowledge valuable help and contributions from 'Bodegas Terras Gauda, S.A.' and all project partners.

References

Ariza-Sentís, M., Valente, J., 2021. Early Botrytis in Terras Gauda 2021. https://doi.org/10.5281/zenodo.5654707

Ariza-Sentís, M., Vélez, S., Baja, H., Valente, J., 2022. IPPS 2022 Conference Book 231.

Bolya, D., Zhou, C., Xiao, F., Lee, Y.J., 2019. YOLACT: Real-Time Instance Segmentation. Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9157-9166.

De Jong, S., Baja, H., Tamminga, K., Valente, J., 2022. APPLE MOTS: Detection, Segmentation and Tracking of Homogeneous Objects Using MOTS. IEEE Robotics and Automation Letters 7, 11418-11425. https://doi.org/10.1109/LRA.2022.3199026

Dharmaraj, V., Vijayanand, C., 2018. Artificial Intelligence (AI) in Agriculture. Int. J. Curr. Microbiol. App. Sci 7, 2122-2128. https://doi.org/10.20546/ijcmas.2018.712.241

Fraga, H., Malheiro, A.C., Moutinho-Pereira, J., Santos, J.A., 2012. An overview of climate change impacts on European viticulture. Food and Energy Security 1, 94-110. https://doi.org/10.1002/fes3.14

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2017. Mask R-CNN. Presented at the Proceedings of the IEEE International Conference on Computer Vision, pp. 2961-2969.

Hsu, Y.-C., Xu, Z., Kira, Z., Huang, J., 2018. Learning to Cluster for Proposal-Free Instance Segmentation.

International Organisation of Vine and Wine, 2021. State of the World Vitiviniculture Sector in 2020. Paris, France.

Kim, J., Kim, S., Ju, C., Son, H.I., 2019. Unmanned Aerial Vehicles in Agriculture: A Review of Perspective of Platform, Control, and Applications. IEEE Access 7, 105100-105115. https://doi.org/10.1109/ACCESS.2019.2932119

Liang, X., Lin, L., Wei, Y., Shen, X., Yang, J., Yan, S., 2018. Proposal-Free Network for Instance-Level Object Segmentation. IEEE Trans Pattern Anal Mach Intell 40, 2978-2991. https://doi.org/10.1109/TPAMI.2017.2775623

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dolí, P., 2014. Microsoft COCO: Common Objects in Context, in: Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, Vol 8693. Springer, Cham. pp. 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

Loy, G., Zelinsky, A., 2003. Fast radial symmetry for detecting points of interest. IEEE Transactions on Pattern Analysis and Machine Intelligence 25, 959-973. https://doi.org/10.1109/TPAMI.2003.1217601

Neven, D., De Brabandere, B., Proesmans, M., Van Gool, L., 2019. Instance Segmentation by Jointly Optimizing Spatial Embeddings and Clustering Bandwidth. https://doi.org/10.48550/arXiv.1906.11109

Nuske, S., Achar, S., Bates, T., Narasimhan, S., Singh, S., 2011. Yield estimation in vineyards by visual grape detection. Presented at the IEEE International Conference on Intelligent Robots and Systems, pp. 2352-2358. https://doi.org/10.1109/IROS.2011.6095069

Nuske, S., Wilshusen, K., Achar, S., Yoder, L., Narasimhan, S., Singh, S., 2014. Automated Visual Yield Estimation in Vineyards. Journal of Field Robotics 31, 837-860. https://doi.org/10.1002/rob.21541

Vélez, S., Ariza-Sentís, M., Valente, J., 2023. Mapping the spatial variability of Botrytis bunch rot risk in vineyards using UAV multispectral imagery. European Journal of Agronomy 142, 126691. https://doi.org/10.1016/j.eja.2022.126691

Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B., 2019. MOTS: Multi-Object Tracking and Segmentation. https://doi.org/10.48550/arXiv.1902.03604

Xu, Z., Zhang, W., Tan, X., Yang, W., Huang, H., Wen, S., Ding, E., Huang, L., 2020. Segment as Points for Efficient Online Multi-Object Tracking and Segmentation, in: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), Computer Vision – ECCV 2020, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 264-281. https://doi.org/10.1007/978-3-030-58452-8_16