Computers and Electronics in Agriculture 211 (2023) 108051

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Object detection and tracking on UAV RGB videos for early extraction of grape phenotypic traits

Mar Ariza-Sentís [a,1], Hilmy Baja [b,*,1], Sergio Vélez [a], João Valente [a]

[a] Information Technology Group, Wageningen University & Research, 6708 PB Wageningen, the Netherlands
[b] Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, 6708 PB Wageningen, the Netherlands

ARTICLE INFO

Keywords:
Viticulture
Instance segmentation
MOTS
PointTrack
YOLACT
Spatial Embeddings
UAV
Video

ABSTRACT

Grapevine phenotyping is the process of determining the physical properties (e.g., size, shape, and number) of grape bunches and berries. Grapevine phenotyping information provides valuable characteristics to monitor the sanitary status of the vine. Knowing the number and dimensions of bunches and berries at an early stage of development provides relevant information to the winegrowers about the yield to be harvested. However, the process of counting and measuring is usually done manually, which is laborious and time-consuming. Previous studies have attempted to implement bunch detection on red bunches in vineyards with leaf removal and surveys have been done using ground vehicles and handled cameras. However, Unmanned Aerial Vehicles (UAV) mounted with RGB cameras, along with computer vision techniques offer a cheap, robust, and timesaving alternative. Therefore, Multi-object tracking and segmentation (MOTS) is utilized in this study to determine the traits of individual white grape bunches and berries from RGB videos obtained from a UAV acquired over a commercial vineyard with a high density of leaves. To achieve this goal two datasets with labelled images and phenotyping measurements were created and made available in a public repository. PointTrack algorithm was used for detecting and tracking the grape bunches, and two instance segmentation algorithms - YOLACT and Spatial Embeddings - have been compared for finding the most suitable approach to detect berries. It was found that the detection performs adequately for cluster detection with a MODSA of 93.85. For tracking, the results were not sufficient when trained with 679 frames.This study provides an automated pipeline for the extraction of several grape phenotyping traits described by the International Organization of Vine and Wine (OIV) descriptors. The selected OIV descriptors are the bunch length, width, and shape (codes 202, 203, and 208, respectively) and the berry length, width, and shape (codes 220, 221, and 223, respectively). Lastly, the comparison regarding the number of detected berries per bunch indicated that Spatial Embeddings assessed berry counting more accurately (79.5%) than YOLACT (44.6%).

1. Introduction

Viticulture is relevant in many countries in Europe because of its large contribution to the European socioeconomic sector (Fraga et al., 2012). Of the 7.3 million hectares devoted to vineyards worldwide, 45% of that, 3.3 million hectares, are located in Europe (International Organisation of Vine and Wine, 2021). In the last years, with the growing importance of precision agriculture and specifically precision viticulture, worldwide winegrowers are applying the newest advances in technology to their vineyards to increase accuracy in crop monitoring, precise fertilization and pesticide application, and yield forecasting, among other activities (Matese & Di Gennaro, 2015).

To this extent, phenotyping is an important tool in agriculture, usually made through field inspections, which are time-consuming and laborious (Rahaman et al., 2015). However, advances in remote sensing, such as the usage of Unmanned Aerial Vehicles (UAVs) with multiple types of sensors onboard offer a time-saving alternative to traditional phenotyping. In this sense, computer vision techniques, such as object detection and tracking, come into play as analysis tools of the datasets acquired with the UAVs or Unmanned Ground Vehicles (UGVs).

Recent studies have focused on phenotyping, mostly on 3D point clouds. Rose et al. (2016) used a vehicle that had multiple cameras

[*] Corresponding author.
E-mail address: hilmy.baja@wur.nl (H. Baja).
[1] Authors contributed equally.

https://doi.org/10.1016/j.compag.2023.108051
Received 11 April 2023; Received in revised form 5 July 2023; Accepted 6 July 2023
Available online 13 July 2023

mounted on it, capturing 3D data by reconstructing the stereo images of the grapes using point clouds in a vineyard setting to then obtain semantic data of the berry phenotype. Milella et al. (2019) used a similar method to obtain the data using an RGB-Depth camera, which reached an accuracy of 91% of semantically segmenting grapes using the VGG19 neural network architecture (Simonyan & Zisserman, 2014). Rist et al. (2019) utilized predictive modeling and 3D field phenotyping with 360° lab scans of grape bunches as Ground Truth (GT).

Many studies have applied object detection in the field of woody crops, using one-stage or two-stage detection algorithms. For mangoes, Stein et al. (2016) deployed a Faster R-CNN detection algorithm, and Wang et al. (2019) deployed a YOLO-based detection algorithm. A study by Bargoti & Underwood (2017) looked into apples, mangoes, and almonds using Faster R-CNN. Apolo-Apolo et al. (2020) focused on orange detection using images captures from a UAV implementing Faster R-CNN. Fruit identification on canopies is difficult; occluded fruit-on-fruit and fruit-on-leaves is a scenario that simple bounding boxes may not be able to handle. Consequently, extra information is required for precise classification. Adding masks on top of the bounding boxes can significantly increase accuracy as demonstrated by Santos et al. (2020) in a study about grapes. It was found that using Mask R-CNN achieved an F1 score of 0.84 compared to 0.65 for YOLOv2 (considering an intersect over union, or IoU, of 0.5). In addition, Tian et al. (2019) found that while detecting apples, simple bounding boxes cannot precisely retrieve shape and contour information, which are important additional features for the recognition of fruits. In accordance, Jia et al. (2020) have also achieved similar results to Santos et al. while using Mask R-CNN for apple detection. A recent study from Li et al. (2023) focused on multitask-aware network for fruit bunch detection and region segmentation, obtaining promising results for assisting cherry tomato harvesting in greenhouses.

Several studies in the agricultural field about the detection and tracking of fruits have utilized the Hungarian algorithm or Kalman filter (Kalman, 1960) to track different fruits, such as seedlings, mangoes, apples, and oranges (Jiang et al., 2019; X. Liu et al., 2018; Wang et al., 2019) with positive results. However, the methods they use are not end-to-end trainable, since they add an additional tracking branch to the detection algorithm (Voigtlaender et al., 2019) (L. Yang et al., 2019). A standardized detection and tracking framework with an end-to-end trainable algorithm is needed in order to evaluate performance for different objects and research fields.

Multi Object Tracking (Leal-Taixé et al., 2015) is a popular computer vision task that has several existing state-of-of-the-art algorithms. However, the MOT framework is an object detection task, so it uses simple bounding boxes to track objects. Multi Object Tracking and Segmentation (MOTS) paved the way to much more accessible computer vision research pertaining to object tracking with instance segmentation. Voigtlaender et al. (2019) developed this computer vision task alongside the first novel end-to-end trainable MOTS detection and tracking framework, called TrackR-CNN.

There are several state-of-the-art MOTS algorithms that have been developed such as ViP-DeepLab (Qiao et al., 2020), ReMOTS (F. Yang et al., 2021), and PointTrack (Xu et al., 2020). These algorithms have been tested on KITTI MOTS, a dataset of cars and pedestrians that has been annotated with the MOTS standard. ViP-DeepLab utilizes 3D point clouds (Nguyen & Le, 2013) to predict spatial location, temporal class, and a consistent temporal location for each 3D cloud. This temporal consistency helps increase tracking performance for the algorithm. ReMOTS uses a simple self-supervising refining of tracklets from predicted masks. PointTrack learns instance embeddings by converting images into 2D point cloud representations (Neven et al., 2019). These 2D point clouds allow a tracking-by-points system that achieves quite accurate results.

There have been previous studies on MOTS for woody crops. De Jong et al. (2022) implemented additional tracking branches on TrackR-CNN. The additional tracking branches are the Kalman filter, and optical flow (Horn & Schunck, 1981). Moreover, PointTrack (Xu et al., 2020) was also implemented showing promising results and potential in apple yield estimation. Nevertheless, they also revealed many challenges in using MOTS for fruit counting and tracking, such as the homogeneity of fruits, the size of the fruits, and the challenging orchard environment. Ariza-Sentís et al. (2022) showed the potential of PointTrack algorithm for grape bunch detection and tracking using UAV RGB videos. Nevertheless, they also faced the same problems of fruit homogeneity and complex environment illumination.

A common technique used in the studies mentioned is the usage of 3D input data of grape bunches for accuracy. Santos et al. (2020) did a study about instance segmentation with grape bunches. The dataset used is a very well-made grape bunch annotated dataset called the Embrapa Wine Grape Instance Segmentation Dataset (WGISD), composed of 300 images showing around 4000 grape bunches. The WGISD is a dataset composed of images from vineyards with a trellis system-based wine grape production, taken with two cameras. Hence, the images taken were very clear and close, with a 1-meter distance from the grapes. The clear and clean images of the grapes bring questions as to whether a model trained with this dataset will be robust enough for images acquired from different platforms, e.g., UAVs. So far, there is a lack of datasets that were taken from UAVs. It is therefore interesting to test images acquired from UAVs, considering the many "all-in-one" uses they have (Tsouros et al., 2019), and their increasing research and use in agriculture (Rejeb et al., 2022).

With respect to the berry counts, Nuske et al. (2011) explored the computer vision field with the Radial Symmetry Transform (Loy & Zelinsky, 2003), which employed the transform to find berry candidates in images. This is further filtered with a machine learning technique (K nearest neighbor classifier), which then finally performed linear regression on the detected berries. In a further study, Nuske et al. (2014) relayed the difficulty of berry and bunch association due to touching bunches adjacent grape bunches. Hence, a deep learning method that first detects bunches and subsequently detects berries from that bunch could potentially solve this problem.

The main aims of this article are the following: 1) to detect and track green grape bunches and berries over UAV RGB videos recorded on a commercial vineyard, presenting challenging lighting conditions and leaf occlusion, and 2) provide phenotypic traits such as the bunch and berry length, width, and shape at a relatively early stage of bunch development, which is critical in viticulture.

2. Materials and methods

The workflow followed in this research is summarized in Fig. 1. The procedure started by acquiring the UAV RGB videos, with the posterior data cleaning and annotation with grape bunch and berry labels. Afterward, the workflow is subdivided into two main branches, the first, in red, devoted to detecting and tracking bunches, for which the PointTrack algorithm was trained. The second main branch, in blue, aimed at detecting the berries within the already identified grape bunches. Finally, the outputs of the research are the automatically-extracted the International Organization of Vine and Wine (OIV) descriptors for bunches and berries. Further details of each step are provided in the following subsections.

2.1. Data acquisition

The flights were carried out on June 28th, 2021, over four rows of a 1.06 ha and 8.1% slope commercial vineyard *Vitis vinifera* cv. Loureiro. The vineyard, property of "Bodegas Terras Gauda, S.A." is located in Tomiño, Spain (X: 516989.02, Y: 4644806.53; ETRS89 / UTM zone 29 N) (Fig. 2). The vines were planted in 1990 with a NE-SW orientation, and grafted on 196.17C rootstock. Spontaneous vegetation species, such as mint, were present between rows. The plants are trained in vertical shoot positioning and managed in vertical trellis system. The distance

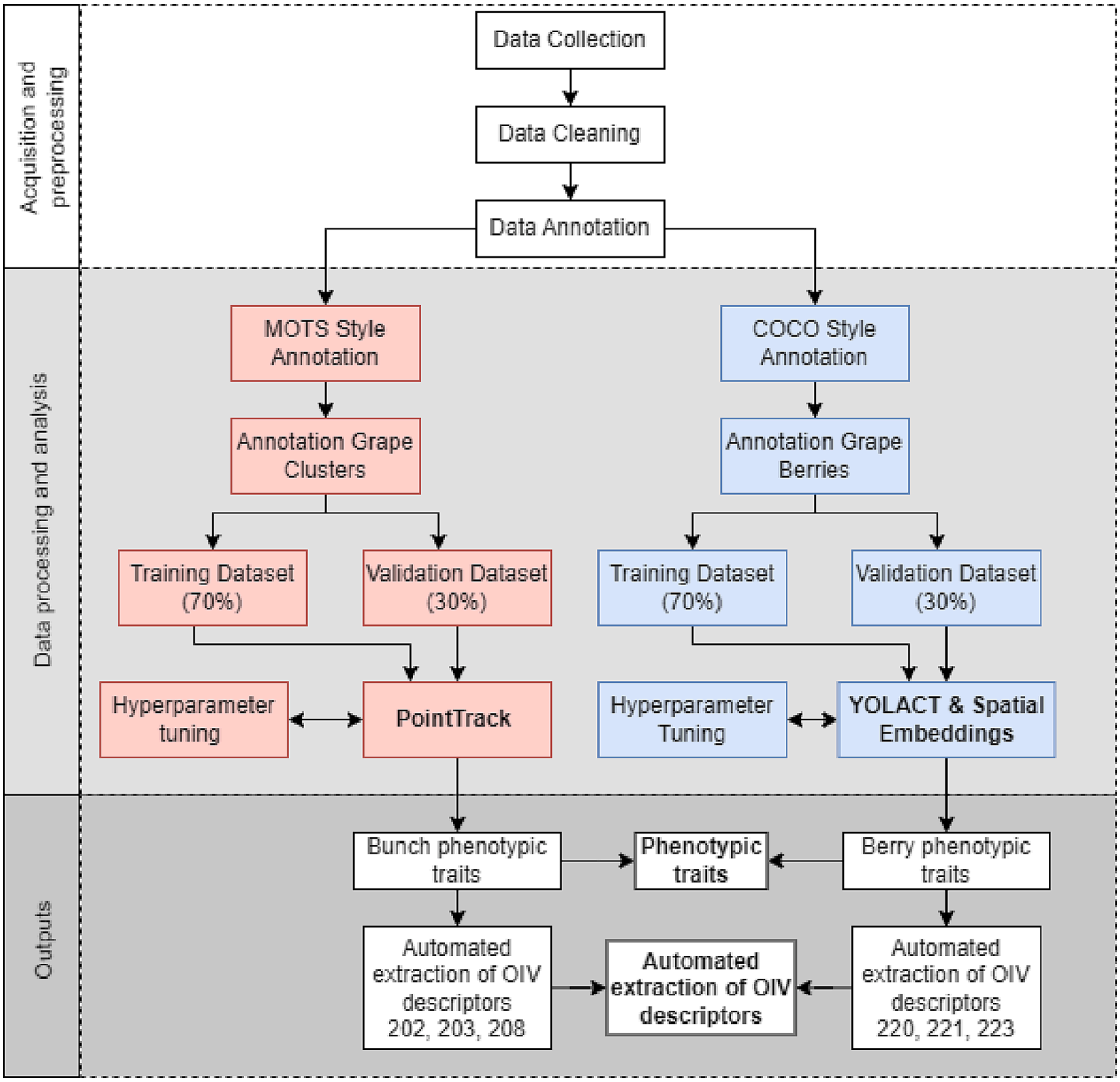


Fig. 1. Workflow diagram of this study. The common branch for all procedures, in white, consists of acquiring the dataset with the UAVs and the posterior cleaning and annotation of the bunches and berries. In red, is the grape bunch detection and tracking procedure. In blue, the detection of berries. Finally, in white again since it is a common branch, the outputs of the study, being the bunch and berry phenotypic traits and the extraction of OIV descriptors automatically.

between plants and rows was 2.5 × 3 m, respectively. The vineyard is part of the “Rías Baixas AOP” (Appellation of Origin) and hence, the vineyard is managed following the protocol and legislation of the AOP. No leaf removal was carried out and therefore, the videos present leaf occlusion.

The RGB videos were recorded using the UAV platform DJI Matrice 210 (DJI Sciences and Technologies Ltd., Shenzhen, Guangdong, China) at a flight speed of 0.7 m per second and a flight altitude of 3 m above ground level. The flights were carried out on a sunny day, with wind velocity lower than 0.5 m/s. The camera mounted on the UAV was a DJI Zenmuse X5S. The specifications of the camera are shown in Table 1.

In total, 40 videos were recorded over the four rows in which the flights were executed, reaching 7.49 gigabytes of video sequences. The rows over which the UAV flew were selected according to the ripening stage of the bunches, to have a representative sample of the ripening phase over the four rows. It can be observed in Fig. 2-right that no row 5 was flown over. The main reason was that many plants of that row were affected by esca disease and hence, the vines did not count with many bunches.

2.2. Annotation procedure

During this research two annotation types were used: MOTS to detect and track grape bunches, and COCO to detect berries (Fig. 3). All the annotations were labeled using CVAT software2 (CVAT.ai Corporation, 2022).

2.2.1. Grape bunch dataset

The grape bucnhes were annotated with a per-pixel accuracy, making sure that only the grapes were annotated, without the peduncle of the bunch. A grape bunch was annotated if it was visible on the camera, even when it was under a shade. In total, 29 video sequences were labeled to detect and track bunches, with a total number of 679 annotated frames. From those videos, an approximate 70/30 split was used for training and testing purposes. For reproducibility and to extend the research done in this field, the dataset and the MOTS labels of the grape bunches were made available (Ariza-Sentís et al., 2023).

2.2.2. Berry dataset

Each visible berry was annotated in each bunch, so occluded berries

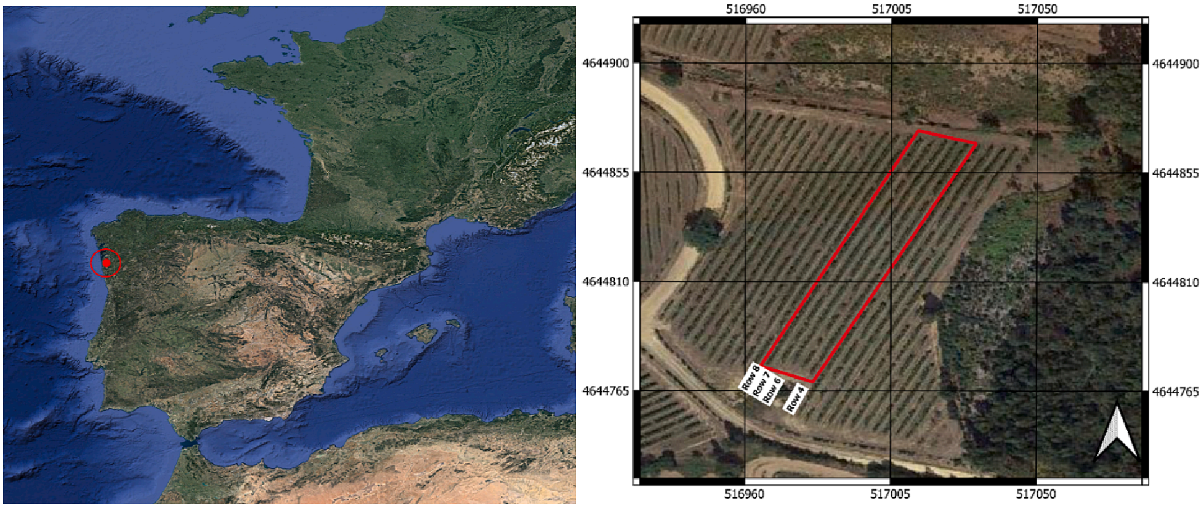


Fig. 2. Left: Location of the vineyard over the Iberian peninsula (coordinates in WGS84). Right: Location of rows 4, 6, 7, and 8 within the vineyard (coordinates in ETRS89 / UTM zone 29 N). .
Adapted from Ariza-Sentís et al. (2023)

Table 1
Camera specifications of the DJI Zenmuse X5S.

Camera characteristics	Values
Focal aperture range	f1.7 - f.16
Shutter speed	1/8000
Frame width	4096
Frame height	2160
Frame rate	59.94 frames/second

were ignored in the annotation process. The berries were not annotated across frames because they were not meant to be tracked, only detected. Hence, the dataset was composed of selected frames from the training sequences from the grape bunch dataset. The berry dataset consists of a total of 33 images including 4905 annotated berry masks. From those, an approximate 70/30 split was implemented for training/testing.

2.3. *Algorithms and model evaluation*

2.3.1. *PointTrack for bunch detection and tracking*

For the detection and tracking of the bunches, PointTrack was implemented in two steps: (1) training the instance segmentation model (Spatial Embeddings), and (2) training the PointTrack model, for instance, embedding association.

To train the instance segmentation model, an Adam optimizer (Kingma & Ba, 2017) was used with a learning rate of 5 × 10^{-5}, and the finetune training used a learning rate of 5 × 10^{-6}. These learning rates were the best values used in the original implementation of Spatial Embeddings.

To start the pre-training, the image crops of the instance annotations were generated first, so the algorithm could learn from the instance crops. In practice, the authors of PointTrack used the KINS dataset (Qi et al., 2019) that was annotated in the COCO format to produce these instance crops. However, using a custom dataset to generate these instance crops required quite some work to convert them to the right COCO format files.

Other parameters that needed to be defined at the start of the training session were (1) batch size and (2) epochs. A batch size of 20 was used to train the instance crops, considering the high number of available instance crops, and the limitation of memory. A number of at least 50 epochs is needed to let the network learn all the instance crops, in accordance with the number of instance crops available and the batch size. Xu et al. (2020) also used this number when training KITTI MOTS. However, for the grape bunches, this number was not enough to show any meaningful improvement in segmentation performance, therefore, training in increments of 200 was done, then further increased until the performance was stagnating, or overfitting was observed.

Transfer learning (Torrey & Shavlik, 2010) was used to train the

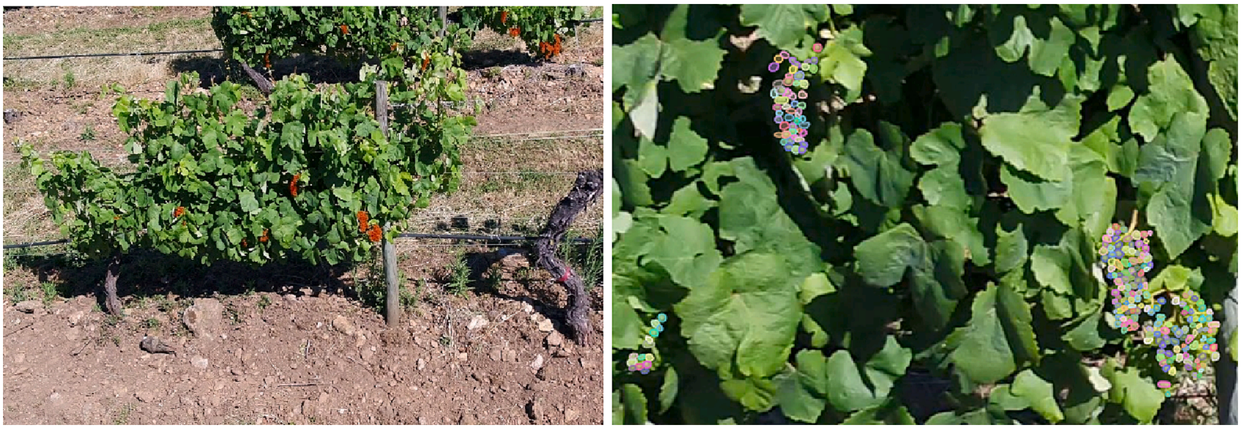


Fig. 3. Example of the annotations produced with the CVAT software. Left: grape bunch dataset used for detection and tracking. Right: berry dataset used for detection.

Spatial Embeddings model. Hence, pre-trained weights from the KITTI MOTS dataset were implemented to boost the identification of more general features such as shapes, edges, and textures (Neuhold et al., 2017).

Finally, to train PointTrack, Adam optimizer was also used with a learning rate of 2 × 10^{-3}, in accordance with Xu et al. (2020). The batch size used was 64, considering the memory limitations of the hardware used.

Two metrics were used to evaluate the tracking performance of PointTrack: MOTS (Multiple Object Tracking and Segmentation Accuracy) and sMOTSA (soft MOTSA) (Eq 1, 2).

$$MOTSA = \frac{|TP| - |FP| - |IDS|}{|M|} \tag{1}$$

$$sMOTSA = \frac{\widetilde{TP} - |FP| - |IDS|}{|M|} \tag{2}$$

where:

TP are true positives, number of masks mapped to ground truth masks (where IoU > 0.5).

TP are soft true positives, sum of the IoU of all true positives.

FP are false positives, the number of masks that are not mapped to a ground truth mask.

IDS are id switches, ground truth mask in which its ID was switched in a previous frame.

M is the number of ground truth masks.

Concerning the detection performance, MOTSP (Multiple Object Tracking and Segmentation Precision) and MODSP (Multiple Object Detection and Segmentation Precision) were calculated (Eq. (3), 4).

$$MOTSP = \frac{\widetilde{TP}}{|TP|} \tag{3}$$

$$MODSP = \frac{TP}{|TP|} \tag{4}$$

2.3.2. YOLACT and Spatial Embeddings for berry detection

The implementation of YOLACT was straightforward since it was declared in the config file that contained various configurations such as backbone network, iterations, batch size, and dataset path, among others.

The COCO detection metrics used mAP (mean average precision) as its ultimate metric, to determine how precise an instance segmentation model could predict the masks compared to a ground truth annotation. The mAP was calculated using Precision and Recall (Eq. (5), 6, 7).

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$AP = i \sum_{Recall_i} Precision(Recall_i) \tag{7}$$

where FN is the false negatives, which is the number of segments from the precision-recall curve.

2.4. Phenotyping assessment

The International Organization of Vine and Wine (OIV) has many standards for the vineyard ecosystem, which include the classification of grape bunches and berries for several purposes, such as phenotyping.. One of these standardsis a characteristic that defines phenotyping of bunches and berries and is represented with an OIV code (International Organisation of Vine and Wine, 2009). OIV numbers are useful for winegrowers to determine the intrinsic characteristics of their varieties. The OIV codes can represent quantitative or qualitative characteristics, such as the number of consecutive tendrils or the degree of resistance to a certain disease, respectively.

In this study, several OIV characters were extracted from the identified bunches and berries. For bunches, the length, width, and shape of the bunch are defined as OIV codes 202, 203, and 208, respectively. For berries, the length, width, and shape are defined as OIV codes 220, 221, and 223, respectively. Fig. 4 visually indicates how OIV codes 202 and 203 are measured in the bunch. The guidelines to measure the rest of the OIV characters can be found in the descriptor list of the OIV (International Organisation of Vine and Wine, 2009).

The OIV establishes that to determine descriptors, 10 bunches, and 30 berries should be considered and therefore, a total of ten bunches and thirty berries were considered to extract their respective OIV descriptors.

To automatize the extraction of the OIV descriptors, the length, and width of the bunch were obtained using two methods. The first one consisted of cropping the image to the mask size and extracting the length and width of the image properties to convert them to OIV 202 and 203 descriptors. However, this method considered that the bunch was oriented downwards, which was not the case in all bunches and also at an early stage of development since the weight of the bunch was still not sufficient to drive the bunch in a downwards position. Hence, a second method was considered. This second method consisted of identifying the largest distance within the mask and rotating the mask so that it had a 0-degree-angle to the vertical axis. Afterward, the width was detected as the second largest distance perpendicular to the previous one identified. To obtain the OIV descriptors of the berries, because of their spherical nature, the first method mentioned for the bunch was implemented. To validate, all metrics were compared to the ones visually assessed and measured in the video frames that were annotated, mentioned as

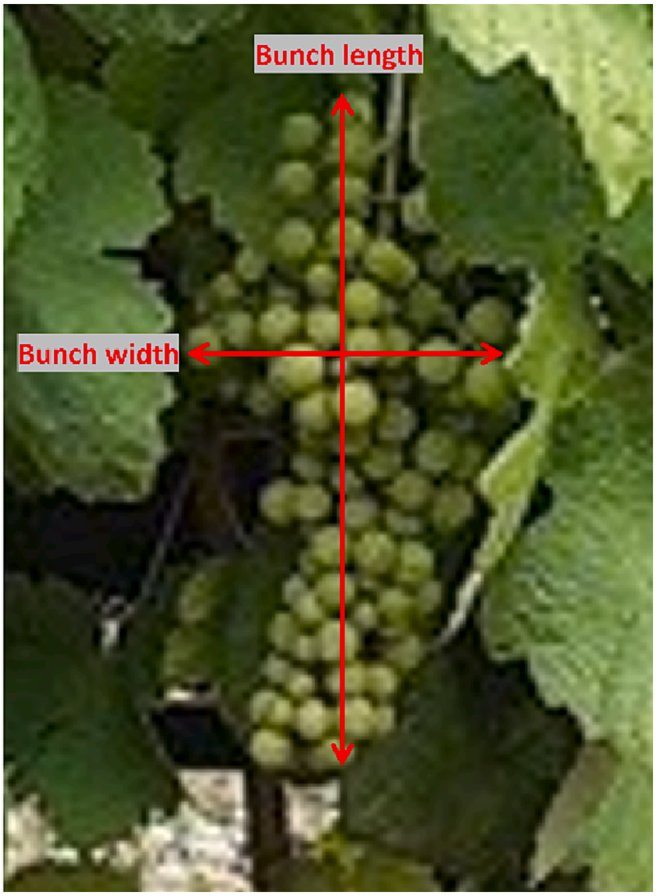


Fig. 4. OIV codes 202 (bunch length) and 203 (bunch width).

Ground Truth data in the rest of the document.

The conversion from pixels to cm was done by information on the pole width of each video sequence. Since the flights were performed in manual mode, each sequence had a slightly different length from the grape bunches, the ratio of conversion was different for each video sequence. The poles from the vineyard had a fix width of 9 cm. Hence, a ratio for each sequence was defined in Eq. (8).

$$Ratio\ cm\Big/pixel = \frac{Pole\ width\ (cm)}{Pole\ width\ (pixel)} \tag{8}$$

To assess the bunch shape, the OIV establishes that the focus should be located between the third and fourth and fifth of the bunch. To automatize it, the already downward-oriented mask was cropped into five pices and the third and fourth starting from the top were selected. Afterward, the ratio between the top and bottom width was used to classify the shape of each bunch (Fig. 5). If the ratio was below 1.1, it was classified as level 1, bigger than 1.3 was level 2. Lastly, between 1.1 and 1.3 was considered level 3.

Regarding the berry shape, a visual inspection was performed firstly to corroborate that all berries of the Loureiro variety were spherical. Because of that, they could only belong to levels 1 to 4. The ratio between the length and the width was calculated to categorize the berries. If the ratio was below 0.95 it was classified as level 1, between 0.95 and 1.05 level 2, between 1.05 and 1.25 level 3 and finally, larger than 1.25 was considered level 4. These values were selected to quantify the qualitative levels of the OIV regulations.

Finally, a comparison between the number of the berries annotated inside each bunch and the amount of berries identified by both YOLACT and Spatial Embeddings was calculated to asses the feasibility of berry counting for each algorithm.

2.5. Hardware

A high-performance computer (HPC) was used to implement the models of PointTrack, YOLACT, and Spatial Embeddings. It was equipped with two Nvidia RTX Titan GPUs with 24 GB of GDDR6 memory, running on Linux, Ubuntu 20.04.1 LTS. Furthermore, it was also equipped with 64 GB of memory and an Intel© Core™ i9-10940X CPU @ 3.30 GHz × 28 to support the training and testing process of the algorithm.

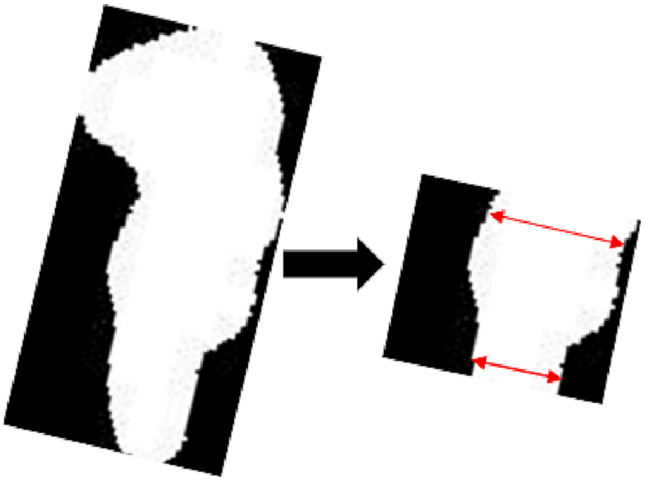


Fig. 5. Selection of the third and forth fifth of the grape bunch to obtain OIV 208. The red arrows represent the top and bottom width to calculate the ratio and classify it within an OIV level.

3. Results

3.1. Bunch detection and tracking

In total, five models were generated with PointTrack (Table 2). In addition to the crop instance, the number of epochs were changed in the PointTrack training. The training schemes are shown, with detail on how they were deployed. The last model shown in Table 2, "*BoxApp128_200 + 1200*", was trained with transfer learning from the apple dataset used by de Jong et al. (2022).

The results of the bunch tracking with the PointTrack algorithm are shown in Table 3. The tracking metrics have a negative value due to false positives and hence, for the goal of phenotyping with the masks, the detection metrics are given more emphasis than tracking metrics. Across the models, the results of detection are quite consistent, with a 66% performance. The model "*Rec128_800 + 3200*" performed the best in the MODSA metrics, achieving a 10% increase compared to the second best. The second-best model was also the model trained with rectangular shaped crop instance, "*Rec64_600 + 1200*". The model trained with transfer learning from APPLE MOTS, "*BoxApp128_200 + 1200*", was one of the worst-performing models, indicating transfer learning did not improve inference for grape bunch detection. Additionally, "*Rec128_800 + 3200*" also had the least amount of ID switches, meaning that tracking of that model worked better than the rest.

This grape bunch dataset presented a real, but very challenging environment for object detection and tracking due to the severe sunny conditions, which are common in traditional vineyard regions, such as southern Europe. Fig. 6 presents the grape bunch prediction using the model "*Rec128_800 + 3200*", the one with the highest MODSA metrics. The white rectangles indicate false positives in the form of leaves detected as bunches. As can be observed, it is also quite difficult for the human eye to distinguish between grape bunches and surrounding vegetation. Moreover, in the example provided in Fig. 6, there are multiple vine rows, complicating the algorithm's detection of bunches in the closest row.

3.2. Berry detection

The detection results on the berries are displayed in Table 4. Four models were trained for berry detection. The first two, starting with "*YO*" were trained with YOLACT, whereas the remaining two ("*SE*") were trained with Spatial Embeddings. The numbers provided after the model name indicate the total number of training epochs, which were 80.000, 1.500 and 2.300, respectively. It can be observed that the required time to train the models varied significantly from YOLACT to Spatial Embeddings models. Because of resource limitations, YOLACT models were trained with lower batch sizes (2 and 8, respectively).

Table 5 presents the mean average precision of all the models trained. The low metrics can be explained by the challenging environment that surrounds each berry (Fig. 7). The metrics shown in Table 5 indicate that SE models outperform YOLACT. There is a lack of "Box"

Table 2
Configuration of the five models generated with PointTrack.

No.	Model Name	Time (h)	Epoch	Sizes
1	Rec64_600 + 1200	~ 33	600SE + 1200FT	64 × 128 → 128 × 256
2	Rec128_800 + 3200	~ 60	800SE + 3200FT	128 × 256 → 256 × 512
3	Box80_800 + 2400	~ 48	800SE + 2400FT	80 × 80 → 160 × 160
4	Box160_1000 + 2400	~ 55	1000SE + 2400FT	160 × 160 → 320 × 320
5	BoxApp128_200 + 1200	~ 26	200SE + 1200FT	128 × 128 → 256 × 256

Table 3
Results of grape bunch tracking and detection with PointTrack. The model with the highest metrics, “Rec128_800 + 3200”, is highlighted in bold.

No.	Model Name	sMOTSA	MOTSA	MOTSP	IDS	MODSP
1	Rec64_600 + 1200	−14.37	−7.61	65.18	80	81.60
2	**Rec128_800 + 3200**	**−9.51**	**−8.17**	**66.58**	**19**	**93.85**
3	Box80_800 + 2400	−28.43	−21.97	63.47	71	82.54
4	Box160_1000 + 2400	−75.78	−66.42	64.75	129	80.08
5	BoxApp128_200 + 1200	−55.12	−46.70	65.11	155	79.92

evaluation metrics for Spatial Embeddings since the algorithm does not generate bounding boxes.. Comparing the two YOLACT models, it can be observed that the model led to lower mAP50 results due to the downsizing of the images.

Fig. 7 depicts how the model “*YO_80000_original*“ detected berries from a full-size image (4096 × 2160). The predictions contained false positives, which are depicted with the dots sprinkled throughout the image. However, there were also high quality detections, which are depicted by the small bounding boxes around the grape bunches. The main challenge of berry detection was largely due to the size of the detections compared to the images. Each berry represented 4–12 pixels, compared to the 4096 × 2160 pixels of the full image. Overall, the model was able to correctly detect the berries in the bunch. However, there was a significant number of false detections, which lowered the model’s performance.

Fig. 8 indicates the workflow that Spatial Embeddings conducted to determine the berry predictions. Spatial Embeddings was trained to detect grape bunches only in the lowest part of the canopy since it is the area in which bunches develop (Reynolds, 2015). There were false positives in the lowest part of the canopy because the algorithm predicted every pixel that had a similar seed map size as a berry. Since the environment was similar in color to the berries, and the berries had a small size compared to the whole image, there were parts of the image that were mistaken for berries.

3.3. Grape bunch phenotyping

A comparison of the bunches as seen in the video and the bunch mask as detected by the algorithm is provided in Fig. 9. The phenotyping traits of the shown bunches were automatically extracted, which are discussed in the rest of this section. It can be observed that for Bunch 3, two bunches were identified as one and hence, the phenotyping measurements obtained differed from the ground truth values.

The algorithm is able to properly detect the grape bunches, thus, the next steps correspond to extracting phenotyping traits of each bunch, starting from a comparison between their predicted and labeled size, and followed by the obtention of OIV levels for each of them. Fig. 10 shows the comparison between the ground truth measurements of each bunch dimensions (length, width, and shape) and the predicted measurements obtained with the two methods already explained (mask size and rotating the mask). It can be observed that the latter methodology proposed had a higher correlation and lower RMSE ($R^2 = 0.62$, RMSE = 32.5) compared to the method of extracting the phenotyping traits with the mask size ($R^2 = 0.47$, RMSE = 37.7).

Once it is identified that the second proposed methodology leads to better results in bunch size compared to ground truth values, the comparison between OIV levels is assessed. Since the mask-rotation methodology obtained more accurate results than the first process, that method was considered in the rest of the study. A comparison table between the ground truth OIV category of each grape bunch and the predicted levels following the mask-rotation methodology is provided in Table 6, where every row represents each single bunch shown in Fig. 9.

Table 4
Configurations of the four models trained with YOLACT (models 1 and 2) and Spatial Embeddings (models 3 and 4). The YOLACT models are trained in two stages. The SE models are trained in three stages.

No.	Model Name	Time (h)	Batch size	Epoch
1	YO_80000_original	~ 672 (26 days)	2	~ 80,000
2	YO_80000_downsized	~ 437 (18 days)	8	~ 80,000
3	SE_1500	~ 40	32	400 + 600 + 500
4	SE_2300	~ 80	32	600 + 900 + 800

Table 5
Berry detection metrics for both YOLACT and Spatial Embeddings models. In the case of Spatial Embeddings, bounding boxes are not available and hence, only mask evaluations are provided. The best model of YOLACT and Spatial Embeddings are highlighted in bold.

No.	Model Name	mAP50	
		Box	Mask
1	**YO_80000_original**	**0.68**	**0.41**
2	YO_80000_downsized	0.02	0.01
5	SE20_1500		1.82
6	**SE20_2300**		**2.42**

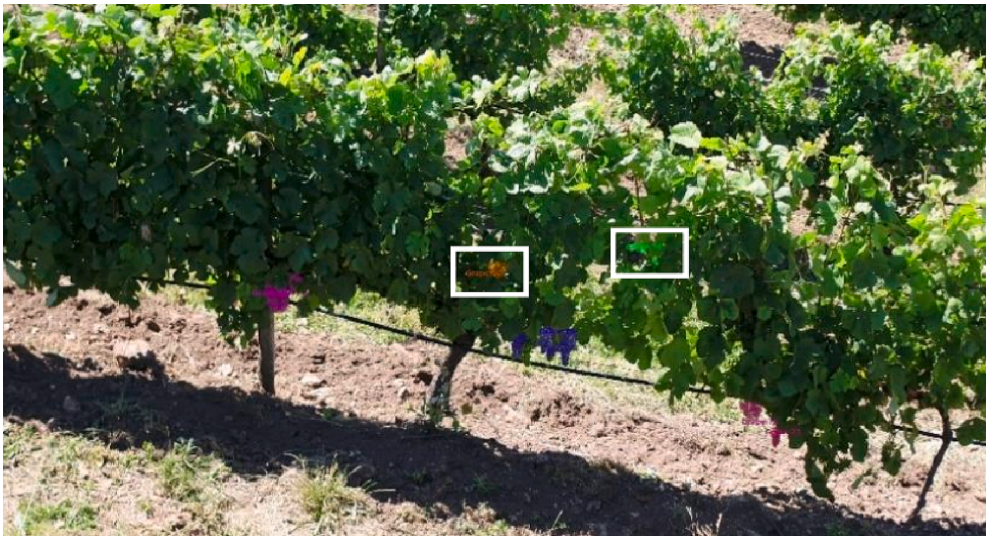


Fig. 6. Grape bunch predictions using the model Rec128_800 + 3200. The white rectangles indicate false positives.

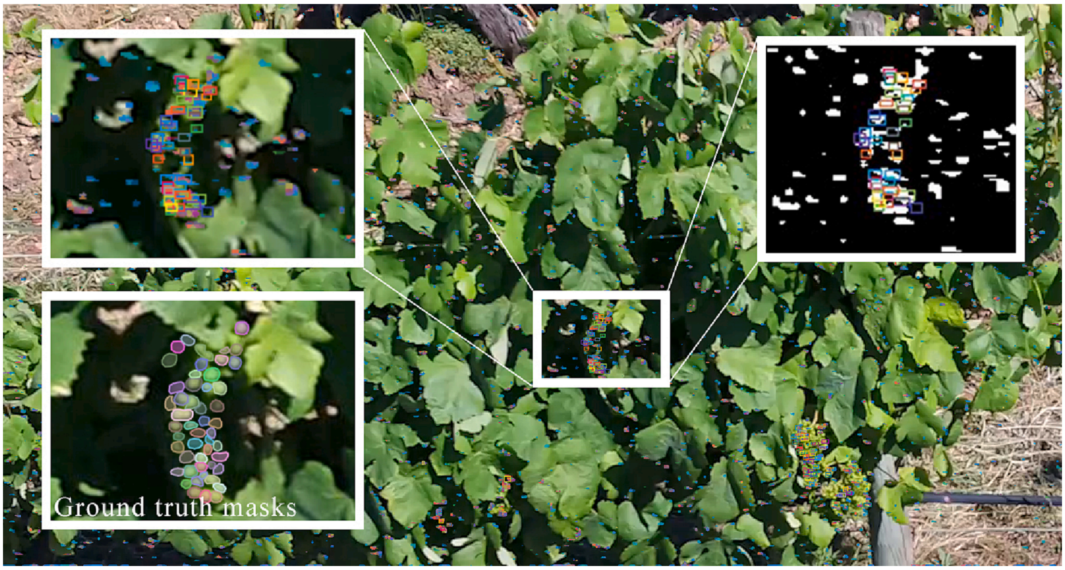


Fig. 7. YOLACT detection of berries. On the top left it is zoomed in with the predicted grape masks, and in the bottom left with ground truth masks.

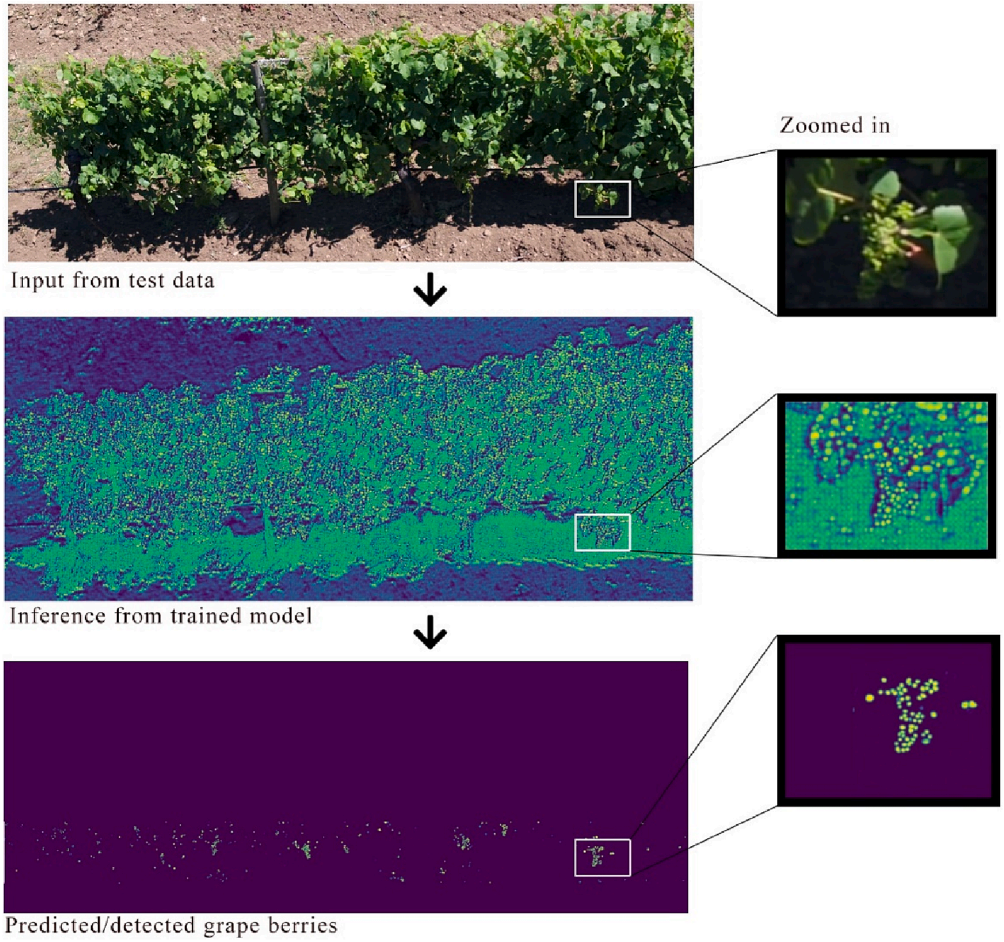

Zoomed in
Input from test data
Inference from trained model
Predicted/detected grape berries

Fig. 8. Spatial Embeddings process to properly detect berries. On the right side, zoomed-in snapshots to better visualize each of the steps that take place in a bunch to detect the berries it contains.

The GT values correspond to the number of berries that were annotated and hence counted inside each grape bunch. However, it is important to remark that only visible grapes, meaning that they are seen from a front view, were annotated.

It can be observed that for most of the cases, the bunch dimensions were properly classified, meaning that both ground truth and prediction have the same OIV level, for instance, level 3 in both cases (3/3). OIV 202 and OIV 203 have 5 levels (International Organisation of Vine and

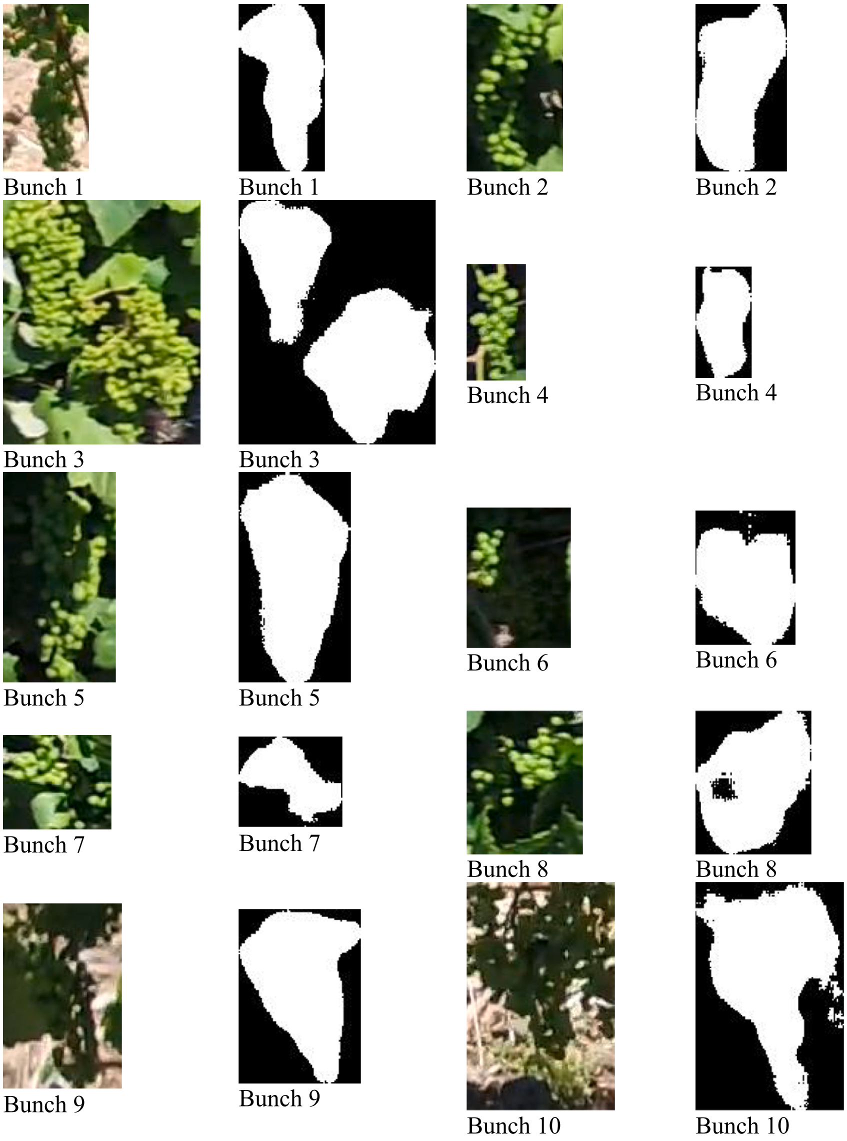


Fig. 9. Grape bunch video snapshots (first and third column) and bunch mask detections obtained with PointTrack (second and forth column).

Wine, 2009) and hence, the legend is split into 5 categories to indicate the number of in-between misclassified levels (1 to 5). For the case of OIV 208, there are only 3 levels, and therefore, a misclassification of 1 level was already penalized as if they had 2 misclassified levels, for instance being GT level 2 and the prediction category 3 (2/3) is shown in orange instead of light yellow.

3.4. Berry phenotyping

The individual berries within each of the ten grape bunches shown in Fig. 9 were manually annotated and the two instance segmentation algorithms were used to identify and count the number of berries present per bunch. Table 7 provides a visual representation of the detected berries using YOLACT and Spatial Embeddings, along with the ground truth number of berries that were manually labelled. It can be observed that, in general, Spatial Embeddings provides a better estimation of the berries inside each bunch, and also a better reconstruction of the shape of the bunch filled with berries. Bunch 3 includes two bunches in the same instance and hence, it can be observed that the berry count is divided into the left and right bunches to provide a better comparison between models. To determine how accurate the models were compared to the ground truth counts, an estimation ratio was calculated. Spatial Embeddings is the most accurate model for berry detection. All berry counts were better estimated using Spatial Embeddings compared to YOLACT. Bunches 3 right, 4, and 8 had less than 5% deviation from the ground truth values. Nevertheless, the highest accuracy for YOLACT was on Bunch 7, with a 7% deviation from the annotated value. Spatial Embeddings proved to assess the berry counting more accurately (79.5%) than YOLACT (44.6%). It has been observed that Spatial Embeddings better predicts the amount of berries per bunch and hence, those predictions were used in the rest of the study.

Once it is corroborated that the algorithm, especially Spatial Embeddings, can identify each individual berry from the detected bunches, several traits such as the length, width, and shape of each berry is extracted and compared with the ground truth values, which were manually counted in the image and then labelled. The correlation between the ground truth measurements of the number of berries and the predicted values is provided in Fig. 11. There is a high correlation ($R^2 = 0.85$) and low RMSE of 0.65, indicating that the dimensions were accurately predicted.

The OIV characteristics of each individual berry are provided in Table 8, which provides a comparison table between the ground truth OIV levels of the berries and the predicted levels. Each row represents an individual berry. In 84% of the cases, the OIV level predicted was the

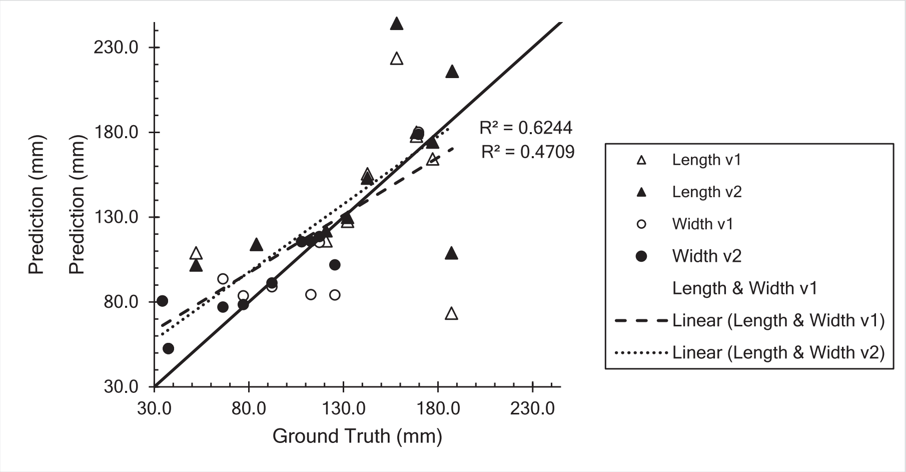


Fig. 10. Correlation between the ground truth measurements of the grape bunch dimensions (length, width, and shape) and the predicted measurements using two methodologies (1-obtaining the phenotyping traits based on the mask size, and 2- rotating the mask to the largest length of the bunch and obtaining the dimensions after the rotation).

Table 6
Comparison table between the Ground Truth and the predicted OIV levels of each grape bunch. Each row represents one grape bunch, corresponding to Fig. 9. In green, the cells that were classified as the same level for GT and prediction. The scale informs about the difference in level between ground truth and prediction. In the case of OIV 208, because there are only 3 possible levels, a difference of 1 level is classified as 2 missed levels.

	Ground Truth / Predicted levels		
	OIV 202	OIV 203	OIV 208
Bunch 1	7 / 7	5 / 5	1 / 1
Bunch 2	5 / 5	3 / 3	2 / 2
Bunch 3	5 / 7	7 / 7	1 / 1
Bunch 4	3 / 3	1 / 3	1 / 1
Bunch 5	7 / 5	5 / 5	2 / 2
Bunch 6	1 / 3	1 / 5	2 / 2
Bunch 7	7 / 3	7 / 5	3 / 2
Bunch 8	5 / 5	3 / 3	2 / 2
Bunch 9	5 / 5	5 / 5	2 / 2
Bunch 10	7 / 5	5 / 5	2 / 3

Levels
0
1
2
3
4

same as the ground truth level allocated. In the other 16% of the cases, there was only a 1- level difference between ground truth and predictions, for instance, both being classified as level 1 (1/1). Therefore, it can be concluded that Spatial Embeddings provides accurate predictions for berry OIV categorization.

4. Discussion

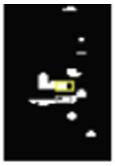
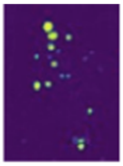
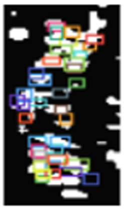
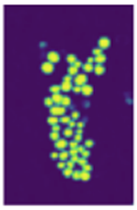
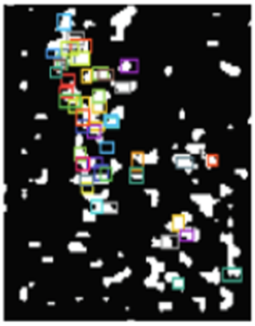
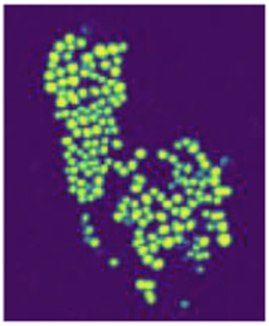
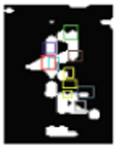
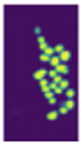
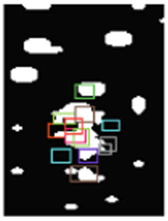
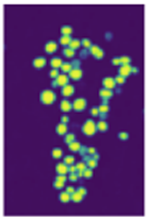
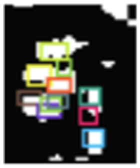
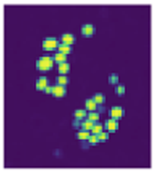
This study was successful in reaching its objectives: 1) to detect and track grape bunches using PointTrack and to detect berries within the identified bunches using two state-of-the-art instance segmentation algorithms (YOLACT and Spatial Embeddings), and 2) to extract phenotypic characteristics of the bunches and berries.

4.1. Object detection and tracking

The tracking performance for bunches was insufficient using PointTrack, a state-of-the-art MOTS algorithm. The tracking metrics were negative due to the switches in ID and/or false positives. The model used in the evaluation had a relatively low number of ID switches, which is comparable to the results of Xu et al. (2020), the original authors of PointTrack. Hence, the low metrics problem lies in the predicted false positives. Based on the review and findings of instance segmentation deep learning techniques by Hafiz & Bhat (2020), there are two reasons why this is the case: (1) the challenging environment, and (2) the large image size compared to the grape bunches.

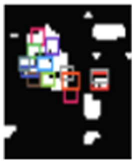
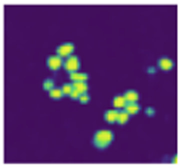
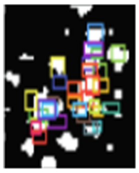
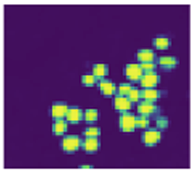
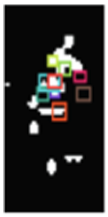
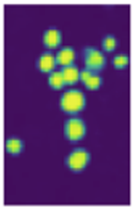
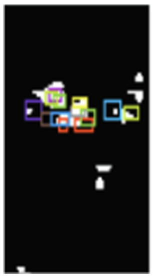
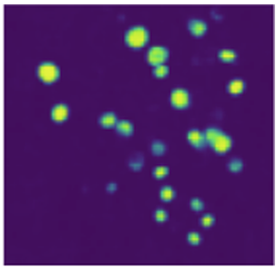
Concerning the first possibility, the results of this study differed from the findings of de Jong et al. (2022), who tested PointTrack on a dataset of apples on an orchard, which also had the datasets obtained from UAVs. One factor that is quite crucial in why the apple dataset is more accurate is due to the colour of the objects compared to its surrounding environment (Bullinger et al., 2017). Apples have a distinct red colour, moreover, the PointTrack network emphasizes a data modality that is based on differentiating the colour of the target object (Xu et al., 2020). On the other hand, the bunches in this dataset are green, with leaves that have a similar shade of green surrounding them. Many studies have brought up the importance of good visual features in distinguishing objects, which puts this as a primary importance for object detection, especially instance segmentation (Garcia-Garcia et al., 2018; Girshick et al., 2014; Zhu et al., 2012). This lack of colour distinction between the target object and the environment hindered the model from correctly detecting grape bunches, despite the many different strategies applied to

Table 7
Berry detection and count within each grape bunch. The second column indicates the number of berries that were manually annotated per bunch and the following two columns are the prediction of the number of berries per bunch using the two instance segmentation algorithms (YOLACT and Spatial Embeddings) along with the percentage of berry count accuracy.

ID	Ground Truth berry count	YOLACT count and estimated amount	Spatial Embeddings count and estimated amount
Bunch 1	33	2 (6%) 	10 (30%) 
Bunch 2	43	35 (81%) 	47 (109%) 
Bunch 3	Left: 68 Right: 56 Total: 124	Left: 39 (57%) Right: 7 (13%) Total: 46 	Left: 79 (116%) Right: 58 (104%) Total: 135 
Bunch 4	22	10 (45%) 	23 (105%) 
Bunch 5	45	13 (29%) 	39 (87%) 
Bunch 6	31	11 (35%) 	21 (68%) 

(*continued on next page*)

Table 7 (*continued*)

ID	Ground Truth berry count	YOLACT count and estimated amount	Spatial Embeddings count and estimated amount
Bunch 7	15	14 (93%) 	16 (107%) 
Bunch 8	21	31 (148%) 	22 (105%) 
Bunch 9	21	9 (43%) 	11 (52%) 
Bunch 10	46	13 (28%) 	19 (41%) 

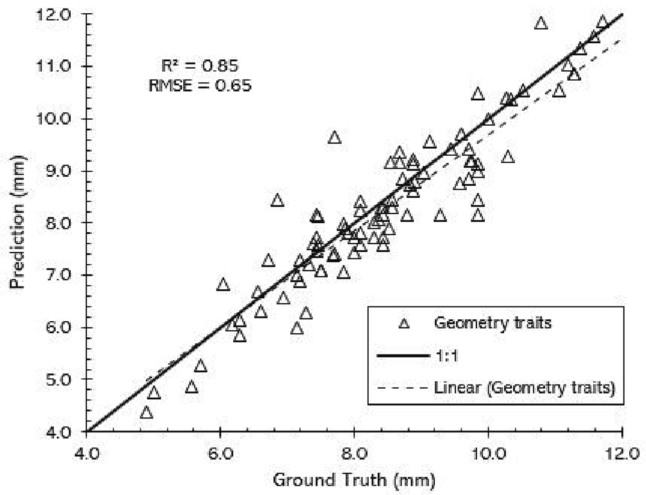
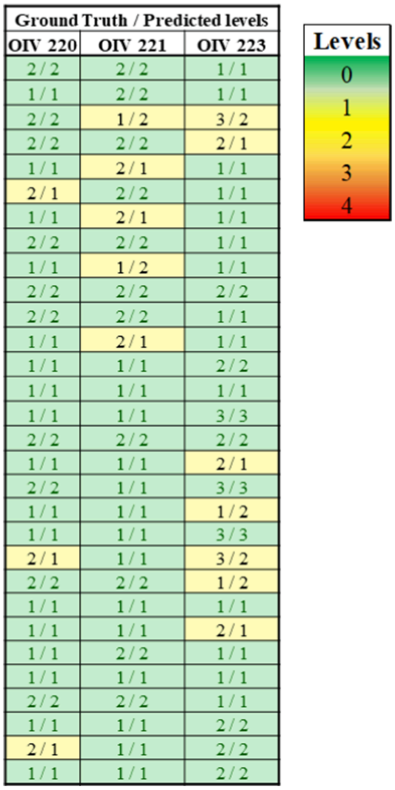


Fig. 11. Correlation between the ground truth measurements of the berry dimensions (length, width, and shape) and the predicted measurements.

train the dataset. Most of the current studies implementing deep learning instance segmentation techniques for grape bunch and berry detection use red grape varieties (S. Liu & Whitty, 2015; Torres-Sánchez et al., 2021; Zhang et al., 2022), which eases the computer vision task of detecting objects based on color instead of shape. Moreover, many studies work with leaf removal, fully observing the shape of the bunches (Nuske et al., 2014; Rose et al., 2016; Santos et al., 2020; Shen et al., 2023). Nevertheless, in this study, the grape variety is white, complicating the detection of the bunches in such a homogeneous environment. It should be tested if the trained algorithms are capable of detecting grape bunches and berries in a less challenging environment and with red grape varieties. Nevertheless, red grapes are also green before veraison, so algorithms that work properly with white grapes are completely necessary for early assessment. Another strength of this study is that the hardware used were commercial-grade UAVs and camera, easing winemakers and farm managers to use these technologies to monitor the bunch growth along the season. Lastly, working with UAVs permits the analysis of a bigger area than with UGV within the same amount of time, which is relevant in big vineyards to reduce the time window from robot inspection to result extraction and decision-making.

In this study, images are of great size compared to the target objects: bunches. As explained by Hafiz & Bhat (2020), training object inference in CNNs is still an issue, due to the inherent way the layers are trained. In fully convolutional networks, higher CNN layers have lower resolutions but more robustness in different illuminations and poses, and on the other hand, lower CNN layers have higher resolutions but are less sensitive to semantic detail (Long et al., 2015). This approach of creating

Table 8
Comparison table between the Ground Truth and the predicted OIV levels of berries. Each row includes the characteristics of an individual berry. In green, the cells that were classified as the same level for GT and prediction. The scale informs about the difference in level between ground truth and prediction. In the case of OIV 223, because there are only 3 possible levels, a difference of 1 level is classified as 2 missed levels.



Ground Truth / Predicted levels		
OIV 220	OIV 221	OIV 223
2 / 2	2 / 2	1 / 1
1 / 1	2 / 2	1 / 1
2 / 2	1 / 2	3 / 2
2 / 2	2 / 2	2 / 1
1 / 1	2 / 1	1 / 1
2 / 1	2 / 2	1 / 1
1 / 1	2 / 1	1 / 1
2 / 2	2 / 2	1 / 1
1 / 1	1 / 2	1 / 1
2 / 2	2 / 2	2 / 2
2 / 2	2 / 2	1 / 1
1 / 1	2 / 1	1 / 1
1 / 1	1 / 1	2 / 2
1 / 1	1 / 1	1 / 1
1 / 1	1 / 1	3 / 3
2 / 2	2 / 2	2 / 2
1 / 1	1 / 1	2 / 1
2 / 2	1 / 1	3 / 3
1 / 1	1 / 1	1 / 2
1 / 1	1 / 1	3 / 3
2 / 1	1 / 1	3 / 2
2 / 2	2 / 2	1 / 2
1 / 1	1 / 1	1 / 1
1 / 1	1 / 1	2 / 1
1 / 1	2 / 2	1 / 1
1 / 1	1 / 1	1 / 1
2 / 2	2 / 2	1 / 1
1 / 1	1 / 1	2 / 2
2 / 1	1 / 1	2 / 2
1 / 1	1 / 1	2 / 2

Levels
0
1
2
3
4

weights in inference directly affects the ability to train on smaller size objects. Due to the small object sizes, the resolution in lower CNN layers is smaller, which results to higher CNN layers having even smaller resolutions, leading to inferior robustness compared to inference with bigger objects.

Regarding the berry detection metrics, it can be observed in Table 5 that the mAP50 of Spatial Embeddings is 5.9 times higher than the YOLACT's value. Spatial Embeddings is a proposal-free instance segmentation model that employs a sigma function that could resize instance learning boundaries based on its value, whether it is large or small. As Neven et al. (2019) points out, they treat instance segmentation as a pixel assignment problem, a so-called context aware detector, which is done by learning a seed map that locates the objects centre to learn an optimal clustering region for each object. This is practically achieved through the training of crop instances. The instance crops let the model learn features of the instance and the surrounding background. However, this convention does not let Spatial Embeddingslearn the surrounding environment of the berries. Hence, the many false positives that come around the image is due to the model never being exposed to the background, which in small part defeats the purpose of the 'context-aware' detection system. This problem is not present for the training of other larger objects, i.e., cars, because cars are larger, and its surrounding environment is normally on an urban road. An argument could be made to increase the crop size, so more of the surrounding environment could be learned by the model. However, the training on smaller crops gave a better advantage in shorter training time. To further improve the berry detection, it would be relevant to experiment on training bigger crops, to try and reduce the occurrence of false positives.

4.2. Grape bunch and berry phenotyping

With respect to the berry count, Spatial Embeddings performed better than YOLACT, with an average count accuracy of 79.5 and 44.6%, respectively. YOLACT counts range from 0% to 148%, undercounting the berries in most of the bunches. Nevertheless, Spatial Embeddings' range from 30% to 116%. It is observed in Table 7 that berries inside bunch number 7 are as well counted by YOLACT than SE, which is due to the bunch having very well-defined berries. In general, Spatial Embeddings could segment and detect the berries quite well, except in cases where the berries are totally located under the shade. When the bunch is shaded (Grape 9 and 10), the algorithms do not properly detect the whole bunch. However, when the bunch is partly shaded (Grape 5), the algorithm performs better in the detection of the berries within the whole bunch. It can be argued that the detected bunches that do not have visible berries to count are not valid bunches, since it is also difficult for a person to count them by looking at the image. However, those bunches were included to boost robustness of the algorithm. It is observed in Table 7 that the outer shape of the detections with Spatial Embeddings is more similar to a grape bunch shape than the detections of YOLACT, which have irregular shapes in each prediction. Even if the inside detections of Spatial Embeddings might be missing, having an accurate perimeter of the bunch allows for future possibilities such as object reconstruction, which has widely been applied in medical disciplines (Lin et al., 2021; Singh et al., 2020).

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

Concerning the phenotypic traits extracted, the main method is by obtaining pixel counts of the bunch measurements and subsequently converting those numbers to centimetres. Several studies reported reasonable accuracy when phenotyping using pixel conversions to metric (Cabrera-Bosquet et al., 2016; Komyshev et al., 2017; Zhang et al., 2018). The videos taken from a UAV are stable, however, the distance taken from the rows could produce a small variation of angles between different rows. Since the vine poles in the images were taken at different angles, there is a possibility that the pixel measurements are also skewed, generating slight errors between the conversion of different rows. Seethepalli et al. (2020) describe that if the images the pixel conversions would have up until millimeter accuracy if it had at least 10 pixels/millimetre, which is not the case for this study. Hence, a range of ~ 3 cm deviation from the real measurements is expected. In future studies, this issue can be overcome by flying the UAV following a path that respects a constant distance to the target row to decrease this error.

This study offered two methodologies to automatically extract OIV descriptor 202 (bunch length) and 203 (bunch width). The first methodology consisted on extracting the length and width of the mask without considering the orientation of the bunch, with resulted in an R^2 of 0.47 compared to the GT dimensions. Nevertheless, the second methodology proposed, which involves calculating the maximum distance within the bunch mask and rotating the bunch to the vertical angle had a higher R^2 value of 0.62. This second methodology does not need to extract the dimensions of other fruits, for instance, with apples and oranges, because of their spherical nature. Nonetheless, because of the non-spherical shape of grape bunches, this methodology was relevant to increase the accuracy of those two OIV descriptors.

This study offers great potential for object detection and tracking for automatizing the extraction of bunch and berry OIV descriptors. There were two more OIV descriptors (OIV 204 – bunch density, and OIV 222 –

uniformity of berry size) that could have been obtained if there was a lower lack of berry detections. OIV 204 consists of defining 5 levels for bunch density (from very loose to very dense). With the amount of detected berries per bunch, the length and width of the berries and the bunch, some threshold can be stablished to define in which level should the bunch be classified. However, due to the missed berry detections, that OIV descriptor was not provided. Moreover, OIV 222 refers to the uniformity of the berry size and it has two levels: not uniform and uniform. With the results from OIV 223 (berry shape), it can be automatized that if the OIV 223′s level is the same for the majority of berries inside the bunch, the result is uniformity in berry size (level 2 of OIV 223). Otherwise, the bunch receives the level 1. Nevertheless, because of the missed detections, this OIV descriptor was also skipped. For future work, when the berry detection metrics are higher, these two OIV descriptors can be provided and automatized.

4.3. *Future recommendations*

Most of the current studies on object detection focus on the computer vision task itself. Nevertheless, the first important step before training the state-of-the-art algorithms is to actually acquire the datasets in the most efficient way focusing on the future purpose of the dataset. For instance, in the case of grape bunch and berry detection, the location of the bunches is crucial to properly plan the path to be followed by the UAV. The size and shape of the leaves, as well as the development of the bunches in vineyards, depends on their position along the stem since it is a function of the node in which it is positioned (Reynolds, 2015). Thus, bunches are always inserted in front of a leaf, up to the tenth bud position or even only up to the eighth, depending on the vine variety, due to various factors such as the inhibitory influence of the apical meristem (Keller, 2020, p. 2). However, since bunches are generated the year before harvest, pruning systems not only regulate vine fruitfulness but also regulate the position of bunches (Eltom et al., 2014), limiting their position to the lower part of the stem or canopy. In a context of commercial vineyards on trellises, the bunches would be located in the bottom part of the canopy, most probably in the first bottom half or the first bottom third of vegetation.

As observed in Fig. 6, there are videos which count with multiple vine rows, which complicates bunch detection and increases the ratio in size from the target object and the whole frame. Therefore, for future work, it would be important to focus on recording videos in which only one vine row is present, enhancing the recording of the bottom part of the vegetation.

Some authors have already optimized path planning for general purposes (Balampanis et al., 2016, 2017; Raptis et al., 2023; Valente et al., 2013). However, they focus on nadir flights, which have some limitations in agricultural purposes such as disease monitoring and bunch detection. Hence, future work should focus on optimizing image and video acquisition for computer vision purposes. In that way, it would facilitate grape bunch and berry detection in the areas in which the likelihood of finding bunches and berries is higher.

5. Conclusions

The objective of this study was to obtain measurements of the phenotyping traits of grape bunches and berries within detected bunches at an early stage by applying instance segmentation models with RGB videos obtained with a UAV. This study was carried out in a commercial vineyard presenting leaf-occlusion. The homogeneous background - green berries over green vegetation wall - and the high sunny conditions are challenging factors for bunch and berry detection. The proposed workflow outputs the detected bunch and berry masks along with their dimensions measurements (length, width, and shape) and the International Organization of Vine and Wine (OIV) levels of those descriptors, which are important for early assessment of yield prediction. Moreover, an evaluation of the berry count compared to the Ground Truth measurements is provided. Spatial Embeddings proved to assess the berry counting more accurately (79.5%) than YOLACT (44.6%). This work is interesting for early vineyard assessment since red and white grape are very similar at early stages. For future work, it would be relevant to focus not only on the computer vision task but also on data acquisition to optimize its collection. The dataset containing the UAV RGB videos and the MOTS grape bunch annotations used in this study are available online to boost reproducibility and future work in the field.

CRediT authorship contribution statement

Mar Ariza-Sentís: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization, Visualization. **Hilmy Baja:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Sergio Vélez:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – review & editing, Visualization, Supervision. **João Valente:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset is available in Data in Brief (mentioned in the manuscript).

Acknowledgements

This work has been carried out in the scope of the H2020 FlexiGroBots project, which has been funded by the European Commission in the scope of 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 partners of the project.

References

Apolo-Apolo, O.E., Martínez-Guanter, J., Egea, G., Raja, P., Pérez-Ruiz, M., 2020. Deep learning techniques for estimation of the yield and size of citrus fruits using a UAV. Eur. J. Agron. 115, 126030 https://doi.org/10.1016/j.eja.2020.126030.

Ariza-Sentís, M., Vélez, S., Baja, H., Valente, J., 2022. International Plant Phenotyping Symposium 2022 Conference Book. 231. https://www.plant-phenotyping.org/lw_resource/datapool/systemfiles/elements/files/e1ab35a9-3eae-11ed-9086-dead53a91d31/current/document/Conference_Book_IPPS_2022_DRUKKER.pdf.

Ariza-Sentís, M., Vélez, S., Valente, J., 2023. Dataset on UAV RGB videos acquired over a vineyard including bunch labels for object detection and tracking. Data Brief 46, 108848. https://doi.org/10.1016/j.dib.2022.108848.

Balampanis, F., Maza, I., Ollero, A., 2016. Area decomposition, partition and coverage with multiple remotely piloted aircraft systems operating in coastal regions. International Conference on Unmanned Aircraft Systems (ICUAS) 2016, 275–283. https://doi.org/10.1109/ICUAS.2016.7502602.

Balampanis, F., Maza, I., Ollero, A., 2017. Coastal areas division and coverage with multiple UAVs for remote sensing. Sensors 17 (4), Article 4. https://doi.org/10.3390/s17040808.

Bargoti, S., Underwood, J., 2017. Deep fruit detection in orchards. IEEE International Conference on Robotics and Automation (ICRA) 2017, 3626–3633. https://doi.org/10.1109/ICRA.2017.7989417.

Bullinger, S., Bodensteiner, C., Arens, M., 2017. Instance Flow Based Online Multiple Object Tracking.

Cabrera-Bosquet, L., Fournier, C., Brichet, N., Welcker, C., Suard, B., Tardieu, F., 2016. High-throughput estimation of incident light, light interception and radiation-use efficiency of thousands of plants in a phenotyping platform. New Phytol. 212 (1), 269–281. https://doi.org/10.1111/nph.14027.

CVAT.ai Corporation, 2022. Computer Vision Annotation Tool (CVAT) (2.2.0). https://github.com/opencv/cvat.

de Jong, S., Baja, H., Tamminga, K., Valente, J., 2022. APPLE MOTS: detection, segmentation and tracking of homogeneous objects using MOTS. IEEE Rob. Autom. Lett. 7 (4), 11418–11425. https://doi.org/10.1109/LRA.2022.3199026.

Eltom, M., Winefield, C.s., Trought, M.c.t., 2014. Effect of pruning system, cane size and season on inflorescence primordia initiation and inflorescence architecture of Vitis vinifera L. Sauvignon Blanc. Aust. J. Grape Wine Res. 20 (3), 459–464. https://doi.org/10.1111/ajgw.12097.

Fraga, H., Malheiro, A.C., Moutinho-Pereira, J., Santos, J.A., 2012. An overview of climate change impacts on European viticulture. Food Energy Secur. 1 (2), 94–110. https://doi.org/10.1002/fes3.14.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., Garcia-Rodriguez, J., 2018. A survey on deep learning techniques for image and video semantic segmentation. Appl. Soft Comput. 70, 41–65. https://doi.org/10.1016/j.asoc.2018.05.018.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 580–587. https://doi.org/10.1109/CVPR.2014.81.

Hafiz, A.M., Bhat, G.M., 2020. A survey on instance segmentation: State of the art. Int. J. Multimedia Inform. Retrieval 9 (3), 171–189. https://doi.org/10.1007/s13735-020-00195-x.

Horn, B.K.P., Schunck, B.G., 1981. Determining optical flow. Artif. Intell. 17 (1), 185–203. https://doi.org/10.1016/0004-3702(81)90024-2.

International Organisation of Vine and Wine. (2021). State of the World Vitiviniculture Sector in 2020.

International Organisation of Vine and Wine, 2009. OIV Descriptor List for Grape Varieties and Vitis Species (2nd ed.).

Jia, W., Tian, Y., Luo, R., Zhang, Z., Lian, J., Zheng, Y., 2020. Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot. Comput. Electron. Agric. 172 https://doi.org/10.1016/j.compag.2020.105380.

Jiang, Y., Li, C., Paterson, A.H., Robertson, J.S., 2019. DeepSeedling: Deep convolutional network and Kalman filter for plant seedling detection and counting in the field. Plant Methods 15 (1), 141. https://doi.org/10.1186/s13007-019-0528-3.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. J. Basic Eng. 82 (1), 35–45. https://doi.org/10.1115/1.3662552.

Keller, M., 2020. The Science of Grapevines, 3rd ed. Academic Press.

Kingma, D.P., Ba, J., 2017. Adam: A Method for Stochastic Optimization (arXiv:1412.6980). arXiv. https://doi.org/10.48550/arXiv.1412.6980.

Komyshev, E., Genaev, M., Afonnikov, D., 2017. Evaluation of the seedcounter, a mobile application for grain phenotyping. Front. Plant Sci. 7. https://www.frontiersin.org/articles/10.3389/fpls.2016.01990.

Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K., 2015. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking (arXiv:1504.01942). arXiv. https://doi.org/10.48550/arXiv.1504.01942.

Li, Y., Feng, Q., Liu, C., Xiong, Z., Sun, Y., Xie, F., Li, T., Zhao, C., 2023. MTA-YOLACT: Multitask-aware network on fruit bunch identification for cherry tomato robotic harvesting. Eur. J. Agron. 146, 126812 https://doi.org/10.1016/j.eja.2023.126812.

Lin, D.J., Johnson, P.M., Knoll, F., Lui, Y.W., 2021. Artificial intelligence for MR image reconstruction: an overview for clinicians. J. Magn. Reson. Imaging 53 (4), 1015–1028. https://doi.org/10.1002/jmri.27078.

Liu, X., Chen, S. W., Aditya, S., Sivakumar, N., Dcunha, S., Qu, C., Taylor, C. J., Das, J., Kumar, V., 2018. Robust Fruit Counting: Combining Deep Learning, Tracking, and Structure from Motion (arXiv:1804.00307). arXiv. https://doi.org/10.48550/arXiv.1804.00307.

Liu, S., Whitty, M., 2015. Automatic grape bunch detection in vineyards with an SVM classifier. J. Appl. Log. 13 (4, Part 3), 643–653. https://doi.org/10.1016/j.jal.2015.06.001.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation (arXiv:1411.4038). arXiv. https://doi.org/10.48550/arXiv.1411.4038.

Loy, G., Zelinsky, A., 2003. Fast radial symmetry for detecting points of interest. IEEE Trans. Pattern Anal. Mach. Intell. 25 (8), 959–973. https://doi.org/10.1109/TPAMI.2003.1217601.

Matese, A., Di Gennaro, S.F., 2015. Technology in precision viticulture: a state of the art review. Int. J. Wine Res. 69 https://doi.org/10.2147/IJWR.S69405.

Milella, A., Marani, R., Petitti, A., Reina, G., 2019. In-field high throughput grapevine phenotyping with a consumer-grade depth camera. Comput. Electron. Agric. 156, 293–306. https://doi.org/10.1016/j.compag.2018.11.026.

Neuhold, G., Ollmann, T., Bulò, S.R., Kontschieder, P., 2017. The mapillary vistas dataset for semantic understanding of street scenes. IEEE International Conference on Computer Vision (ICCV) 2017, 5000–5009. https://doi.org/10.1109/ICCV.2017.534.

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

Nguyen, A., Le, B., 2013. 3D Point Cloud Segmentation: A survey. In: 6th IEEE Conference on Robotics, Automation and Mechatronics (RAM), 225–230.

Nuske, S., Achar, S., Bates, T., Narasimhan, S., & Singh, S. (2011). Yield estimation in vineyards by visual grape detection. 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. J. Field Rob. 31 (5), 837–860. https://doi.org/10.1002/rob.21541.

Qi, L., Jiang, L., Liu, S., Shen, X., Jia, J., 2019. Amodal Instance Segmentation With KINS Dataset. 3014–3023. https://openaccess.thecvf.com/content_CVPR_2019/html/Qi_Amodal_Instance_Segmentation_With_KINS_Dataset_CVPR_2019_paper.html.

Qiao, S., Zhu, Y., Adam, H., Yuille, A., Chen, L.-C., 2020. ViP-DeepLab: Learning Visual Perception with Depth-aware Video Panoptic Segmentation (arXiv:2012.05258). arXiv. https://doi.org/10.48550/arXiv.2012.05258.

Rahaman, M.M., Chen, D., Gillani, Z., Klukas, C., Chen, M., 2015. Advanced phenotyping and phenotype data analysis for the study of plant growth and development. Front. Plant Sci. 6. https://www.frontiersin.org/articles/10.3389/fpls.2015.00619.

Raptis, E.K., Krestenitis, M., Egglezos, K., Kypris, O., Ioannidis, K., Doitsidis, L., Kapoutsis, A.C., Vrochidis, S., Kompatsiaris, I., Kosmatopoulos, E.B., 2023. End-to-end precision agriculture UAV-based functionalities tailored to field characteristics. J. Intell. Rob. Syst. 107 (2), 23. https://doi.org/10.1007/s10846-022-01761-7.

Rejeb, A., Abdollahi, A., Rejeb, K., Treiblmaier, H., 2022. Drones in agriculture: a review and bibliometric analysis. Comput. Electron. Agric. 198, 107017 https://doi.org/10.1016/j.compag.2022.107017.

Reynolds, A.G., 2015. Grapevine breeding programs for the wine industry: Traditional and molecular techniques, 1st ed. Woodhead Publishing.

Rist, F., Gabriel, D., Mack, J., Steinhage, V., Töpfer, R., Herzog, K., 2019. Combination of an automated 3D field phenotyping workflow and predictive modelling for high-throughput and non-invasive phenotyping of grape bunches. Remote Sens. (Basel) 11 (24), Article 24. https://doi.org/10.3390/rs11242953.

Rose, J.C., Kicherer, A., Wieland, M., Klingbeil, L., Töpfer, R., Kuhlmann, H., 2016. Towards automated large-scale 3D phenotyping of vineyards under field conditions. Sensors 16. https://doi.org/10.3390/s16122136.

Santos, T.T., de Souza, L.L., dos Santos, A.A., Avila, S., 2020. Grape detection, segmentation and tracking using deep neural networks and three-dimensional association. Comput. Electron. Agric. 170, 105247 https://doi.org/10.1016/J.COMPAG.2020.105247.

Seethepalli, A., Guo, H., Liu, X., Griffiths, M., Almtarfi, H., Li, Z., Liu, S., Zare, A., Fritschi, F. B., Blancaflor, E. B., Ma, X.-F., York, L. M., 2020. RhizoVision crown: an integrated hardware and software platform for root crown phenotyping. Plant Phenomics, 2020. https://doi.org/10.34133/2020/3074916.

Shen, L., Su, J., He, R., Song, L., Huang, R., Fang, Y., Song, Y., Su, B., 2023. Real-time tracking and counting of grape clusters in the field based on channel pruning with YOLOv5s. Comput. Electron. Agric. 206, 107662 https://doi.org/10.1016/j.compag.2023.107662.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 1–14.

Singh, R., Wu, W., Wang, G., Kalra, M.K., 2020. Artificial intelligence in image reconstruction: the change is here. Phys. Med. 79, 113–125. https://doi.org/10.1016/j.ejmp.2020.11.012.

Stein, M., Bargoti, S., Underwood, J., 2016. Image based mango fruit detection, localisation and yield estimation using multiple view geometry. Sensors, 16(11), Article 11. https://doi.org/10.3390/s16111915.

Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E., Liang, Z., 2019. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. Comput. Electron. Agric. 157, 417–426. https://doi.org/10.1016/j.compag.2019.01.012.

Torres-Sánchez, J., Mesas-Carrascosa, F.J., Santesteban, L.-G., Jiménez-Brenes, F.M., Oneka, O., Villa-Llop, A., Loidi, M., López-Granados, F., 2021. Grape cluster detection using UAV photogrammetric point clouds as a low-cost tool for yield forecasting in vineyards. Sensors 21 (9), Article 9. https://doi.org/10.3390/s21093083.

Torrey, L., Shavlik, J., 2010. Transfer Learning. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques (transfer-learning; pp. 242–264). IGI Global.

Tsouros, D.C., Bibi, S., Sarigiannidis, P.G., 2019. A review on UAV-based applications for precision agriculture. Information, 10(11), Article 11. https://doi.org/10.3390/info10110349.

Valente, J., Del Cerro, J., Barrientos, A., Sanz, D., 2013. Aerial coverage optimization in precision agriculture management: A musical harmony inspired approach. Comput. Electron. Agric. 99, 153–159. https://doi.org/10.1016/j.compag.2013.09.008.

Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B., 2019. MOTS: multi-object tracking and segmentation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019, 7934–7943. https://doi.org/10.1109/CVPR.2019.00813.

Wang, Z., Walsh, K., Koirala, A., 2019. Mango fruit load estimation using a video based mangoYOLO—kalman filter—hungarian algorithm method. Sensors, 19(12), Article 12. https://doi.org/10.3390/s19122742.

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 (pp. 264–281). Springer International Publishing. https://doi.org/10.1007/978-3-030-58452-8_16.

Yang, L., Fan, Y., Xu, N., 2019. Video Instance Segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 5188–5197.

Yang, F., Chang, X., Dang, C., Zheng, Z., Sakti, S., Nakamura, S., Wu, Y., 2021. ReMOTS: Self-Supervised Refining Multi-Object Tracking and Segmentation (arXiv:2007.03200). arXiv. https://doi.org/10.48550/arXiv.2007.03200.

Zhang, C., Si, Y., Lamkey, J., Boydston, R.A., Garland-Campbell, K.A., Sankaran, S., 2018. High-Throughput phenotyping of seed/seedling evaluation using digital image analysis. Agronomy 8. https://doi.org/10.3390/agronomy8050063.

Zhang, C., Ding, H., Shi, Q., Wang, Y., 2022. Grape cluster real-time detection in complex natural scenes based on YOLOv5s deep learning network. Agriculture 12 (8). https://doi.org/10.3390/agriculture12081242. Article 8.

Zhu, X., Vondrick, C., Ramanan, D., Fowlkes, C. (2012). Do we need more training data or better models for object detection? Procedings of the British Machine Vision Conference 2012, 80.1-80.11. https://doi.org/10.5244/C.26.80.