Computers and Electronics in Agriculture 232 (2025) 110025

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Yield estimation in precision viticulture by combining deep segmentation and depth-based clustering

Rosa Pia Devanna [a,1], Laura Romeo [a], Giulio Reina [b], Annalisa Milella [a,*]

[a] Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing (STIIMA), National Research Council of Italy (CNR), 70126, Bari, Italy
[b] Department of Mechanics, Mathematics and Management, Polytechnic University of Bari, Via Orabona 4, 70125, Bari, Italy

ARTICLE INFO

Keywords:
Agricultural robotics
Precision agriculture
Grape bunch detection
Grape bunch volume and weight estimation
Semantic segmentation

ABSTRACT

Grapevine phenotyping, that is the process of determining the physical properties (e.g., size, shape, and number) of grape bunches, provides valuable information for growth and health monitoring, yield estimation and efficient crop management in precision viticulture. Currently, grape bunch counting and sizing is done manually, which is labor intensive and often impractical for large-scale field applications. This paper describes a novel framework to automatically detect, count and estimate the volume/weight of grape bunches using RGB and depth data acquired in the field by a farmer robot. The proposed pipeline starts with the semantic segmentation of RGB images based on a pre-trained MANet architecture with EfficientnetB3 backbone to separate fruit from non-fruit regions. The segmented fruit mask is then projected onto the co-registered depth image to recover a depth mask, allowing for three-dimensional (3D) data association. After a pre-processing step to correct anomalies, such as corrupted and missing values, and to remove outliers, a depth gradient-based clustering algorithm is applied that detects individual grape bunch clusters. This enables the separation of adjacent and partially overlapping bunches. In addition, a method to reconstruct the whole 3D shape of a bunch is introduced, so as to provide an estimate of volume and weight. Experiments performed in a commercial vineyard in Italy are presented showing that, despite the low quality and high variability of the input images, the proposed approach is able to count grape bunch clusters with an average error of about 12% with respect to visual ground-truth and an average error less than 30% with respect to manual weight measurements. It is also shown that the processing framework can be applied to geo-referenced image sequences acquired by the farmer robot while traversing vineyard rows, thus providing an automated pipeline for the generation of high-resolution yield maps for precision viticulture applications.

1. Introduction

Yield estimation has been historically a major issue in vine and winery production systems. Knowing vineyard productivity before harvesting may help making informed decisions regarding a multitude of operations such as irrigation, nutrient management, harvest timing and workforce scheduling. Nonetheless, yield estimation largely remains an open issue. Current methods to estimate vineyard productivity include visual inspection by experts or destructive sampling of limited areas, mainly aimed at determining the number of grape bunches and measuring their shape parameters and weight. These methods are labor-intensive and do not guarantee reliable and accurate extrapolations, especially for large-scale fields. Grape bunch detection and characterization is also crucial for several other precision farming applications, such as growth monitoring, spraying, and harvesting tasks (Lytridis et al., 2023).

As a result, a vast body of recent research in the field of agricultural automation has been devoted to develop new approaches to automate fruit counting and sizing using diverse combinations of sensing and Artificial Intelligence (AI) techniques (Mohimont et al., 2022). In general, computer vision approaches applied to RGB images have demonstrated good results for bunch detection; nevertheless, the outcomes of these algorithms are significantly influenced by image acquisition settings such as background effects and light conditions, as well as inherent grape canopy characteristics, such as dense fruit distribution, leaf and branch occlusions, overlapping fruits, and other factors. To address these problems, the use of depth data may be helpful by providing complementary information such as 3D descriptors (Tao and Zhou, 2017) or for background exclusion (Fu et al., 2020).

* Corresponding author.
E-mail address: annalisa.milella@cnr.it (A. Milella).
[1] R.P. Devanna is pursuing her Ph.D. in AI for Environment and Agriculture at the University of Naples Federico II, Italy.

https://doi.org/10.1016/j.compag.2025.110025
Received 7 December 2023; Received in revised form 9 December 2024; Accepted 21 January 2025
Available online 11 February 2025

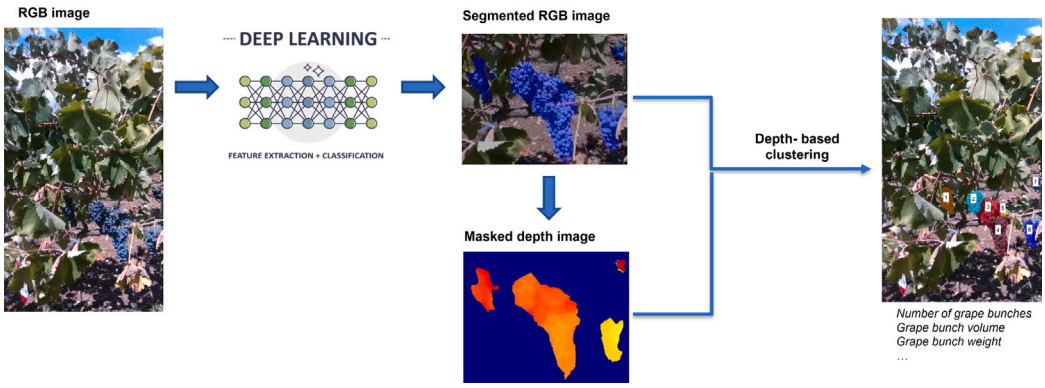

Fig. 1. Overview of the proposed approach.

This paper presents a novel framework to automatically count grape bunches and estimate their volume and weight in the field using RGB and depth data acquired by a farmer robot. A scheme of the overall approach is displayed in Fig. 1. First, RGB images are segmented using a pre-trained neural network that undergoes transfer learning for grape bunch recognition. Then, the fruit labeling mask obtained from segmentation is used to identify the areas of interest in the co-located depth image. Once the depth is appropriately pre-processed to remove outliers, individual clusters are identified based on depth gradient information. The number of detected clusters provides a measure of the number of fruits on the plant. In addition, a method to reconstruct the full 3D shape of a bunch is proposed, which subsequently allows for volume and weight estimation of individual bunches.

In the rest of the manuscript, first, related research is presented in Section 2. Materials and methods used in this study are described in Section 3, then experimental results obtained in a commercial vineyard in Southern Italy are presented in Section 4. Finally, relevant discussion and conclusions are drawn in Section 5.

2. Related research

Accurate in-field fruit detection is crucial for a number of precision farming applications, including automated fruit counting and measurement for yield estimation, robotic harvesting, crop assessment and growth monitoring (Meng et al., 2023; Chen et al., 2024). In this respect, the advent of inexpensive off-the-shelf RGB cameras and computer vision techniques have been shown to provide affordable and versatile solutions. Early work employed machine learning approaches such as Bayesian classifiers, support vector machines and clustering, using human-engineered descriptors based on color, geometric and texture features (Liu and Whitty, 2015; Dunn and Martin, 2008). Although these approaches achieve accurate results for the specific crop/data set they are designed for, they cannot be easily extended to different crops and environmental conditions. More recently, classical machine learning and pattern recognition have been replaced by deep learning techniques, which are able to deal with the high variability of object appearance in field conditions (Devanna et al., 2022). First attempts, such as (Bargoti and Underwood, 2016; Chen et al., 2017), employed Convolutional Neural Networks (CNNs) to perform semantic segmentation to separate fruit from non-fruit regions. Since then, multiple semantic networks have been developed and demonstrated for the task of agricultural image segmentation. With specific reference to vineyards, in Marani et al. (2020), a comparative study on the performance of four pre-trained network architectures, namely the AlexNet (Krizhevsky et al., 2012), the GoogLeNet (Szegedy et al., 2015), the VGG16, and the VGG19 (Simonyan and Zisserman, 2015), is presented. In Kalampokas et al. (2020) eleven CNN models, derived from combining three different types of feature learning sub-networks with five meta-architectures, are examined to separate grape bunch clusters and leaves from background as part of the sensing system of an autonomous agricultural robot. In Casado-García et al. (2022), multiple network architectures, including, among others DeepLabV3+ (Chen et al., 2018) and MANet (Li et al., 2021), are compared to find the best solution for vineyard image segmentation in low quality images acquired by a consumer-grade camera onboard an agriculture vehicle. Semi-supervised strategies are also investigated to reduce the burden of manual labeling.

The output of the semantic segmentation methods above consists of class probability maps or binary masks indicating the location of fruit pixels in the image. To translate this to agronomically relevant information, such as fruit counts, shape characteristics, and yield maps, a subsequent step must be performed to separate and characterize individual fruits. To this end, algorithms able to separate individual fruits in clusters and connect regions of the same fruit that have been disjointed due to occlusions by leaves and branches need to be applied. For instance in Bargoti and Underwood (2016), watershed segmentation and circular Hough transform are used to separate and count fruits in apple orchards.

When the fruit shape or arrangement are complex, the separation of single fruit instances may be particularly challenging. In such cases, the adoption of deep learning-based object detectors such as Faster-R CNN (Ren et al., 2015) and YOLO (Redmon et al., 2016) architectures may be helpful, as they directly provide individual objects in the form of rectangular bounding boxes, thus not requiring additional clustering strategies for counting and precision farming applications in general (Sozzi et al., 2022; Li et al., 2024). However, bounding boxes do not allow for extraction of other phenotypic parameters especially for irregularly shaped fruits like grape bunches, where rectangular boxes would not properly adjust to the berries. More recently, instance segmentation approaches have been introduced to combine fruit/non-fruit pixel with instance assignment. In Santos et al. (2020), an instance segmentation approach, namely, the Mask R-CNN network (He et al., 2017) is compared with a box-based YOLO detector showing the superior performance of instance segmentation in vineyard field applications. However, instance segmentation methods present some drawbacks, such as low detection efficiency for low-resolution objects and slow detection speed in presence of complex backgrounds (Wang et al., 2022) that may limit their application in case of low quality natural images acquired by moving platforms. In addition, instance-based image annotation for network training may be challenging in the presence of grape bunch occlusions that make particularly difficult to accurately delineating object boundaries and handling overlapping instances. As an alternative solution, end-to-end learning strategies such as (Olenskyj et al., 2022) have been also proposed to estimate grape yield from imagery without the need for hand labeling. These methods have been demonstrated effective in determining the expected productivity; nonetheless, they may not allow for direct extraction of additional phenotypic features.

While RGB images provide a flat 2D representation of the targets, sensors like lidars, stereo and RGB-D cameras are able to produce three-dimensional (3D) colored models of the crops, which can be exploited for the task of fruit monitoring and counting. 3D information may be particularly useful in case of large agglomerations of clusters and occlusions that make cluster segmentation based on 2D information only challenging even to the human annotators. Most 3D vineyard modeling works make use of point cloud data recovered from tripod mounted laser scanning systems (Keightley and Bawden, 2010; Mack et al., 2017), close-range photogrammetry and structure-from-motion (SFM) (Herrero-Huerta et al., 2015; Woo et al., 2023) with high-resolution imaging. The use of consumer-grade RGB-D cameras to recover geometric parameters and volume measurements in vineyards has been recently investigated. Preliminary tests to estimate the grape mass using a Microsoft Kinect V1 sensor are presented in Marinello et al. (2016). An application of the Kinect sensor for three-dimensional characterization of vine canopy is presented in Marinello et al. (2017), using the sensor statically positioned in the vineyard. 2D RGB and 3D RGB-D yield estimation techniques using a Kinect V1 are compared in Hacking et al. (2019), under both laboratory and field conditions, showing that the 3D approach outperforms the 2D technique only in ideal laboratory conditions. However, the Kinect uses infrared structured light, which does not perform well in sunlight conditions due to possible saturation of the projected infrared (IR) pattern, therefore the application of this sensor remains mostly limited to indoor contexts and close-range monitoring applications. Alternatively, the Intel RealSense consumer depth cameras have been adopted by a few authors. These cameras are based on infrared stereoscopy, which makes them suitable in outdoor settings (Liu et al., 2024; Condotta et al., 2020). For instance, in Milella et al. (2019) different computational geometry methods are applied and evaluated for canopy volume estimation starting from a 3D point cloud generated by an Intel RealSense R200 mounted onboard an agricultural vehicle. In Xu et al. (2023) a RealSense D455 is used for trellis grape recognition and picking point positioning of a robotic arm in grapevine harvesting operations. A method for vine yield estimation from an Intel D435 camera mounted on a mobile robotic platform is presented in Kurtser et al. (2020). It uses, first, color information to detect grape bunches based on K-means and then 3D data to estimate the grape cluster size based on different geometric modeling approaches. However, the grape detection algorithm includes a number of parameters whose values are fitted to the gathered data, thus limiting the generalization capability of the proposed workflow. Recent studies have investigated the integration of depth data to enhance semantic segmentation in viticulture. For instance, Casado-García et al. (2023) examined the fusion of RGB and depth information for segmenting vineyard elements, demonstrating that combining these data modalities improves segmentation accuracy compared to using RGB data alone, thus underscoring the potential benefits of incorporating depth information in complex agricultural environments.

A novel framework is proposed in this research, which combines RGB and depth data acquired in the field by a farmer robot equipped with a consumer-grade infrared stereoscopic sensor to detect and characterize grape bunches. The proposed system exploits highly accurate deep learning semantic segmentation to separate fruit regions from the background. By adopting a semantic segmentation approach, the requirement for precise delineation of each grape bunch instance for the training of the network can be relaxed and pixel labeling without distinguishing between different instances of the same class can be performed, making the training phase less challenging. Depth information provided by the sensor is employed to refine the segmentation results, discarding far away points that do not pertain to the foreground plant. A depth-gradient based technique is then applied to the masked depth image to separate individual clusters, acting as an additional layer of information to identify boundaries even in poor visibility conditions. The method also extends to estimating the volume and weight of each grape bunch, hence providing a comprehensive solution for grape yield prediction.

In summary, the main contributions of this research are:

- a novel framework that combines deep learning semantic segmentation and three-dimensional data to detect and count grape bunches for yield estimation and crop monitoring purposes. While much work has been done in the context of 2D image segmentation for the task of fruit detection, the integration of 3D data in the image segmentation process still represents an open challenge, especially for complex-shaped fruits like grape bunches. On the other hand, the use of 3D data proved to be fundamental to properly separate and count touching grape clusters in highly occluded scenarios, as is the case of the work presented in this research. To the best of our knowledge, the combined use of color segmentation and depth-based cluster separation has not been proposed before;
- an automated system for in-field geo-localized data acquisition using a robotic platform. The use of a robotic farmer is essential to automate image acquisition and to guarantee continuous crop production monitoring directly in the field. In this research, an integrated robotic platform is used to automatically deploy the visual sensor throughout the vineyard rows and collect geo-localized color and depth data. The platform features a tracked architecture and an articulated suspension system that guarantee efficient robot navigation on rough terrain, as well as high vibration isolation which is beneficial to reduce the influence of camera vibration on image quality. Compared to traditional manual data acquisitions and measurements, the use of an automated platform significantly relieves labor burden and time;
- once grape bunches are extracted from the scene, they can be characterized in terms of volume and weight. As such, the proposed approach encompasses segmentation, counting, and modeling issues with limited human intervention.

3. Materials and methods

3.1. Robotic acquisition system

A custom-built farmer robot, named Polibot (see Fig. 2), specifically designed to provide high mobility over challenging terrains (Ugenti et al., 2023), was adopted for in-field data gathering. It features an articulated suspension system capable of handling heavy loads, isolating vibrations, and navigating rough terrain, much like a multi-legged insect. Isolation of the vehicle body from irregularities transmitted by the ground is especially beneficial to reduce disturbances and vibrations that may generate motion artifacts during data acquisition from the onboard sensors. The control and acquisition systems are implemented under the Robot Operating System (ROS), providing a flexible and robust framework for the operation. The robot is equipped with an Intel RealSense D435 (Santa Clara, CA, USA) imaging device, which is used for the collection of visual data. It consists of a left–right infrared (IR) stereo pair, a color sensor and an IR projector. The IR stereo system has a field of view of 87(H) × 58(V) deg, maximum depth resolution of 1280 × 720 px, and frame rate up to 90 fps, with an ideal perception range of 0.3 m up to 3 m. The stereo frames are spatially calibrated and time synchronized with the color stream provided by a FullHD (1920 × 1080) CMOS camera, with nominal field of view of 69(H) × 42(V) deg and 30 fps at full resolution. As can be seen in Fig. 2, the camera was mounted on a metal frame and tilted by 90 deg to have data in portrait mode. Multiple cameras can also be integrated to extend the total field of view (Vulpi et al., 2022). The onboard sensor suite also includes a U-Blox ZED-F9P RTK-GPS providing centimeter-level accuracy position information for robot localization and image geo-referencing that allows for autonomous robot navigation throughout the vineyard rows (Galati et al., 2022).

Fig. 2. The farmer robot, used for data acquisition, equipped with a side-view Intel RealSense D435 camera (red framed) and a GPS sensor. Additional sensors shown in the picture, such as a 2D lidar and frontal cameras, are not used in this work.

3.2. Datasets

Data collection was performed in a commercial farm located in San Donaci (BR), Italy, on a field with Negroamaro (red wine) grape variety. A Google Earth view of the test site is shown in Fig. 3. The plants were acquired laterally at a frame rate of 6 Hz and at a distance ranging between 0.8 and 1 m from the row canopy, as the vehicle autonomously traversed the vineyard rows at an average speed of about 0.5 m/s.

Three sets of images were captured at three stages of the seasonal grapevine phenological development, i.e., at fruit set, before harvesting, and at the beginning of leaf fall. This choice guarantees good variability among the images of the dataset in terms of both grape color and shape. A total of about 30,000 RGB images and corresponding depth maps with 1280 × 720 pixel resolution were collected. GPS data for robot localization and image geo-referencing were also registered at a frequency of 1 Hz.

As it would have been impractical to manually label all data, a subset of 315 images was selected from the overall dataset and was used for train (90%) and validation (10%) of the segmentation network. Image selection was aimed at covering as much as possible all typical scenarios encountered during the experimental tests, including different grape maturation stages and varying sky and lighting conditions. Manual annotation of the selected images was performed to generate a segmentation ground-truth of three classes: canopy (high vegetation other than the trunk), grape bunches, and background (the remaining pixels). However, for the purpose of this work, only the grape bunch class was considered, whereas all other classes were included in the background class. The dataset used for network training is publicly available on GitHub (https://github.com/ispstiima/ECSDVineyardDataset).

Thirteen plants were selected in three different regions of the field, shown as pinpoints in the Google Earth view of Fig. 3. For these plants, 29 images, acquired just before harvesting and not previously included in the train/validation set, were selected and used for testing. Grape bunches from these plants were harvested and individually counted and weighted to provide ground-truth measurements.

3.3. Grape bunch segmentation

As a first step, grape bunches are separated from non-fruit regions (background) using a deep learning segmentation network. Specifically, a MANet architecture combined with EfficientnetB3 backbone is adopted (Li et al., 2021). The network is pre-trained based on the ImageNet (Deng et al., 2009) database and fine-tuned via transfer learning using a subset of the images acquired in the field and manually labeled to provide ground-truth masks, as described in Section 3.2. The MANet architecture combines a multi-scale feature extraction with attention mechanisms, which allows the model to focus on relevant parts of the analyzed image, even under challenging environmental conditions. Such network was selected due to its capability to handle the complex texture and variability of field images, particularly regarding grape bunches. The integration of EfficientNetB3 backbone further enhances its performance, demonstrating its efficiency in extracting discriminant information from natural images.

The choice of this network is based on its robust feature extraction ability using a multi-attention mechanism and superior context comprehension, which significantly contribute to the precision and overall efficacy of semantic segmentation. In previous work by the authors (Casado-García et al., 2022), it has been shown that this architecture provides high-accuracy pixel-wise classification results, effectively segmenting grape bunch pixels in low-quality field images acquired by a ground vehicle. The model leverages the extensive knowledge gained during pre-training to showcase robust generalization capabilities during transfer learning. In this phase, the model exploits pre-existing knowledge to improve the learning efficiency and predictive accuracy of grape bunches, ensuring precise grape bunch identification with minimal performance degradation under low-lighting conditions. The result of segmentation for a sample image extracted from the dataset used in this work is reported in Fig. 4, showing the original RGB image (a) and the result of segmentation (b), where pixels identified as belonging to the grape bunch class are indicated in blue.

3.4. Depth gradient-based clustering

An algorithm is proposed that uses depth information to separate individual grape clusters. The rationale behind the proposed approach is that color information is not sufficient to separate touching and overlapping grape bunches, whereas depth data can be exploited to identify regions where the depth gradient changes significantly and are most likely to represent grape bunch separation zones. Based on this assumption, the algorithm proceeds as follows:

1. *depth mask generation*: the color segmentation mask is projected onto the co-located depth image. This leads to a masked depth image that contains valid depth values only for pixels pertaining to the grape bunch class and null values for background regions;

2. *depth mask refinement*: the depth mask is refined to remove outliers and/or corrupted and missing values. To this end, points lying at a distance greater than a threshold are removed as possible outliers or as potentially pertaining to grape bunches lying on far away rows. Then, an interpolation strategy is applied to reconstruct and complete missing or inconsistent depth data that may occur in the depth mask, particularly at border regions. Specifically, each point with a valid value in the color mask and a null value in the depth mask is assigned the depth value given by interpolating in the nearest neighborhood;
3. *depth gradient computation*: a Canny edge detector (Canny, 1986) is applied to the depth mask to compute depth gradient and detect strong edges, which most likely define grape bunch separation lines. Detected borders are further refined via interpolation;

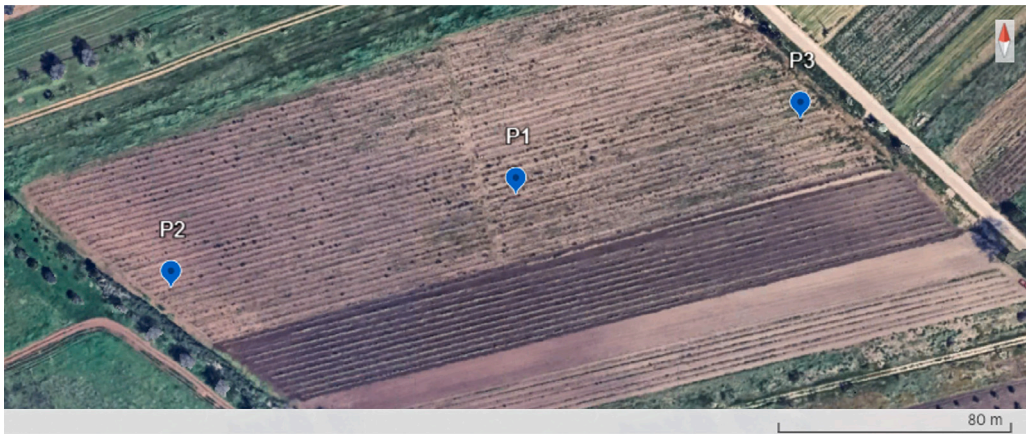

Fig. 3. Satellite image of the vineyard test field at San Donaci (BR), Italy. Pinpoints P_1 (lat.: 40.454259, long.: 17.907422), P_2 (lat.: 40.453981, long.: 17.906072), P_3 (lat.: 40.454485, long.: 17.908534) denote the position of three regions of interest identified in the field for the purpose of fruit counting.

Fig. 4. Image labeling for a sample case: (a) original RGB image; (b) segmented image. The blue points indicate pixels identified as part of grape bunches. All other pixels have been identified as belonging to the background class.

4. *grape bunch separation*: individual grape bunch instances are detected as 8-connected components.

An example of depth-based cluster separation is shown in Fig. 5 for the case of Fig. 4. In detail, from left to right, it can be seen the masked depth image, the result of the application of edge detection and the final clusters. Note that clusters whose area is below a threshold are discarded, as possible outliers or highly occluded bunches.

3.5. Volume and weight estimation

Once a grape cluster has been detected, its 3D shape can be reconstructed, using the point cloud corresponding to its segmented depth mask portion. As a first step, the 3D point cloud pertaining to the bunch region is extracted. Since the acquisition is performed from one side only of the canopy, a method to generate a complete 360 deg representation of the grape bunch cluster is needed. To this aim, only the first three quartiles (75th percentile) of the full depth value distribution are retained to remove possible outliers. Then, the resulting point cloud is mirrored over a plane that is parallel to the focal plane of the camera and passing through the maximum distance point (Fig. 6). The reflection process relies on the assumption that the grape bunch clusters are symmetric with respect to the identified plane, an approximation that was found reasonable given the overall shape and structure of the grape bunch clusters.

The point cloud obtained from the reflection process is finally modeled using a computational geometry approach, namely the alpha-shape (Edelsbrunner et al., 1983). This technique allows for transforming an unorganized point cloud into a geometrically defined shape based on 3D Delaunay triangulation. The volume of the grape bunch is subsequently calculated as the volume of the 3D alpha-shape generated from the grape bunch point cloud. This volume estimation process treats the grape bunch cluster as a solid object. This approximation is justified by the high density of berries within a cluster. Under the hypothesis that the density of a grape bunch is uniform and the grape bunch cluster can be modeled as a solid object, the weight W of each grape bunch cluster can be finally inferred as:

$$W = \rho V \tag{1}$$

where V is the estimated bunch volume and ρ is the average density of the bunch that can be recovered either from literature or from field measurements.

4. Results

In this section, experimental tests performed using the dataset described in Section 3.2 are reported. Specifically, first, the results obtained for a subset of selected plants are presented and compared with ground-truth manual measurements. Then, the application of the counting approach to automatically generate a yield map along a vineyard row using the robotic platform is discussed.

4.1. Individual plant analysis

The proposed approach was applied to 13 distinct plants, chosen in three parcels of the experimental farm. Specifically, four plants were selected in parcels P_1 and P_3 and five in P_2. Given the camera setup and field of view and the plant horizontal extension ranging from approximately 1.0 to 1.5 m, a single image was not sufficient to frame the entire plant. Instead, for each plant, two to three frames acquired every 1 s, were selected among the images acquired before harvesting, for a total of 29 images and were used for the purpose of fruit counting and sizing. This time span was chosen so as to minimize the overlap among the frames pertaining to the same vine, thus avoiding double counting of grape bunches.

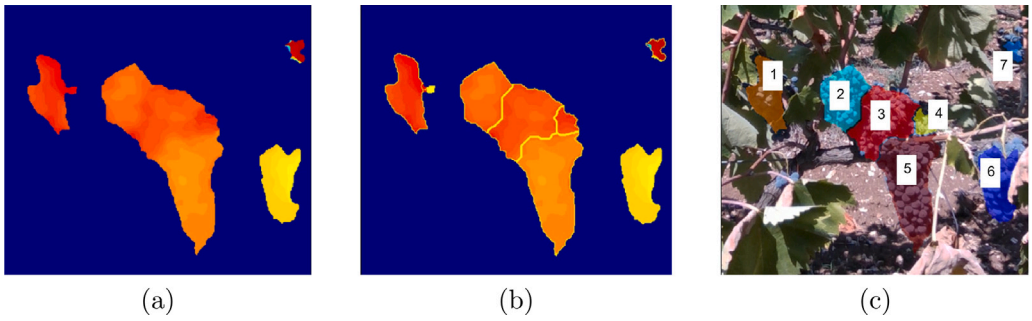

Fig. 5. Result of the depth-based clustering algorithm for the sample case of Fig. 4. From left to right: (a) masked depth image; (b) edge detection for grape bunch boundaries identification; (c) detected grape bunch clusters, with each cluster represented with a different color and numerically labeled.

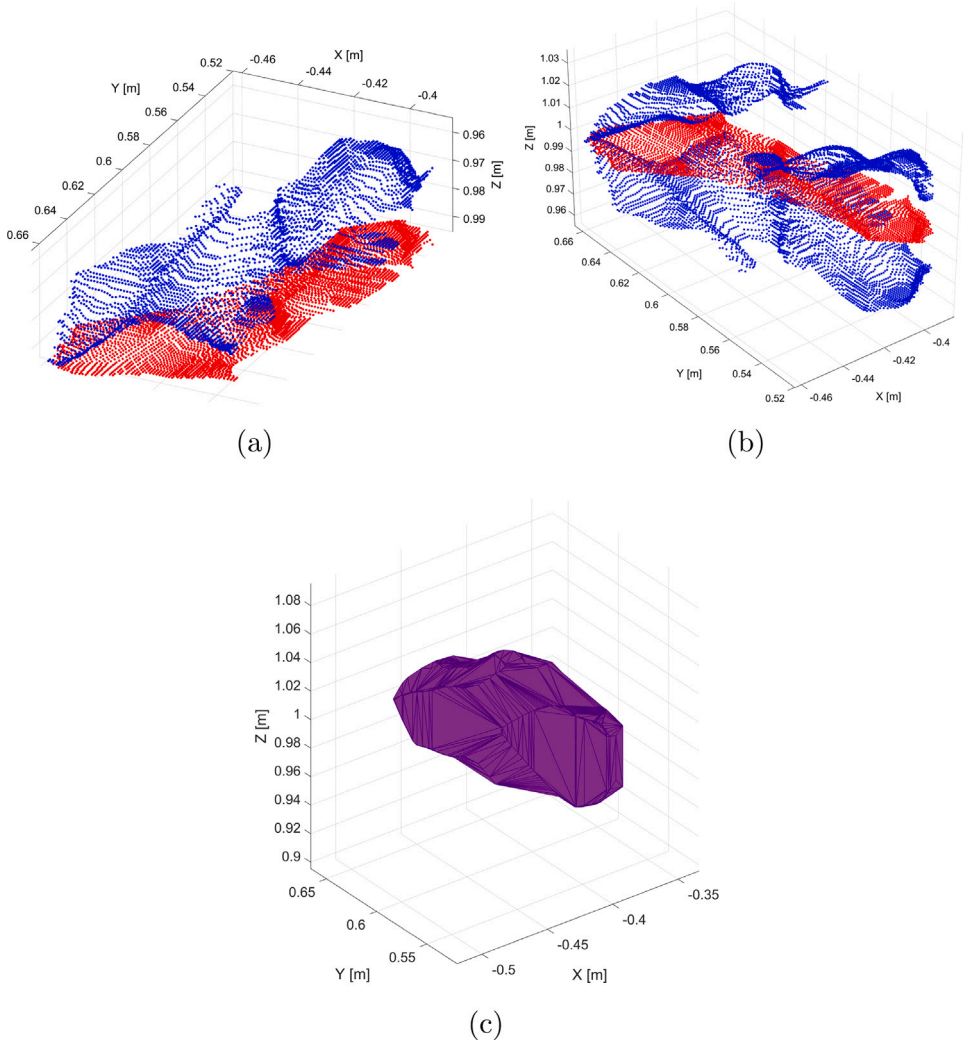

Fig. 6. Stages of 3D grape bunch cluster reconstruction. (a) Original point cloud of the grape bunch cluster (blue points) with the plane of reflection at the 75th percentile depth (red points); (b) final point cloud representation of the grape bunch cluster after reflection; (c) alpha-shape based bounding volume of the blue points ($\alpha = 0.05$).

4.1.1. Semantic segmentation

The MANet architecture with EfficientNetB3 backbone was used to segment RGB images into two classes, namely grape bunches and background. First, the network was fine-tuned using transfer learning from the 315 images of the whole dataset. Fig. 7 presents the convergence between the training and the validation curves. The graph presents the loss function over batches, which gives a quantitative measure of how well the predictions match the actual segmentation. Lower loss values indicate better performance. More specifically, the Training Loss (in blue) shows a sharp decrease initially, which is typical as the model begins to learn from the data. As training progresses, the loss continues to decrease but at a slower rate, suggesting that the model is starting to converge. The Validation Loss (in orange) decreases alongside the Training Loss, which indicates that the model is generalizing well to new data. Such convergence indicates that there is no sign of overfitting during the training. The loss values reaching a plateau suggest that

Table 1
Percentage values of Accuracy, IoU and mean F-score metrics of the semantic segmentation obtained using MANet architecture, with EfficientnetB3 as backbone, for 29 images acquired from 13 selected plants.

	Accuracy [%]	IoU [%]	Mean F-score [%]
Grape Bunches	80.44	68.94	80.29
Background	99.52	98.95	94.04

additional training may not lead to significant improvements and that the MANet architecture has captured the essential patterns required for the segmentation task.

Once trained, the network was used to segment the 29 test images of the plants of interest. Results for some sample cases are shown in Fig. 8, with the original images depicted in the left column and the results of segmentation in the central column. Blue pixels denote points belonging to the grape bunches, whereas all other pixels have been classified as pertaining to the background class. The classifier performance for the semantic segmentation was validated considering the metrics of Accuracy, Intersection over Union (IoU), and F-Score. Such metrics were selected since they allow a comprehensive and clear evaluation of the segmentation task. Specifically, the Accuracy measures the percentage of correctly classified pixels within the image, thus comparing the predicted segmentation to the ground truth. It can be defined according to the following equation:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

where TP represents the True Positive pixels, TN represents the True Negative pixels, FP represents the False Positive pixels, and FN represents the False Negative pixels.

As for the IoU metric, it measures the overlap between the predicted segmentation and the ground truth, and it is computed as follows:

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{3}$$

Finally, the F-score evaluates the average performance of the segmentation model across the classes, and it is defined by the following equation:

$$\text{F-score} = \frac{2TP}{2TP + FP + FN} \tag{4}$$

The F-score is particularly useful for managing the imbalance between the number of positive and negative pixels.

Table 1 reports the classifier performance for the semantic segmentation. The network achieves a mean F-score of 80.29% for the grape bunch class and 94.04% for the background class, despite the low quality and high variability of the images. Such values are comparable with the ones in the literature (Casado-García et al., 2022), and thus they can be considered a reliable input to the successive grape bunch clustering phase. As for the IoU, it is evident that the model does not have any problem in properly localizing the background within the images, reaching a value of 98.95%. On the other hand, the grape bunch class presents an IoU of 68.94%. This is mainly due to uncontrolled illumination variations and occlusions. Both issues have an impact on the segmentation outcomes, especially the lighting, since it can cause color shifts in the images. These alterations may be particularly critical given the complex form and texture of the grape bunches. To confirm such a statement, the IoU was re-calculated, excluding those images that were significantly affected by uneven lighting conditions or shadows, thus considering a new test set composed of 15 images. Fig. 9 depicts an example of such images. In this case, as expected, the IoU reached a higher score of 73.13% for the grape bunch class.

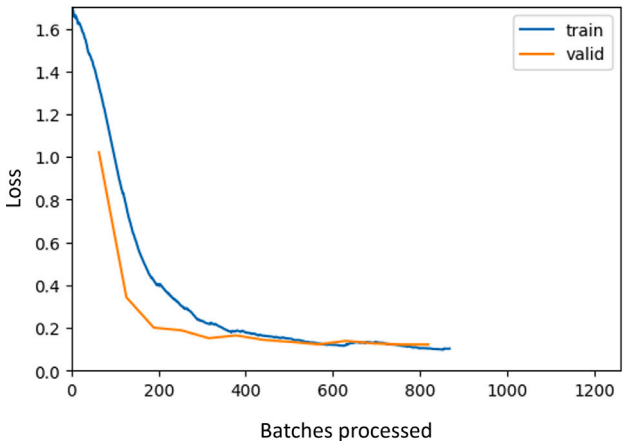

Fig. 7. Convergence between the training and the validation curves, while fine-tuning the MANet architecture with EfficientNetB3 as backbone. The graph shows the decline in loss values throughout the processing of the batches until convergence, which has been achieved after 819 batches. Such convergence indicates successful model optimization on grape bunch and background classification tasks.

4.1.2. Clustering and counting

The segmented plant images were used as input to the depth-based clustering algorithm to separate individual grape bunch instances from the overall grape bunch class. Some sample cases are shown in the right column of Fig. 8. Numerical results obtained for all plants are reported in Table 2 and in the bar plot of Fig. 10. The number of grape bunches per vine obtained from the proposed method (N_{est}) is compared with the ground-truth number of grape bunches manually counted after harvesting (GT). In addition, grape bunch clusters were counted by visual inspection of the plant images for direct comparison of the algorithm results with the expected visual ground-truth (VGT). Table 2 also displays the absolute counting errors relative to ground-truth (e_{GT}) and visual ground-truth (e_{VGT}) for each vine, which were computed as:

$$e_{GT} = |N_{est} - GT| \tag{5}$$

$$e_{VGT} = |N_{est} - VGT| \tag{6}$$

Percentage errors of estimated counts with respect to GT and VGT can be also recovered as:

$$e_{GT\%} = \frac{e_{GT}}{GT} \times 100 \tag{7}$$

$$e_{VGT\%} = \frac{e_{VGT}}{VGT} \times 100 \tag{8}$$

Mean values and standard deviations of the absolute and percentage errors are reported in Table 3. Results show that for most plants, the estimated count is close to the visual ground truth with an average absolute error of 1.5 clusters and standard deviation of 1.0 corresponding to an average percentage error of 12.2% with standard deviation of 8.6%. A higher discrepancy is observed compared to the field ground truth with an absolute average error of 6.7 and standard deviation of 4.8 on the grape bunch count, corresponding to a mean percentage error of 33.9% and standard deviation of 18.4%. In Fig. 11 two scatter plots display the dispersion of the estimated cluster counts relative to the GT and VGT counts. Count estimates correlates well with the visual ground truth reaching a correlation coefficient R^2 of 0.79, whereas poor correlation is found with the manual ground truth (R^2= 0.073), further confirming the discrepancy between post-harvesting counting and visual counting in images. This might be explained by a high occurrence of bunches visually occluded by the canopy. In addition, it should be noted that acquisition was performed only from

Fig. 8. Sample results obtained from the clustering algorithm. Note the high variability of the environmental lighting conditions. Left: original RGB image; center: semantic segmentation results; right: clustering results. (a) Example of correct cluster separation; (b) occlusions by leaves induce an incorrect separation of the third bunch from left into two clusters (4 and 5); (c) poor lighting conditions lead to poor segmentation results.

Fig. 9. Example images illustrating the impact of lighting conditions in framed plants. Left and right: low-light scenarios; center: direct sunlight.

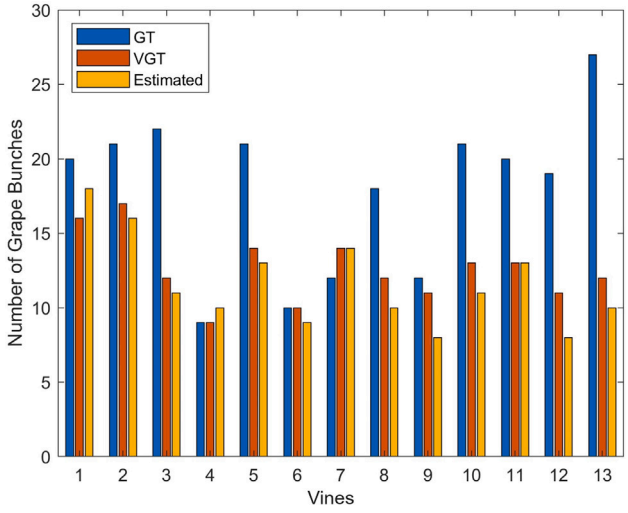

Fig. 10. Bar plot comparing estimated count of grape bunches with Visual Ground Truth (VGT) and manual Ground Truth (GT).

one side of the row, which may lead to further underestimation. It is also generally challenging to determine whether clusters located at the image borders belong to a given plant or to the adjacent one, thereby bringing to additional discrepancies between VGT and GT measurements. For these reasons, in future data collection campaigns, it will be necessary to optimize the acquisition phase, implementing active strategies for the temporary removal of leaves that can occlude the bunches, e.g., through the use of blower systems.

4.1.3. Volume and weight estimation

Once a grape bunch has been detected, its volume and weight can be estimated. Specifically, for each vine, first the volume of detected grape bunches is computed; then their weight is recovered based on Eq. (1) assuming a known density. In this respect, it should be noticed that, in general, the density depends on mass, volume, and chemical composition. It can be different among different grape varieties and can change significantly during growth (Letchov and Roychev, 2017). In this study, a single grape variety is considered at the harvesting stage. Hence, it is reasonable to assume a uniform density value and highly compact bunches. Specifically, a density of $\rho = 1.005$ g/mL was used, based on measurements performed on one hundred berries after harvesting by laboratory instruments (i.e., a scale for mass estimation and a graduated cylinder for volume estimation).

Table 2
Grape bunch counting results as obtained by: manual counting after harvesting (GT), visual inspection of images (VGT), depth-based clustering (N_{est}). The absolute errors between estimated and ground-truth measures are reported in the last two columns for ground-truth (e_{GT}) and visual ground-truth (e_{VGT}), respectively.

Vine	GT	VGT	N_{est}	e_{GT}	e_{VGT}
1	20	16	18	2	2
2	21	17	16	5	1
3	22	12	11	11	1
4	9	9	10	1	1
5	21	14	13	8	1
6	10	10	9	1	1
7	12	14	14	2	0
8	18	12	10	8	2
9	12	11	8	4	3
10	21	13	11	10	2
11	20	13	13	7	0
12	19	11	8	11	3
13	27	12	10	17	2

Table 3
Mean and standard deviation of absolute and percentage errors of estimated grape bunch number with respect to ground truth and visual ground truth.

	e_{GT}	e_{VGT}	$e_{GT\%}$	$e_{VGT\%}$
Mean	6.7	1.5	33.9	12.2
Standard deviation	4.8	1.0	18.4	8.6

Weight estimates provided by the proposed algorithm for all vines were compared with manual weight measurements, which were considered as the ground-truth. The latter were performed for two grape bunches per vine, which were chosen as the most representative of the bunch distribution for the given plant. An average of three bunches per vine were instead selected among the bunches detected by the algorithm, choosing only those that were almost fully visible (i.e., not occluded). A bar plot comparing average weight values per vine is reported in Fig. 12. The discrepancy between the estimated average weight (W^k_{est}) and measured weight (W^k_{GT}) for the *k-th* vine can be defined in terms of mean absolute percentage error (MAPE) as:

$$MAPE = \frac{1}{K}\sum_{k=1}^{K}\frac{|W^k_{est} - W^k_{GT}|}{W^k_{GT}} \times 100 \quad (9)$$

resulting in an average discrepancy on all $K = 13$ vines of 29.2% and standard deviation of 17.3%.

The distribution of the data is visualized in the box plots of Fig. 13 showing both the distribution of data and the related summary statistics.

As apparent from these graphs, the vision-based approach tends to underestimate the weights, as a result of grape bunch occlusion or inaccurate segmentation. There is a considerable amount of variability

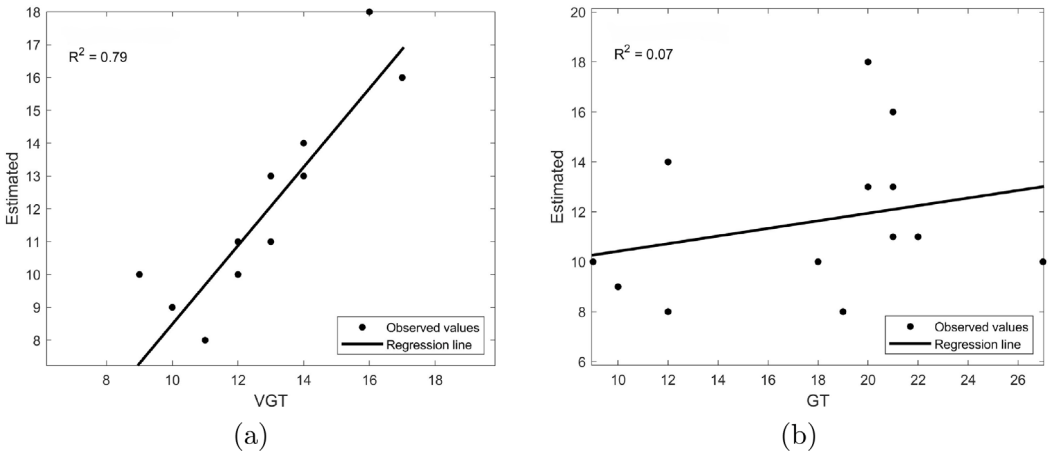

Fig. 11. Scatter plots showing the dispersion of the estimated grape bunch counts with respect to visual ground-truth (VGT) (a) and ground-truth (GT) (b).

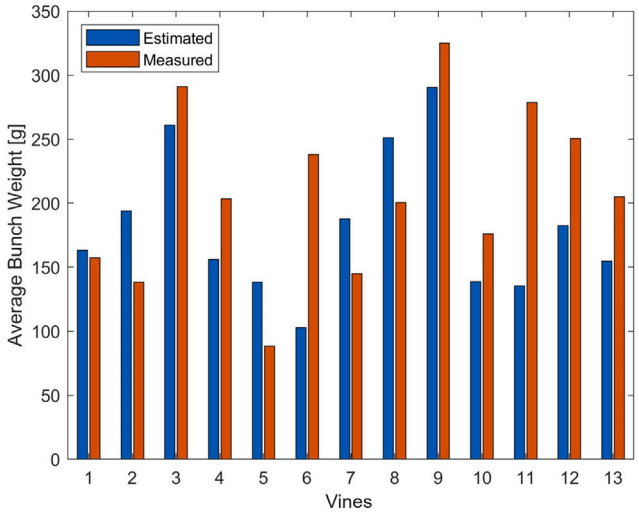

Fig. 12. Bar plot of average estimated weights and measured grape bunch weights per vine.

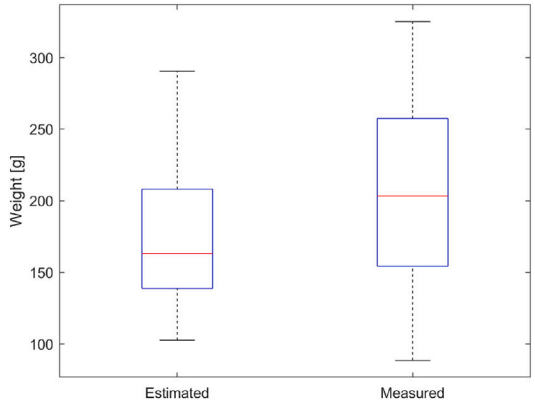

Fig. 13. Box plots of estimated weights and measured weights, representing the data distribution and related statistics, including the median (red lines inside the box), the 75th and 25th percentiles (upper and lower lines of the box), and maximum and minimum values (upper and lower whiskers).

in both the vision-based estimates and the real weights, as evidenced by the spread of the box plots along the *y*-axis. This variability is expected, due to natural variations in grape bunch cluster weights. Despite the tendency to underestimate, the proposed approach is able to provide weight estimates that fall within the range of the real weights. It is also worth to note that the proposed weight estimation approach provides a non-destructive and automated way to predict the weight of grape bunch clusters in a vineyard that can be easily applied to extensive datasets of grape bunch cluster images. However, the relatively high average discrepancy suggests the need for further research and algorithm improvements.

4.2. Yield estimation map

The proposed framework can be applied to geo-referenced image sequences with the aim of recovering yield information along vineyard rows. Fig. 14 shows the output of the system for a test along a row of the experimental site, with greener points denoting a higher number of detected clusters. In this experiment, the robot traversed a row of about 100 m while acquiring images and GPS data. A total of 357 images were synchronized with RTK-GPS position information at a frequency of 1 Hz and were processed offline to detect the number of grape bunches per image and eventually generate a yield map along the traversed row. This map demonstrates the potential of the proposed approach towards automated in-field yield estimation.

5. Discussion and conclusions

In this study, a novel approach was developed for the counting and sizing of grape bunch clusters. The proposed method integrates RGB-based semantic segmentation, depth-based clustering, and 3D representation and weight estimation techniques. The results reveal a promising potential for the application of the proposed framework for automatic yield estimation in precision viticulture. This method demonstrated reasonably accurate grape bunch cluster counting, with an average error of 1.5 grape bunches (12%) compared to the visual ground truth. These results suggest that the depth-based clustering algorithm can effectively identify and count grape bunch clusters in vineyard images. Once grape bunches are extracted from the scene, they can also be characterized in terms of volume and weight, leveraging the power of both image segmentation and 3D geometric reconstruction. Results show that the proposed approach exhibits a tendency to underestimate the actual weight of grape bunch clusters. Nevertheless, the estimated weights fall within a reasonable range, showing that the system is capable of providing estimates that capture the overall trend correctly, despite the high complexity of the real-world scenario. The system's ability to provide automatic bunch number and weight estimations could have significant implications for precision viticulture, enabling the farmers to monitor and manage their crops more efficiently.

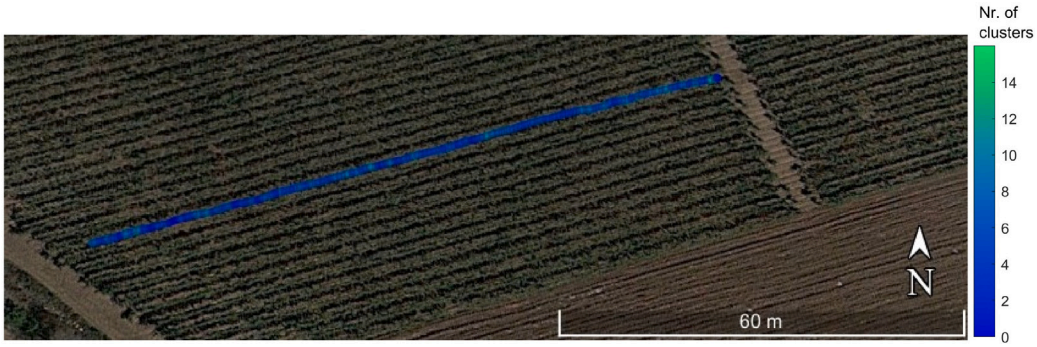

Fig. 14. Google Earth projection of the robot trajectory with associated grape bunch clusters. Greener points denote images with a higher number of clusters.

The challenges encountered in this study, including the complexity of grape bunch cluster morphology, the heterogeneity of grape bunch sizes within a cluster, lighting variations, and issues related to depth data-based volume estimation, point to areas where the automated system could benefit from refinement. Specifically, results suggest that occlusions of grape bunch clusters and image segmentation failures mainly due to uncontrolled illumination variations may limit the system performance. One additional limitation refers to the use of a consumer-grade camera worth a few hundred euros that handles two separate imaging streams, one for color and one for depth data. This may result in frame loss and latency issues. The adoption of higher-end sensing technologies, such as high-resolution stereocameras or combinations of RGB sensors and 3D Lidars, could likely lead to an improvement in accuracy, albeit at higher costs. The current system implementation restricts the use of 3D data to the clustering and volume modeling phases. Future research will investigate the incorporation of the depth channel in the deep learning models also for the task of semantic segmentation. Efforts will also be devoted to refining the volume modeling and weight estimation algorithms to better handle complexities, such as reconstruction errors and occlusions, and improve the overall accuracy. The potential of the proposed depth-based clustering technique will be further investigated through a comparative analysis with instance segmentation approaches. In this respect, recent zero-shot learning networks will be specifically investigated to reduce the need for manual labeling and improve grape clustering. Finally, in this investigation, the proposed techniques were tested at a specific growing stage, i.e., mature grapes, and on a single grape variety. In addition, a relatively limited number of vines was used for testing. In future research, the study results should be confirmed targeting a wider field extension, as well as different growing stages and cultivars.

CRediT authorship contribution statement

Rosa Pia Devanna: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Laura Romeo:** Writing – original draft, Software, Methodology, Investigation, Formal analysis. **Giulio Reina:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Annalisa Milella:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

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.

Acknowledgments

This work was partially funded by the projects: AgRibot-Harnessing Robotics, XR/AR, and 5G for a New Era of Safe, Sustainable, and Smart Agriculture, European Union's Horizon Europe research and innovation programme under grant agreement (No. 101183158); E-crops - Technologies for Digital and Sustainable Agriculture, Italian Ministry of University and Research (MUR) under the PON Agrifood Program (No. ARS01_01136); giving Smell sense To Agricultural Robotics (STAR), ERA-NET COFUND ICT AGRI-FOOD (Grant No. 45207); CNR DIITET project DIT.AD022.207, STRIVE-le Scienze per le TRansizioni Industriale, Verde ed Energetica (FOE 2022), sub-task activity Agro-Sensing2. The authors are grateful to Cantina San Donaci agricultural farm, Sysman Progetti & Servizi S.r.l, and to all partners of the E-crops project taking part to experimental setup and design and to ground truth data gathering. The administrative support of Giuseppe Bono, Michele Attolico and Paola Romano is also gratefully acknowledged.

Data availability

Data will be made available on request.

References

Bargoti, S., Underwood, J.P., 2016. Image segmentation for fruit detection and yield estimation in apple orchards. J. Field Robot. 34, URL https://api.semanticscholar.org/CorpusID:7524678.

Canny, J., 1986. A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. (6), 679–698.

Casado-García, Á., Heras, J., Marani, R., Milella, A., 2023. Taking advantage of depth information for semantic segmentation in field-measured vineyards. In: Lecture Notes in Computer Science, vol. 13955, Springer, pp. 3–15. http://dx.doi.org/10.1007/978-3-031-62799-6_1.

Casado-García, A., Heras, J., Milella, A., Marani, R., 2022. Semi-supervised deep learning and low-cost cameras for the semantic segmentation of natural images in viticulture. Precis. Agric. 1–26.

Chen, M., Chen, Z., Luo, L., Tang, Y., Cheng, J., Wei, H., Wang, J., 2024. Dynamic visual servo control methods for continuous operation of a fruit harvesting robot working throughout an orchard. Comput. Electron. Agric. 219, 108774. http://dx.doi.org/10.1016/j.compag.2024.108774, URL https://www.sciencedirect.com/science/article/pii/S0168169924001650.

Chen, S.W., Shivakumar, S.S., Dcunha, S., Das, J., Okon, E., Qu, C., Taylor, C.J., Kumar, V., 2017. Counting apples and oranges with deep learning: A data-driven approach. IEEE Robot. Autom. Lett. 2 (2), 781–788.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV.

Condotta, I.C., Brown-Brandl, T.M., Pitla, S.K., Stinn, J.P., Silva-Miranda, K.O., 2020. Evaluation of low-cost depth cameras for agricultural applications. Comput. Electron. Agric. 173, 105394. http://dx.doi.org/10.1016/j.compag.2020.105394, URL https://www.sciencedirect.com/science/article/pii/S0168169919325037.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. http://dx.doi.org/10.1109/CVPR.2009.5206848.

Devanna, R.P., Milella, A., Marani, R., Garofalo, S.P., Vivaldi, G.A., Pascuzzi, S., Galati, R., Reina, G., 2022. In-field automatic identification of pomegranates using a farmer robot. Sensors 22 (15), http://dx.doi.org/10.3390/s22155821, URL https://www.mdpi.com/1424-8220/22/15/5821.

Dunn, G.M., Martin, S.R., 2008. Yield prediction from digital image analysis: A technique with potential for vineyard assessments prior to harvest. Aust. J. Grape Wine Res. 10, 196–198, URL https://api.semanticscholar.org/CorpusID:96400234.

Edelsbrunner, H., Kirkpatrick, D., Seidel, R., 1983. On the shape of a set of points in the plane. IEEE Trans. Inform. Theory 29 (4), 551–559. http://dx.doi.org/10.1109/TIT.1983.1056714.

Fu, L., Majeed, Y., Zhang, X., Karkee, M., Zhang, Q., 2020. Faster R–CNN–based apple detection in dense-foliage fruiting-wall trees using RGB and depth features for robotic harvesting. Biosyst. Eng. 197, 245–256. http://dx.doi.org/10.1016/j.biosystemseng.2020.07.007, URL https://www.sciencedirect.com/science/article/pii/S1537511020302002.

Galati, R., Mantriota, G., Reina, G., 2022. RoboNav: An affordable yet highly accurate navigation system for autonomous agricultural robots. Robotics 11 (5).

Hacking, C., Poona, N., Manzan, N., Poblete-Echeverría, C., 2019. Investigating 2-D and 3-D proximal remote sensing techniques for vineyard yield estimation. Sensors 19 (17), http://dx.doi.org/10.3390/s19173652, URL https://www.mdpi.com/1424-8220/19/17/3652.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision. ICCV, pp. 2980–2988. http://dx.doi.org/10.1109/ICCV.2017.322.

Herrero-Huerta, M., González-Aguilera, D., Rodriguez-Gonzalvez, P., Hernández-López, D., 2015. Vineyard yield estimation by automatic 3D bunch modelling in field conditions. Comput. Electron. Agric. 110, 17–26. http://dx.doi.org/10.1016/j.compag.2014.10.003, URL https://www.sciencedirect.com/science/article/pii/S0168169914002506.

Kalampokas, T., Tziridis, K., Nikolaou, A., Vrochidou, E., Papakostas, G.A., Pachidis, T., Kaburlasos, V.G., 2020. Semantic segmentation of vineyard images using convolutional neural networks. In: Iliadis, L., Angelov, P.P., Jayne, C., Pimenidis, E. (Eds.), Proceedings of the 21st EANN (Engineering Applications of Neural Networks) 2020 Conference. Springer International Publishing, Cham, pp. 292–303.

Keightley, K.E., Bawden, G.W., 2010. 3D volumetric modeling of grapevine biomass using Tripod LiDAR. Comput. Electron. Agric. 74 (2), 305–312. http://dx.doi.org/10.1016/j.compag.2010.09.005, URL https://www.sciencedirect.com/science/article/pii/S0168169910001638.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25.

Kurtser, P., Ringdahl, O., Rotstein, N., Berenstein, R., Edan, Y., 2020. In-field grape cluster size assessment for vine yield estimation using a mobile robot and a consumer level RGB-D camera. IEEE Robot. Autom. Lett. 5 (2), 2031–2038. http://dx.doi.org/10.1109/LRA.2020.2970654.

Letchov, G., Roychev, V., 2017. Growth kinetics of grape berry density (Vitis vinifera L. "Black Corinth"). Vitis: J. Grapevine Res. 56, 155–159, URL https://api.semanticscholar.org/CorpusID:55193377.

Li, H., Gu, Z., He, D., Wang, X., Huang, J., Mo, Y., Li, P., Huang, Z., Wu, F., 2024. A lightweight improved YOLOv5s model and its deployment for detecting pitaya fruits in daytime and nighttime light-supplement environments. Comput. Electron. Agric. 220, 108914. http://dx.doi.org/10.1016/j.compag.2024.108914.

Li, R., Zheng, S., Zhang, C., Duan, C., Su, J., Wang, L., Atkinson, P., 2021. Multiattention network for semantic segmentation of fine-resolution remote sensing images. IEEE Trans. Geosci. Remote Sens. PP, 1–13. http://dx.doi.org/10.1109/TGRS.2021.3093977.

Liu, X., Jing, X., Jiang, H., Younas, S., Wei, R., Dang, H., Wu, Z., Fu, L., 2024. Performance evaluation of newly released cameras for fruit detection and localization in complex kiwifruit orchard environments. J. Field Robot. http://dx.doi.org/10.1002/rob.22297, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.22297, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.22297.

Liu, S., Whitty, M., 2015. Automatic grape bunch detection in vineyards with an SVM classifier. J. Appl. Log. 13 (4, Part 3), 643–653. http://dx.doi.org/10.1016/j.jal.2015.06.001, URL https://www.sciencedirect.com/science/article/pii/S1570868315000531.

Lytridis, C., Bazinas, C., Kalathas, I., Siavalas, G., Tsakmakis, C., Spirantis, T., Badeka, E., Pachidis, T., Kaburlasos, V.G., 2023. Cooperative grape harvesting using heterogeneous autonomous robots. Robotics 12 (6), http://dx.doi.org/10.3390/robotics12060147, URL https://www.mdpi.com/2218-6581/12/6/147.

Mack, J., Lenz, C., Teutrine, J., Steinhage, V., 2017. High-precision 3D detection and reconstruction of grapes from laser range data for efficient phenotyping based on supervised learning. Comput. Electron. Agric. 135, 300–311. http://dx.doi.org/10.1016/j.compag.2017.02.017, URL https://www.sciencedirect.com/science/article/pii/S0168169916308602.

Marani, R., Milella, A., Petitti, A., Reina, G., 2020. Deep neural networks for grape bunch segmentation in natural images from a consumer-grade camera. Precis. Agric. 1–27.

Marinello, F., Pezzuolo, A., Cillis, D., Sartori, L., et al., 2016. Kinect 3d reconstruction for quantification of grape bunches volume and mass. Eng. Rural. Dev. 15, 876–881.

Marinello, F., Pezzuolo, A., Meggio, F., Martínez-Casasnovas, J., Yezekyan, T., Sartori, L., 2017. Application of the Kinect sensor for three dimensional characterization of vine canopy. Adv. Anim. Biosci. 8 (2), 525–529. http://dx.doi.org/10.1017/S2040470017001042, URL https://www.sciencedirect.com/science/article/pii/S2040470017001042.

Meng, F., Li, J., Zhang, Y., Qi, S., Tang, Y., 2023. Transforming unmanned pineapple picking with spatio-temporal convolutional neural networks. Comput. Electron. Agric. 214, 108298. http://dx.doi.org/10.1016/j.compag.2023.108298, URL https://www.sciencedirect.com/science/article/pii/S0168169923006865.

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. http://dx.doi.org/10.1016/j.compag.2018.11.026, URL https://www.sciencedirect.com/science/article/pii/S0168169918307580.

Mohimont, L., Alin, F., Rondeau, M., Gaveau, N., Steffenel, L.A., 2022. Computer vision and deep learning for precision viticulture. Agronomy 12 (10), http://dx.doi.org/10.3390/agronomy12102463, URL https://www.mdpi.com/2073-4395/12/10/2463.

Olenskyj, A.G., Sams, B.S., Fei, Z., Singh, V., Raja, P.V., Bornhorst, G.M., Earles, J.M., 2022. End-to-end deep learning for directly estimating grape yield from ground-based imagery. Comput. Electron. Agric. 198, 107081. http://dx.doi.org/10.1016/j.compag.2022.107081, URL https://www.sciencedirect.com/science/article/pii/S0168169922003982.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28.

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. http://dx.doi.org/10.1016/j.compag.2020.105247, URL https://www.sciencedirect.com/science/article/pii/S0168169919315765.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations.

Sozzi, M., Cantalamessa, S., Cogato, A., Kayad, A., Marinello, F., 2022. Automatic bunch detection in white grape varieties using YOLOv3, YOLOv4, and YOLOv5 deep learning algorithms. Agronomy 12 (2), http://dx.doi.org/10.3390/agronomy12020319, URL https://www.mdpi.com/2073-4395/12/2/319.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE Computer Society, Los Alamitos, CA, USA, pp. 1–9. http://dx.doi.org/10.1109/CVPR.2015.7298594, URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298594.

Tao, Y., Zhou, J., 2017. Automatic apple recognition based on the fusion of color and 3D feature for robotic fruit picking. Comput. Electron. Agric. 142, 388–396. http://dx.doi.org/10.1016/j.compag.2017.09.019, URL https://www.sciencedirect.com/science/article/pii/S0168169917302764.

Ugenti, A., Galati, R., Mantriota, G., Reina, G., 2023. Analysis of an all-terrain tracked robot with innovative suspension system. Mech. Mach. Theory 182, 105237. http://dx.doi.org/10.1016/j.mechmachtheory.2023.105237, URL https://www.sciencedirect.com/science/article/pii/S0094114X23000113.

Vulpi, F., Marani, R., Petitti, A., Reina, G., Milella, A., 2022. An RGB-D multi-view perspective for autonomous agricultural robots. Comput. Electron. Agric. 202, 107419. http://dx.doi.org/10.1016/j.compag.2022.107419, URL https://www.sciencedirect.com/science/article/pii/S016816992200727X.

Wang, G., Lin, J., Lei, C., Dai, Y., Zhang, T., 2022. Instance segmentation convolutional neural network based on multi-scale attention mechanism. PLoS ONE 17, URL https://api.semanticscholar.org/CorpusID:246359529.

Woo, Y.S., Li, Z., Tamura, S., Buayai, P., Nishizaki, H., Makino, K., Kamarudin, L.M., Mao, X., 2023. 3D grape bunch model reconstruction from 2D images. Comput. Electron. Agric. 215, 108328. http://dx.doi.org/10.1016/j.compag.2023.108328, URL https://www.sciencedirect.com/science/article/pii/S0168169923007160.

Xu, Z., Liu, J., Wang, J., Cai, L., Jin, Y., Zhao, S., Xie, B., 2023. Realtime picking point decision algorithm of trellis grape for high-speed robotic cut-and-catch harvesting. Agronomy 13 (6), http://dx.doi.org/10.3390/agronomy13061618, URL https://www.mdpi.com/2073-4395/13/6/1618.