In-field grape cluster size assessment for vine yield estimation using a mobile robot and a consumer level RGB-D camera

Polina Kurtser[1], Ola Ringdahl[2], Nati Rotstein[3], Ron Berenstein[4], Yael Edan[3]

***Abstract*—Current practice for vine yield estimation is based on RGB cameras and has limited performance. In this paper we present a method for outdoor vine yield estimation using a consumer grade RGB-D camera mounted on a mobile robotic platform. An algorithm for automatic grape cluster size estimation using depth information is evaluated both in controlled outdoor conditions and in commercial vineyard conditions. Ten video scans (3 camera viewpoints with 2 different backgrounds and 2 natural light conditions), acquired from a controlled outdoor experiment and a commercial vineyard setup, are used for analyses. The collected dataset (GRAPES3D) is released to the public. A total of 4542 regions of 49 grape clusters were manually labeled by a human annotator for comparison. Eight variations of the algorithm are assessed, both for manually labeled and auto-detected regions. The effect of viewpoint, presence of an artificial background, and the human annotator are analyzed using statistical tools. Results show 2.8-3.5 cm average error for all acquired data and reveal the potential of using low-cost commercial RGB-D cameras for improved robotic yield estimation.**

***Index Terms*—Field Robots, RGB-D Perception, Agricultural Automation, Robotics in Agriculture and Forestry**

I. Introduction

YIELD estimation is of critical importance to vineyard growers for optimizing growth, harvesting preparations, crop shipment, storage scheduling, and marketing purposes. Vineyard yield can be estimated from the average number of clusters per vine and the average cluster weight [1]. Typically yield estimation is performed manually by sampling part of the vineyard and extrapolating to the rest of the vineyard [2]. Since manual practice is labor intensive and expensive, the sampled portion of the vineyard is very limited, yielding inaccurate and biased yield estimation [2]. Automation of the procedure, using remote sensing and mobile robots, introduces non invasive methods that can cover larger portions of the vineyard in timely manner leading to improved yield estimation. To achieve this task, appropriate sensor selection and data driven algorithm development are necessary.

Fig. 1. The Greenhouse Spraying Robot (GSR) mobile robot platform equipped with an Intel Realsense D435 RGB-D sensor in a commercial vineyard setup.

A. Related work

Current efforts in automation of yield estimation include autonomous mobile robots navigating through the vineyard for monitoring larger portions of it [3], [4], [5]. The robot is usually equipped with a RGB camera scanning the vineyard for grape vines while moving along the rows. Detection is complex due to the high variability in fruit sizes, shape, and colors, high occlusion rates and varying illumination conditions [6]. In some cases, depending on the fruit variety, the lack of proper colour cues (e.g., distinguishing green grapes from green foliage, [4]) further complicates the detection. Nevertheless, detection algorithms that are based on color [3], [4], [5], [7], shape [8] and texture [9] or a subset of these three features [2] are continuously developed and enhanced and show promising success rates (with 88% precision and recall obtained for state-of-the-art DNN [10]), so detection has become less of a prominent problem, but still has not been fully solved.

Once a grape cluster is detected, yield estimation is predicted based on the number of detected clusters and number of berries in a cluster [2], [10]. Detecting each single berry in a cluster in a commercial vineyard setup is rather complex due to light conditions and occlusion, and therefore some publications [3], [11] suggest grape cluster area pixel count as an alternative. Estimation based on berry count is prone to errors [2] even for ideal berry detection, since berry size varies significantly between varieties, growing conditions, and even within the vineyard. Nevertheless, these measures are currently

Manuscript received: September 10, 2019; Revised December 11, 2019; Accepted January, 8, 2019.

This paper was recommended for publication by Editor Jonathan Roberts upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported by the Swedish Knowledge Foundation (KKS) under the Semantic Robots research profile, from the Israeli Ministry of Science Grant Number 20187 and from Ben-Gurion University of the Negev through the Helmsley Charitable Trust, the Agricultural, Biological and Cognitive Robotics Initiative, the Marcus Endowment Fund, and the Rabbi W. Gunther Plaut Chair in Manufacturing Engineering

[1]Centre for Applied Autonomous Sensor Systems, Örebro University, 701 82 Örebro, Sweden polina.kurtser@oru.se

[2]Department of Computing Science, Umeå University, 901 87 Umeå, Sweden ringdahl@cs.umu.se

[3]Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer Sheva, Israel yael@bgu.ac.il(YE); natirot@post.bgu.ac.il (NR)

[4]Institute of Agricultural Engineering, Agricultural Research Organization, The Volcani Center, Rishon Lezion, Israel, ronb@volcani.agri.gov.il

Digital Object Identifier (DOI): see top of this page.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2020.2970654, IEEE Robotics and Automation Letters

used in practice to estimate cluster weight since they are the only ones possible to extract from solely visual RGB data.

Other measures considered by research groups for yield estimation include grape cluster volume, length, and width [11], [12], [13]. These measures require depth data for real-world size estimation. In their very recent work, Hacking et al. [11] showed that vine yield estimation is more accurate when based on cluster volume measured using RGB-D data, compared to estimation based on cluster area measured using RGB data only. However, the Kinect sensor, used in their experiments, that performed very well in laboratory conditions failed to provide sufficient accuracy in outdoor conditions. Therefore, for realistic field scenarios, the authors stress the need for a better RGB-D sensor. The lack of proper RGB-D sensors is pointed out by several other research groups exploring RGB-D usage in agricultural applications, e.g [14], [15], [16]. Due to this limited performance, authors commonly limit their applications to controlled indoor laboratory conditions [12], [13], [17]. With the recent penetration of new affordable RGB-D cameras operating relatively reliably in field conditions, [15], [16] it is important to evaluate these sensors in the agricultural and viticulture settings [10], [11], [15].

B. Our contributions

The contribution of this paper is threefold. First, we investigate the performance of state-of-the-art RGB-D cameras to determine if sufficient for accurate grape cluster size estimation in outdoor conditions. Secondly, we present a novel single frame RGB-D algorithm for estimating grape cluster size. The method is evaluated using several different sensor viewpoints and different backgrounds that might influence the estimation [18], [19], [20], in controlled outdoor and commercial vineyard conditions.

As mentioned in the related work section, development of an algorithm for detecting grape clusters is a well researched complex problem, which is out of scope of this paper. Since the problem has not been fully solved yet we chose not to implement complex detection algorithms. Instead, we focused on evaluating the influence of cluster detection accuracy on the cluster size estimation. This was conducted by comparing a basic unsupervised grape detection algorithm to manually labeled clusters, which serve as the golden standard for cluster detection.

Finally, the collected dataset (GRAPES3D[1]) is released to the public. Real-world agriculture datasets are scarce and hard to obtain, especially with a novel RGB-D sensor. This dataset provides an important contribution to the research community, which we believe can enhance future developments of perception algorithms in viticulture robotic applications. The data includes RGB-D scans of grape clusters on real vine plants in various conditions (more details in Sections II-A, III-A).

II. Methods

A. Data collection

An Intel Realsense D435 depth camera (an RGB-D sensor with reported promising outdoor performance [16]) was fixed to the front left corner of a Greenhouse Spraying Robot (GSR) platform at 625 mm above ground level (Fig. 1). The Realsense D435 has a FOV of $87°(\pm 3°) \times 58°(\pm 1°) \times 95°(\pm 3°)$, 1280×720 active stereo depth resolution. The GSR platform is a skid steer, 4-wheeled robotic platform designated for greenhouse spraying purposes with a 160 kg payload and max velocity of 2.5 m/s. The GSR is equipped with a NVIDIA Jetson Nano computing unit, running ROS Melodic. The GSR can be manually controlled using an Xbox controller, or programmed to do autonomous tasks.

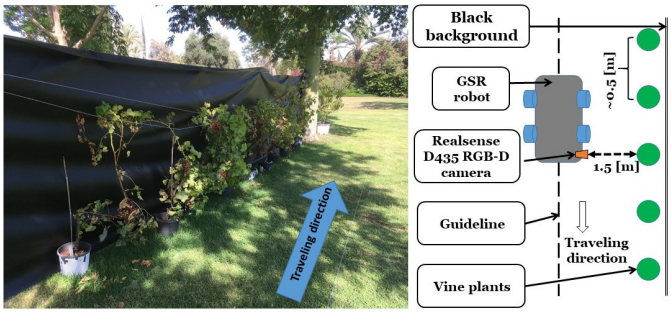

Fig. 2. Experimental setup with a black background including five vine plants.

Two experiments were performed. The *controlled outdoor setup* (Fig. 2) included 10 vine plants in pots that were placed outdoors 0.5 meter apart. A total of 17 grape clusters harvested from a commercial vineyard were placed on the vines (1-3 clusters on each vine with a higher amount of clusters on the high foliage vines). The visibility of the clusters depends on the acquisition angle. Therefore, they were placed in a way to have an equal amount of hidden and visible clusters from each sensor pose. The *commercial vineyard setup* (Fig. 1) included operation in a commercial vineyard in south Israel at the end of the growing season (Nov. 2019, one week before harvest) and represents a real-world environment. Two middle lanes were selected (one facing the sun and one opposite to the sun direction, at time of acquisition) with 5 randomly selected plants scanned in each lane (8-19 grape clusters on each vine). The plants were 1.5-2m apart from each other. Distance between growing lanes is about 3.5m.

Ground truth measures were acquired for both experiments by measuring the width and the length of the cluster at the widest and longest points respectively with a caliper (Fig. 3). In the controlled outdoor setup, the clusters were attached to the stem in a free hanging form. In the commercial vineyard setup the clusters were measured while on the stem, and an additional measure of cluster depth was acquired by measuring the widest point along the z axis. Since only the surface facing the camera is detected in a single frame, only X and Y were used in the algorithms.

In both experimental setups, the robot was moved manually parallel to the plants at 1.5 meter distance while continuously acquiring RGB-D images at 30 fps. In the controlled outdoor setup six videos were acquired, 30-90 s each, from three different viewpoints (left/front/right) with and without background (Table I). In the left and right viewpoints the camera was rotated $45°$ around the y axis. A background

[1]https://sites.google.com/view/grapes3d/home

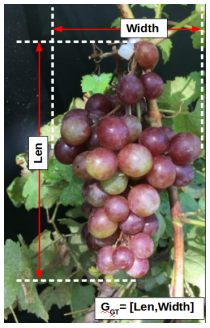

Fig. 3. Ground truth acquisition procedure. $\boldsymbol{G}_{GT}$ includes two measures - length and width of the grape cluster at the longest and widest points as measured when the cluster is hanging free.

consisting of a black cover was hanged behind the plants to investigate how an even background color and limited depth influences performance. Figure 5 shows the influence of having a background on both RGB and the point cloud images. The different viewpoints were investigated to analyse the RGB-D data sensitivity to the sensor pose.

In the commercial vineyard experiment four videos were acquired, each between 45-80 s from two viewpoints (front and right) for each of the two lanes. No background was used in this setup.

B. Grape cluster detection and size estimation

The algorithm developed for grape cluster size estimation from an RGB-D video stream consists of three major steps (Fig. 4): 1) Pre-processing, 2) Automatic detection and manual labeling of grape clusters, and 3) Estimation of grape cluster size.

1) Pre-processing: In the pre-processing phase, every fifth frame of the video is parsed into a point cloud (PC) and RGB image using Intel's Realsense Software Development Kit[2] in order to down sample the data for manual labeling purposes. The extracted PCs are then visually reviewed to locate frames in which the plant of interest is located at x = [-0.5, 0.5]. Each chosen PC is then filtered to remove the background. This is achieved by keeping points between 0.3 in 2m in z (depth) and between -1 and 1m in x (right-left) based on to the placement of the plants described in Section II-A. This pre-processing results in a PC as presented in Fig. 5 (bottom).

Images acquired in the commercial vineyard experiment, were found to be exceptionally dark and required an additional pre-processing step of lighting correction to make it easier for the annotator to find the clusters. The lighting correction includes de-hazing of the inverted image according to the method of Dong et al., [21], and the results are presented in Figure 6.

2) Manual labeling and automatic detection: From the pre-processed PCs, 3D regions of interest (ROIs) of grape clusters are segmented. The process is a four step procedure: first, a rough location ROI_R of the grape cluster is found either through manual labeling ROI_R^M (1) or through an automatic detection algorithm ROI_R^A (2). The rough location is either

[2]https://github.com/IntelRealSense/librealsense

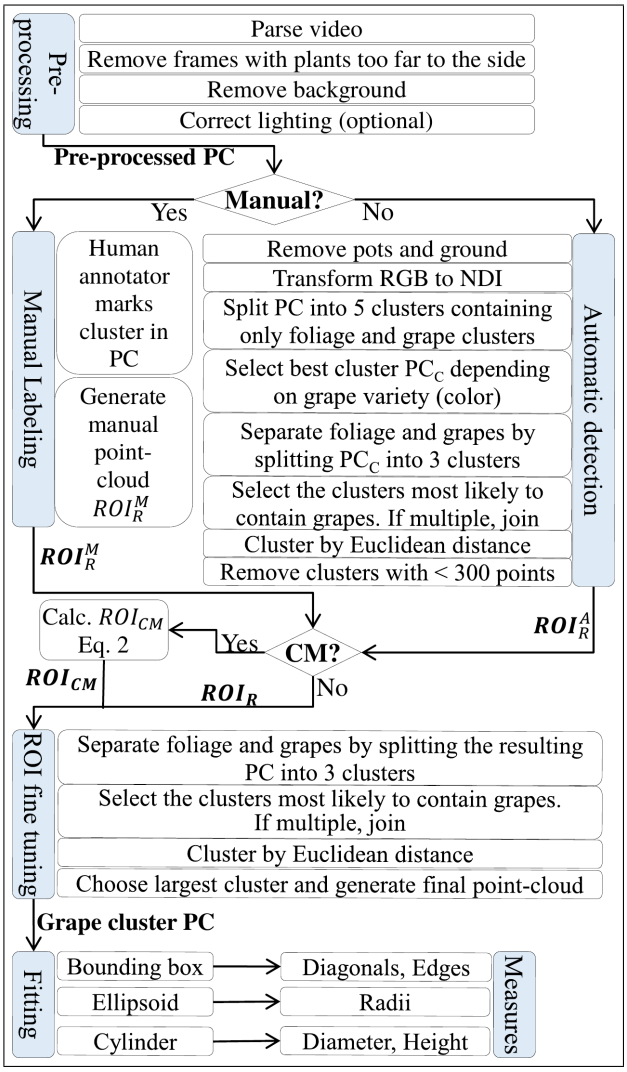

Fig. 4. Overall algorithm flowchart for estimating grape cluster size from a recorded RGB-D sequence.

fined tuned directly or goes through another step that computes ROI_{CM} based on the center of mass and radius of ROI_R (3) before being fed to the fine tuning step as ROI (4).

Manual labeling, rough location (ROI_R^M): a human annotator is presented with overview and closeup images of the plants and the location of the attached clusters. Additionally, the annotator is presented with an RGB image and a point cloud PC of the scene (Fig. 5), as acquired by the robot. The annotator is asked to draw a rectangle around a grape cluster in a XY projection of the PC in RGB color space. By including the points from the PC that correspond to the marked 2D region ROI_R^M is created. This procedure is repeated for each extracted frame from the video. Thereby the annotator gets temporal information about the clusters as well since they move only slightly between frames.

Automatic detection, rough location (ROI_R^A): The automatic detection procedure includes 4 steps: (1) Filtering out unwanted objects based on spatial location; (2) Detecting cluster and foliage based on color; (3) Extraction of clusters from foliage; (4) Disregarding possible false positives by cluster size. Since, as mentioned in the introduction, this paper does

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2020.2970654, IEEE Robotics and Automation Letters

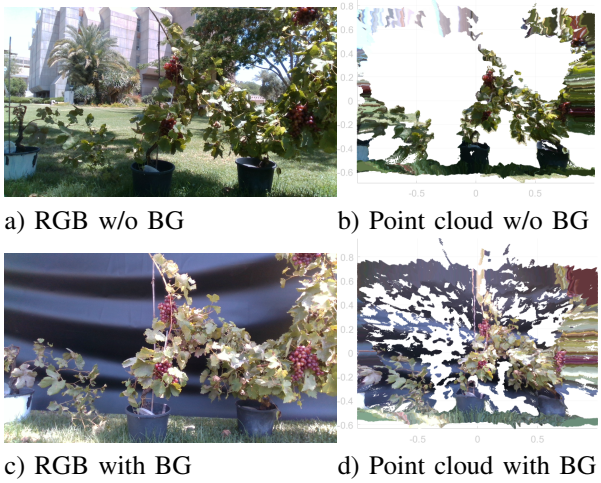

Fig. 5. Acquired RGB and point cloud with and w/o background (BG) using an Intel Realsense D435 RGB-D camera for detection and labeling of grape clusters.

Fig. 6. Light correction of data obtained in the commercial vineyard setup. (a) Original RGB, (b) Enhanced RGB, (c) Enhanced Point-cloud.

not aim to provide a new detection algorithm, the suggested algorithm presented below, includes a number of parameter values that are fitted to the gathered data. The parameters were adjusted based on 5-10 randomly selected frames from the dataset, and the location of objects in the scene as described in the experimental setup (Figure 2).

Filtering out unwanted objects based on spatial location was done by filtering the pre-processed PC to exclude the pots and parts of the ground (Y<-0.3). Next, for detecting areas that include only foliage and clusters, a color based K-means clustering [22] is performed in the NDI color space. The NDI color-space is superior to RGB for detection tasks in agricultural and viticulture settings [7], [23]. Therefore, the RGB channel of the filtered PC is transformed according to Eq. 1.

$$N = \frac{R-G}{R+G};\ D = \frac{R-B}{R+B};\ I = \frac{B-G}{B+G} \tag{1}$$

where N,D,I,R,G and B are the three channels of the NDI and RGB color-spaces respectively. The three NDI channels are used as the 3D features space fed into a K++ algorithm [24], an improved version of the classical K-means implemented in the default MATLAB function *kmeans*[3], to segment the point-cloud. For the controlled outdoor dataset a suitable cluster is obtained by clustering the pointcloud into 5 clusters and choosing the cluster with the centroid with largest N in the NDI color space. The choice of highest N centroid allows for objects with highest contrast of red and green (Eq. 1) to be chosen as the suitable cluster. Since in the gathered dataset, the grape variety was of high contrast of red and green, these clusters, and their close surrounding foliage were selected. Once the suitable cluster is selected, the pre-processed PC is filtered to include only the spatial points corresponding to the chosen cluster.

Next, given that the cluster chosen includes foliage and grapes only, an additional color clustering is performed, to separate the grape cluster from the surrounding foliage. The red grapes, as mentioned, are expected to have high N values, but to discriminate them from the green leaves they are also expected to have higher D values (contrast of red to blue). Therefore, the step is similar to the one suggested above and was performed through NDI based k-means clustering. For the controlled outdoor dataset, best results were achieved by clustering into 3 clusters and choosing the cluster(s) with centroid values $N > 0.2$ and $D > 0.2$. By allowing multiple clusters to be chosen, the algorithm provides freedom to join together clusters wrongly split (in case too little foliage is present in the region).

Finally, filtration of possible false positive small clusters is done through clustering the point-cloud according to spatial location. The color filtered PC is segmented into clusters, with a minimum Euclidean distance of $0.01m$ between points from different clusters. Clusters with less than 300 points are disregarded. Each remaining cluster corresponds to ROI_R^A.

ROI from center of mass (ROI_{CM}): consists of all points of the pre-processed PC satisfying Eq. 2, derived from the center of mass CM of ROI_R (either ROI_R^M or ROI_R^A). Optionally this step can be skipped by feeding ROI_R to the fine-tuning step directly.

$$\begin{aligned} CM_x - \mathcal{R} < x < CM_x + \mathcal{R} \\ CM_y - \mathcal{R} < y < CM_y + \mathcal{R} \\ CM_z - \mathcal{R} < z < CM_z + \mathcal{R} \end{aligned} \tag{2}$$

$\mathcal{R}$ is a constant radius of interest around the CM in which the grape cluster is suspected to be. For the analysis in this paper the value was selected as $\mathcal{R} = 0.15$m (about double the expected average radius of a grape cluster of the variety chosen). Accuracy measures extracted from ROI_{CM} provide insights into the robustness of the fine tuning step to the size of the labeled or auto-detected region.

Fine tuning: Each ROI generated according to one of the four aforementioned methods (ROI_R or ROI_{CM} derived from manual labeling or automatic detection) is then subject to a fine tuning process, to segment the grape cluster based on color and spatial cues.

The fine tuning step makes two assumptions: the ROI consists of grapes and foliage only and the cluster containing a grape is the largest in size. This leads to a fine tuning procedure that resembles the second step of auto-detection - separation of the grape cluster from the surrounding foliage. The fine tuning is therefore redundant for ROI_R^A for the current auto-detect algorithm, but is kept for the possibility to substitute auto-detect with a window based grape detection, as described in the introduction, and also for generalization of the overall algorithm.

[3] https://se.mathworks.com/help/stats/kmeans.html#bues5gz

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2020.2970654, IEEE Robotics and Automation Letters

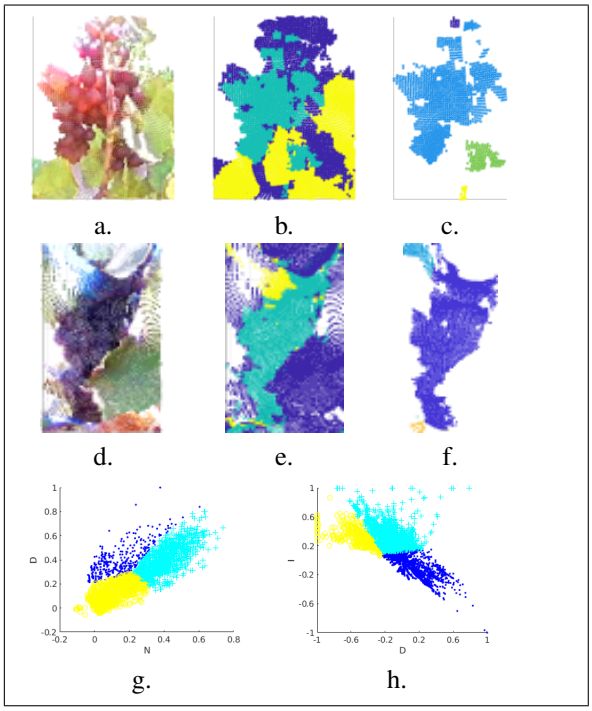

Fig. 7. Fine tuning. (a-c): controlled outdoor setup, (d-f): commercial vineyard setup. The point clouds (a) and (d) are clustered by NDI into 3 groups (b and e). The chosen clusters (cyan in (b), cyan and yellow in (e)) are clustered spatially with minimum distance $\geq 0.01m$, resulting in 4 clusters found in (c) and 3 in (f). The largest cluster is classified as grape (cyan in (c), blue in (f)). (g): N-D scatter for controlled outdoor setup, (h): D-I scatter for commercial vineyard setup. Clusters colored according to clustering in (b) and (e) respectively.

Fine tuning consists of color clustering by a K-means clustering algorithm using the NDI color space with 3 clusters (Fig. 7). As mentioned above, for the controlled outdoor setup the cluster chosen as the one representing the grape cluster is the one with centroid values $N > 0.2$ and $D > 0.2$ in the NDI color space (Figure 7g). For the commercial vineyard setup, due to the dominance of the blue color in the specific grape color variety, the cluster chosen as the one representing the grape cluster is the one with centroid values $D < 0.2$ and $I > -0.05$ (Eq. 1: higher blue contrast leads to lower D values and higher I values, Figure 7h). If more than one cluster satisfies these conditions, they are joined into a single cluster. Then, spatial clustering takes place with a minimum Euclidean distance of $0.01m$ between points from different clusters (Fig. 7). The cluster with the maximum number of points is chosen to represent the grape cluster.

3) Estimation of grape cluster size: Several methods to estimate the grape cluster size from the obtained ROI are evaluated. All methods are based on fitting a geometrical form to the ROI and then extracting measures from the fitted geometrical object.

The ***Percentile bounding box*** fitting returns a bounding box according to Algorithm 1. The percentile was chosen empirically as $\alpha = 0.02$ (i.e., the algorithm derives the 20th and 80th percentile for the pointcloud's x,y, and z values respectively). The measures of grape size cluster include $\hat{\boldsymbol{G}}_E$ and $\hat{\boldsymbol{G}}_D$ - the two largest edges and diagonals (representing X and Y) of the fitted box respectively.

Algorithm 1 Percentile bounding box.

function GETBOUNDINGBOX($PtCloud, \alpha$)
  $(minX, maxX) \leftarrow$ Percentiles(PtCloud.X,α,$1-\alpha$)
  $(minY, maxY) \leftarrow$ Percentiles(PtCloud.Y,α,$1-\alpha$)
  $(minZ, maxZ) \leftarrow$ Percentiles(PtCloud.Z,α,$1-\alpha$)
  $B \leftarrow [minX\ maxX\ minY\ maxY\ minZ\ maxZ]$
  return B
end function

The ***Ellipsoid fitting*** method was inspired by [8] who fitted ellipsoids for grape berries detection. The ellipsoid fitting is implemented using Petrov's[4] function for fitting an ellipsoid, sphere, paraboloid or hyperboloid to a surface. Important to note, that the fitting is limited due to the single viewpoint approach taken. The measure of grape size cluster is $\hat{\boldsymbol{G}}_R$ - the two largest radii of the fitted ellipsoid.

The ***Cylinder fitting*** method fits a cylinder to the PC using M-estimator SAmple Consensus (MSAC) algorithm [25] implemented with MATLAB's *pcfitcylinder*[5] function. The measure of grape size cluster is $\hat{\boldsymbol{G}}_C$, the diameter and the height of the fitted cylinder.

C. Analysis

The influence of viewpoints and the different variations of the detection algorithm were evaluated as follows.

1) Accuracy compared to ground truth: Statistics on the difference between the ground truth (GT) measures $\boldsymbol{G}_{GT}$ and each of the proposed estimation measures $\hat{\boldsymbol{G}}_E$, $\hat{\boldsymbol{G}}_D$, $\hat{\boldsymbol{G}}_R$, and $\hat{\boldsymbol{G}}_C$ (see Sec. II-B3) are provided. For each ROI_R and ROI_{CM} (auto-detected or manual), that went through the fine tuning process, a measure of accuracy is defined as the average absolute error, according to Eq. 3.

$$Err_x = |\hat{\boldsymbol{G}}_x - \boldsymbol{G}_{GT}|\boldsymbol{w}_e \tag{3}$$

where $\hat{\boldsymbol{G}}_x$ is one of the estimation measure vectors $\hat{\boldsymbol{G}}_E$, $\hat{\boldsymbol{G}}_D$, $\hat{\boldsymbol{G}}_R$, or $\hat{\boldsymbol{G}}_C$, and $\boldsymbol{w}_e$ is the weights given to the first and second size measure. The error is reported in cm with $\boldsymbol{w}_e = \begin{bmatrix}0.5 & 0.5\end{bmatrix}'$.

2) Identifying sources of error: To evaluate the statistical significance of the main sources of error Err_x, an ANOVA model is applied. The model accounts for the fixed effect of cluster position (top/center/bottom), viewpoint (front/right), and background presence (yes/no) and random effect of plant number.

3) Detection results: To be able to compare the error generated by the auto-detected label to the ground truth, a process of pairing the detected labels and the manually marked one was conducted. In this algorithm, an overlap between auto-detected areas and labeled areas are calculated for each frame. The regions with the highest overlap are considered a "pair", implying the measures for the auto-detected area will be compared to the ground truth associated with the labeled cluster.

[4]https://ww2.mathworks.cn/matlabcentral/fileexchange/24693-ellipsoid-fit?s_tid=FX_rc2_behav

[5] https://se.mathworks.com/help/vision/ref/pcfitcylinder.html

TABLE I
NUMBER OF ROI_R^A/ROI_R^M FOR DIFFERENT VIEWPOINTS (VP), PRESENCE OF BACKGROUND (BG), AND CLUSTER POSITION FOR CONTROLLED OUTDOOR SETUP.

		Cluster position		
BG	VP	Bottom	Center	Top
No	Left	44/74	16/23	27/32
	Front	34/320	59/141	154/238
	Right	18/382	119/183	95/206
Yes	Left	121/285	129/137	80/121
	Front	73/542	98/197	185/211
	Right	0/419	111/249	242/312

TABLE II
Err_x FOR SIZE ESTIMATION METHODS IN CONTROLLED OUTDOOR SETUP: AVERAGE ±STD CM (NO. OF ROI_R KEPT POST-FILTER). NO. OF UNFILTERED ROI_R^M: 4072, ROI_R^A: 1605.

		Automatic		Manual	
		ROI_R^A	ROI_{CM}^A	ROI_R^M	ROI_{CM}^M
Percentile box diag. $\hat{G}_D$	All	4.6 ±4.6	4.4 ±4.1	2.9 ±2.8	2.9 ±2.8
	Filter	3.2 ±2.5 (1386)	3.3 ±2.5 (1402)	2.6 ±2.0 (3909)	2.6 ±2.0 (3913)
Percentile box edges $\hat{G}_E$	All	4.9 ±3.4	4.8 ±3.2	3.3 ±2.4	3.4 ±2.4
	Filter	4.3 ±2.6 (1417)	4.3 ±2.6 (1434)	3.3 ±2.2 (4027)	3.3± 2.2 (4004)
Cylinder diam. & height $\hat{G}_C$	All	10.9 ±12.9	10.9 ±12.9	10.6 ±20.9	10.3 ±13.3
	Filter	5.6 ±2.8 (570)	5.6 ±2.8 (570)	6.0 ±2.7 (1442)	6.0 ±2.6 (1471)
Ellipsoid radii $\hat{G}_R$	All	22.0 ±134.8	19.9 ±97.1	25.2 ±329.9	25.7 ±408.7
	Filter	5.4 ±2.7 (937)	5.4 ±2.7 (935)	5.3 ±2.7 (2705)	5.3 ±2.7 (2688)

III. RESULTS

A. Description of the dataset acquired

In the controlled outdoor setup, all 17 clusters were identified by the human annotator and a total of 4072 manual labels were collected (Table I). The difference in number of labeled frames for different combinations of viewpoints and background is due to the fact that each combination is acquired in a separate session with the robot and therefore the number of frames can vary slightly depending on robot speed and distance to plants. Additionally, the annotators reported that background made it easier to locate the cluster, which explains the larger number of frames in the acquisitions with background. Ground truth measurements of the 17 clusters yielded an average grape cluster of 0.12m (± 0.02) X 0.14m (± 0.02). In the commercial vineyard, the annotators were able to identify 32 clusters, and 470 manual labels were marked. Ground truth measurements of the identified clusters yielded an average grape cluster of 0.14m (± 0.05) X 0.10m (± 0.04). The labels and ground truth measures were then analyzed and the following accuracy measures were extracted.

B. Estimation method accuracy compared to GT

For the controlled outdoor setup, accuracy measures of each of the proposed methods are outlined in Table II. The results reveal that the fitting method providing the best accuracy is the diagonals of the percentile box $\hat{\boldsymbol{G}}_D$ (2.9 cm average absolute error) with percentile box edges $\hat{\boldsymbol{G}}_E$ slightly behind (3.4 cm average error). When manual detection was substituted with the detection algorithm the average absolute error rises to 4.6-4.9 cm suggesting the algorithm can benefit from a better auto-detection algorithms than the one proposed (detection rates in Table I). Manual labeling also showed significantly lower standard deviation of absolute error $\hat{\boldsymbol{G}}_D$ as compared to the the auto-detected areas (~2.8 cm compared to ~4.6 cm) and $\hat{\boldsymbol{G}}_E$ (~2.4 cm compared to ~3.4 cm). Estimation based on ROI_{CM} compared to directly extracting from the rough location ROI_R showed very similar results, suggesting the algorithm is robust to the size of the region detected or labelled. Fitting of a cylinder and ellipsoid showed a great number of outliers. Filtering detection with higher than 10 cm absolute error results reveal that the ellipsoid and cylinder methods yield an average error of ~5.3 cm, however with a large number of outliers (30-60%). Fitting a percentile box to manual labeling resulted with very little outliers (2-3%) with edges being slightly more robust but slightly less accurate than diagonals.

For the commercial vineyard setup, slightly higher error was observed (3.6 ±3.0 cm for $\hat{\boldsymbol{G}}_E$). $\hat{\boldsymbol{G}}_D$ showed significantly weaker performance with 4.6 ±3.9 cm average error. To avoid reporting misleading results affected mostly by the detection algorithm performance, rather than the evaluated sensor or the cluster size estimation algorithm, evaluation in the commercial vineyard setup was performed only for the ROI_R^M. As defined in the introduction, the detection algorithm presented is rather simplistic and is not fitted for the highly occluded dataset acquired at the end of the growing season.

Comparison of the grape cluster volume in vineyard conditions, calculated by (ground truth) width x length x depth, to the estimated volume acquired by multiplying all 3 bounding box edges $\hat{G}_E$ showed an average absolute error of 67% and average error of -6% ($\pm 95\%$). These results suggest that the estimation on average slightly underestimates the real volume of the cluster, as expected from a single frame estimation that can only detect the face area of the cluster facing the camera. The very high standard deviation suggests this measure to be inaccurate for volume estimation and therefore the volume was not analyzed. Therefore, further analysis is focused on $\hat{G}_E$ (percentile box edges). The analysis is conducted on the unfiltered data (including outliers), for manual labeling, with ROI_R^M fed directly into the fine tuning process (no ROI_{CM}).

C. Sources of error identification

In the controlled outdoor setup, error as function of viewpoint (front/right/left) and background presence (yes/no) are presented in Fig 8. Results show statistically significant ($p.val < 2e-16$) lower error for images with the background present (3.06-3.18 cm compared to 3.57-3.76 cm). Results also show statistically significant ($p.val = 0.004$) lower error for images taken from the right with no background compared to front and left views. These differences between viewpoints are not observed for images taken with a background. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2020.2970654, IEEE Robotics and Automation Letters

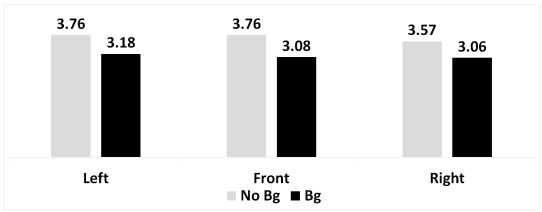

Fig. 8. Estimation accuracy as function of viewpoint (front/left/right) and background presence (yes/no) for the controlled outdoor setup.

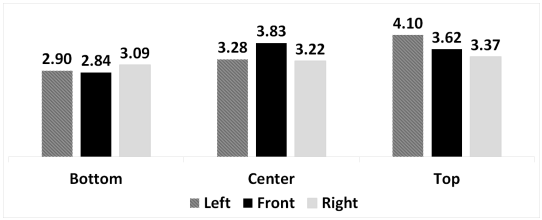

Fig. 9. Estimation accuracy as function of viewpoint (front/left/right) and cluster position (top/center/bottom) for the controlled outdoor setup.

differences may be explained by the sun direction (similar to [18]), which made detection, annotation and depth information extraction to be more complex, while presence of background seemed to simplify both the annotation and the depth information extraction.

Error rates as function of viewpoint (front/right/left) and cluster position (top/center/bottom) are presented in Fig 9. Results show statistically significant ($p.val < 2e{-}16$) increase in error with increase in vertical cluster position, with very high error rates for top clusters. This could be explained both by distortion differences in depth errors at different locations in the image [15], and the complexity of annotation of non central areas (Fig. 5). Hence, analysis of those clusters could be either disregarded a-priori or performed with the camera higher up. Results also suggest that a right viewpoint leads to statistically significant ($p.val = 7.2e-09$) lower error rates, unless the cluster is located at the bottom of the plant, in this case a front view is preferable. This can be explained by less direct sun exposure in bottom clusters leading to less distortion of the data acquired from the front viewpoint.

For the commercial vineyard setup, error as function of sun direction (facing/back) and viewpoint (front/right) is presented in Fig 10. Results show no statistically significant differences in natural light direction ($p.val > 0.05$), this could be partially attributed to the imbalanced number of clusters identified by the human annotator in the lane facing the light versus the lane with back to the light (26 clusters and 6 accordingly). The differences in viewpoint were borderline ($p.val = 0.097$), most probable due to the imbalanced dataset mentioned above.

IV. Conclusions

In this paper we present a method for outdoor grape size estimation using a commercial RGB-D camera mounted on a mobile robotic platform. The best algorithm resulted in a ~2.9 and 3.6 cm average absolute error in length and width estimation compared to ground truth (Eq. 3) for a controlled outdoor setup and commercial vineyard setup respectively. The

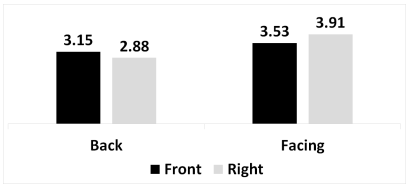

Fig. 10. Estimation accuracy as function of viewpoint (front/right) and sun direction (facing/back) for the commercial vineyard setup.

algorithm is based on fitting a bounding box to fine tuned data through color based k-means clustering. The algorithms showed robustness to the size of the detected or labelled region.

Due to the lack of available RGB-D sensors operating reliably in the field, there are a very limited number of published work to compare the reported results to. Yield estimation efforts often focus on the end mean - report of yield accuracy, while not reporting accuracy of the intermediate steps such as detection accuracy, or volume estimation accuracy. This paper focuses on estimation of grape cluster size accuracy, under the assumption that better size estimation accuracy will lead to better yield estimation (supported by previous findings [11]). As a result, no grapes were harvested to estimate yield.

Size estimation of crops is also relevant to the robotic harvesting domain, where several relevant work can be mentioned for comparison. Luo et al., [26] reported an error of 1.6-1.9 cm for a dataset of 12 grape clusters. However, the selected clusters were captured in closeup images without occlusion, and no sensitivity analyses were performed. For apple harvesting, Gongal et al., [27] reported 69-84% size estimation accuracy in outdoor conditions. Laboratory conditions improve the results significantly with reported 0.4-0.6 cm error for post harvest measurement of mango fruits [28] and 0.8-1% error for size estimation of olives in controlled photo box conditions [29].

The results of this paper suggest that high accuracy vine cluster size estimation is possible in outdoor conditions using consumer level RGB-D cameras. The presented size estimation with depth cameras can lead to improved yield estimation results. The differences in error obtained from different viewpoint stresses previous work that sensor pose is an important parameter to be considered in design of automation procedures in outdoor conditions [18]. The higher error rates in the top clusters compared to bottom clusters suggest future scanning should be performed at several heights. The experiment in a commercial vineyard confirmed that grapes and foliage in the next lane are successfully filtered out and therefore do not influence the results. The reason for the slightly lower performance in vineyard conditions compared to the controlled outdoor setup can be affiliated to the higher occlusion rate in the vineyard setup at the end of the growing season, when the experiment was held. Finally, presence of background showed significantly better accuracy, suggesting, incorporation of such a background during scanning is preferable. However, the decision to do so depends on logistical and financial considerations that should be taken into account.

Introduction of an auto-detection algorithm in the controlled outdoor setup increases the average absolute error to 4.4-

4.6 cm. Therefore, the system will benefit from applying more sophisticated detection algorithms based on advanced machine learning and deep learning techniques (e.g [10]). These require large datasets which can be collected using the mobile robotic platform and protocols presented in this paper. The released GRAPES3D dataset, will hopefully contribute to further development of such algorithms. Additionally, our results show that fitting of ellipsoid or cylinders do not provide good estimation despite previously published result [2]. Future work could also include other more sophisticated geometrical fitting forms.

The reported results could additionally be valuable to the harvesting robotics community that are also interested in grape cluster size estimation, as mentioned above. For greater cluster size estimation robustness, the research should be validated in additional vineyard conditions including different seasons, growing conditions, and cultivars.

REFERENCES

[1] J. Wolpert and E. Vilas, "Estimating vineyard yields: Introduction to a simple, two-step method," *American Journal of Enology and Viticulture*, vol. 43, no. 4, pp. 384–388, 1992.

[2] S. Nuske, K. Wilshusen, S. Achar, L. Yoder, S. Narasimhan, and S. Singh, "Automated visual yield estimation in vineyards," *Journal of Field Robotics*, vol. 31, no. 5, pp. 837–860, 2014.

[3] M.-P. Diago, C. Correa, B. Millán, P. Barreiro, C. Valero, and J. Tardaguila, "Grapevine yield and leaf area estimation using supervised classification methodology on rgb images taken under field conditions," *Sensors*, vol. 12, no. 12, pp. 16 988–17 006, 2012.

[4] R. Berenstein, O. B. Shahar, A. Shapiro, and Y. Edan, "Grape clusters and foliage detection algorithms for autonomous selective vineyard sprayer," *Intelligent Service Robotics*, vol. 3, no. 4, pp. 233–243, 2010.

[5] G. M. Dunn and S. R. Martin, "Yield prediction from digital image analysis: A technique with potential for vineyard assessments prior to harvest," *Australian Journal of Grape and Wine Research*, vol. 10, no. 3, pp. 196–198, 2004.

[6] K. Kapach, E. Barnea, R. Mairon, Y. Edan, and O. Ben-Shahar, "Computer vision for fruit harvesting robots–state of the art and challenges ahead," *International Journal of Computational Vision and Robotics*, vol. 3, no. 1/2, pp. 4–34, 2012.

[7] E. Zemmour, P. Kurtser, and Y. Edan, "Automatic parameter tuning for adaptive thresholding in fruit detection," *Sensors*, vol. 19, no. 9, p. 2130, 2019.

[8] G. Rabatel and C. Guizard, "Grape berry calibration by computer vision using elliptical model fitting," in *European Conference on Precision Agriculture*, vol. 6, 2007, pp. 581–587.

[9] M. Grossetete, Y. Berthoumieu, J.-P. Da Costa, C. Germain, O. Lavialle, G. Grenier *et al.*, "Early estimation of vineyard yield: site specific counting of berries by using a smartphone," in *International Conference of Agricultural Engineering—CIGR-AgEng*, 2012.

[10] A. Milella, R. Marani, A. Petitti, and G. Reina, "In-field high throughput grapevine phenotyping with a consumer-grade depth camera," *Computers and electronics in agriculture*, vol. 156, pp. 293–306, 2019.

[11] C. Hacking, N. Poona, N. Manzan, and C. Poblete-Echeverría, "Investigating 2-d and 3-d proximal remote sensing techniques for vineyard yield estimation," *Sensors*, vol. 19, no. 17, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/17/3652

[12] E. Ivorra, A. Sánchez, J. Camarasa, M. P. Diago, and J. Tardáguila, "Assessment of grape cluster yield components based on 3d descriptors using stereo vision," *Food Control*, vol. 50, pp. 273–282, 2015.

[13] A. Kicherer, R. Roscher, K. Herzog, W. Förstner, and R. Töpfer, "Image based evaluation for the detection of cluster parameters in grapevine," in *XI International Conference on Grapevine Breeding and Genetics 1082*, 2014, pp. 335–340.

[14] F. Marinello, A. Pezzuolo, D. Cillis, and L. Sartori, "Kinect 3d reconstruction for quantification of grape bunches volume and mass," *Engineering for Rural Development*, vol. 15, pp. 876–881, 2016.

[15] O. Ringdahl, P. Kurtser, and Y. Edan, "Performance of rgb-d camera for different object types in greenhouse conditions," in *2019 European Conference on Mobile Robots (ECMR)*. IEEE, 2019, pp. 1–6.

[16] A. Vit and G. Shani, "Comparing rgb-d sensors for close range outdoor agricultural phenotyping," *Sensors*, vol. 18, no. 12, p. 4413, 2018.

[17] F. Rist, K. Herzog, J. Mack, R. Richter, V. Steinhage, and R. Töpfer, "High-precision phenotyping of grape bunch architecture using fast 3d sensor and automation," *Sensors*, vol. 18, no. 3, p. 763, 2018.

[18] P. Kurtser and Y. Edan, "Statistical models for fruit detectability: spatial and temporal analyses of sweet peppers," *Biosystems Engineering*, vol. 171, pp. 272–289, 2018.

[19] P. Kurtser and Y. Edan, "The use of dynamic sensing strategies to improve detection for a pepper harvesting robot," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2018, pp. 8286–8293.

[20] J. Hemming, J. Ruizendaal, J. W. Hofstee, and E. van Henten, "Fruit detectability analysis for different camera positions in sweet-pepper," *Sensors*, vol. 14, no. 4, pp. 6032–6044, 2014.

[21] X. Dong, G. Wang, Y. Pang, W. Li, J. Wen, W. Meng, and Y. Lu, "Fast efficient algorithm for enhancement of low lighting video," in *2011 IEEE International Conference on Multimedia and Expo*. IEEE, 2011, pp. 1–6.

[22] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[23] B. Arad, P. Kurtser, E. Barnea, B. Harel, Y. Edan, and O. Ben-Shahar, "Controlled lighting and illumination-independent target detection for real-time cost-efficient applications. the case study of sweet pepper robotic harvesting," *Sensors*, vol. 19, no. 6, p. 1390, 2019.

[24] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[25] P. H. Torr and D. W. Murray, "The development and comparison of robust methods for estimating the fundamental matrix," *International journal of computer vision*, vol. 24, no. 3, pp. 271–300, 1997.

[26] L. Luo, Y. Tang, X. Zou, M. Ye, W. Feng, and G. Li, "Vision-based extraction of spatial information in grape clusters for harvesting robots," *Biosystems Engineering*, vol. 151, pp. 90–104, 2016.

[27] A. Gongal, M. Karkee, and S. Amatya, "Apple fruit size estimation using a 3d machine vision system," *Information Processing in Agriculture*, vol. 5, no. 4, pp. 498–503, 2018.

[28] Z. Wang, A. Koirala, K. Walsh, N. Anderson, and B. Verma, "In field fruit sizing using a smart phone application," *Sensors*, vol. 18, no. 10, p. 3331, 2018.

[29] J. M. Ponce, A. Aquino, B. Millan, and J. M. Andújar, "Automatic counting and individual size and mass estimation of olive-fruits through computer vision techniques," *IEEE Access*, vol. 7, pp. 59 451–59 465, 2019.