PointNet and geometric reasoning for detection of grape vines from single frame RGB-D data in outdoor conditions

Polina Kurtser*[1], Ola Ringdahl[2], Nati Rotstein[3], and Henrik Anderson[1]

[1]Centre for Applied Autonomous Sensor Systems, Örebro University, Örebro, Sweden
[2]Department of Computing Science, Umeå University, Umeå, Sweden
[3]Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer Sheva, Israel

Abstract

In this paper we present the usage of PointNet, a deep neural network that consumes raw un-ordered point clouds, for detection of grape vine clusters in outdoor conditions. We investigate the added value of feeding the detection network with both RGB and depth data, contradictory to common practice in agricultural robotics which to-date relies only on RGB data. A total of 5057 pointclouds (1033 manually annotated and 4024 annotated using geometric reasoning) were collected in a field experiment conducted in outdoor conditions on 9 grape vines and 5 plants. The detection results show overall accuracy of 91% (average class accuracy of 74%, precision 48% recall 53%) for RGBXYZ data and a significant drop in recall for RGB or XYZ data only. These results suggest the usage of depth cameras for vision in agricultural robotics is crucial for detection in crops where the color contrast between the crop and the background is complex. The results also suggest geometric reasoning can be used to increase training set size, a major bottleneck in the development of agricultural vision systems.

1 Introduction

The interest in development of *agrobots*, robots supporting or replacing humans in agricultural operations (e.g. harvesting and weeding), has grown lately due to the lack of workforce and growing food demands [3]. One of the challenges preventing commercialization of agrobots is the low detection rates of the targets (e.g. fruit) due to occlusions (foliage) and accurate localization of the target [3].

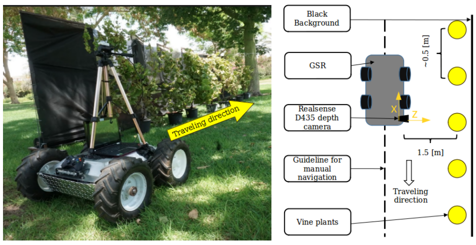


Figure 1: Experimental setup including five vine plants with a black background behind.

RGB cameras are low priced and show reliable performance in outdoor conditions [3, 4]. As a result, current detection efforts in agrobots are mainly based on color information extracted from RGB cameras [2, 22, 12]. Color based detection algorithms shown remarkable results for objects with high color-background contrast (e.g. red fruit on green leaves) but perform poorly for low contrast objects (e.g. green peppers and grapes on green leaves) [2, 20, 22, 6] or in complex lighting [2]. Usage of hyperspectral, multispectral, [1], or thermal cameras [9] for detection were also explored, but often overruled in agrobots integration due to high weight and costs. Depth sensors have not been used extensively for crop detection in agriculture [13]. This can partly be attributed to the high cost of reliable outdoor 3D depth sensors, or because of the too low point density provided by even high-end 3D LiDaR sensors for typical agricultural operation (e.g. detection of crops, leaves, stems). Detection based on a color-depth combination, in outdoor conditions, has been problematic to-date

*Corresponding Author: polina.kurtser@oru.se

https://doi.org/10.7557/18.5155

due to the lack of RGB-D sensors reliably performing outdoors [21].

Reasonably priced commercial RGB-D cameras showing unprecedented performance in outdoor conditions [18, 21] have recently started to enter the market. So far they were mostly used for localisation in world coordinates [4] of RGB based detected items [17], an operation that was previously done through visual servoing (e.g. [5, 17]). Presence of accurate RGB-D sensors with high pointcloud (PCD) density, allows development of detection algorithms based on both color and depth data for agrobots. Since reliable indoor RGB-D sensors are on the market for a while now (e.g. Kinect, RealSense) the literature provides a variety of detection algorithms developed for dense PCD data [7]. These algorithms provide a starting point for development for outdoor detection, but for it to be relevant for agricultural robotics the following challenges should be addressed:

Single frame detection. Due to the dynamic and highly occluded conditions in agriculture, registration of multiple frames is challenging and introduces additional error and therefore not applicable. Furthermore, most agrobots are equipped with an eye-in-hand sensor configuration [3, 4, 15], leading to increased cycle times for collection of multiple frames of the same object, and limiting commercialization of the agrobot [3, 14, 15]. Therefore, single frame detection is critical for fast and reliable operation of agrobotic vision.

Low resolution detection. Commercially available sensors to date still have sparse depth density and low registered resolution [18, 21]. Therefore, the detection must work well on sparse data.

Limited training data. Labeling is a significant bottleneck preventing rapid acquisition and learning procedures [19]. Acquiring data and ground truth in agri-robotics poses an extra challenge due to the harsh conditions. Furthermore, there is large variability between objects [3] and growing conditions, seasons and even within the same crop [3, 4, 15]. As an evident, current agricultural vision datasets often lack proper labeling [12].

To address these challenges, we chose *PointNet* [7] - a deep neural network developed for classification and semantic segmentation that consumes raw point clouds. PointNet has been widely used in indoor applications [23, 8] for dense multi-view object scans, and has shown good classification results [7, 23]. Limited work was applied in the agricultural domain for LiDar data [10], but to the best of our knowledge the network has not been tested for sparse PCDS and single frame detection.

We use PointNet for detection of grape vine clusters in outdoor conditions. Due to the lack of open access RGB-D agricultural dataset, we performed a data collection where grapes were scanned using a commercial RGB-D camera mounted on a mobile robot. Only **single framed detection** was employed. A subset of the frames were labeled manually. The rest were labeled by geometrical reasoning to address the **limited training data** problem.

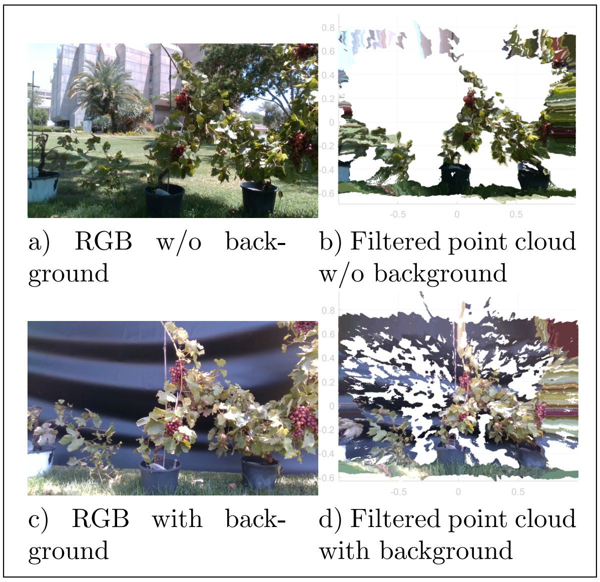
2 Methods

To achieve the goals defined above, data collection is performed in outdoor conditions on vineyard plants. Next, the data is processed and labeled by human annotators and geometrical reasoning methods. Finally, the labeled dataset is fed into the network for detection of grapes and evaluation of the algorithm performance.

2.1 Data collection

An Intel Realsense D435 depth camera 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 is an active stereo vision camera which is equipped with two RGB sensors and an active IR pattern projector [11]. This technology allows depth reading in outdoor lighted conditions.

The experimental setup (Fig. 1) included 5 vine plants in pots, placed outdoors 0.5 meter apart. A total of 9 grape clusters just harvested from a commercial vineyard were placed on the vines in both easily visible and occluded locations. A ratio of 1:1 between the occluded and non occluded grapes for a front view (camera centered to plant). Due to the camera's continues movement the visibility in each frame is dependant on camera viewpoint. The clusters were attached to the stem in a free hanging form. The robot was moved manually parallel to the plants at 1.5 meter distance while continuously acquiring RGB-D images with 1280×720 resolution at 30 fps. Six videos were acquired, each between 30-90 s, from three differ-



a) RGB w/o background

b) Filtered point cloud w/o background

c) RGB with background

d) Filtered point cloud with background

Figure 2: Acquired RGB and point cloud with and w/o background using an Intel Realsense D435 RGB-D camera.

ent viewpoints (left/front/right) with and without background. The background consisted of a black cover that was hanged behind the plants. The background is expected complicate the depth based detection by having similar range readings for the target and the background, which are difficult to filter, while the RGB detection is expected to be simpler due contrasting background. Fig. 2 presents the influence of the background on RGB and PCDs.

2.2 Data pre-processing and labeling

In the pre-processing phase, every single frame of the video is parsed into a single PCD using Intel's Realsense SDK[1]. The extracted PCDs are visually reviewed to locate frames where the plant of interest is centered ($x = [-0.5, 0.5]$). Each such PCD is filtered to remove background ($0.3 < Z < 2 \cap -1 < X < 1$), based on the experimental setting (Fig. 1). Results of pre-processing in Fig. 2 (right).

From the pre-processed PCDs, 3D labeled grape clusters are segmented in a **three step process**.

First, a subset of frames is generated by extracting every 5th frame. For each such frame, a human annotator, who first observed the approximate location of the clusters in overview images of the scene, draws a rectangle around the grape cluster in XY projection of the PCD (Fig. 3). Then, the marked labels are interpolated for the 4 frames between two consecutive labeled frames (Fig. 3), by linear interpolation of the XY coordinates of the rectangle corners initially drawn by the annotator. Finally, by including the points from the PCD that correspond to the marked 2D region, a labeled grape cluster PCD is created.

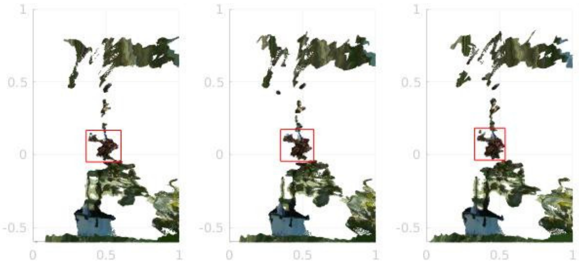


Figure 3: Example of manual and interpolated labels. Left and right - labeled manually; Middle - one of the 4 interpolated frames. While changes are subtle, the reader is advised to pay attention to the changes in pot location.

2.3 Train-test for grape detection

As mentioned, we use the deep-net PointNet [7] for detection (Fig. 4)[2]. The network is designed to input un-ordered PCDs and can be trained for PCDs with or without RGB. To comply with the network data formatting requirements, the PCDs are down-sampled to 2 cm voxel resolution and the 4096 most central voxels in X and Y are used. Only frames with at least one labeled cluster were included. A total of 5057 such PCDs (1033 manually annotated and 4024 interpolated) were analyzed according to the train-test splits in Table 1. The train-test split was done in order of acquisition, to ensure plants used for training were not used for testing. The training dataset always included the interpolated dataset (the labels generated using geometrical reasoning) while the test set was examined on both interpolated and non-interpolated labeled data. Additionally, results were examined for frames with and without background (Fig 2).

The network was evaluated under 3 different data configurations - PCD consisting of XYZ data only,

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

[2] We used the open source Python implementation- https://github.com/charlesq34/pointnet

Table 1: Detection results for training-testing splits and types of sensory data. Reported results for best combined P and R as extracted from P-R curves (Fig. 5)

		RGBXYZ				XYZ				RGB			
Training Set	Test Set	A	AC	P	R	A	AC	P	R	A	AC	P	R
Interpolated all data (2757 PCDs)	Non interpolated all data (482 PCDs)	91	74	48	53	92	62	53	25	93	61	77	24
Interpolated all data (2757 PCDs)	Interpolated all data (2300 PCDs)	92	69	59	40	92	61	57	24	93	61	82	22
Interpolated bg (1467 PCDs)	Non interpolated bg (331 PCDs)	91	65	65	33	90	61	61	23	91	60	74	21
Interpolated no bg (1290 PCDs)	Non interpolated no bg (151 PCDs)	92	77	30	60	96	60	44	21	97	66	72	34

PCD=point-cloud; A = Accuracy; AC=Average Class Accuracy; P = Precision; R = Recall

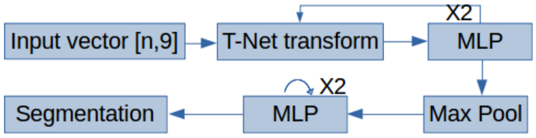


Figure 4: Schematic architecture of PointNet. For original architecture refer to [7]

RGB data only, or both (XYZRGB). Each detection evaluated by overall accuracy (A), average class (grape/background) accuracy (AC), precision (P), and recall (R). Each train-test and data configuration was trained max 50 epochs.

3 Results

The best precision-recall (P-R) rates for each data configuration (XYZ, RGB and RGBXYZ) are reported in Table 1. Since grape detection can be used for several agri-robotic applications, there is no clear priority for lower FP or lower FN and therefore the chosen *best* combination is the one maximizing both precision and recall as shown in the P-R curves in Fig. 5.

For RGBXYZ data, the highest average accuracy is achieved for the no-background dataset (77%). This training-testing split provides the lowest precision rate (30%) and the highest recall rate (60%), suggesting that for no background data the vines are detected to a high degree but many non-vine regions were falsely detected as grapes. The data *with* background yields lowest average accuracy (65%), highest precision (65%) and lowest recall (30%) suggesting with background a lot of the grapes were missed but non grape regions were less likely to be mis-classified.

The most variable dataset that includes PCDs acquired both with and without a background provides a good balance between the two detection rates with average class accuracy of 74%, precision of 48% and recall of 53%. The *red* P-R curve in Fig. 5 presents rather constant levels of precision for a range of recall values.

The results in Table 1 also show that inclusion of interpolated data in the test sets slightly harms the results (69% average accuracy compared to 74%), while the gap between precision and recall slightly increases. This might be explained by errors included in the dataset by interpolation of labels, where regions that were not grape, and were classified as non grapes were mistakenly labeled as grape.

For XYZ data the results show similar overall accuracy but significantly lower average class accuracy then for RGBXYZ (60-62% compared to 65-77%). The *blue* curve in Fig. 5 shows similar precision was achieved for the XYZ data compared to the RGBXYZ with significantly lower recall rates. This suggest that the use of XYZ retained the number of FPs but significantly increased the FNs, which means the XYZ network detected significantly less grapes then the RGBXYZ.

The RGB data showed similar results in average accuracy to the XYZ data (~60%) but a significantly higher precision for the same recall levels (74-82% precision compared to 57-61% for 20-35% recall). This suggests that for the same ratio of undetected grapes the number of falsely detected grapes was significantly lower and in general the true positive rate is high. This means that RGB data is a more vital cue for grapes detection as com-

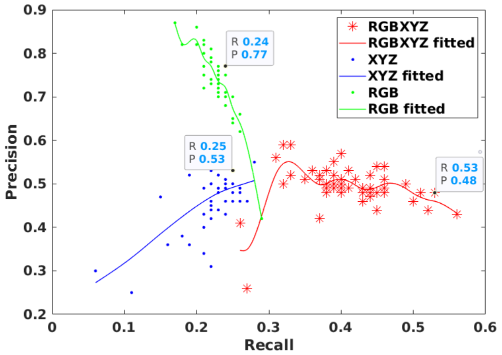


Figure 5: Precision-Recall curve for RGBXYZ, XYZ and RGB based detection for full dataset. The marked data points are the chosen best configuration reported(first row Table 1)

pared to XYZ data. Comparing the RGB data to the RGBXYZ data using the P-R curves (*green* and *red* respectively in Fig. 5) shows a decrease in precision with the usage of XYZ in addition to RGB but increase in recall. This suggests that the RGBXYZ network is less conservative than the RGB network (RGBXYZ assigns more regions to be grapes then RGB) with lower FNs but higher FPs.

4 Conclusions

This paper presents the use of a commercial grade RGB-D camera for detection of grape vines in outdoor conditions, for agrobotics applications. Detection results show overall best accuracy of 91% (average class accuracy of 74%) when using both RGB and XYZ data. The results were obtained from a **single frame detection** configuration with limited manually annotated data (1033 PCDs), based on low spatial resolution data. The results also show an expected increase in accuracy when both RGB and XYZ data are used compared to solely RGB or XYZ. Usage of uniform background yielded higher precision but lower recall for RGBXYZ based detection. For XYZ detection, recall remained similar for both background and no background data but a higher precision for frames with background. Hence XYZ info contributes more for detection when using a background. For RGB detection, increase in recall was observed for no background data suggesting lower number of FN when no background is present. These conclusions contradict with the initial assumption that placing a uniform background will increase the complexity for depth detection and decrease the complexity of RGB detection and should be investigated further.

To give the results a context, a review of previous research on grape detection shows similar or slightly better results. Nuske et al. [16] used K-NN and obtained 63.7% precision and 98% recall. Berenstein et al. [6] used morphological operations resulting with an accuracy of 90% (no precision or recall reported). Zemmour et al. [22] used adaptive thresholds and achieved 89% recall. The authors did not report precision or accuracy but reported 33% TPR. Compared to other field crops with a more prominent color contrast to background (e.g. [2, 3, 20, 22]) grape detection has overall lower detection results which shows the task complexity.

Our results show that addition of depth information to RGB data can lead to improved grape vines detecting results with respect to some metrics. With more sensors providing better RGB-D data, detection rates are expected to rise. Furthermore, we present the importance of automatic annotation. Nevertheless, the detection rates presented are similar or slightly lower then the reported numbers in the literature, that are based on RGB only. This is most probable due to an important aspect of RGB that is lost with the analysis of PCD data: the spatial order of the image. A possible continuation of the work would include usage of *ordered* PCD detection networks.

References

[1] T. Adão, J. Hruška, L. Pádua, J. Bessa, E. Peres, R. Morais, and J. Sousa. Hyperspectral imaging: A review on uav-based sensors, data processing and applications for agriculture and forestry. *Remote Sensing*, 9(11):1110, 2017.

[2] 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*, 19(6):1390, 2019.

[3] C. W. Bac, E. J. van Henten, J. Hemming, and Y. Edan. Harvesting Robots for High-value

Crops: State-of-the-art Review and Challenges Ahead. *JFR*, 31(6):888–911, 2014.

[4] W. Bac, J. Hemming, B. van Tuijl, R. Barth, E. Wais, and E.J. Van Henten. Performance evaluation of a harvesting robot for sweet pepper. *JFR*, 34(6):1123–1139, 2017.

[5] R. Barth, J. Hemming, and E. J. van Henten. Design of an eye-in-hand sensing and servo control framework for harvesting robotics in dense vegetation. *Biosystems Engineering*, 146:71–84, 2016.

[6] R. Berenstein, O. B. Shahar, A. Shapiro, and Y. Edan. Grape clusters and foliage detection algorithms for autonomous selective vineyard sprayer. *Intelligent Service Robotics*, 3(4):233–243, 2010.

[7] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CCVPR*, pages 652–660, 2017.

[8] R. Cong, H. Chen, H. Zhu, and H. Fu. Foreground detection and segmentation in rgb-d images. In *RGB-D Image Analysis and Processing*, pages 221–241. Springer, 2019.

[9] Y. Edan, S. Han, and N. Kondo. Automation in agriculture. In *Springer handbook of automation*, pages 1095–1128. Springer, 2009.

[10] D. A. Eroshenkova and V. I. Terekhov. Automated determination of forest-vegetation characteristics with the use of a neural network of deep learning. In *Advances in Neural Computation, Machine Learning, and Cognitive Research III*, volume 2, page 295. Springer Nature, 2019.

[11] A. Grunnet-Jepsen and D. Tong. Depth post-processing for intel realsense d400 depth cameras. *New Technologies Group, Intel Corporation*, 2018.

[12] A. Kamilaris and F. X. Prenafeta-Boldú. Deep learning in agriculture: A survey. *COMPAG*, 147:70–90, 2018.

[13] M. Kragh, R. N. Jørgensen, and H. Pedersen. Object detection and terrain classification in agricultural fields using 3d lidar data. In *International conference on computer vision systems*, pages 188–197. Springer, 2015.

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

[15] P. Kurtser and Y. Edan. The use of dynamic sensing strategies to improve detection for a pepper harvesting robot. In *IROS*, pages 8286–8293, 2018.

[16] S. Nuske, S. Achar, T. Bates, S. Narasimhan, and S. Singh. Yield estimation in vineyards by visual grape detection. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2352–2358. IEEE, 2011.

[17] O. Ringdahl, P. Kurtser, and Y. Edan. Evaluation of approach strategies for harvesting robots: Case study of sweet pepper harvesting. *Journal of Intelligent and Robotic Systems*, pages 1–16, 2018.

[18] O. Ringdahl, P. Kurtser, and Y. Edan. Performance of rgb-d camera for different object types in greenhouse conditions. In *ECMR*, pages 1–6. IEEE, 2019.

[19] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, 2008.

[20] I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. McCool. Deepfruits: A fruit detection system using deep neural networks. *Sensors*, 16(8):1222, 2016.

[21] A. Vit and G. Shani. Comparing rgb-d sensors for close range outdoor agricultural phenotyping. *Sensors*, 18(12):4413, 2018.

[22] E. Zemmour, P. Kurtser, and Y. Edan. Automatic parameter tuning for adaptive thresholding in fruit detection. *Sensors*, 19(9):2130, 2019.

[23] C. Zhao, L. Sun, P. Purkait, T. Duckett, and R. Stolkin. Dense rgb-d semantic mapping with pixel-voxel neural network. *Sensors*, 18(9):3099, 2018.