

Creating a Segmented Pointcloud of Grapevines by Combining Multiple Viewpoints Through Visual Odometry

Michael Adlerstein¹, Angelo Bratta¹, João Carlos Virgolino Soares¹, Giovanni Dessy¹, Miguel Fernandes¹, Matteo Gatti², Claudio Semini¹

Abstract—Grapevine winter pruning is a labor-intensive and repetitive process that significantly influences the quality and quantity of the grape harvest and produced wine of the following season. It requires a careful and expert detection of the point to be cut. Because of its complexity, repetitive nature and time constraint, the task requires skilled labor that needs to be trained. This extended abstract presents the computer vision pipeline employed in project Vinum, using detectron2 as a segmentation network and keypoint visual odometry to merge different observation into a single pointcloud used to make informed pruning decisions.

I. INTRODUCTION

Winter pruning is a critical, labor-intensive task in vineyards, crucial for balancing yield and grape quality to maximize income. It requires 80 to 120 man-hours per hectare annually. The process becomes even more significant considering labor shortages in agriculture. Skilled workers in vineyards must know plant anatomy to make the correct pruning cuts. Identifying the correct pruning regions is a non-trivial task that requires a trained eye in order to distinguish between the different organs, creating a need for a reliable and accurate system to identify such regions and pruning points. With the rapid advancements in machine learning and neural networks, autonomous methods have been explored to bridge this gap.

In the agricultural sector neural networks have been used for weeds detection [1] [2] and for fruit detection and segmentation, as demonstrated by Borianne et al. [3], and Santos et al. [4]. Further applications include plant phenotyping where Grimm et al. [5] utilized neural networks for detailed plant analysis.

In the context of vineyard pruning, Gentilhomme et al. [6] presented ViNet, a network specifically designed to identify nodes within a grapevine plant. Their method relies on a stacked hourglass network [7] for detailed reconstruction. The paper also introduces the 3D2Cut dataset, which provides annotated data with node information against synthetic backgrounds. Guadagna et. al. [8] demonstrated that a R-CNN model is successful in achieving organ segmentation (cordón, arm, spur, cane and node are the visible organs on a grapevine) demonstrating promising results in terms of recall, highlighting a detection performance for visible co-planar simple spurs with high precision and recall. Notable contributions also came from Fernandes et al. [9] which

demonstrated that it is possible to detect the potential pruning points based on 3D scans of the pruning regions, and Botterill et al. [10] and Silwal et al. [11] which developed robot prototypes towards the automation of pruning.

Despite the promising advancements showcased, neural networks are not always able to give a one-shot complete segmentation of the plant. This is because the dataset is not diverse enough and the labelling of the different organs is a non-trivial task that requires a specialized person. In addition, some organs might be occluded from a certain point of view, requiring more than one image; as a result the neural networks are sometimes unable to give an accurate result from just one image. This extended abstract builds upon the previous work done inside the Vinum project¹ (e.g. [8], [9]). In particular, we describe here an approach for 2-dimensional visual odometry used to combine different viewpoints into one segmented pointcloud, resulting in a detailed 4D structure (x y z position and class label), which can be used by a robot to accurately prune the grapevines.

II. METHOD

A convolutional neural network using the detectron2 backbone is trained on a dataset containing segmented images of grapevines, enabling the detection of the five main organ classes (cordón, arm, spur, cane and node) [8]. This training allows for precise inferences on new images, as illustrated in Fig. 1, showcasing a recall of 81% and precision of 97%. To construct a more accurate representation of the plant, pictures from different viewpoints are taken using a robotic manipulator equipped with an RGB-D camera, covering the entire plant. This process ensures that the same part of the plant appears in multiple frames, facilitating the extraction of keypoints. The keypoints are then matched across different instances and projected into 3D space. Following this, a transformation matrix is calculated, enabling the transformation of the entire pointcloud into the same frame of reference as the robot, thereby merging multiple distinct viewpoints into a single pointcloud.

A. Keypoint Matching

Traditional descriptor methods, such as ORB [12], were not suitable for the task due to the homogeneous colour and texture of the plant. Advancements in the state of the art have led to the adoption of SuperPoint [13], which offers a balanced approach between detecting a substantial

¹ Dynamic Legged Systems Laboratory, Istituto Italiano di Tecnologia (IIT), Genova, Italy. Email: name.surname@iit.it

² Department of Sustainable Crop Production (DI.PRO.VE.S.), Università Cattolica del Sacro Cuore, Via Emilia Parmense 84, 29122 Piacenza, Italy.

¹<https://vinum-robot.eu/>



Fig. 1. Output inference of detectron2, giving precise segmentation labels of the five classes.

number of keypoints and maintaining manageable inference times. LightGlue [14] was chosen as the matcher due to its speed and matching reliability. These novel approaches, utilizing neural networks and attention, represent a significant improvement over traditional methods by offering a superior trade-off between the speed of inference and the number of detected features. The improvements in the detection and the matching allow Visual odometry to merge pointclouds, which are spatially distant from each other reducing the number of frames used whilst also maximising the amount of new information detected, as shown in Fig. 2.

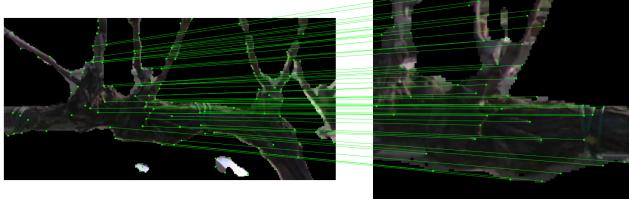


Fig. 2. Keypoint detection and matching using Superpoint and Lightglue.

B. 3D Projection and Registration

Once the keypoints are matched, the registration is achieved through the minimization in the Orthogonal Procrustes problem. This method seeks to optimally map a set of keypoints between two distinct frames of reference. Allowing for the alignment of two different pointclouds into the same frame. Mathematically, the problem is formulated as follows:

$$\min_{\mathbf{Q}} \|\mathbf{A} - \mathbf{B}\mathbf{Q}\|_F \quad \text{subject to} \quad \mathbf{Q}^T \mathbf{Q} = \mathbf{Q} \mathbf{Q}^T = I, \quad (1)$$

where \mathbf{A} and \mathbf{B} are vectors containing the positions of the keypoints to be matched. The term $\|\cdot\|_F$ denotes the Frobenius norm, and \mathbf{Q} represents the rotation matrix that minimizes the equation.

To determine the quality of alignments, two metrics are used: RMSE (Root Mean Square Error) and fitness. If the fitness value is too low or the RMSE value is too high, the match is discarded, and a new set of frames is selected for alignment. After aligning all pointclouds to the same frame of reference (the camera frame), a final transformation then merges pointclouds into the base frame, in order for

the pointcloud to be in the same frame as the robotic manipulator.

For global optimization, pose graph optimization is employed to refine locally optimized transformations from the final matched pointclouds. This approach uses a "pose graph" structure to achieve optimal overlap between pointclouds. The pose graph comprises nodes representing the positions of the pointclouds and edges that represent transformations between these nodes.

C. Clustering

Once the entire scan set of RGB-D images is obtained and correctly aligned, they are merged into a single pointcloud representing the entire structure of the plant. HDBSCAN [15] is used to address the clustering issue, as it works well with different densities of points allowing for the creation of comprehensive clusters. The process of clustering points is crucial for distinguishing between different instances of the same object type, such as individual organs in a plant as shown in Fig. 3. This enables detailed analysis of the plant, which is essential for making informed decisions regarding which parts to keep or remove. The HDBSCAN merges all the points belonging to the same class from different scans, discarding duplicate points and selecting the instance which has the highest probability.

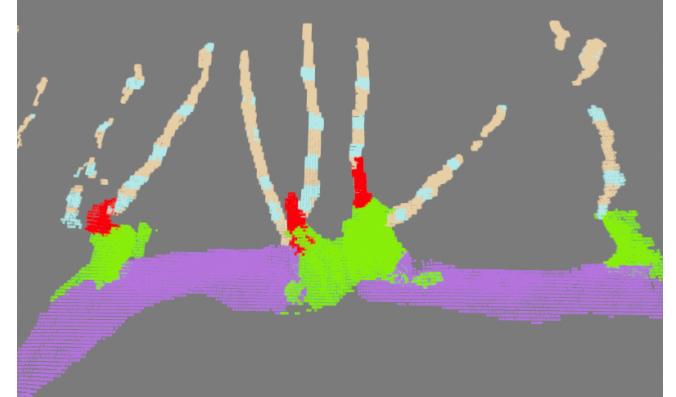


Fig. 3. Final pointcloud with segmented organs.

III. CONCLUSION AND FUTURE WORK

This extended abstract presented a novel approach for scanning grapevines, allowing the creation of an accurate pointcloud representation that can be used by a robotic manipulator to plan for accurate pruning. This approach is able to produce a 3D representation of the plant and its segmented organs in matter of minutes. The merging of pointclouds from several frames enables the system to account for neural network failures, making the system more robust. The system occasionally fails to cluster regions depending on the morphology of the plant due to the nature of HDBSCAN. Furthermore, due to the resolution of the camera some ghosting can be observed on the canes resulting in a less precise pruning region calculation. To account for these shortcomings, more work has to be done to have a better clustering algorithm.

REFERENCES

- [1] M. Di Cicco, C. Potena, G. Grisetti, and A. Pretto, “Automatic model based dataset generation for fast and accurate crop and weeds detection,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5188–5195.
- [2] A. Milioto, P. Lottes, and C. Stachniss, “Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns,” in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 2229–2235.
- [3] P. Borianne, F. Borne, J. Sarron, and É. Faye, “Deep mangoes: from fruit detection to cultivar identification in colour images of mango trees,” *arXiv preprint arXiv:1909.10939*, 2019.
- [4] T. T. Santos, L. L. De Souza, A. A. dos Santos, and S. Avila, “Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association,” *Computers and Electronics in Agriculture*, vol. 170, p. 105247, 2020.
- [5] J. Grimm, K. Herzog, F. Rist, A. Kicherer, R. Töpfer, and V. Steinhage, “An adaptive approach for automated grapevine phenotyping using vgg-based convolutional neural networks,” *arXiv preprint arXiv:1811.09561*, 2018.
- [6] T. Gentilhomme, M. Villamizar, J. Corre, and J.-M. Odobez, “Towards smart pruning: Vinet, a deep-learning approach for grapevine structure estimation,” *Computers and Electronics in Agriculture*, vol. 207, p. 107736, 2023.
- [7] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in *Computer Vision–ECCV*. Springer, 2016, pp. 483–499.
- [8] P. Guadagna, M. Fernandes, F. Chen, A. Santamaria, T. Teng, T. Frioni, D. Caldwell, S. Poni, C. Semini, and M. Gatti, “Using deep learning for pruning region detection and plant organ segmentation in dormant spur-pruned grapevines,” *Precision Agriculture*, vol. 24, no. 4, pp. 1547–1569, 2023.
- [9] M. Fernandes, A. Scaldaferri, G. Fiameni, T. Teng, M. Gatti, S. Poni, C. Semini, D. Caldwell, and F. Chen, “Grapevine winter pruning automation: On potential pruning points detection through 2d plant modeling using grapevine segmentation,” in *IEEE Cyber*, 2021.
- [10] T. Botterill, S. Paulin, R. Green, S. Williams, J. Lin, V. Saxton, S. Mills, X. Chen, and S. Corbett-Davies, “A robot system for pruning grape vines,” *Journal of Field Robotics*, vol. 34, no. 6, pp. 1100–1122, 2017.
- [11] A. Silwal, F. Yandun, A. Nellithimaru, T. Bates, and G. Kantor, “Bumblebee: A path towards fully autonomous robotic vine pruning,” *Field Robotics*, 2022.
- [12] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: An efficient alternative to sift or surf,” in *IEEE International Conference on Computer Vision*, 2011, pp. 2564–2571.
- [13] D. DeTone, T. Malisiewicz, and A. Rabinovich, “Superpoint: Self-supervised interest point detection and description,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.
- [14] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, “Lightglue: Local feature matching at light speed,” in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17627–17638.
- [15] G. Stewart and M. Al-Khassawneh, “An implementation of the hdbSCAN* clustering algorithm,” *Applied Sciences*, vol. 12, no. 5, p. 2405, 2022.