

3D Robotics and LMM for Vineyard Inspection

Samuele Facenda¹, Paweł Trybała¹, Luca Morelli¹, Nazanin Padkan^{1,2}, Fabio Remondino¹

¹ 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy

² Dept. Mathematics, Computer Science and Physics, University of Udine, Italy

Web: <http://3dom.fbk.eu> - Email: (sfacenda, ptrybala, lmorelli, npadkan, remondino)@fbk.eu

Keywords: Mobile robotics, Precision agriculture, Autonomous inspection, Planning, 3D, Large Multimodal Models.

Abstract

Autonomous mobile robotic solutions are increasingly being explored in precision agriculture to aid human workers in labour-intensive or repetitive tasks. Moreover, the emergence of foundation models in vision-based AI domain presents an opportunity to perform automated interpretation of in-field collected data. This study presents a cost-effective mobile robotic research platform designed for autonomous vineyard inspection: it integrates mission planning, real-world navigation and a post-processing pipeline of multimodal data. The system, based on the Leo rover, is equipped with LiDAR, RGB cameras and GNSS-visual-inertial positioning, ensuring reliable operation in GNSS-degraded vineyard environments. We propose a novel methodology for automating several stages of the workflow using various open and in-situ collected data. The robotic platform and processing pipeline were validated through simulation and field experiments, demonstrating its capability for autonomous navigation, 3D reconstruction, AI-based fruit detection and an initial plant health assessment through Large Multimodal Models (LMM). Results show that while 3D mapping provides high-resolution spatial data, AI-driven object detection and vision models require further domain adaptation for reaching reliable and trustable operation. The study highlights the feasibility of cost-effective mobile robotic solutions in vineyard monitoring and the potential of integrating AI to enhance agricultural automation.

1. Introduction

The agricultural sector represents a highly promising domain for the deployment of mobile robotic systems, particularly due to the scale and repetitive nature of farming operations throughout the year (Gao et al., 2018; Hrabar et al., 2021). These characteristics create significant potential for reducing manual labour through autonomous robotic solutions. In many agricultural environments, with unobstructed sky, it is possible to boost automation with Unmanned Aerial Vehicles (UAVs) and ground-based robots, that can fully rely on Global Navigation Satellite System (GNSS) positioning for navigation and task execution. Current robotic technologies already deliver robust solutions for applications such as e.g., precision plant spraying and large-scale remote sensing monitoring (Neupane and Baysal-Gurel, 2021; Hanif et al., 2022; Di Gennaro 2023; Wang et al., 2024).

However, certain agricultural tasks, particularly those requiring close-proximity inspection, centimeter-level high spatial resolution data or physical interaction, demand higher levels of positioning precision and reliability (Liu and Liu, 2024). Achieving these capabilities often necessitates the integration of additional sensors to compensate for the limitations of GNSS-based systems, particularly in complex environments. Such challenging conditions can be found in pergola or guyot vineyard cultivations, where dense grapevine foliage forms archways that degrade the quality of GNSS signals, making global positioning systems unreliable.

At the same time, the production of high-quality wine grapes relies on frequent and detailed inspections to monitor plant growth and detect early signs of disease or damage. Timely and accurate detection is critical for maintaining crop health and ensuring optimal yield. Autonomous mobile robots capable of operating effectively in partially GNSS-denied, natural environments offer a compelling solution to this challenge (Fasiolo et al., 2023). By leveraging intelligent mission planning, 3D mapping technologies and artificial intelligence (AI)-assisted image analysis (Mendes et al., 2022), robotics solutions can support efficient inspections and management of vineyards, addressing the precision requirements and adaptability needed for such tasks (Hrabar et al., 2021; Izquierdo-Bueno et al., 2024).

In recent years, mobile robotics applications specifically targeting ground-based applications in the vineyards started to appear, showing growing affordability and feasibility of employing autonomous solutions in these settings. Many of these scientific works leveraged the advantages of AI methods and 3D measurements. A study by Roure et al. (2018) presented a mobile robot designed to distribute pheromone dispensers in the spring across the vineyard, focusing also on a low-cost positioning solutions dedicated for the vineyard setting. Their lesson learnt reported the importance of dense 3D reconstruction of grapevines (Roure et al., 2020). Similarly, Williams et al. (2023) investigated stereo-based 3D reconstruction of plants in a vineyard, jointly with deep learning methods for panoptic segmentation, to perform autonomous cane pruning. Iberraken et al. (2022) presented a high-resolution 3D digitization of a real vineyard, acting as a very realistic simulation environment in *Gazebo* to develop and test a novel navigation approach for vineyard monitoring robots. Lastly, Stavridis et al. (2024) proposed two separate mobile robotic platforms, enabling autonomous execution of tasks related to inspection or harvesting, respectively. Although presenting impressive abilities, such advanced and complex solution required on-board employment costly hardware, like manipulators and hyperspectral cameras, limiting its applicability on a large scale.

1.1 Aim of the work

In this work, we present an in-house assembled ground mobile robotic platform, based on the *Leo* rover¹, for autonomous 3D mapping and AI-assisted inspection of vineyards. We focus on automation capabilities in different stages of the proposed framework: from the mission planning (before entering the field), through verification in the simulation, to autonomous mission in real outdoor conditions and post-processing of the acquired image and LiDAR data. Our aim is to demonstrate how, and to what extent, current state-of-the-art methods from robotics, computer science and geomatics communities are ready to be integrated into a unique and reliable framework able to:

- support the planning, testing and deployment of a robotic platform in real use-case scenarios for 3D mapping purposes;

¹ <https://www.leorover.tech/>

- enable periodic autonomous monitoring of agricultural environments;
- provide AI-based data interpretation to support human experts in automatically identifying relevant location with potential issues (e.g. diseases).

We propose a cost-effective framework to support these challenging tasks and we demonstrated it in a test site in Trento, Italy. With respect to the literature, our work presents:

- a technology integration on a low-cost robotic platform, including real-time navigation and obstacle avoidance;
- tests of AI components on autonomously acquired data in the field to support decision-making.

Therefore this study exhibits a range of robotic and AI-driven solutions with potential applications in vineyard management, including fruit detection and mapping, 3D digital twin creation and vision-based inspection. The ultimate goal of the proposed framework is not to replace qualified human experts in the field but to optimize and support their workflow by automating non-critical, repetitive tasks, in an effort to finally enhance efficiency in vineyard monitoring and management.

2. Multi-sensor mobile robotic research platform

The proposed robotic solution is based on a wheeled rover *Leo*, equipped with sensors selected specifically for such task (Figure 1). The small size of the robot (a footprint of approx. 0.5 x 0.5 m) allows it to traverse the field with minimal impact on the surroundings and manoeuvre easily even in narrow lanes, still allowing a good payload. The sensors carried by the rover includes: a MandEye device (3D Livox LiDAR with an integrated IMU; Będkowski, 2024), a 2D laser scanner RPLidar A2, two RGB cameras, a RGB-D Intel RealSense D435i sensor and a Fixposition Vision-RTK2² global positioning system (internally fusing GNSS, inertial and visual data).



Figure 1. The FBK-3DOM mobile robot *Leo* in the field with its sensors.

These sensors enable acquiring different types of data for various, dedicated purposes, optimizing the needs of the mission and the limited real-time computational capability. Namely, the 3D LiDAR data is used only in post-processing for obtaining a dense point cloud of the surveyed site, as the Fixposition sensor already provides a reliable positioning solution for real-time navigation. On the other hand, the 2D laser scanner and depth camera are used for onboard real-time obstacle avoidance as they are less suitable for high-quality 3D mapping (and their data are not stored on the robot). Finally, images from the two RGB cameras pointing forward are used in the post-processing stage for object detection and Large Multimodal Models (LMM) queries.

3. Methodology for autonomous vineyard inspection

3.1 Mission planning

In general, mobile inspection tasks are expected to be repeatable to allow capturing the data in a similar manner, in turn helping to

establish the data associations between different epochs (Maset et al., 2022). Due to that, the mission trajectory should be planned beforehand, in contrast to free autonomous exploration approach, to yield more predictable, consistent results and optimize the execution time. Moreover, the planning phase can be automatized, utilizing remote sensing data.

Because of that, we extended and modified the method of Hassanein et al. (2019) for crop row detection in UAV imagery. The proposed method is used to compute a general path that goes across all the vineyard lanes in the geofenced area. For the robot, this path, consisting of waypoints defined by geographical coordinates, will constitute a global plan, that it is supposed to roughly follow. During the mission execution, the local planner is used to actively control the short-term robot motion, detecting obstacles in the vicinity of the robot.

The described mission planning method is designed to work well on a wide range of typical vineyard layouts. As input data, we use a georeferenced RGBI aerial orthoimage of the area, in addition to the user-defined geofence of the field. The first step is the detection of the average azimuthal angle of the lanes in the image. A small vertical section of the image, of a width similar to the expected row width, is rotated and analysed for every candidate angle; a step of 0.25° was empirically found to yield good results for vineyards with an area below ca. 10 km².

For every section, a Principal Component Analysis (PCA) of the raster values is performed on the greyscale image strip to identify the optimal rotation angle through calculating two first principal components. Several PCA strategies were considered. The sum of the principal components was found to be the strongest indicator for a cross-section not representing a line; the variance is excessively noisy on some fields. The second principal component often manifests sinusoidal shapes that conflicts with the peaks at the row inclination angle. Thus, the angle with the lowest sum of principal components is taken as the main azimuth of the grapevine rows. An example comparison of those indicators is presented in Figure 2.

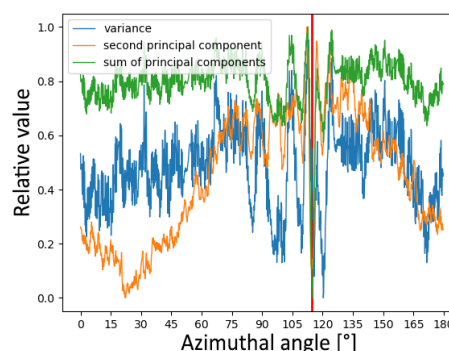


Figure 2. Different variance indicators in the data for detecting the field image rotation angle. The minimum of the sum of principal components is shown as a red vertical line.

The next step involves detecting grapevine rows' locations. Using the computed azimuthal angle, the image is rotated to align the plant rows to appear vertical in the rotated image. Then, for every column, a mean NDVI value is computed. In this step, we use an NDVI raster, calculated using an RGBI orthophoto, as it gave clearer results across different tested vineyards designs and inclinations. Furthermore, to increase the robustness of the method, a two-step refinement procedure is carried out to reduce the probability of detecting wrong, doubled lanes, and missing correct ones. The peaks above a certain prominence threshold of the NVDI are chosen and passed through a simple outlier filter

² <https://www.fixposition.com/pages/our-product>

(see Table 1). The filter creates a distribution of all distances between rows detected up to the analyzed row. Given this distribution, a standard score of the newly calculated distance is computed. If it is higher than a predefined threshold (e.g., 2σ), the lane position is substituted by the position of the previous lane shifted by the median distance between the lanes. Then, the empty spaces between the lanes are identified, selecting the distances higher than two times the median multiplied by a tolerance factor (<1). Those gaps are filled with a lane with a distance to the previous one equal to the median distance between the lanes. This step is carried out in both directions. In the end, the duplicates are removed considering as duplicate lanes the ones that have a distance lower than the median distance multiplied by the tolerance factor. The process is repeated iteratively until no more changes can be made.

Finally, having obtained the georeferenced line equations, they are intersected with the geofencing polygon. The points of intersections, taken in an alternating way between northmost and southmost points for each subsequent line, define a global inspection path for the robot.

3.2 In-field inspection

Preparing the complete software stack for an autonomous mobile robot, despite utilizing many off-the-shelf components available in the Robot Operating System (ROS) environment, is a complex process that requires extensive testing, particularly time-consuming when conducted in real-world environments. To limit the time spent in the field, the development and testing processes can be greatly simplified and speeded up with a faithful simulation of the use case environment. For this purpose, we adopted an open-source Gazebo world of a vineyard (Hroob et al., 2021) with our simulated *Leo* rover.

Through this software and our configuration, all onboard sensors (excluding the 3D scanner, which acquires and stores the data independently) are simulated and seamlessly interact with the running robot software. The sensors output is computed in the simulation world and then modified adding some noise. Within the developed simulation environment, different algorithms and approaches for navigation and mapping were tested, to finally develop a use case specific robot configuration.

This setup was specifically tailored to our needs, i.e. a mobile robot with low processing power that can navigate through uneven outdoor terrain. As the precise, large-scale consistent 3D mapping and navigation would greatly increase the complexity and require much more computational power, we decided it was not feasible to run it online on the rover. Thus, we exploited the features of the standard 2D ROS navigation stack to obtain a lightweight and simple yet effective setup.

Leveraging the combination of precise 3D localization from the Vision-RTK2 sensor and 3D point clouds from the RealSense camera, a 3D obstacle map was created with a consistency and resolution enabling successful navigation. For this step, we used a spatio-temporal-voxel-layer (Macenski et al., 2020), as it is specifically designed to efficiently save 3D data and build a 2D map from them. Instead of using a full SLAM solution for the map refinement, we utilized only the pose from the Vision-RTK2 as the single source of pose estimates. The spatio-temporal-voxel-layer simply projects the point cloud data from the 2D laser scanner and the RealSense, filtered by height to remove the ground and the canopy ceiling, into the 3D space to create the obstacle map. This map is then used to generate a 2D cost-map by projecting occupied voxels onto a planar grid, with gradually inflating cost values towards the obstacle locations.

Since the Vision-RTK2 provides a GNSS-based global positioning, rapid shifts can be expected in a situation when the

high-quality satellite-based position is available after a period of dead reckoning based only on inertial and visual data. As this can cause issues for the robot control, a pose smoothing is applied to remove such sudden pose changes. Effectively, this is done by calculating the robot pose in two reference frames: one that provides a continuous, smooth trajectory, and its parent, that maintains a valid global positioning in relation to the ENU frame. The global planner then uses the planar projection of the spatio-temporal-voxel-layer and the list of waypoints' UTM coordinates as an input to provide a global plan. The plan is then realized by the local planner. We selected *teb_local_planner* (Rösmann et al., 2017) due to its suitability for car-like movements, in an Ackermann steering scheme. It is worth notice that such planner improves the robot performance in our scenario of carrying out an inspection mission in an uneven ground environment, as it avoids issues with sensor shaking, slipping or temporarily too high grip that can be experienced with differential drive planners.

3.3 Data post-processing

Part of the data collected during the autonomous mission is not processed in real-time for the autonomous navigation and execution of the mission presented in the previous sections. This data is, however, crucial to provide qualitative and quantitative insights about the vineyard through AI-assisted methods, and it is collected and stored to allow some aspects of the inspection to be carried out in post-processing.

First, the data acquired with MandEye LiDAR sensor are processed with a LiDAR-inertial SLAM method, described in Trybała et al. (2023). The pose graph-based approach generates a dense 3D representation of the vineyard, enabling for example volumetric calculations and 3D change detection. Although not providing georeferenced results directly, thanks to the time synchronization of all sensors on the robot, the trajectory is indirectly georeferenced through alignment with the georeferenced trajectory generated by Vision-RTK2. Since both sensors estimate the robot pose with a high frequency, in metric reference frames, the consistency and coherence of positioning provided by them is evaluated. As in our experiments no ground truth trajectory is available, the deviations in global trajectory and relative poses (i.e., differences in subsequent poses reported by both systems at the same time) computed independently with both devices were analyzed.

Next, image analysis and object detection methods are investigated to identify and localize the fruit growth in the field for purposes of inventory and monitoring. YOLO (Redmon et al., 2016) was employed as one of the state-of-the-art neural networks for object detection and capable to achieve both high precision and speed in various object detection tasks. Despite its age, continuous developments of new YOLO models keep it one of the currently best performing methods. For our use case of grape detection, we used YOLOv8 (Jocher, 2023). Since in the originally released pretrained model grape class does not exist, we leveraged transfer learning on our dataset to adapt it for our needs. The retraining was performed only on a limited sample of the images, captured by the robot in the vineyard with the forward-looking cameras. The dataset was then split into train, validation and test sets, and the performance evaluation was performed only using a dedicated test subset.

Finally, we deployed a suit of state-of-the-art LMMs on a Jetson Orin Nano Super (8GB VRAM) to test the feasibility of using them in real time for detecting clues of unhealthy growth of grapevines based on the images collected by the robot.


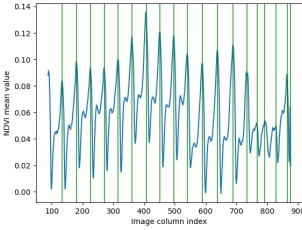


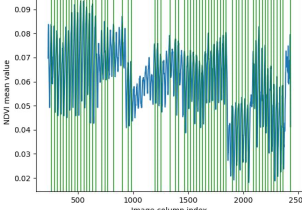
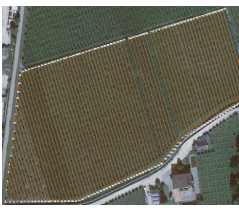

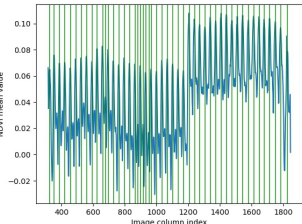
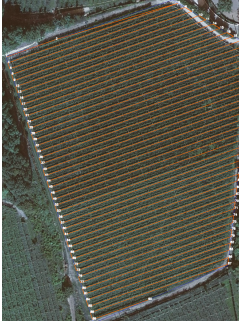
Vineyard #	Cropped RGBI orthomosaic	Peaks of mean NVDI values per a raster section	Row width σ before refinement [m]	Row width σ after refinement [m]	Generated mission plan
1			0.97	0.32	
2			2.66	0.35	
3			0.66	0.11	

Table 1. The three test cases used to evaluate the proposed mission planning method: employed orthomosaic, identified vineyard lanes and generated inspection plan.

These clues could include discoloured leaves, dried out fruits, or other visual signs of plants' distress or disease. Due to the limitations of the image resolution, the low ground sampling distance (GSD) on the grapes and the use of only the general knowledge of the models (i.e., employing them as zero-shot solutions), only clearly identifiable symptoms were taken into account. From a full image sequence in the field, 12 images of 2 MP resolution were manually selected: 4 without any noticeable issues, and 8 with some worrying symptoms. However, due to overall good health status of the vineyard, it was not possible to test the ability to recognize severe diseases or damages. Thanks to the synchronization of the two RGB cameras with the continuously geolocalized robot, the LMM solutions have the potential to directly provide geographic locations of detected signs of degraded plant's health. Nevertheless, these should always be only an indication for a human expert to investigate the issues by themselves and make the final verdict.

In the evaluation phase, different LMMs were evaluated, instructing them to detect the issues described above. We finally used *ollama*³, an open-source and user-friendly platform for running large language or multimodal models (LLMs/LMMs) directly on a local machine. We used low model temperature whenever possible (to make the model's output more deterministic and repetitive) and forcing the output to be in JSON format with a Boolean classification result. Not significant changes to the prompts had to be made to adapt it to specific requirements of some models. We ran the inference 10 independent times for each considered image to take into account

the randomness of the models. Additionally, we used selected commercial, closed-source models to test if there is a noticeable difference when offloading the task to a much more complex model in an offline manner. These tests, however, were not repeated multiple times, thus constitute only a rough indication of their performance.

4. Results

All components were deployed in a single vineyard: from the phase of mission planning, through verification in the simulation, to mission execution in the field and offline data post-processing. The outdoor tests were performed in September 2024. Additionally, we validated the mission planning (Section 3.1) on two other vineyards of different size and characteristics.

4.1 Mission Planning

The robot path plan generation was tested on 3 distinct vineyards (Table 1) in the Italian Trentino region, where RGBI orthomography with a 10 cm/px resolution is available⁴. The first analyzed vineyard was selected as a test site for executing the actual inspection mission. The second and third cases are examples of vineyards of greater size and row width (the former) or located on steeper hills and not aligned well with the image axes (the latter).

The proposed method produces consistent and accurate results in all examined vineyards with different layouts and sizes. The final

³ <https://ollama.com/>

⁴ <https://www.comune.trento.it/Aree-tematiche/Cartografia>

refinement enables robustly producing high quality outputs even from problematic intermediate results, like the second case study in Table 1. Although the refinement step consistently reduces the variability of the estimated row widths (i.e., change in standard deviation σ values), its effect is most evident in this case, reducing the standard deviation over 7 times and adding most of the rows missed by the NDVI peak detection process.

The other noticeable problems are experienced on steep vineyards with low sun angle (#3, Table 1), where the path between the lanes is too dark. In some cases, the plant rows can be confused with the space between them by the algorithm, outputting a path that follows the vines instead of the space between them. Using this path in an actual mission might not be a critical problem, as the robot's local planner should adapt the actual rover goal trajectory considering actual obstacles, still following the inspection pattern. Nevertheless, some of the lanes can be skipped and in the worst case only half of the inter-rows could be explored. Thus, a further work is needed in identification of such cases to correctly disambiguate plant rows and inter-row spacing.

4.2 Simulation tests

The simulated mission was performed in Gazebo environment, mimicking the expected real mission procedure. It was especially useful for designing and building the ROS transformation frame system to ensure reliable positioning and navigation. It involves a global geographic reference frame and two navigation reference frames, and depending on the node some of them uses one or the other. To implement this, we publish the waypoints using UTM coordinates. Another node, that stands in between the Vision-RTK2 driver and the navigation stack, translates the poses from the ECEF coordinate system (as provided directly by the Vision-RTK2), smooths out the pose shifts, and publishes the transformations between the frames.

Simulations proved to be essential for a fast integration testing on the different robot modules, especially to select and tune the local planner and the cost map layer creation process. The visualizations of the simulated mission are presented in Figure 3: a) and c) depict the realistic simulation view in Gazebo, b) and d) show the actual perception of the rover with the cost-maps and the travelled trajectory.

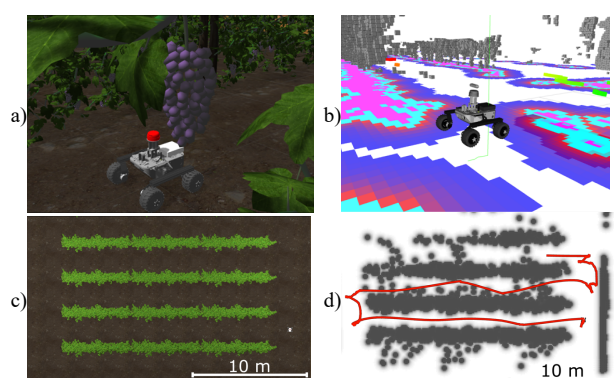


Figure 3. Inspection simulation: Gazebo environment (a) and the live data view (global voxel cost-map in grey, its planar projection in colors from pink to violet) (b). Corresponding top views of the Gazebo environment (c) and the cost-map, with a trajectory in red (d).

4.3 Test mission execution

The simulated robot configuration (sensors and algorithms) was then deployed in the field. The study site is a vineyard with long

(approx. 100 m) and wide (approx. 4 m) lanes, located on a slightly inclined hill in Trento, Northern Italy. The mission was performed in bright and covered sky conditions, with relatively low grass and with mature grapes on the vines. The vineyard canopy reached ca. 3 m height, creating arches and severely covering the lanes from the top, hence significantly obstructing satellite view for the GNSS receivers.

The test mission plan was prepared to inspect 4 lanes of the grapevines of approximately 100 m length and was executed autonomously by the robot, which correctly executed it and returned to the starting point after finishing the plan. Some local path deviations were observed, when the robot properly avoided small obstacles, such as tall grass patches. The rover maintained the expected average speed of around 30 cm/s (with its maximal achievable 50 cm/s). Due to this low speed, the full mission execution took around 30 minutes.

4.4 Post-processing results

4.4.1 SLAM-based 3D reconstruction

The LiDAR-inertial sequence was collected by the MandEye sensor, attached to the robot for the autonomously performed mission, and processed later in an offline mode. The results contained 10 610 frames of the dense pose graph, corresponding to a final point cloud size of 81.8M points (Figure 4a). A noticeable quality of the produced data is the correlation of foliage and fruits locations with the laser beam return intensity (Figure 4b), which potentially could later alleviate the complexity of their identification in the process of point cloud segmentation.

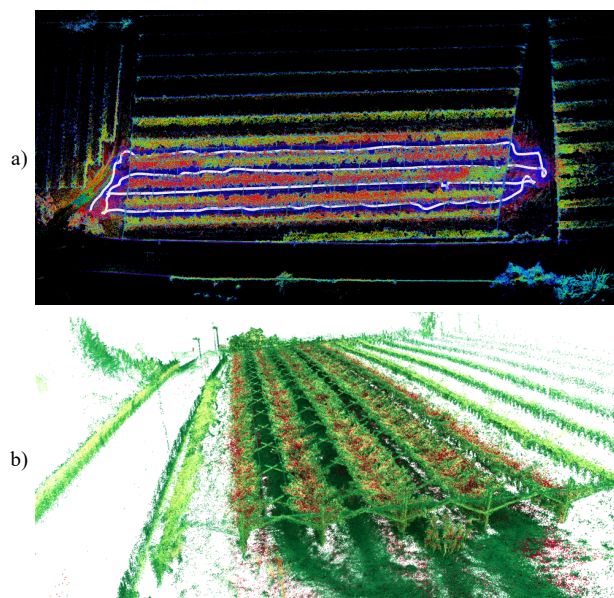


Figure 4. 3D point cloud of the vineyard obtained from a LiDAR-inertial SLAM: a view from the top, with trajectory in white (a); a perspective view with points coloured by laser intensity (b).

In the next step, the trajectory obtained from LiDAR-inertial SLAM and Vision-RTK2 systems were compared. Because the latter provides georeferenced poses, the LiDAR data was aligned to its global coordinate system. A threshold of 0.1s of time difference was set for pose timestamp matching. The differences of absolute and relative poses were calculated and the visualizations of the results are plotted in Figures 5 and 6. As seen in Figure 6, both systems provided reliable and highly coherent trajectories. The highest deviations were obtained for the spaces in the middle of the vine rows, as expected due to the weak

quality of GNSS signal, affecting Vision-RTK2, and high repeatability and linearity of structures seen by the LiDAR. Nonetheless, neither the mean average difference of absolute poses equal to 29 cm nor the mean relative pose difference of 5 cm negatively impacted the reliability of robot's positioning or the quality of the provided 3D data.

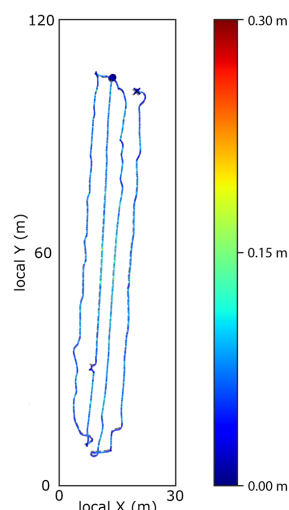


Figure 5. Top view of the trajectory from LiDAR-inertial SLAM (colored by the relative pose deviation) compared to GNSS-visual-inertial solution of Fixposition (grey, dashed line).

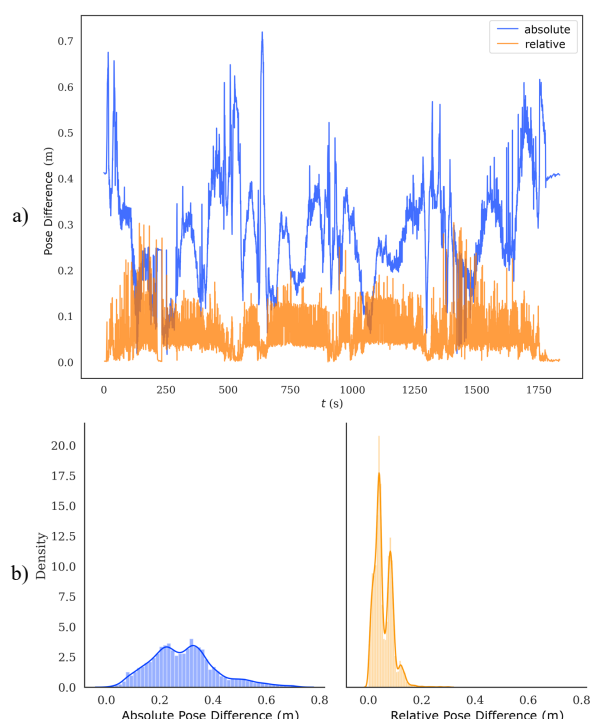


Figure 6. Absolute and relative pose differences between LiDAR-inertial SLAM and Vision-RTK2, plotted in time (a) and as distributions (b).

4.4.2 Object detection

The retrained YOLOv8 model was used to detect the location of grape bunches. All the data used for the model adaptation was collected by the robot during the autonomous mission. Initially, we annotated 128 images, that, after data augmentation, increased to a total of 236 images. Brightness and rotation were

applied during the augmentation process. The dataset was split as follows: 69% for training, 9% for validation, and 22% for testing. The best performance was achieved at 200 epochs of training, with a recall of 55.2%, precision of 51.1%, and a mean average precision (mAP) of 57.3%. According to the testing results (Figure 7), the model demonstrated higher accuracy in detecting grapes at closer distances compared to those farther away. Qualitatively judging the detection results, the object detection in the foreground worked to a satisfactory extent. Therefore, we plan to further address this issue in future studies, using a dedicated, higher resolution camera for grape detection and monitoring.

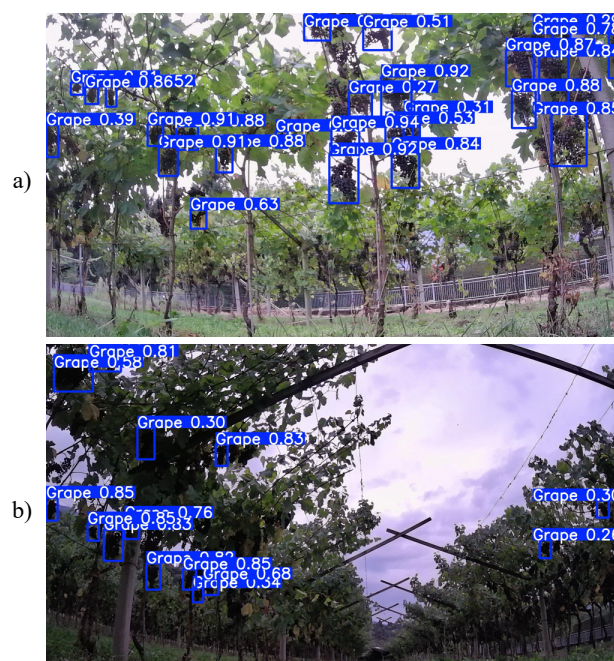


Figure 7. YOLO-based model detections of grape bunches (in blue, with confidence scores).

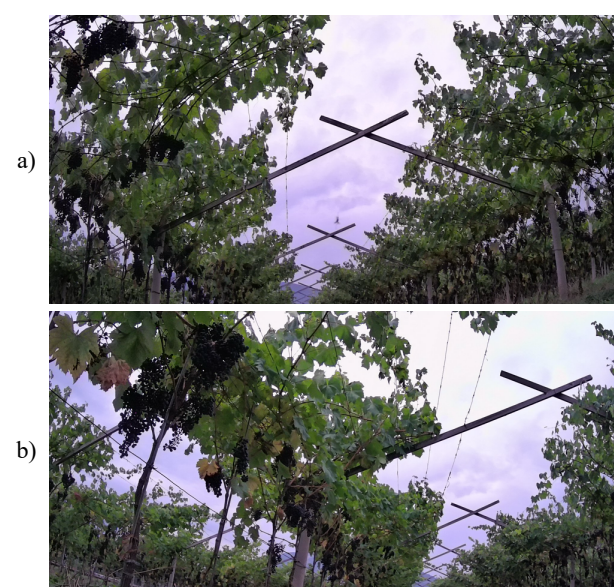


Figure 8. Examples of LMM input images: with only healthy growth (a) and symptomatic plants, namely yellow and brown leaves (b).

4.4.3 LMM-based anomaly detection

Using the same dataset of images as for the object detection, we performed the test of vision AI model capabilities at a higher abstraction level. We selected 4 open-source models for the tests: MiniCPM-V 2.6 (Yao et al., 2024), VILA1.5 8B (Lin et al., 2024), Llava 7B (Liu et al., 2024) and Llama3.2-vision 11B (Meta, 2024). All models were fed by 12 images, selected as clearest examples of two classes: healthy and symptomatic plants. The models' task was to identify any visible symptoms of possible improper grapevine growth, and flag images for which they were sure such symptoms are clearly present. Each image was processed independently 10 times to reduce the influence of model output randomness on the results. The example images from a negative and positive sample can be seen in Figure 8.

The details of the results, shown in Figure 9, clearly indicate a high complexity of the task and generally inconsistent performance of all models. These are further backed by the statistics, outlined in Table 2. Although the feasibility of deploying these models locally, on the edge device directly on a mobile robot, is quite high. Apart from the largest tested model, llama3.2-vision, all models were able to process the image within a few seconds, generating more than 1 token per second. However, models often struggled to correctly follow the prompt and enforced output format.

The final verdict on the accuracy of the symptomatic plant detection is inconclusive. MiniCPM-V 2.6 achieved the highest recall, but more frequently reported false positives. All other examined models obtained similar, low accuracy ca. 55%, with Llama3.2-vision obtaining distinctly higher precision, but lower recall. It is clear, that without domain adaptation and additional training on highly specific datasets, these models are not ready to be deployed in a zero-shot manner in the examined scenario.

A control, single run of commercial models, resulted in mediocre results as well. Large flagship models of the market leaders, such as Google Gemini 2.0 Flash and OpenAI Chat-GPT 4o, reached accuracy of 75% (correctly classifying 3/4 healthy, 6/8 symptomatic scenes), and poor 58% (only 1/4 healthy, 6/8 symptomatic correct detections), respectively. The former can be seen as the current best-obtainable baseline of a zero-shot method: potentially useful but still leaving large room for improvements.

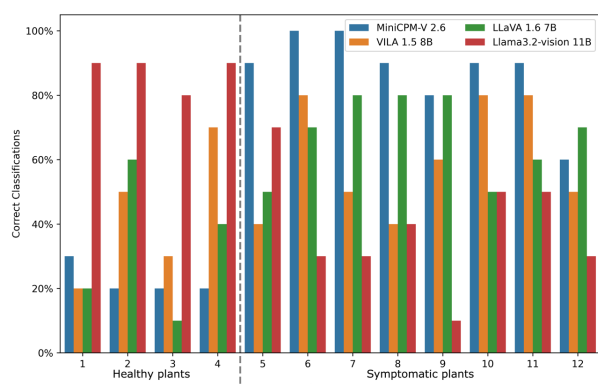


Figure 9. Percentages of correct classifications of grapevine images by locally deployed LMMs.

Model	Precision	Recall	Accuracy	Tokens per sec.
MiniCPM-V 2.6	69%	88%	66%	4
VILA1.5 8B	68%	60%	54%	16
Llava 7B	67%	68%	56%	4
Llama3.2-vision 11B	86%	39%	55%	1

Table 2. Performance metrics of all tested local LMMs.

5. Conclusions

This research presented an approach for using cost-effective robotics and AI solutions for autonomous vineyard monitoring and inspection. We leveraged on various data sources and procedures to provide automated aid in the inspection mission plan generation, mission execution, 3D data generation and data interpretation. The reported field experiments prove that the proposed framework is suitable for successfully performing such a mission using a low-cost rover. In turn, it indicates that the adoption of such procedures could greatly democratize access to similar mobile robotic solutions as reliable and affordable platforms for vineyard monitoring. We foresee its applicability and readiness especially for application in automated high-quality visual and 3D data collection.

The maturity of the tested AI-based components (object detection and LMM) greatly varies. It is clear that learning-based methods applied in this study need dedicated retraining or other type of adaptations to yield results of quality adequate for daily-uses and significantly aiding human workers in the vineyard environments. However, some of the results show a potential to provide meaningful and more accurate output in the near future, given the rapid progress of foundation vision-AI models and LMMs in recent years, especially if more detailed, domain-specific datasets would be exploited. An important aspect of our feasibility study was to reach good, close-to real-time performance of selected open-source LMMs on edge computing units, which was demonstrated as shown in Section 4.4.3.

As future works, we plan to further work on improving accuracy and reliability of AI-based methods in the agricultural scenarios and with data acquired by a robot in the field. The possibilities of providing deeper insights into the vineyard status using 3D data, e.g., with semantic segmentation or joint 3D object detection and mapping, will be investigated.

Acknowledgements

The project was partially supported by the EU FEROX project (GA 101070440). Authors would like to thank Fixposition AG for providing access to the Vision-RTK 2 sensor for research purposes.

References

- Będkowski, J., 2024. Open source, open hardware hand-held mobile mapping system for large scale surveys. *SoftwareX*, 25, 101618.
- Di Gennaro, S. F., Vannini, G. L., Berton, A., Dainelli, R., Toscano, P., Matese, A., 2023. Missing plant detection in vineyards using UAV angled RGB imagery acquired in dormant period. *Drones*, 7(6), 349.
- Fasiolo, D. T., Scalera, L., Maset, E., Gasparetto, A., 2023. Towards autonomous mapping in agriculture: A review of supportive technologies for ground robotics. *Robotics and Autonomous Systems*, 104514.
- Gao, X., Li, J., Fan, L., Zhou, Q., Yin, K., Wang, J., Song, C., Huang, L., Wang, Z., 2018. Review of wheeled mobile robots' navigation problems and application prospects in agriculture. *IEEE Access*, 6, 49248-49268.
- Hanif, A. S., Han, X., Yu, S. H., 2022. Independent control spraying system for UAV-based precise variable sprayer: A review. *Drones*, 6(12), 383.

- Hassanein, M., Khedr, M., El-Sheimy, N., 2019. Crop row detection procedure using low-cost UAV imagery system. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42, pp. 349-356.
- Hrabar, I., Goričaneč, J., Kovačić, Z., 2021. Towards autonomous navigation of a mobile robot in a steep slope vineyard. In: *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, pp. 1119-1124.
- Hroob, I., Polvara, R., Molina, S., Cielniak, G., Hanheide, M., 2021. Benchmark of visual and 3D lidar SLAM systems in simulation environment for vineyards. In: *22nd TAROS Conference, 2021*, pp. 168-177.
- Iberraken, D., Gaurier, F., Roux, J. C., Chaballier, C., Lenain, R., 2022. Autonomous vineyard tracking using a four-wheel-steering mobile robot and a 2D LiDAR. *AgriEngineering*, 4(4), pp. 826-846.
- Izquierdo-Bueno, I., Moraga, J., Cantoral, J. M., Carbú, M., Garrido, C., González-Rodríguez, V. E., 2024. Smart Viniculture: Applying Artificial Intelligence for Improved Winemaking and Risk Management. *Applied Sciences*, 14(22), 1027.
- Jocher, G., Chaurasia, A., Qiu, J. 2023. YOLO by Ultralytics (Version 8.0.0). Computer software. GitHub. <https://github.com/ultralytics/ultralytics>.
- Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S., 2024. Vila: On pre-training for visual language models. *Proc. CVPR*, pp. 26689-26699.
- Liu, H., Li, C., Li, Y., Lee, Y. J., 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296-26306.
- Liu, J., Liu, Z., 2024. The Vision-Based Target Recognition, Localization, and Control for Harvesting Robots: A Review. *International Journal of Precision Engineering and Manufacturing*, 25(2), pp. 409-428.
- Macenski, S., Tsai, D., Feinberg, M., 2020. Spatio-temporal voxel layer: A view on robot perception for the dynamic world. *International Journal of Advanced Robotic Systems*, 17(2), 1729881420910530.
- Maset, E., Scalera, L., Beinat, A., Visintini, D., & Gasparetto, A., 2022. Performance investigation and repeatability assessment of a mobile robotic system for 3D mapping. *Robotics*, 11(3), 54.
- Mendes, J., Peres, E., Neves dos Santos, F., Silva, N., Silva, R., Sousa, J. J., Cortez, I., & Morais, R., 2022. VineInspector: The Vineyard Assistant. *Agriculture*, 12(5), 730.
- Meta, 2024. Llama-Vision (3.2 version) [Large multimodal model].
- Neupane, K., Baysal-Gurel, F., 2021. Automatic identification and monitoring of plant diseases using unmanned aerial vehicles: A review. *Remote Sensing*, 13(19), 3841.
- Redmon, J., 2016. You only look once: Unified, real-time object detection. *Proc. CVPR*, pp. 779–788.
- Rösmann, C., Hoffmann, F., Bertram, T., 2017. Kinodynamic trajectory optimization and control for car-like robots. In: *2017 IEEE/RSJ IROS*, pp. 5681-5686.
- Roure, F., Bascetta, L., Soler, M., Matteucci, M., Faconti, D., Gonzalez, J. P., Serrano, D., 2020. Lessons Learned in Vineyard Monitoring and Protection from a Ground Autonomous Vehicle. *Advances in Robotics Research: From Lab to Market: ECHORD++: Robotic Science Supporting Innovation*, pp. 81-105.
- Roure, F., Moreno, G., Soler, M., Faconti, D., Serrano, D., Astolfi, P., ..., Matteucci, M., 2018. GRAPE: ground robot for vineyard monitoring and protection. In: *ROBOT 2017: Third Iberian Robotics Conference: Volume 1*, pp. 249-260. Springer International Publishing.
- Stavridis, S., Droukas, L., Doulgeri, Z., Papageorgiou, D., Dimeas, F., Soriano, Á., ..., Tzovaras, D., 2024. Robotic Grape Inspection and Selective Harvesting in Vineyards: A Multisensory Robotic System with Advanced Cognitive Capabilities. *IEEE Robotics & Automation Magazine* [Early Access], doi: 10.1109/MRA.2024.3487324.
- Trybała, P., Kujawa, P., Romańczukiewicz, K., Szrek, A., Remondino, F., 2023. Designing and Evaluating a Portable LiDAR-based SLAM system. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 191-198.
- Williams, H., Smith, D., Shahabi, J., Gee, T., Nejati, M., McGuinness, B., ..., MacDonald, B. A., 2023. Modelling wine grapevines for autonomous robotic cane pruning. *biosystems engineering*, 235, pp. 31-49.
- Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., ... Sun, M., 2024. Minicpm-v: A GPT-4v level MLLM on your phone. *arXiv preprint arXiv:2408.01800*