YOLO-based Multi-Modal Analysis of Vineyards using RGB-D Detections

T Clamens, J Rodriguez, M Delamare, L Lew-Yan-Voon, E Fauvet, David Fofi

5th International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI' 2023), 7-9 June 2023, Tenerife (Canary Islands), Spain

Type of Presentation:

Oral: ☐ In-person: ☒

Poster: ☐ Virtual in Zoom: ☐

The same: ☒

Topic:

< Artificial Intelligence Tools & Applications >

YOLO-based Multi-Modal Analysis of Vineyards using RGB-D Detections

T. Clamens [1], J. Rodriguez [1], M. Delamare [2], L. Lew-Yan-Voon [1], E. Fauvet [1] and D. Fofi [1]

[1] VIBOT, ImViA EA 7535, Université de Bourgogne, 12 rue de la fonderie 71200 Le Creusot, France

E-mail: thibault.clamens@u-bourgogne.fr

[2] CESI Engineering School, 80 avenue Edmund Halley Rouen Madrillet Innovation, 76800 Saint-Etienne-du-Rouvray, France

Summary: Agricultural robotics is a rapidly growing research area due to the need for new practices that are more environmentally responsible. It involves a range of technologies including autonomous vehicles, drones and robotic arms. These systems can be equipped with sensors and cameras to gather data and perform tasks autonomously or with minimal human intervention. For robot navigation and manipulation, and plant monitoring and analysis, perception is of prime importance and is still a challenging task today. For instance, visual perception using color images only for disease detection in vineyards, such as Mildew in which the symptoms manifest as small spots on or beneath the leaves, is still a hard task that does not allow to achieve high detection accuracy. To extract more representative features to improve the detection accuracy, other modalities must be used in addition to the Red Green and Blue (RGB) information of color images. In this paper, we present first a multimodal acquisition system that we have developed. It is composed of a multi-spectral (MS) camera and an RGB-D camera that are mounted on a mobile robot for data acquisition in a vineyard. Next, we describe the multi-modal dataset that we have built based on the data acquired with our system in a commercial vineyard. Finally, we implemented an Early RGB and depth data fusion technique together with the YOLOv5m Deep Learning network to detect the main parts of the vine: leaves, branches, and grapes using our dataset. The results that we have obtained, compared to those obtained using RGB images only with the YOLOv5m architecture, demonstrate the benefits of adding multi data fusion techniques to the object detection pipeline. These results are encouraging and show that multi-sensor data fusion is a technique that is worth considering as it can be useful for improving grapevine disease recognition technologies.

Keywords: viticultural robotics, vineyard analysis, multi-modal dataset, RGB-D camera, multi-spectral camera, RGB-D fusion, object detection.

1. Introduction

Nowadays, in modern agriculture, farmers must deal with several trade-offs every day. They must satisfy a rising public demand while maintaining the quality of their products and of their land for the good health of the consumers of their products. The new environmental considerations imply an evolution of agricultural and viticultural practices [1] to increase sustainability and farmers' health and safety. For these reasons, robotic and precision agricultures are developing at a large pace as they can contribute to a reduction in the use of phytosanitary products such as pesticides, herbicides, and fungicides in farming as well as a reduction of human labor in harsh working conditions.

Robotic agriculture and precision agriculture are two different concepts but they are related. Robotic agriculture involves the use of robots or automated machines to perform specific tasks in the agricultural process, such as planting, harvesting, or spraying crops. On the other hand, precision agriculture is a farming method that uses technology to optimize crop production by analyzing and managing various factors, such as soil characteristics, weather patterns, and crop health, with high precision and accuracy. Although both concepts are different, they nevertheless have one common point in that they both rely heavily on visual perception and more particularly on one important application of visual perception that is object detection [2]. Indeed, object detection allows robots to navigate autonomously while avoiding obstacles in robotic agriculture. In precision agriculture it allows to identify and locate objects of interest, such as weeds, crops, and pests for weed and pest management and crop health monitoring. It is typically achieved through machine learning and deep learning algorithms, which are trained on large datasets of annotated images [3]. Once trained, these algorithms can accurately detect and classify objects in real-time, enabling robots to perform tasks such as crop monitoring [4], weed control [5, 6], and fruit picking [7].

Object detection using visual cues is challenging due to factors that are outside human control such as illumination, geometric properties of agricultural fields, weather conditions, and plant structure uncertainty. Appropriate detection and localization of plants, fruits and weeds are the backbones for inspection, robotics, and autonomous systems for agriculture. It helps farmers in several ways: to monitor crop health more efficiently, to identify and to respond to potential pest or disease outbreaks before they become severe resulting in a reduction in the need for phytosanitary products, and to estimate production yield and increase product quality.

Our work is targeted towards vineyard inspection and analysis. In this field, standard RGB cameras and computer vision can provide affordable and versatile solutions for object detection such as leaves, branches and grapes. The Faster R-CNN architecture provides

accurate results for grape detection. Recognition results can be integrated through data association approaches that use object tracking or mapping to perform fruit counting in the vineyard. An alternative to the two-stage object detection algorithm is the one-stage algorithms such as the YOLO (You-Only-Look-Once) architectures. YOLOv5 model has outstanding performance in terms of speed and accuracy [8, 9]. To further improve object detection, multi-modal data fusion is crucial. The use of depth cameras can provide 3D information and highlight the unique geometry of objects present in the scene [10].

In the next section, we describe the multimodal acquisition system that we have developed as well as the multimodal dataset that we have built with data acquired with it. In section 3, we present the data fusion method that we have developed and used together with the YOLOv5m Deep Learning network to detect leaves, branches and grapes in a vineyard. In section 4, we present the results obtained and a comparison of the results with and without multi-sensor data fusion. Finally, we conclude in section 5.

2. Data acquisition and multimodal dataset

In this study, a multimodal acquisition system was developed to acquire vineyard data in several modalities. It is composed of a SILIOS CMS-V multi-spectral camera sensitive to eight different wavelengths bands from 550 to 830 nm and of a Microsoft Kinect V2 camera for RGB and depth information. Both cameras have been mounted on a Summit XL mobile robot as shown in **Fig. 1** for ease of acquisition in a vineyard field which can have an irregular or an uneven shape. The positions of the cameras on the robot have been carefully chosen for accurate capture of the fine details of the scene, i.e., leaves, branches, and berries. Also, the cameras are fixed on the mobile robot so that their relative positions are perfectly known and do not vary during acquisition.

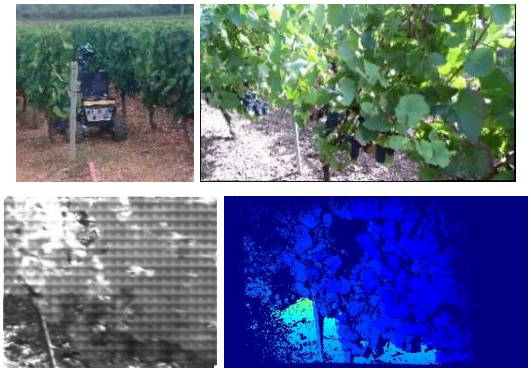

Fig. 1. Top-left: Summit XL mobile robot instrumented with the multi-sensors system; Top-right: RGB image; Bottom left: MS image; Bottom right: Depth image.

To acquire data with our acquisition system, the cameras must be calibrated, and the acquired images registered. Camera calibrations are achieved using the calibration method described in [11] for the multi-spectral camera and the iai_kinect2 package by [12] for the Microsoft Kinect V2 camera. Regarding image registration, the feature-based image registration algorithm is used [13]. This algorithm uses the Harris corner detector to extract the corner features from the reference images, the RGB images in our case, and the target images, the multispectral images. Then, it matches the corner features using the Scale Invariant Feature Transform (SIFT) algorithm to find the best affine transformation that aligns the images. Only the RGB and the multispectral images need to be registered. The depth image is already registered with the RGB image by the Microsoft Kinect V2 camera that is used to acquire both images.

Image acquisition is done at a frame rate of 15 images per second and with the mobile robot moving at a speed of 0.6 meters per second to reduce motion blur [14]. An example of the three types of images produced by our acquisition system, namely RGB images, depth images, and multispectral images is shown in **Fig. 1**. All the acquired images depict vine plants and 300 of them have been manually labelled with four classes to build our multimodal dataset. The four classes are branches, leaves, grapes, and background which represents any area that is not assigned to anyone of the other three classes.

3. Object detection with multi-sensor data

The method presented in this paper aims to improve object detection in vineyards by combining RGB and depth information using an Early Fusion architecture together with the YOLOv5 Deep Learning network [15]. The whole network architecture is shown in **Fig. 2**. YOLOv5 is a popular convolutional neural network (CNN) for real-time object detection with high accuracy [8, 9]. The architecture employs a single neural network to analyze the entire image, subsequently dividing it into regions and predicting bounding boxes and class probabilities for each. The network comprises a backbone, which consists of convolutional layers that extract and generate image features at multiple scales; a neck, which generates feature pyramids to facilitate scale-invariant object detection; and a head, which utilizes anchor boxes to produce final output vectors containing class probabilities, objectness scores, and bounding box coordinates. Compared to its predecessor YOLOv4, YOLOv5 is 88% smaller in size and 180% faster in performance while maintaining comparable accuracy on the same task.

Since the YOLOv5 model is designed to process images in the RGB format, it can only accept an image with the three color channels (Red, Green and Blue) as input. However, our RGB-D data contains four channels: three channels for the RGB image and one channel for the depth image. Thus, the four channel RGB-D image must be transformed to a three-channel image to be able to be processed by YOLOv5. This is achieved by first transforming the depth image into a three-channel image so that it is of the same dimension as the RGB image and can thus be fused with it channel by channel. We have considered two types of transformation: replicating the raw depth image two

times to obtain a depth image with three identical channels and colorizing the depth image using the Jet color palette to obtain a colored depth image in the RGB format. The effect of both transformations on the detection accuracy will be studied in the experiment section.

After transforming the depth image to a three-channel image, we now have a six-channel RGB-D image. To transform this six-channel image to the three-channel input of the YOLOv5 network, we have defined an Early data fusion network. It consists in first stacking the RGB and the depth information into a single tensor. Next, a convolution operation with six input filters, and three output filters is performed. The filter size has been chosen to be 3x3, and by doing zero-padding, the input and the output image sizes are the same. Finally, a batch normalization function is added to prevent vanishing gradients during training. Thus, at the output of this fusion module, a three-channel image is obtained that is then input to the YOLOv5 pipeline for training, test, and validation.

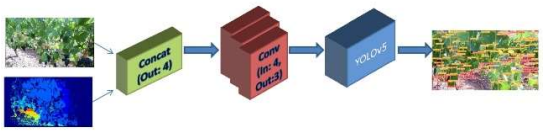

Fig. 2. Early fusion network architecture together with YOLOv5 network.

4. Experimental Results

In all our experiments, we have used the multimodal dataset that we have built with 300 manually labeled images of three classes of objects: leaves, grapes, and branches and a fourth background class which represents all areas that are none of the three object classes. The images of our dataset are images of vine plants where the "leaves" object is abundantly present. Moreover, the leaves overlap with each other such that it is very difficult to label all of them to generate the ground truth data. Thus, in the images many leaves are left unlabeled. In supervised Deep Learning networks, unlabeled data are automatically assigned to the no-class or the background class. In our case, they are assigned to the no-class. In the general case, if a valuable information (leaves, grapes, or branches) is not labeled, then during training the network might get confused because this information will be considered as pertaining to both the object class and the no-class. This will result in a degradation of the performance. For example, if a leaf is unlabeled, there will be the same type of object in two different classes, the "leaves" object class and the no-class. Then, during training similar features will be used to identify the leaf and the no-class object. Consequently, during testing, detected leaves will be assigned a lower probability than if all the leaves in the dataset were correctly labeled. To overcome this problem, a pre-processing step has been implemented to clear (i.e., set to zero) all the non-labeled regions. An example of the resulting images of this pre-processing step is shown in **Fig. 3**.

Fig. 3. Example of a labeled image processed to clear out no labeled regions. Vineyard RGB images (left) and Image with no labeled regions clear out (right).

Three training and test cases have been considered using two Deep Learning architectures: a Data Fusion network that we have defined and the YOLOv5m, a medium size YOLOv5 network, with the Ranger optimizer and pretrained with the MS COCO dataset. We have used and compared both the Adam and the Ranger optimizers [16]. Finally, we have chosen the Ranger optimizer, which is a combination of RAdam and Lookahead, because it offers improved training efficiency, faster convergence, and better performance compared to Adam. The three cases are:

- YOLOv5m network trained, validated and tested with the RGB images only.
- Data fusion and YOLOv5m networks trained, validated, and tested with RGB and raw depth images.
- Data fusion and YOLOv5m networks trained, validated, and tested with RGB and colored depth images.

For the first training and test case where only the RGB images are needed, the other modalities of our multimodal dataset are just ignored and not used. In the two other cases, the depth information is used in addition to the RGB ones. For the experiments, the dataset has been randomly split into 70% of training, 15% of validation and 15% of testing data. Regarding the network parameters, the number of epochs is 100, the batch size is 4 and the image size is set to 640×640 pixels.

To evaluate our model, we have computed the precision and recall scores, and the Mean Average Precision (mAP) at 50% and in the interval [50-95%]. mAP@0.5 is a measure used to evaluate the overall performance of an object detection model by considering a prediction as correct if its Intersection over Union (IoU) with manual annotation is greater than or equal to 0.5, and by taking the average of the mean accuracies for each object class.

The quantitative results are summarized in **Table 1**. They show that the network with data fusion and multi-sensor data achieves better precision, and especially better robustness. The combination of data fusion with the YOLOv5m network allows for better adaptability to different vineyards, which is our ultimate goal. We can clearly see in **Table 1** that for the mAP@0.5:0.95, we have a gap of almost double in the robustness of the system. We can therefore conclude that the RGB-D system allows for better robustness and therefore better adaptation to different configuration changes such as luminosity and seasonality.

Table 1. Quantitative results of the combined data fusion and YOLOv5m model trained with the multimodal dataset.

	P	R	mAP@0.5	mAP@0.5:0.95
RGB only	0.82862	0.68724	0.76331	0.35314
RGB + D	0.82562	0.69211	0.76367	0.3593
RGB + D colored	0.83125	0.68456	0.75909	0.34817

Fig. 4 represents an example of the output of the combined data fusion and YOLOv5m network trained with RGB images only, and with RGB-D images. The red, green and blue boxes are respectively the leaves, the grapes, and the branches classes.

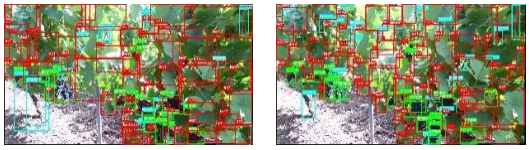

Fig. 4. Example of detection using the combined data fusion and YOLOv5m network trained with RGB images (left) and with RGB-D images (right).

5. Conclusions

We have developed a multimodal acquisition system that is particularly suitable for data acquisition in a vineyard. Using our acquisition system, we have acquired data in a commercial vineyard and built a multimodal dataset with 300 manually labelled images of vine leaves, grapes, and branches. We then used this dataset to study the effects of using depth information in addition to RGB information and data fusion techniques to detect object in vineyards. To do so, we have trained a combined data fusion and medium size YOLOv5 network, denoted by YOLOv5m, to detect leaves, grapes, and branches in vineyard images. The results that we have obtained show that the use of multimodal data allows to increase the detection accuracy while reducing false negatives.

These results are encouraging, and we intend in future work to either use our dataset or create a new dataset with even more modalities, and data fusion techniques to address viticultural challenges such as the rapid and accurate detection of plant pathologies for vine plant health monitoring and disease management, berry detection for automatic harvesting or production yield estimation, and weed detection for autonomous weed removal in robotic agriculture. For example, in disease detection and recognition, the detection result mixed with the localization of the robot, can help to create a map of where the different diseases are. In this way, the vineyard owners can localize problems earlier, and solve them before they become a real problem.

Acknowledgements

We gratefully acknowledge the support of the French government's Plan France Relance initiative via the European Union, which provided funding for this project under contract: ANR-21-PRRD-0047-01.

References

[1] Pörtner H.O., Roberts D.C., et al. Climate change 2022: impacts, adaptation and vulnerability; *IPCC*, 2022.

[2] Sharma V., Mir R.N., A comprehensive and systematic look up into deep learning based object detection techniques: A review, Computer Science Review, Volume 38, November 2020, 100301.

[3] Mohimont L., Alin F., Rondeau M., Gaveau N., Steffenel L.A., Computer, Vision and Deep Learning for Precision Viticulture. Agronomy 2022, 12(10), 2463.

[4] Kishan Das Menon H., Mishra D., Deepa D., Automation and integration of growth monitoring in plants (with disease prediction) and crop prediction, Proceedings of Materials Today, Volume 43, Part 6, 2021, Pages 3922-3927. https://doi.org/10.1016/j.matpr.2021.01.973.

[5] A S M Mahmudul Hasan, Ferdous Sohel, Dean Diepeveen, Hamid Laga, Michael G.K. Jones, A survey of deep learning techniques for weed detection from images, Computers and Electronics in Agriculture, Volume 184, May 2021, 106067, https://doi.org/10.1016/j.compag.2021.106067.

[6] Nitin Rai, Yu Zhang, Billy G. Ram, Leon Schumacher, Ravi K. Yellavajjala, Sreekala Bajwa, Xin Sun, Applications of deep learning in precision weed management: A review, Computers and Electronics in Agriculture Volume 206, March 2023, 107698, https://doi.org/10.1016/j.compag.2023.107698.

[7] Guan Zhaoxin, Li Han, Zuo Zhijiang, Pan Libo, Design a Robot System for Tomato Picking Based on YOLO v5, IFAC-PapersOnLine, Volume 55, Issue 3, 2022, Pages 166-171. https://doi.org/10.1016/j.ifacol.2022.05.029.

[8] Glenn Jocher, YOLOv5 by Ultralytics, 5 2020

[9] Bochkovskiy A., Wang C. Y., Liao H. Y. (2021). YOLOv5: Improved real-time object detection. arXiv preprint arXiv:2103.14030.

[10] Wu Z., Allibert G., Stolz C., Demonceaux C., Depth-Adapted CNN for RGB-D cameras, Proc. of the Asian Conference on Computer Vision (ACCV), Nov. 2020.

[11] ROS, Camera calibration, http://wiki.ros.org/camera_calibration. 2020.

[12] Wiedemeyer T., IAI Kinect2, https://github.com/code-iai/iai_kinect2. 2014-2015.

[13] Islam Md B., Kabir Mir Md J., A new feature-based image registration algorithm, Computer Technology and Application 4 (2013), No. 2.

[14] Clamens T., Alexakis G., Duverne R., Seulin R., Fauvet E., Fofi D., Real-time Multispectral Image Processing and Registration on 3D Point Cloud for Vineyard Analysis, In Proceedings of the VISIGRAPP (4: VISAPP), 2021, pp. 388–398.

[15] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, Wolfram Burgard, Multimodal deep learning for robust RGB-D object recognition, 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 28 September - 02 October 2015, https://doi.org/10.1109/IROS.2015.7353446.

[16] Wright Less, Ranger - a synergistic optimizer, GitHub repository, Github, 2019, https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer.