Taking Advantage of Depth Information for Semantic Segmentation in Field-Measured Vineyards

Ángela Casado-García, Jónathan Heras
Department of Mathematics and Computer Science
University of La Rioja, La Rioja, Spain
{angela.casado,jonathan.heras}@unirioja.es

Annalisa Milella, and Roberto Marani
Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing
National Research Council of Italy, Bari, Italy
{annalisa.milella, roberto.marani}@stiima.cnr.it

1. Introduction

Analysing natural images captured by moving robotic platforms is a key point for yield monitoring at the plant level [10]. In this context, convolutional neural models have been widely used to automatically segment crop elements based on their color and texture from RGB images [1], and depth information can reduce the uncertainty of the segmentation of objects having similar appearance [3]. However, it is not clear what is the optimal way of fusing RGB and depth information. Several works suggest that depth information can help the segmentation of classes of close depth, appearance and location [4]. On the contrary, it is better to use only RGB information to recognize object classes containing high variability of their depth values [4].

Despite the benefits of using RGB-D images for segmentation in the agricultural setting, RGB-D cameras remain relatively expensive, posing a significant barrier to their widespread adoption in agricultural applications. This challenge could be faced by automatically estimating depth information from RGB images. Currently, this task, known as monocular depth estimation, is mainly tackled by means of deep learning models able to understand the relationships between objects in the scene and the corresponding depth information [7, 11].

In this work, we aim to address two questions related to the usage of depth information for the segmentation of different elements in a vineyard. Namely, we first investigate whether the fusion of RGB and depth data can enhance the segmentation accuracy in viticulture compared to using RGB data alone. Moreover, we investigate how segmentation models trained with RGB-D images behave when depth information is automatically generated in order to mitigate the reliance of specialized hardware.

2. Materials and methods

The dataset used in this work was acquired in a vineyard in San Donaci (Italy) with an Intel Realsense D435 camera mounted on a moving robot. The camera acquired lateral views of the line of the grape plants at a distance of 0.8 to 1 m, see Figure 1(a). These images were taken at three different times of the year (July, September and October). The dataset consists of 265 RGB color images together with the depth of each image in the RAW format. The images were manually annotated to produce the segmentation masks with the regions corresponding to the grape bunches and canopy, see Figure 1(b). The dataset was divided into a training set (212 images) and a test set (54 images)[1].

There are three versions of the dataset: RGB, RGB-D, and RGB-D-generated. In the RGB version, the images of the dataset were the RGB images. In the RGB-D version, the information from RGB channels and depth channel was combined as follows. The depth RAW images provide information about the depth of objects that are located up to 65 metres away; however, plants are located less than 3 metres away; hence, the depth information related to objects farther than 3 metres away was removed from the image. Finally, such an image was combined with the RGB image obtaining an RGB-A image with four channels where the depth information is used as the alpha channel. In the last version of the dataset (called RGB-D-generated), the images of the training set were generated by using the RGB-D procedure. However, for the images of the test set, their depth information was computed by means of the Dense Prediction Transformer [6]; and, then, the RGB images and the automatically generated depth images were combined.

[1] The dataset is available at the following webpage https://github.com/joheras/ECSDVineyardDataset/

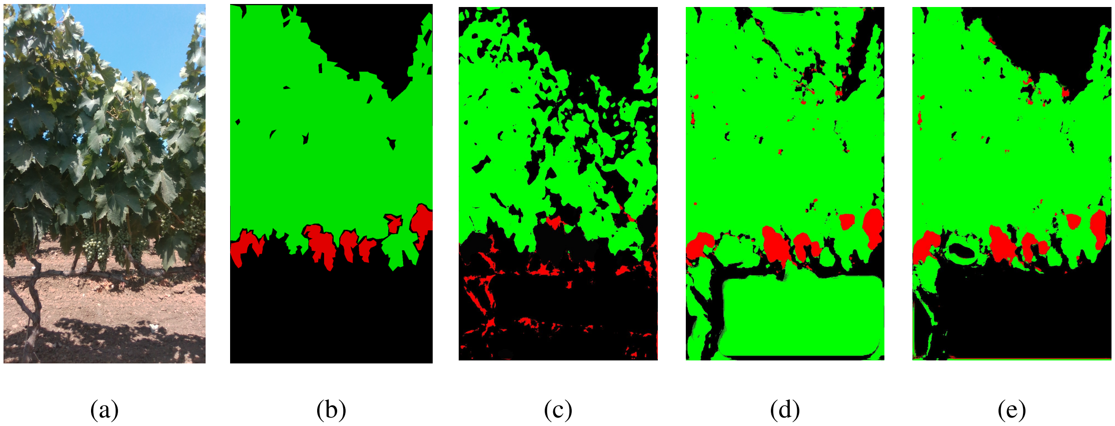

Figure 1. (a) Original Image; (b) Mask; (c) Prediction with the best RGB model; (d) Prediction with the best RGB-D model; (e) Prediction with the best RGB-D model where the depth of the original image was automatically generated.

From the training sets of the RGB and RGB-D datasets, several deep-learning segmentation architectures were fine-tuned [12]. Namely, the Unet++ architecture with a RenNet50 backbone [13], the DeepLabV3 architecture with a ResNext50 backbone [2], and the Manet architecture with EfficientNetB3 and ResNest50 backbones [8] have been employed. The models were trained thanks to the functionality of the FastAI library [5]. The code for training the models is available at the project webpage. After training, all the models were then evaluated on the corresponding test set of 54 annotated images using the mean segmentation accuracy of the $c-th$ class (MSA_c) [9].

3. Results and Discussion

The performance of the trained networks was first evaluated using the RGB version of the dataset. If the segmentation networks are compared, Deeplab-Resnext showed better overall MSA than the other networks. The Unet++-ResNet50 model produced the best results for canopy segmentation with an accuracy of 79.98, whereas the Deeplab-Resnext model, with an accuracy of 94.46, outperformed the others for segmenting grape bunches.

The RBG-D models improved up to a 4% the overall MSA of their RGB counterparts. For this dataset, the best model was built using the Unet++ architecture, that achieved a MSA of 81.91% for the canopy, of 95.83% for grape bunches, and an overall MSA of 95.47%. This shows the positive effect of adding the depth information to the RGB image, since it allows the models to focus on the objects of interest, and discard elements of the background.

Finally, for the RGB-D-generated dataset, the overall MSA of all the models improved up to 0.47%. Again, the best results were achieved with the model built using the Unet++ architecture. Such a model obtained a MSA of 96.09% for grape bunches (an improvement of 0.26% regarding the best previous model), and 86.54% for canopy (an improvement of 4.6% regarding the best previous model). This improvement was due to the fact that automatically generated depth images provide a higher level of detail at close distances than depth images captured with the camera. Hence, in addition to removing regions that are not relevant, images from the RGB-D-generated dataset preserve some information discarded in the RGB-D images.

In addition to the raw numbers, several conclusions can be draw from visually inspecting the results, see Figure 1. As we can see in Figure 1(c), the best RGB segmentation model finds where the leaves are but misses many of them; for the grapes, such a model is not able to find them and gets confused with the pole — this might happen due to the similarity of colors. On the contrary, the RGB-D model, see Figure 1(d), knows where the grapes are and can differentiate the pole (this occurs because the model gains additional information about the scene's geometry and spatial relationships), but mixes the leaves with the background. Finally, when applied the RGB-D model to an image where depth is automatically generated, see Figure 1(e), the model is perfectly capable of detecting where are the leaves and grapes. As we have explained before, this happens because the generated depth image allows us to preserve some information that is removed when the depth from the camera is used.

4. Conclusions and further work

In this work, depth information has been incorporated for automatically segmenting vineyard images. The results show the benefits of working with RGB-D images instead of only using RGB images. Moreover, it has been shown that it is possible to use RGB images as input for models trained with RGB-D images. This is achieved by automatically generating the depth information from RGB images using a Dense Prediction Transformer (DPT) model.

References

[1] A. Casado-García, J. Heras, A. Milella, and R. Marani. Semi-supervised deep learning and low-cost cameras for the semantic segmentation of natural images in viticulture. *Precision Agriculture*, 2022. 1

[2] L-C Chen, Y Zhu, G Papandreou, F Schroff, and H Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2

[3] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. Indoor semantic segmentation using depth information. In *Proceedings of International Conference on Learning Representations*, 2013. 1

[4] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In SH. Lai, V. Lepetit, K. Nishino, and Y. Sato, editors, *Computer Vision – ACCV 2016*, volume 10111 of *Lecture Notes in Computer Science*. Springer, 2017. 1

[5] J. Howard, S. Gugger, and S. Chintala. *Deep Learning for Coders with Fastai and PyTorch: AI Applications Without a PhD*. O'Reilly Media, Incorporated, 2020. 2

[6] D. Kim, W. Ga, P. Ahn, D. Joo, S. Chun, and J. Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*, 2022. 1

[7] D Kim, W Ka, P Ahn, D Joo, C Sehwan, and J Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*, 2022. 1

[8] R Li, S Zheng, C Zhang, C Duan, J Su, L Wang, and P M Atkinson. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021. 2

[9] R. Marani, A. Milella, A. Petitti, and G. Reina. Deep neural networks for grape bunch segmentation in natural images from a consumer-grade camera. *Precision Agriculture*, 22(2):387–413, 2021. 2

[10] L P Osco, K Nogueira, A P M Ramos, M M F Pinheiro, D E G Furuya, W N Gonçalves, et al. Semantic segmentation of citrus-orchard using deep neural networks and multispectral uav-based imagery. *Precision Agriculture*, 1:1–18, 2021. 1

[11] R Ranftl, A Bochkovskiy, and V Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 1

[12] A Sharif Razavian, H Azizpour, J Sullivan, and S Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014. 2

[13] Z Zhou, S Rahman, N Tajbakhsh, and J Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, volume 11045, pages 3–11. Springer, 2018. 2