Using Prior Information to Improve Crop/Weed Classification by MAV Swarms

Federico Magistri[*], Daniele Nardi[†] and Vito Trianni [‡]
ISTC-CNR, Via San Martino della Battaglia 44, 00185 Rome, Italy
DIAG, Sapienza University of Rome, Via Ariosto 25, 00185 Rome, Italy

ABSTRACT

Precision agriculture can benefit from the usage of swarms of drones to monitor a field. Crop/weed classification is a concrete application that can be efficiently carried out through collaborative approaches, whereby the information gathered by a drone can be exploited as prior to improve the classification performed by other drones observing the same area. In this study, we instantiate this concept by exploiting state-of-the-art deep learning techniques. We propose the usage of a shallow convolutional neural network that receives as input, besides the RGB channels of the acquired image, also an additional channel that represents a probability map about the presence of weeds in the observed area. Exploiting a realistic, synthetic dataset, the performance is assessed showing a substancial improvement in the classification accuracy.

1 INTRODUCTION

Use of aerial robots has been steadily increasing over the past decade, thanks to improved remote sensing abilities, better motion control and even onboard manipulation abilities. Such systems constitute a natural fit for tasks related to monitoring and inspection. In particular, small micro aerial vehicles (MAVs) are very well suited to such operations, even indoor, as they can navigate in narrow spaces, get close to the target objects and safely operate around humans. MAVs constitute an extremely attractive option for a number of practical use-cases—e.g., within application areas such as agrifood or infrastructure inspection and maintenance—opening up a wide range of market opportunities. However, for MAVs to realize their potential and get deployed in unstructured environments (outside the lab, without support from any external infrastructure to operate), a number of technical and scientific challenges related to navigation, perception and cognition must be solved. In addition, while MAVs small size is key to operational settings, it also gives rise to a number of limitations, for instance in terms of useful payload and power autonomy. Payload and power limitations do not support the installation onboard of powerful computing devices, high-resolution cameras and heavy optics, and in any case the battery lifetime may be severely limited. Such limitations make it difficult to address inspection and monitoring tasks over extensive areas, a necessary requisite for applications in precision agriculture—both outdoor and within greenhouses—or in large industrial settings.

The above limitations can be gainfully addressed by means of multi-robot systems, and notably MAV swarms, that can improve efficiency through parallel operation over large areas [1]. By exploiting a swarm of small drones, it is possible to acquire data at higher resolution, exploiting their ability to hover close to a given target and to navigate narrow cluttered environments. Additionally, the ability for members of the swarm to actively support each others enables collaborative localisation and collision avoidance [2]. Finally, MAVs in a swarm can collaborate to improve the quality of exteroception and sensory data interpretation, as a result of the collective intelligence of the group.

In our work, we propose the exploitation of drone swarms for precision agriculture applications [3]. Specifically, we consider the problem of identification and mapping of weeds within a crop field. This is a very relevant application in the precision agriculture domain, because the detailed knowledge of the position and type of weeds within a field can support advanced weed control techniques, from variable-rate herbicide application—a practice that can reduce herbicide usage by more than 80%—to mechanical removal of weeds, possibly automatically performed by ground robots. Assuming that the weed distribution within a field is non-homogeneous, inspection of extensive fields by drone swarms can be efficiently performed by means of non-uniform coverage strategies, which deploy resources (i.e., drones) only towards portions of the field with high relevance, while areas of low interest receive much less attention [4]. To this end, an estimation of the utility of each area must be performed first, and on such basis a more or less detailed inspection can be executed. Utility estimation is performed by a high-altitude/low-resolution inspection, while detailed inspection is performed through low-altitude/high-resolution inspection, and the latter can be exploited to continuously update the former. Hence, for non-uniform coverage strategies to be implemented by an autonomous decentralised system, it is necessary that MAVs are capable to communicate and adapt their mission on the

[*]federico.magistri@gmail.com
[†]nardi@diag.uniroma1.it
[‡]vito.trianni@istc.cnr.it

basis of what observed on the field, hence requiring suitable algorithms for online/onboard feature detection.

While high-altitude/low-resolution inspection can be performed by individual drones with standard estimation techniques based on common indexes used in precision agriculture, low-altitude/high-resolution inspection requires the identification and classification of individual plants, so as to determine their type and position within the field. In this study, we address the latter aspect, proposing a framework for collaborative classification of relevant environmental features based on state-of-the-art deep neural networks. We assume here that detailed inspection is performed by a MAV swarm by flying at a relatively low altitude (e.g., 3m from the ground) so that images of the field are taken at a sufficient resolution even with low-end and lightweight cameras. Classification of crop and weed can be carried out with state-of-the-art techniques making use of convolutional neural networks (CNNs) for object detection, which return the position and class type for all relevant objects identified within an image [5]. However, CNNs are computation-hungry methods that are not suitable for the limited devices available onboard MAVs. Therefore, for MAV swarms to be efficient and accurate, it is necessary to reduce the computational complexity of the algorithms running onboard the single MAV, while exploiting collaboration among MAVs that can support each other on the classification task.

We propose to exploit the fact that different MAVs can inspect the same area of the field at different times and from different perspectives, therefore having redundant information about the same plants that can be exploited to improve the classification accuracy. Each MAV is endowed a streamlined version of a deep CNN. On the first passage over an area of the field, a MAV independently makes a classification of the different plants it can perceive. Such classification is geo-localised exploiting onboard devices (e.g., RTK-GNSS) or self-localisation techniques, and then broadcasted to all other MAVs in the swarm, possibly using a simple re-broadcasting protocol to widely diffuse newly available information. Successive passages exploit prior knowledge by building probability maps about the existence of crops and weeds on the current portion of the field. Such probability maps are fed as additional input channels to the CNN (similarly to what proposed in [6] for foreground/background segmentation), so as to improve the classification accuracy on all the relevant elements in the inspected area. To validate this proposal, we developed a realistic 3D simulation of a sugar-beet field in which two types of weed are present. This allowed us to generate a synthetic dataset of field images as gathered from a MAV flying at a low altitude, simulating multiple independent passages over the same area by changing position and illumination parameters. We reduce the depth and complexity of a state-of-the-art CNN and increase the input channels to include also the possible availability of probability maps. We test several training approaches by varying the likelihood of providing the additional probability maps with respect to simple RGB channels. We show that across multiple passages, the performance of the classification substantially improves, validating the proposed concept and calling for further refinements as well as for tests with real-world datasets.

The paper is organised as follows. In Section 2, we briefly review the available techniques for classification in a weed management domain. In Section 3, we describe the experimental setup detailing the synthetic dataset generation, the proposed CNN architecture and the training methods devised. In Section 4, we discuss the testing procedure and the results obtained, comparing our iterative method with one-shot approaches. Section 5 concludes the paper.

2 CROP/WEED METHODS

In recent years, the interest in robotics applications for precision agriculture raised constantly [7]. Among the most important problems tackled through automatic techniques, weed control represents an important case study as it requires both advanced vision to recognise weed type and fine mechanical control to spray or remove the identified plants. As a consequence, several approaches to the crop/weed classification problem have been attempted using both unmanned ground (UGVs) and aerial vehicles (UAV). On the one hand, UGVs are generally large powerful tractors adapted from traditional agricultural machinery, and can be equipped with performing hardware, thus allowing on-board classifications even with modern deep neural networks. However, large UGVs must carefully manoeuvre to avoid damage to the crop field and to reduce soil compaction. Furthermore, the close-up view of a camera mounted on a UGV does not allow to exploit the geometric pattern of a typical field. On the other hand, UAVs have the possibility to quickly cover large crop fields and to perceive a wide area at once. However, due to payload limitations, they cannot exploit hardware with the same computing capability as for the UGV case.

In recent years, efforts have been made to provide reliable crop/weed classification methods. Object-based classification methods exploit the sowing pattern to classify as weed plants that lay outside the crop rows [8]. In [9], a random forest classifier was adapted to UAV imagery, using, as input, a large set of hand computed features including also the main row direction of the crop field. The same authors [10] exploited also a very shallow neural network to classify plant species; before feeding the neural network, a vegetation mask based on the popular NDVI index is computed, and single plants are extracted from this mask and passed to the CNN. In [11], feature learning is exploited for weed classification from UAV images. In [12], a deep auto-encoder architecture composed by 26 convolutional layers (the encoder) and 5 up-sampling layers (the decoder) obtained a pixel-wise semantic segmentation, using as input RGB images with NIR informations. The same approaches have been deployed to UGV based systems [13, 14, 15, 16]. In all these examples, the classification is performed offline after the collecting stage is finished.

3 EXPERIMENTAL SETUP

As described in Section 2, most current approaches tackle the crop/weed classification task by means of semantic segmentation solutions, while our goal is to use state-of-the-art object detection algorithms so that each plant can be individually classified. This will make it possible, once each plant is detected, to take action within the field on a per-plant basis, e.g., with mechanical removal or spraying of individual plants. For this purpose, there is no large publicly available dataset that can be exploited. Additionally, the proposed swarm-based technique requires multiple images of the same portion of the field taken at different times and possibly from slightly different positions. While work is being performed to collect a suitable dataset with the required features, the validation of the concept can be more flexibly performed on synthetic datasets that can be generated through modern computer graphics engines [17]. We describe the dataset generation in Section 3.1. Thanks to such a dataset, we are able to train CNNs for object detection. The chosen CNN architecture and the technical choices to provide prior knowledge from previous passages as input to the CNN are detailed in Section 3.2. Finally, the training methods used to obtain an efficient object detection are discussed in Section 3.3.

3.1 Synthetic dataset generation

The proposed method for crop/weed classification relies on multiple passages over the same area of the field, hence on multiple images with different illumination and possibly different perspective. Given the complexity of acquiring a similar dataset in the field, a synthetic dataset has been generated using the advanced computer graphics features provided by the game engine Unity 3D (https://unity.com). As a bonus, the ground truth labelling is obtained with precision and low effort directly from the simulator, hence removing one of the main difficulties in machine vision research.

Starting from the 2D texture of leaves belonging to the target plant species, it is possible to generate a large variety of individual 3D plants by assembling multiple leaves and realistically bending the texture [17]. In the simulation environment, each plant is generated with several parameters which are individually tuned for each species to resemble as much as possible the aspect of the real counter-part (see Figure 1). To each plant, independently from its species, a vertical growth axis is associated which is slightly perturbed by a random noise. To simulate a uniform growth stage for all the plants that have been generated, each plant has a number of layers up to two. Each plant, moreover, has an associated number of leaves per layer which is different from species to species. At each layer, the leaves are homogeneously spread around the main growth axis, again with a small random disturbance. In order to simulate succesive visits of the same region of the field, once a set of plants has been placed on the scene, the illumination parameters have been randomly changed, moving the position and the intensity of the light sources illuminating the scene, hence also casting different shadows on the ground. As a last step to create a realistic environment, the soil is generated starting from various real world textures. Every time a fragment of terrain is created, two textures are chosen and blended together using Perlin noise based linear interpolation (LERP). Similarly to the plants generation, by changing the parameters of the Perlin noise, it is possible to create a large variety of soil textures.

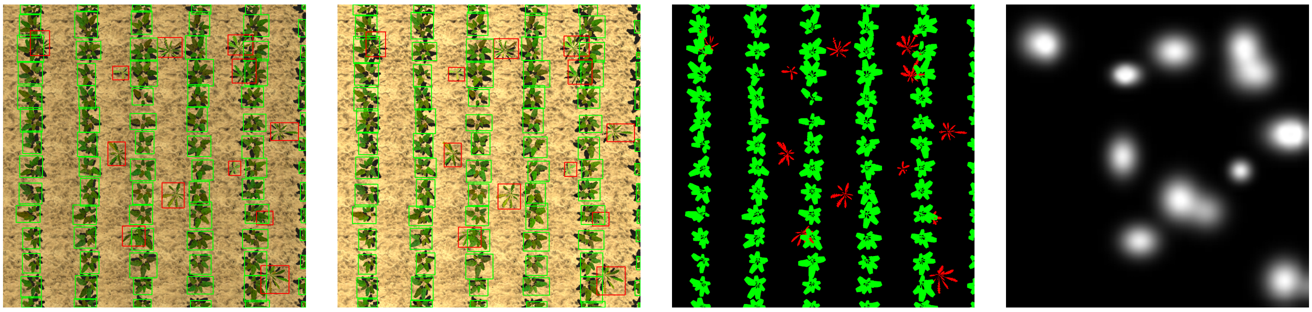
With this method, images of different plants taken at different altitudes can be generated at will. Here, we consider images taken at about 3m altitude, containing about 80 simulated sugar beets as crop, while two types of weed are present in variable number, having up to 20 plants per image (see Figure 1). A training set with 500 images has been generated, while the validation set is composed of 100 images. Regarding the testing set, 10 blocks of 40 images have been generated. Each block is composed of the same 40 plant patterns, but the illumination parameters, such as light source orientation and intensity, randomly change for each image within a block. This results in a testing dataset with overall 400 images that we refer to as dataset A. Additionally, a second testing set has been generated featuring also a small, random perturbation of the camera position within the same image in different blocks. In this way, the position error of the MAVs flying over the same region is simulated. This second dataset is also made of 40 fields and 10 blocks for a total of other 400 images, and is referred to as dataset B.

3.2 CNN architectures

In the literature, several architectures have been proposed for classification purposes, either for image segmentation or for object detection [5, 18]. Competitions and benchmarks have contributed to establish an objective methodology to determine performance and direct choice of the best approach, given the task demands. In our case, to perform the plant detection, Faster R-CNN has been chosen [19], which can

Table 1: CNN Architectures

5 layers	9 layers
7×7, 64, stride 2	
3×3, max pools, stride 2	
$[3\times 3,\quad 64]\times 1$	$[3\times 3,\quad 64]\times 2$
$[3\times 3,\quad 128]\times 1$	$[3\times 3,\quad 128]\times 2$
$[3\times 3,\quad 256]\times 1$	$[3\times 3,\quad 256]\times 2$
$[3\times 3,\quad 512]\times 1$	$[3\times 3,\quad 512]\times 2$



(a) (b) (c) (d)

Figure 1: Synthetic dataset generation. The same field with crop rows and weeds is displayed. Panels (a) and (b) show different illumination conditions on the same field, and also display the automatically generated ground truth. A small difference in the plant position within the image is also present. Panel (c) shows a segmentation on the image obtainable as ground truth directly from the simulation. Finally, panel (d) shows a probability map computed on the ground truth data.

be considered the state-of-the-art method for object detection tasks. In its standard version, it is composed of two stages, in which the first stage—referred to as the backbone—is a deep CNN responsible for generating bounding boxes around objects to be proposed to the second stage as potentially containing relevant features. Considering the computing capabilities of a MAV, it is not possible to use the standard Faster R-CNN backbone such as ResNet50 [20], which is way too demanding in terms of computational power. Therefore, we have chosen to implement two shallow networks with much reduced demands, removing several layers from the standard backbone. In both cases, the first initial layers are the same as the ResNet architectures, namely a 7×7 convolutional layer with 64 filters and stride 2, followed by a 3×3 max-pooling layer. After that, a sequence of 3×3 convolutional layers is presented as described in Table 1. We will refer to the first as FCN5 and to the latter as FCN9. Here, FCN stands for Fully Convolutional Network.

In order to exploit detections previously made by other agents, the input of the CNN is composed of a fresh RGB image together with an auxiliary channel encoding a probability map based on previous classifications. Well-known object detection algorithms usually output 6 values for each detection i, that is, the class of the detected object c_i, a confidence score s_i, and 4 values representing bounding box coordinates encoding the coordinates x_i, y_i of the center and the width w_i and height h_i of the bounding box. From this values a probability $P(x, y)$ for each point x, y is computed as follows:

$$P(x,y) = \sum_{c_i=W} s_i \cdot e^{-\left(\frac{(x-x_i)^2}{2w_i^2}+\frac{(y-y_i)^2}{2h_i^2}\right)}. \quad (1)$$

In other words, each detection belonging to the class $c_i = W$—standing for weed—provides a probability increment proportional to the confidence score s_i, and decaying from the center of the bounding box x_i, y_i as a 2D gaussian with a spread that depends on the bounding box dimensions w_i and h_i. The resulting probability map is practically null when far from any bounding box, indicating that the probability of finding a weed plant in that position is extremely low. Peaks are visible in correspondence of detected weeds, as shown in Figure 1d. Note that we decided to focus on weed classification only, as it turns out that the performance on crop classification is already very high (see Section 4.1), hence requiring a specific method only for improving the weed detection.

3.3 Training methods

The CNN that we have devised must be capable of performing two tasks at the same time. On the one hand, it must observe the RGB channels alone to identify the presence of crops or weeds. This will output a list of detections that can be used to compute a probability map for subsequent passages by other MAVs. On the other hand, the CNN must prove capable of using—when available—the prior information to improve the classification and reduce errors. Possibly, the NN must also identify and remove conflicts between the newly available RGB image and the prior information encoded in the probability map. This turns out to be an important choice to achieve better results, since the network has to learn to balance the information coming from other agents and the fresh image. Therefore the network will not only rely on information coming from the auxiliary channels but it will be able to make a valuable initial classification and to correct possible misclassifications. A correct training of the network is therefore key to obtain both these abilities within a single CNN.

First and foremost, we have devised three different training strategies in terms of the frequency with which the probability map is presented. We used 25%, 50% and 75% of the training cases, hence pushing more or less towards the usage of the information encoded into the probability map. Additionally, to compute the probability map, instead of using the available ground truth we decided to use realistic labelling as produced from a CNN classification. To this end, we trained a FCN5 architecture with the only RGB channels

Table 2: Crop/weed classification performance with FCN5 and FCN9, with only the RGB input channels.

Dataset A

	FCN5		FCN9	
	Crop	Weed	Crop	Weed
Precision	0.96	0.98	0.96	0.98
Recall	0.87	0.65	0.91	0.73
F1	0.91	0.78	0.93	0.83

Dataset B

	FCN5		FCN9	
	Crop	Weed	Crop	Weed
Precision	0.99	0.98	0.99	0.98
Recall	0.94	0.64	0.96	0.73
F1	0.96	0.78	0.97	0.83

on the available dataset, and we used the detections obtained by the FCN5 to generate the probability map, without filtering out bounding boxes with low confidence. As a consequences the neural network will learn to deal with errors in the auxiliary input as generated by a similar CNN architecture.

4 RESULTS

The trainings and testings are performed with a NVIDIA Quadro P6000, a 24 GB GPU with 3840 CUDA cores. Each training was performed with 50000 iterations, a learning rate of 0.01, weight decay of 0.0001 and batch size 4. Testing of the trained networks has been perfomed on the two testing datasets, with and without position error. In order to evaluate our approach, precision, recall and F1-score have been computed. As in many detection tasks, a detection is considered a true positive if the Intersection Over Union (IoU) between the detected box and the ground truth is above a certain threshold (here: 0.5). Otherwise, it is considered a false positive.

4.1 Crop/weed classification with simple RGB images

First of all, we discuss the classification performance on the synthetic dataset when no a priori information is provided, hence no additional input channel is used besides the RGB channels of the input image. The FCN5 and FCN9 networks have been trained and tested on both datasets A and B. In this case, each testing set is composed of 400 images. The performance for the precision, recall and F1 metrics is shown in Table 2. Note that, not using any a priori knowledge, every image is processed independently and the differences observable between dataset A and dataset B are only due to the 40 different synthetic fields generated for each.

Specifically, the *crop* class achieves high scores even with the shallower FCN5 network, and dataset B appears easier to classify, possibly due to the relative positioning of crop and weeds, or border effects (e.g., a crop line partially included into an image because appearing on the border). The performance on the weed class is instead lower, especially for the recall, meaning that several weed plants go undetected. The FCN9 achieves better results here, meaning that there is room for improvement over the FCN5 results by including prior information with additional channels. Considering that the testing datasets are organised in blocks representing the same field but varying the illumination conditions, it is interesting to analyse how performance varies across different blocks, so as to determine how much the illumination matters on the final results. Figure 2 shows that there is indeed some non-negligible variability in performance among the different blocks, hence further motivating the use of prior information for more stable and reliable classification. Given that the performance on the *crop* class is already very high with the FCN5, we decided to use only one auxiliary channel representing a probability map for the *weed* class obtained from previous classifications.

4.2 Crop/weed classification with probability maps

To evaluate the performance achievable over multiple passages on the same field, we perform 10 classifications in a sequence using the output of the current stage to compute the probability map of the following stage (see Figure 3). As it is possible to note, while the first passage has an empty probability map, successive passages can exploit the prior knowledge to improve the classification of weeds. As a matter of fact, it can be noted that in the successive passages, more plants are correctly detected.

A proper performance evaluation is carried out on dataset A, where no position error is included (corresponding to the same condition experienced during training). Considering that each of the 10 blocks in dataset A have independent illumination conditions, we compute 100 different sequences by random permutation of the 10 blocks, and use them to have

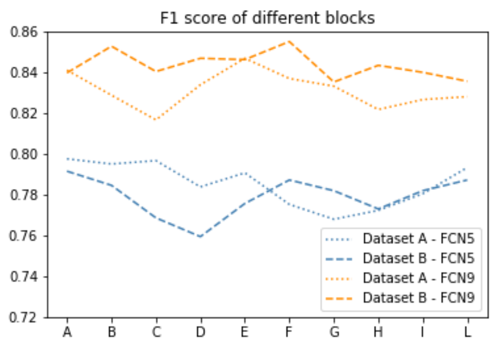


Figure 2: F1 performance of FCN5 and FCN9 across different blocks of images, which differ only in the scene illumination.

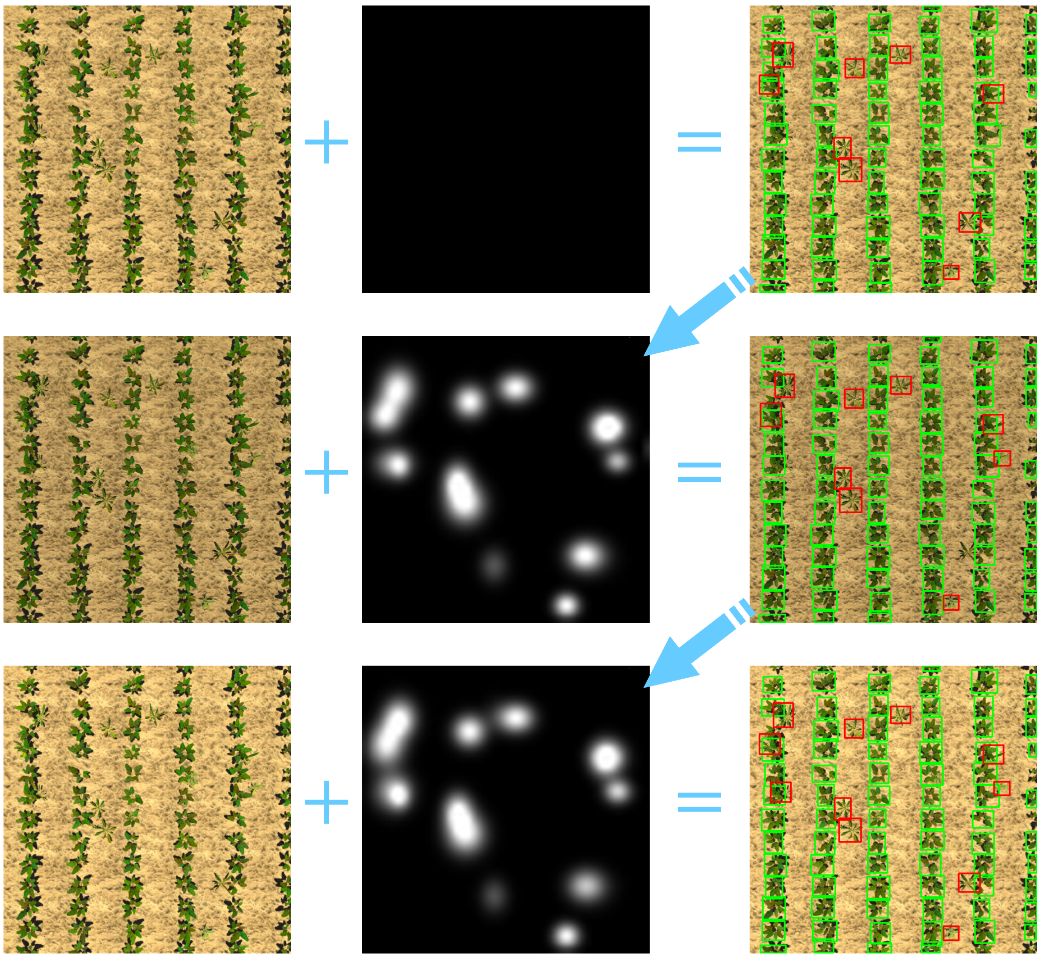


Figure 3: Classification over multiple passages. Each row correspond to a single passage over the same portion of the field, but with different illumination conditions (first column). The RGB image is coupled with a probability map derived from previous passages, when available (second column). The combined input provides an improved object detection across passages (third column). In the second passage, a weed is discovered in the center-right part of the image, but one in the bottom right is lost. In the third passage, the latter weed is detected again, and an additional weed is discovered among the first crop row on the left.

an average performance that is independent as much as possible from the specific sequence observed. The results for precision, recall and F1 on the *weed* class are presented in Figure 4. It can be noted that the overall classification accuracy increases when exploiting the probability maps coming from previous passages. More specifically, the recall is the measure most affected by the auxiliary input, while the precision can undergo a slight degrade, which is observed especially for networks trained with 75% probability of having a probability map in input. The training strategy is indeed very important to obtain a substancial improvement in the classification through multiple observations. When only 25% or 50% of the training examples are provided with a probability map, the improvement in the weed classification is only mild. Instead, with a 75% probability, the neural network learns to properly exploit the additional input when available, reaching comparable levels of performance as the more complex FCN9 network. The proposed approach is intrinsically robust against position errors, as shown by the testing performed on dataset B (see Figure 5). Even though position errors where never presented during the training phase, it is possible to note that a performance improvement is still visible through successive observations of the same region of the field. This improvement is not as considerable as for Dataset A but it is still possible with FCN5 and the auxiliary probability map to approach the performances of FCN9.

5 CONCLUSIONS

We have proposed an approach to exploit knowledge available on portions of the field coming from previous observations to iteratively improve the performance of classification by a shallow neural network, to be executed onboard lightweight MAVs with limited payload and constraints in the computational power. We obtained a substantial improvement in performance, that makes a shallow architecture achieve similar performance of a double-size network.

These results validate the concept proposed here for the first time, and open the way for a thorough analysis of the design space to identify possible improvements that can further boost performance. Future work will be dedicated to this as well as to test the methodology on real-world images. To this end, a dataset with multiple passages on the same area has already been collected, and studies are on the way to provide new grounds for the analysis of the proposed framework.

ACKNOWLEDGEMENTS

This work has been conducted within the project SAGA (Swarm Robotics for Agricultural Applications, see http://laral.istc.cnr.it/saga), an experiment founded within the ECHORD++ EU project (GA: 601116). Vito Trianni acknolwdges the NVIDIA GPU Grant Program providing the Quadro P6000 GPU used in this work.

REFERENCES

[1] K McGuire, M Coppola, C De Wagter, and G de Croon. Towards autonomous navigation of multiple pocket-drones in real-world environments. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems IS* -, pages 244–249, 2017.

[2] Mario Coppola, Kimberly N McGuire, Kirk Y W Scheper, and Guido C H E de Croon. On-board communication-based relative localization for collision avoidance in Micro Air Vehicle teams. *Autonomous Robots*, 42(8):1787–1805, June 2018.

[3] Dario Albani, Daniele Nardi, and Vito Trianni. Field coverage and weed mapping by UAV swarms. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS*, pages 4319–4325. IEEE, 2017.

[4] Dario Albani, Tiziano Manoni, Daniele Nardi, and Vito Trianni. Dynamic UAV Swarm Deployment for Non-Uniform Coverage. In *AAMAS '19: Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*, pages 523–531, Stockholm, Sweden, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

[5] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7310–7311, 2017.

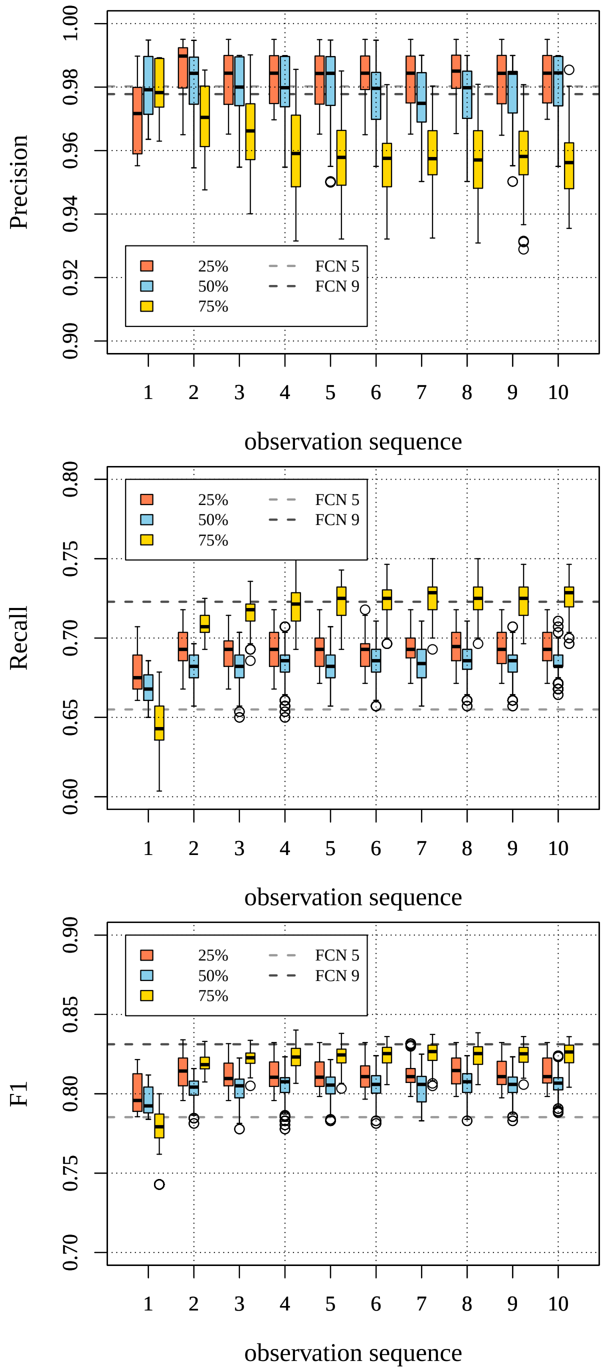


Figure 4: Precision, recall and F1-score on dataset A. The performance achieved with different training strategies (i.e., 25%, 50%, and 75% probability of having a probability map as input) is compared against the average performance of the FCN5 and FCN9 networks using only the RGB inputs.

[6] Kunming Luo, Fanman Meng, Qingbo Wu, Wen Shi, and Lili Guo. A cnn-based segmentation model for segmenting foreground by a probability map. In *2017 In-*

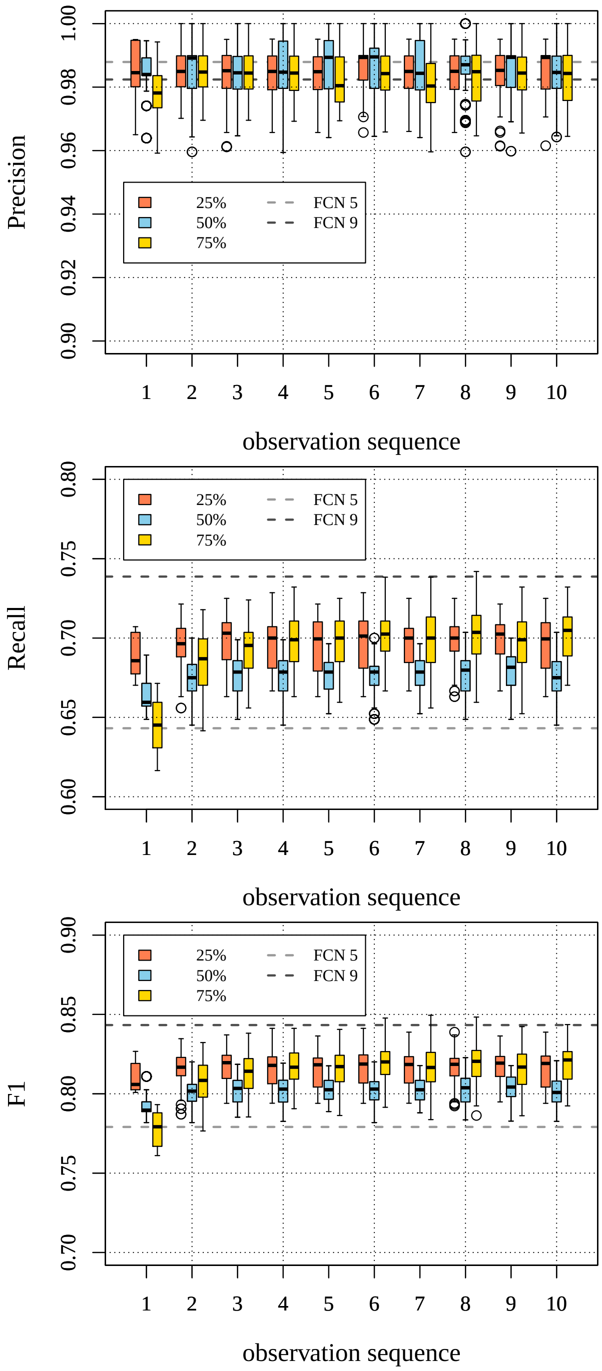


Figure 5: Precision, recall and F1-score on dataset B, where also a positioning error is included.

ternational Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 17–22. IEEE, 2017.

[7] Anthony King. Technology: The Future of Agriculture. *Nature*, 544(7651):S21 EP —-S23, April 2017.

[8] José Manuel Peña, Jorge Torres-Sánchez, Ana Isabel de Castro, Maggi Kelly, and Francisca López-Granados. Weed Mapping in Early-Season Maize Fields Using Object-Based Analysis of Unmanned Aerial Vehicle (UAV) Images. *PLoS ONE*, 8(10):e77151 EP —-11, October 2013.

[9] Philipp Lottes, Raghav Khanna, Johannes Pfeifer, Roland Siegwart, and Cyrill Stachniss. UAV-based crop and weed classification for smart farming. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3024–3031. IEEE, 2017.

[10] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. Real-time blob-wise sugar beets vs weeds classification for monitoring fields using convolutional neural networks. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4:41, 2017.

[11] Calvin Hung, Zhe Xu, and Salah Sukkarieh. Feature Learning Based Approach for Weed Classification Using High Resolution Aerial Images from a Digital Camera Mounted on a UAV. *Remote Sensing*, 6(12):12037–12054, December 2014.

[12] Inkyu Sa, Zetao Chen, Marija Popović, Raghav Khanna, Frank Liebisch, Juan Nieto, and Roland Siegwart. weednet: Dense semantic weed classification using multispectral images and mav for smart farming. *IEEE Robotics and Automation Letters*, 3(1):588–595, 2017.

[13] A T Nieuwenhuizen, L Tang, J W Hofstee, J Müller, and E J van Henten. Colour based detection of volunteer potatoes as weeds in sugar beet fields using machine vision. *Precision Agriculture*, 8(6):267–278, November 2007.

[14] S Haug, A Michaels, P Biber, and J Ostermann. Plant classification system for crop /weed discrimination without segmentation. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 1142–1149. IEEE, 2014.

[15] P Lottes, M Hoeferlin, S Sander, M Müter, P Schulze, and Lammers C Stachniss. An effective classification system for separating sugar beets and weeds for precision farming applications. In *2016 IEEE International Conference on Robotics and Automation (ICRA*, pages 5157–5163. IEEE, 2016.

[16] Ciro Potena, Daniele Nardi, and Alberto Pretto. Fast and Accurate Crop and Weed Identification with Summarized Train Sets for Precision Agriculture. In *Intelligent Autonomous Systems 14*, pages 105–121. Springer, Cham, Cham, July 2016.

[17] 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*, pages 5188–5195, 2017.

[18] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.