POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Robust Camera-Independent Weed and Crop Segmentation through Unsupervised Domain Adaptation Techniques

Laurea Magistrale in Computer Science Engineering
Ingegneria Informatica

Author: Alessio Mazzucchelli

Advisor: Prof. Matteo Matteucci

Co-advisor: Riccardo Bertoglio

Academic year: 2021-2022

1. Introduction

Precision agriculture is a farming management concept based on observing, measuring, and responding to inter and intra-field variability in crops. The goal of Precision Agriculture research is to define decision support systems for whole-farm management optimizing returns on inputs while preserving resources. One important aspect of Precision Agriculture is weed destruction, indeed weeds must be removed to achieve better crop yields. This is usually done by spraying chemicals uniformly all over the field, which poses many environmental and economical concerns. The soil gets contaminated and farmers have economical losses due to the waste of herbicides since only some parts of the field are covered with weeds and a great amount of herbicides gets wasted. One way to implement weed removal is to use a robot that sprays chemicals only in areas where the weed is present. The first difficulty to be faced is the identification and distinction of weeds and crops. Weeds can be detected by RGB cameras mounted on robot platforms. Images are then analyzed by Machine Learning algorithms to classify each pixel of the image into the crop, weed, and background classes. The most successful technique in recent years is the use of Convolutional Neural Networks (CNN) which, thanks to their ability to extract features without human help, can classify images without the need for domain knowledge of the task they are dealing with. To have a CNN system that can distinguish crop and weed we need a series of image-mask pairs where the mask represents the semantic value of each pixel. For every pixel of an image we must know which class it belongs to. During the training, our system learns to associate the value defined by the mask to each pixel of the image. Hopefully, with enough image-mask pairs, our system will also be able to classify images not seen during the training phase. These classification systems typically achieve a great performance when trained classifiers are deployed in the same or at least similar field conditions. However, the performance of a classifier, which has been trained on a particular dataset, i.e., the Source domain, suffers substantially when being deployed in new field environments or under changing conditions, i.e., the Target domain. This gap in performance between Source and Target domains is caused by a so called domain-shift. The domain-shift is due to by a different visual appearance, induced by different weed types, growth stages of plants, soil conditions, and illuminations. It is then required to retrain the system by labeling the new data from the Target domain. However, we are often faced with scenarios where we solely have access to labeled images from the Source domain and only unlabeled image data from the Target domain, for example, when a robot enters a new field environ-

ment or is equipped with a new vision system. The image labeling process is very expensive and it must be done by experts. It is not feasible to expect to have enough labeled images for each new field that needs to be classified, with few labeled data we need to be able to classify different fields.

2. Our Work

We collected all available online datasets suitable for image segmentation for weed detection and we selected a CNN network suitable for supervised segmentation, i.e., where we have labeled images, that performed well on each datasets. Then we segmented images without having their respective labeling using a network trained on a Source dataset with labeled images to classify a Target dataset only with unlabeled images. We tested different methods to transfer the knowledge obtained from the Source dataset to the Target dataset. First, we directly classified the Target images without any modification, this is what we call baseline, second, we use a system based on CycleGAN [10], with modification by [3] and novel additions inspired by [8]. Third we accomplished a style transfer using the Fast Fourier Transform (FFT) [9]. At the time of writing, this is the first time that FFT has been used in agriculture for the weed detection problem. Lastly we also tested our CycleGAN approach's performance between datasets taken in the same field but two years apart and at different growth stages.

2.1. Metrics

The main metric we use to measure our system performance across different datasets is the Inter over Union (IoU) metric.

$$IoU = \frac{TP}{TP + FP + FN}$$

We monitor the IoU obtained on the two classes, crop and weed. IoU is typically used in segmentation activities and essentially quantifies the percentage of overlap between predicted and target segmentations.

2.2. Segmentation system

Our approach to segmentation is a U-Net [6] inspired encoder-decoder architecture using VGG16 [7] as the encoder backbone. This architecture deals start-to-end with the segmentation problem by taking a 352x352x3 image as input, using the three RGB channels, and outputs a 352x352x3 image where each channel represents the probability that a pixel belongs to a certain class, in our case background, crop and weed.

2.2.1 Loss Function

We implemented a Soft-IoU loss function. As suggested by [3], the Soft-IoU Loss is more stable with imbalanced class labels compared to categorical crossentropy and thus well suited for our crop-weed classification.

$$softIoU = \frac{1}{|C|} \sum_{c} \frac{\sum_i p_{ic} \cdot p^*_{ic}}{\sum_i p_{ic} + p_{ic} - p^*_{ic} \cdot p^*_{ic}}$$

p_{ic} is the prediction probability of a given pixel i to be of class c, p^*_{ic} is the ground truth distribution, 1 if the pixel i belongs to class c otherwise 0.

2.2.2 Data Augmentations

We applied to each image at every epoch a flip over the x axis with probability 0.5 and a flip over the y axis with probability 0.5. **Tiling**: depending on resolution and size of the images we applied tiling creating HxW new images after resizing the original images in $256 * H$ x $256 * W$. H is the numbers of time we divide through the y axis, W through the x asix. We also applied padding in order to have more information on the edges of the image and and reach the predetermined size of 352 x 352. During the evaluation phase only the pixel in the 256 x 256 region are kept, the ones in the padding region are discarded.

2.3. CycleGAN

We exploit unpaired image sets from a dataset, that we call Source domain X with labels, and another dataset, that we call Target domain Y, with no labels. Our DA approach is based on CycleGANs [10], an implementation of CycleGAN for plants classification [3] and on how the FFT maintain semantic and style information in its Phase and Amplitude as pointed out by [8], [9]. Our approach consists of two domain-specific Fully Convolutional Neural Networks (FCN) for semantic segmentation, two generator networks for domain adaptation, and two discriminator networks. As an addition to the work of [3] we propose to also add a constraint that encourages the generators to maintain the same phase of the image before and after the transformation, thus maintaining the semantics of the image.

2.4. Fast Fourier Transform Domain Adaptation

We also evaluated an approach that does domain adaptation without the need for deep learning training. It makes use of the FFT and of its inverse, as proposed by [9], to transform pictures from the Source Domain in the style of the Target Domain.

IoU	Mean	Crop	Weed
CWFID	0.81	0.78	0.83
Bonn	0.79	0.96	0.62
ON17	0.63	0.47	0.79
CA17	0.68	0.67	0.68
Rice	0.69	0.66	0.71
Roseau H	0.73	0.73	0.72
Roseau M	0.77	0.83	0.70
Pead H	0.69	0.75	0.62
Pead M	0.58	0.63	0.53
Bipbip H	0.79	0.81	0.77
Bipbip M	0.84	0.88	0.79
Weedelec H	0.82	0.86	0.78
Weedelec M	0.81	0.88	0.74

Table 1: Segmentation results on datasets, datasets are taken from [4], [2], [1], [5] and from the Rose Challenge. In the Rose Challenge four teams Roseau, Pead, Bipbip and Weedelec collected images of both Mais and Haricot plants

We compute the FFT of the Source image and of the Target image, then we swap the amplitude of the Source image with the amplitude of the Target image, we compute the inverse FFT and as a result we get an image with the same semantic of Source but with the style of Target. At this point we can predict the Target dataset using the FCN trained on Source after having transformed the Target images into the style of Source or we can train a new FCN on the Source images, remembering that we have mask available for them, after having transformed them into the style of Target. We will be referring to the first method as FFTtoTarget or FtoT and to the second as FFTtoSource or FtoS.

3. Methods

First we evaluated all the datasets found online, with the system described at Section 2.2. After we tested the performance of a system trained on a dataset, i.e., the Source domain, on another dataset, i.e., the Target domain. The datasets we used were Weedelec and Bipbip. Having for both Mais and Haricot plant, the following 4 combinations have been evaluated: Weedelec Mais on Bipbip Mais, Weedelec Haricot on Bipbip Haricot, Bipbip Mais on Weedelec Mais, Bipbip Haricot on Weedelec Haricot. It is important to remember that both datasets were taken on the same field at the same moment, only the capture methods were different. We will use W for Weedelec, B for Bipbip, H for Haricot and M for Mais. E.g., WH on BH means that we use Weedelec Haricot as the Source dataset and Bipbip Haricot as the Target dataset.

3.1. Comments on Segmentation results

From Table 1 our segmentation system is able to segment all the datasets in a satisfactory way from a quantitative point of view. Even if the metric scores fluctuate in different datasets it is more a problem of how the data was labeled. Indeed, a better score can be obtained with a more accurate labeling where the exact plants' outlines are represented instead of approximate ones.

3.2. Baseline

We evaluated the basic performance directly using the system trained only on the Source dataset. The only changes made to the Target dataset will be the type of tiling performed, to try to match the same zoom effect of the two different cameras, and a 90° rotation, as the images were taken vertically for Weedelec and horizontally for Bipbip.

3.3. CycleGAN

We investigated the effect of the Phase maintenance constraint. The constraint is kept constant until the 8th epoch and then decreased by 0.01 each epoch. The starting weight assigned to the constraint is 0.15. The Phase approach turned out to be the same or slighly better than not using it in all the four cases. In Figure 1, we show the transformation learned by the geneators during the CycleGAN process. The generators are able to maintain the semantic of the image during the transformation and are able to change the image from Source to Target style and vice versa. Only in Bipbip on Weedelec Mais the generators tend to generate false green in order to confuse the discriminator. This has an impact on the scores, as can be seen in Table 2, Bipbip on Weedelec Mais is the one with the lowest IoU scores having high recall but low precision.

3.4. Fast Fourier Transform

In Figures 2 you can see the effect of the Fast Fourier Transform. The change of style takes place with the maintenance of the semantics, but it does not happen in a consistent way. It can be seen that the transformation of Weedelec Mais in Bipbip style is the one that looks the best and this is also reflected in the scores of Table 2 where Weedelec Mais on Bipbip Mais FFTtoTarget is the one with the best IoU score. It can be seen how, in most cases, training the new system on Source images transformed into Target style performs better than using the same system trained on Source images to make predictions on Target images transformed in Source style. It is really important to note that we use only a single amplitude, from a single image from the Target dataset, to

IoU	BH C	BH W	BM C	BM W	WH C	WH W	WM C	WM W
Supervised	0.81	0.77	0.88	0.79	0.86	0.77	0.88	0.79
Baseline	0.82	0.59	0.68	0.69	0.78	0.63	0.85	0.69
FtoTarget	0.82	0.69	0.83	0.73	0.76	0.58	0.85	0.66
FtoSource	0.80	0.63	0.78	0.67	0.76	0.62	0.82	0.64
CycleGAN	0.83	0.70	0.81	0.66	0.83	0.67	0.75	0.55

Table 2: Comparison of Methods, C stand for Crop and W for Weed, the dataset in the table denotes the target dataset, e.g. that means that for BH we evaluate on BH using labeled images of BH for the supervised approach, and unlabeled images of BH and labeled images of WH for the domain adaptation techniques.

Figure 1: Images generated by CycleGAN, order from letf to right is Source, Source to Target, Target, Target to Source. From top to bottom WH on BH, WM on BM, BH on WH, BM on WM

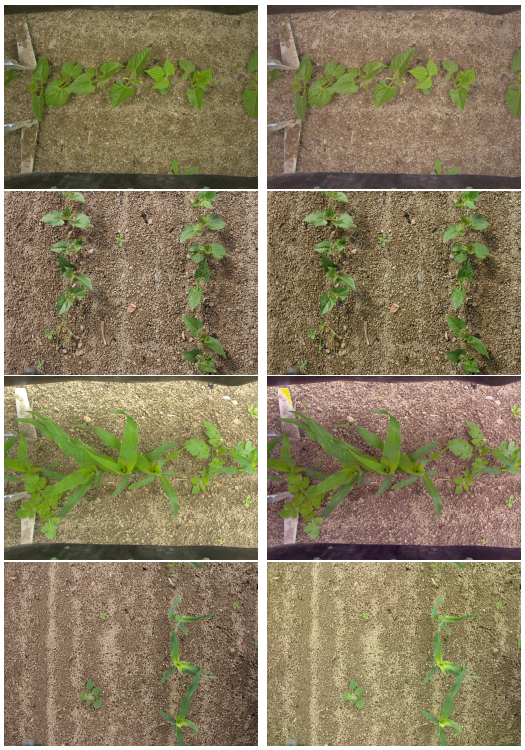

Figure 2: FFT style transfer, from top to bottom BH to WH style, WH to BH style, BM to WM style and WM to BM style.

change the whole Source dataset into Target style.

4. Comparison of Methods

From Table 2 Baseline is improved by CycleGAN on 3/4 of the combinations, the one in which it does not improve is the one, as already mentioned in Section 3.3, where false green was created. In the 3 cases where the baseline improves, it beats the FFT techniques twice. FFT tecniques improves the baseline only in the Weedelec on Bipbip combinations. In the Mais one, where the change of style seems excellent, FFTtoTarget outperforms CycleGAN. Overall FFTtoTarget looks better than FFTtoSource.
From Figure 3 we can see that:
Weedelec on Bipbip Haricot: baseline prediction generates several false positives on weeds, confusing the soil and some parts of the robot. CycleCAN and both FFT techniques seem to solve the problem, FFTtoTarget as per usual performs better than FFTtoSource.
Weedelec on Bipbip Mais: by changing plant we notice the same problem of Haricot and in addition also the black curtains of the robot generate confusion on our baseline. All the three DA methods solve the problem.
Bipbip on Weedelec Haricot: the baseline creates small holes as well as confusing some weed leaves in crop, both FFT methods does not seem to improve the situation too much while CycleGAN fixes some of the confusion and removes the holes from the plants, this is reflected in the metric score.
Bipbip on Weedelec Mais: the only situation in which none of the 3 methods of DA creates an improvement on the baseline, the recall is high but the

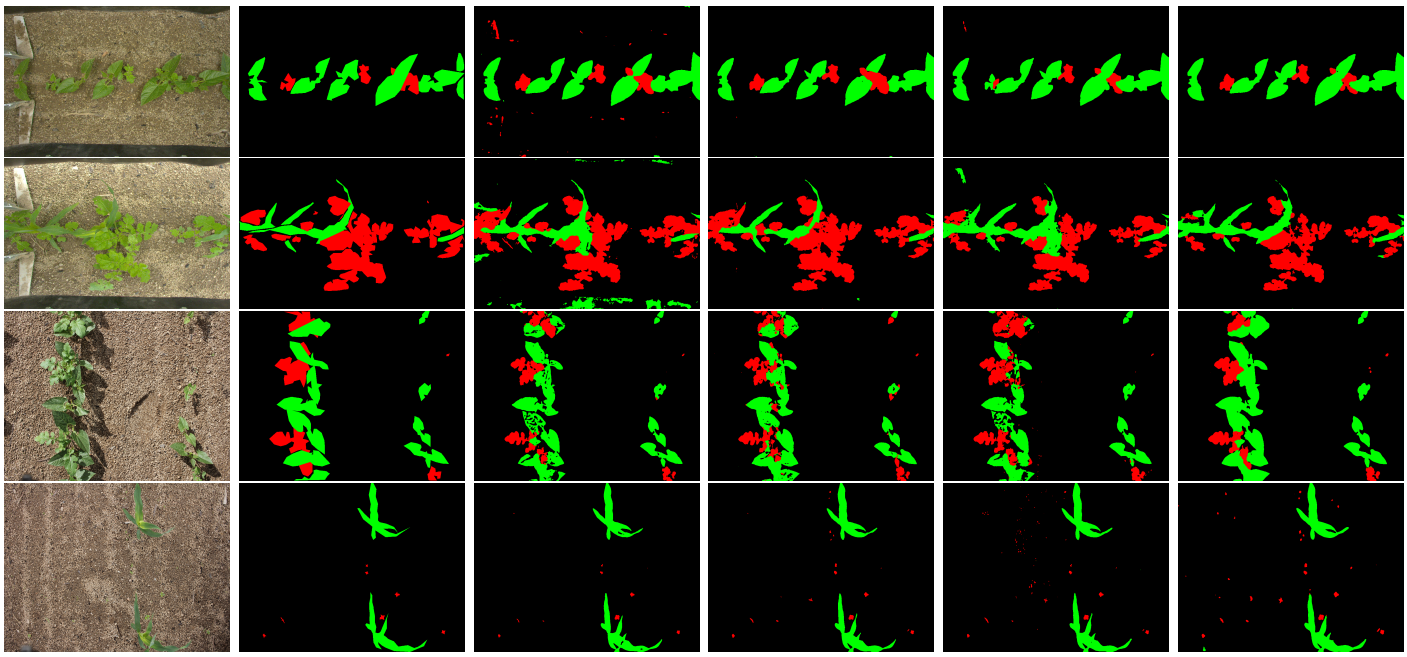

Figure 3: Order is img, truth mask, baseline, FFTtoTarget, FFTtoSource, CycleGAN. From top to bottom WH on BH, WM on BM, BH on WH and BM on WM

problem is the various false positives created from the ground that are marked as weed.

5. CycleGAN Between Fields Of Different Year

We tested the performance of CycleGAN on datasets taken by Weedelec and Bipbip on the same field, again with Mais and Haricot plants, but after two years and at different growth stages. Bipbip used the same acquisition method while Weedelec did not. We call these datasets Bipbip2021 and Weedelec2021. We used as Source the Bipbip and Weedelec datasets and as Target the Bipbip2021 and Weedelec2021 datasets. In all the combinations the generators, in order to confuse the discriminators, created or removed, green as the plants were at different levels of growth and the distribution of green was different. As can be expected, the baseline during the process is worsened obtaining IoU scores close to zero. The CycleGAN framework in these situations and with these hyperparameters was not able to maintain the semantics during the transformations.

6. Conclusions

In this thesis, we presented an approach consisting of encoder-decoder FCN based on [6] to solve the supervised segmentation problem on different datasets with different acquisition methods and different types of plants and weeds. Our segmentation system was able to segment all the datasets in a satisfactory way from a quantitative point of view. Even if the metric scores fluctuate from different datasets it is more a problem of how the data was labeled. Indeed, a better score can be obtained with a more accurate labeling in which the exact plants' outlines are represented instead of approximate ones. The second approach we presented consisted of two different methods involving CycleGAN and FFT techniques to bridge the gap of a domain-shift between Source and Target domain. The datasets used were Weedelec and Bipbip. Having for both Mais and Haricot plant, the following 4 combinations were evaluated: Weedelec Mais on Bipbip Mais, Weedelec Haricot on Bipbip Haricot, Bipbip Mais on Weedelec Mais and Bipbip Haricot on Weedelec Haricot. At the time of writing, this is the first time that FFT has been used in agriculture for the weed detection problem. We also proposed, as a new addition to the CycleGAN framework, a new loss based on the the phase of the FFT which is held constant for the first epochs and then relaxed until it reaches zero. The need for this loss arose to make the generators initialization more stable as CycleGAN, with the addition of the two FCNs for segmentation, did not always converge to an acceptable solution on the first run. On our datasets, since we could not directly compare with the results of [3] as we did not have the same datasets available, the phase approach performed the same or better than the one without it in all the four cases. We argue that the addition of this constraint has the potential to improve the initialization of the process and therefore the possibility of obtaining better scores. During the Source on Target domain evaluations, we obtained

decent results without using any DA technique but only trying to match the same zoom level of the two images. Later with the DA techniques we obtained clear improvements in 3 of the 4 cases studied. It was also important to analyze the qualitative results in addition to the metric scores, often the scores were penalized due to the difference in labeling between Source and Target, a precise pixel perfect labeling when evaluated on a less precise labeling would result in a very low recall score but high precision. We also saw how the use of CycleGAN between fields of different year and with different growth stages did not get as good as a result as in cases where the difference between the datasets was only the acquisition method. In these cases the CycleGAN framework is not able to maintain the semantics during the transformations, therefor the performance is not improved but worsened.
CycleGAN proves to be a technique as powerful as it is unstable and expensive, requiring the use of different losses and 6 deep learning networks, the transformations generated really change the style of the image while maintaining the semantics and it is difficult to distinguish to which domain each image belongs to. FFT demonstrates a much less powerful and versatile method but also much less expensive, almost free compared to the cost of CycleGAN, in situations where the domain shift is minimal and the same use of robot/camera and plant stage is ensured, FFT could unveils to be an excellent but simple solution. In the end, we argue that, when the task to be solved is the same between the two domains, the DA techniques used in the thesis are capable of satisfactorily solving the domain gap between the domains. Our results intrinsically represent this as in the datasets that were taken on the same field at the same time but with different methodology we were able to improve the performance, however in the datasets with different growth stages we were not able to improve the performance.

References

[1] Petra Bosilj, Erchan Aptoula, Tom Duckett, and Grzegorz Cielniak. Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture. *Journal of Field Robotics*, to be determined (published online), 2019.

[2] Nived Chebrolu, Philipp Lottes, Alexander Schaefer, Wera Winterhalter, Wolfram Burgard, and Cyrill Stachniss. Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *The International Journal of Robotics Research*, 2017.

[3] Dario Gogoll, Philipp Lottes, Jan Weyler, Nik Petrinic, and Cyrill Stachniss. Unsupervised domain adaptation for transferring plant classification systems to new field environments, crops, and robots. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2636–2642. IEEE, 2020.

[4] Sebastian Haug and Jörn Ostermann. A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In *European conference on computer vision*, pages 105–116. Springer, 2014.

[5] Xu Ma, Xiangwu Deng, Long Qi, Yu Jiang, Hongwei Li, Yuwei Wang, and Xupo Xing. Fully convolutional network for rice seedling and weed image segmentation at the seedling stage in paddy fields. *PloS one*, 14(4):e0215676, 2019.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[8] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9011–9020, 2020.

[9] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020.

[10] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.