Pixel-wise classification of weeds and crop in images by using a Fully Convolutional neural network

Mads Dyrmann [a,*], Anders Krogh Mortensen [b], Henrik Skov Midtiby[a], Rasmus Nyholm Jørgensen[c]

[a] The Maersk McKinney Moller Institute, University of Southern Denmark, Odense, Denmark

[b] Department of Agroecology, Aarhus University, Aarhus, Denmark

[c] Department of Engineering, Aarhus University, Aarhus, Denmark

* Corresponding author. Email: mady@mmmi.sdu.dk

Abstract

Effective weed control, using either mechanical or chemical means, relies on knowledge of the crop and weed plant occurrences in the field. This knowledge can be obtained automatically by analyzing images collected in the field.

Many existing methods for plant detection in images make the assumption that plant foliage does not overlap. This assumption is often violated, reducing the performance of existing methods. This study overcomes this issue by training a convolutional neural network to create a pixel-wise classification of crops, weeds and soil in RGB images from fields, in order to know the exact position of the plants. This training is based on simulated top-down images of weeds and maize in fields.

The results show an pixel accuracy over 94% and a 100% detection rate of both maize and weeds, when tested on real images, while a high intersection over union is kept. The system can handle 2.4 images per second for images with a resolution of 1MPix, when using an Nvidia Titan X GPU.

Keywords: Deep Learning, Semantic segmentation, weed recognition, Computer vision

1. Introduction

One of the challenges of precision farming is to reduce the usage of herbicides, while keeping a high crop yield. Some precision farming techniques seek to address this challenge by monitoring crops and weeds and target the herbicide only to the weeds. However, monitoring crops and weeds manually is a time consuming and expensive activity. Therefore, automated detection and localization of plants is a prerequisite before such activities can be applied.

An alternative to using herbicide is to use mechanical hoes, which automatically avoid the value crop and only cuts the stems of the weeds. Here, again, it is necessary to know the exact position of the weeds and crops in order to apply this technique.

In addition to use this monitoring to optimize the weed control, the monitoring can also be used to estimate the weed cover and growth of the crops, which is valuable information when applying fertilization to the field.

Previous studies have based plant recognition on a series of steps. The first step is the segmentation, which is about finding out which pixels belong to plants and soil. This has previously been done by using e.g. color index, such as excess green-excessive red, which is based on the green chromaticity. Or it could be based on, NDVI, if the near infrared information is available (Dyrmann et al. 2014). When the pixels are segmented in plants and soil, each plant object is categorized as either weed or crop by using different machine learning techniques.

The challenge of the existing techniques for finding weeds and crops is that they have problems handling overlapping plants.

In this paper, we investigate the problem of plants perception and distinguishing the crops from the weeds growing on the field, even when the plants are overlapping each other. The method is based on a fully convolutional neural network for semantic segmentation (Long, 2015), which will produce images the same size as the input image, in which maize, soil and weeds are classified on a pixel level.

Hundreds or thousands of annotated training images are normally required to make a good training of a convolutional neural network (Oquab et al., 2014). As manual segmentation of this many images is not feasible, a system that simulates top-down images of overlapping plants on soil background has been created. These images are simulated with different soil types and different variations within each plant species. For each of the simulated images, a ground truth segmented image is created in which each pixel is given a label as either plant, weed, or soil.

The output of the networks is an image of the same size as the input, where pixels are labelled according to the predicted class of the given pixel.

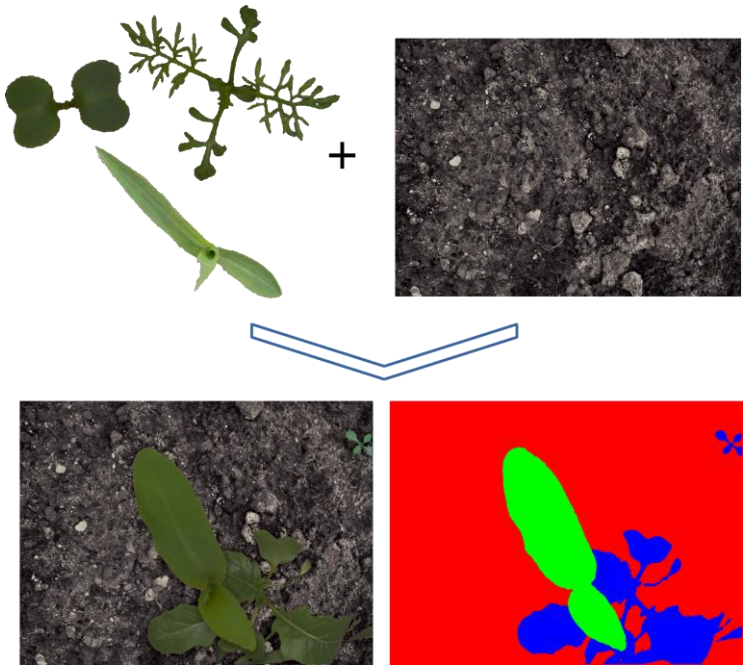

Figure 1. The generation of new training images. Segmented plants are placed randomly on top of images of soil, creating simulated field images. Together with the simulated images, ground truth segmented images are generated for training. In the segmented image: red = soil, blue = weed and green = maize.

2. Materials and Methods

In order to train the neural network to segment images, it is necessary to get training data consisting of images that has already been segmented. This training data can be made by manual hand segmentation of the images, but this is a time consuming affair.

To overcome the problem of getting segmented training images, the training images are modelled from segmented images of single plants rather than hand segmenting full images. This is done by plotting images of segmented plants on top of images of bare soil, which makes it possible to generate as much training data as desired.

The process of this image generation and a sample of one of the generated images is shown in Figure 1 as well as the segmented image. In the segmented image, three labels are used: one for the soil, one for maize, and one for weeds.

A total of 301 images of soil and 8430 images of segmented plants have been used to generate the modelled images. These plants originate from eight different datasets that are acquired both in indoor and outdoor settings and therefore contain variation in terms of both resolution and illumination.

These plants are divided so that 80% of the plants are used to generate the training images, while 20% of the plants are used to generate images for verification. The plants cover 23 different weed species and maize.

A convolutional neural network learns to recognize objects by being presented for a lot of samples that cover the variation of the objects to be classified. In this case, this means that plants with different rotations, scaling, pixel intensities etc. must be used. Likewise, different types of soil should be used in order to make the network able to recognize the soil independent of its type. In order to increase the variations of the plants, different types of augmentation has been applied:

Each plant has been scaled randomly from 80 to 100% of their original size, which helps making the system scale invariant. The plants have also been rotated randomly in one degree increments, which helps in making the system rotation invariant. As the images of the plants originate from a limited set of data, including images acquired under artificial illumination, the hue, saturation and intensity are varied slightly in order to make the system less dependent on illumination and color. Furthermore, random shadows are added on top of the images to simulate shadows cast by other plants.

When the plants are placed in the image, they are placed randomly, with the only requirement that the centers of mass of the plants are placed on top of the soil, which simulates the stem points of the plant. When placing the images this way, some plants will overlap each other, which forces the network to learn to recognize the plants, even in that case.

The images are subsequently cropped in many small images of 800x800 pixels. This size limit has been set, as the Nvidia TitanX GPU, that was used for training, otherwise would run low on memory. After the cropping there is 3463 images for training and 123 images for verification.

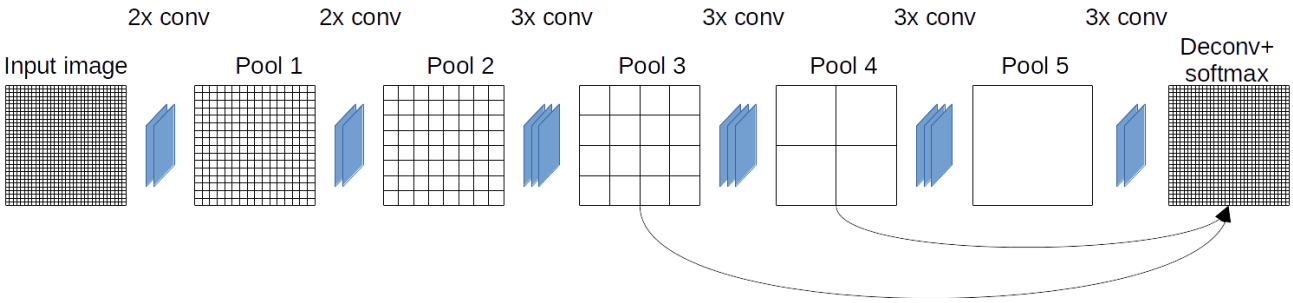

Figure 2. Network architecture. The network consists of 5 max pooling layers and 15 convolutional layers. Shortcuts form pooling layer 3 and 4 to the deconvolution layer is used for restoring smaller details in the segmented images.

2.1. Network Architecture

The convolutional neural network, that is used for training, is based on the work by Long et al. (2015), which is a modified version of the VGG-16 convolutional neural network (Simonyan & Zisserman, 2014). The network is modified so that the output is a convolutional layer instead of a fully connected layer. This enables the network to produce spatial predictions of classes rather than 'per image' classes, as VGG-16 originally was used for. The network is sketched in Figure 2.

The network contains a total of five 2x2 max pooling layers. These pooling layers result in a total downscaling by a factor of 32. We, however, want the output image to be of same size as the input image. Therefore, in order to achieve an output image of the same size as the input image, a deconvolutional layer with a 32 pixel stride is added at the softmax prediction layer. The drawback is that the pooling and subsequent upscaling limits the resolution of the segmented images, which therefore will lack small details. In order to restore some of these details, shortcuts are made to previous layers in the network, at which stages only three and four poolings have been applied, respectively. The output of these shortcuts are upscaled by a factor of 8 and 16, respectively, and afterwards summed with the original output.

2.2. Training

Oquab et al. (2014) have shown that the accuracy of a convolutional neural network can be increased by using transfer learning. Transfer learning is a technique, where a different, but large, dataset is used for pretraining the weights of the network. The large dataset ensures that the features learned in the first layers of the network are general features, such as edge detectors. Here the PASCAL-Context Dataset (Mottaghi et al., 2014) has been used, which is a dataset of segmented images from 59 different categories, which are not related to agriculture.

After the pretraining, the size of the output layer has been changed to reflect the three labels in this study (maize, weeds and soil) and the training has been continued on the images of maize and weeds.

Normally, when training convolutional neural networks, you would run several epochs, where the same training data is reused. However, because of the automated image generation, it is possible to generate as many images as desired. Therefore it is not necessary to run several epochs with the same training images, as new images can be generated for every single update of the network. Even though new images are used for every single update, the network will still be able to overfit the learned model. This is because the training images are all generated from the same plants, which are presented with different poses in different environments.

1. Results and Discussion

The following evaluation is designed to test how well our plant segmentation method performs on real images. In order to test this, two image from two different Danish maize fields have been segmented by hand. The first image is from a healthy maize field with only little plant overlap, while the second image is from a maize filed with smaller maize plants and a higher weed coverage. We call the hand segmented images I_{hand} and the resulting semantic segmented image we called I_{sem}.

1. The images will be evaluated on the following parameters:
2. The ratio of weed objects that has been found out of the total number of weeds objects. A weed objects is said to be found, if there is at least one pixel overlap between the weed in I_{hand} and I_{sem}.
3. The ratio of crop objects that has been found out of the total number of crop plants. A crop objects is said to be found, if there is at least one pixel overlap between the crop in I_{hand} and I_{sem}.
4. Overall accuracy
5. The intersection over union for the weeds, for the crops and for the soil

The intersection-over-union is a measure of the number of pixels of the intersection relative to the number of pixels of the union of each class in I_{hand} and I_{sem}. For weeds, it can be written as:

$$IOU(weeds) = \frac{\sum\left(I_{hand}(weeds) \cap I_{seg}(weeds)\right)}{\sum\left(I_{hand}(weeds) \cup I_{seg}(weeds)\right)}$$

2.3. Image 1: small overlap

This image was taken using a handheld camera, and therefore there has been no shade or artificial lighting, when taking the image. The image is 3840x2304 pixels, which is too large for the image to be processed on the computer's GPU. Therefore the image has been processed in blocks of 768x768 pixels, which subsequently are stitched together.

The image contains a total of 121 weeds, which are group in 116 objects in I_{hand} due to overlap between some weeds. The image contains two maize plants, which also overlap with some of the weeds.

The test image is shown in Figure 3a together with the ground truth hand segmented image in Figure 3b.

The result of the automated semantic segmentation applied to this image is shown in Figure 4a. Here the amount of red indicates how much a pixel is believed to be soil, the amount of green indicates how much a pixels is believed to be maize, and the amount of blue indicates how much a pixel is believed to be a weed. Figure 4b shows the result, where only the dominant class remains.

As can be seen in Figure 4b, the detection rate of the algorithm is 100% for both the weeds and the crops. Some of the weeds are, however, split in multiple parts. This is especially the case in the stem region of the weeds and for the grasses. The reason for this is likely due to the subsequent down and upscaling in the network, which causes the network to lose some of the small details. This lack of small details is, however, not a problem for applications such as precision spraying or mechanical hoeing as long as the weeds can be detected.

Figure 3a. Test image of maize and weeds.

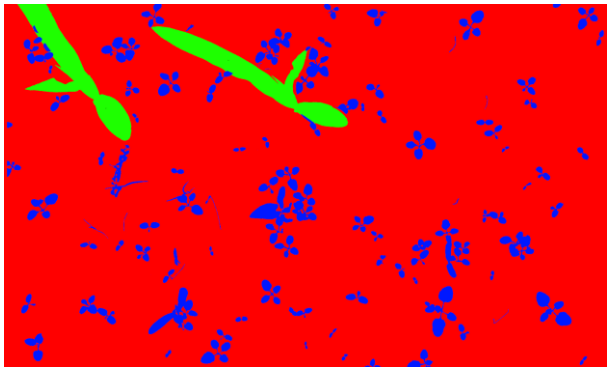

Figure 3b. Hand segmented ground truth image

Figure 3. Input image 1 for test

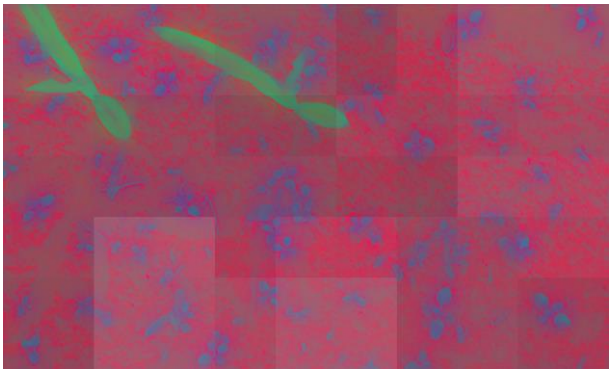

Figure 4a. Output of the network. The amount of red, green and blue indicates the belief of a pixels being soil, maize and weed, respectively.

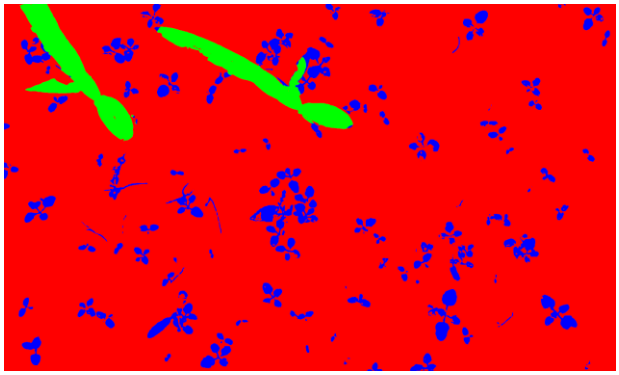

Figure 4b. Non-maximum suppression of network output, where only the dominant color remain.

Figure 4. Result for input image 1

The overall accuracy of the classification is 98.3%. This, accuracy does not take into account, that there are far more soil pixels than weed and crop pixels. Therefore, we will also consider the intersection over union for each class.

The intersection over union for crops is 0.93, for weeds it is 0.79 and for the soil it is 0.98. This means that the areas of the weeds and especially for the maize are close to the real areas of the plants. The method will therefore also be applicable for determine weed coverage and crop sizes.

2.4. Image 2: large overlap

Now we will test how the system performs on a less ideal image. First of all, the maize plant in this image is small and it overlaps with some weeds. Furthermore, some of the weeds are reddish unlike all plants from the training images. The image was taken using a camera that was mounted under a shade on a trailer.

The test image is shown in Figure 5a together with the ground truth hand segmented image. The result of the automated semantic segmentation, applied to this image is shown in Figure 6a. Figure 6b shows the result after non-maximum suppression, where only the dominant class remains.

Figure 5a. Test image of maize and weeds.

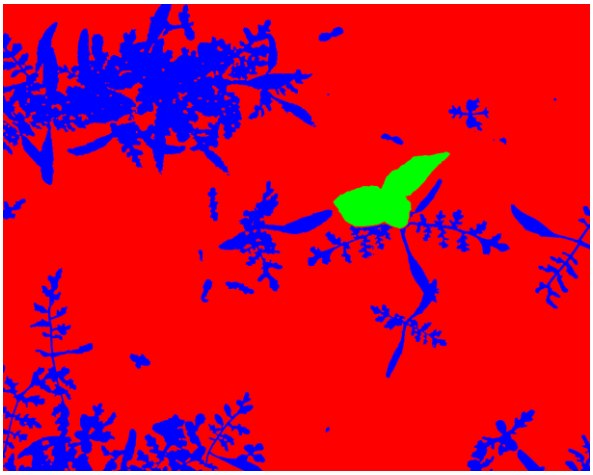

Figure 5b. Hand segmented ground truth image

Figure 5. Input image 2 for test

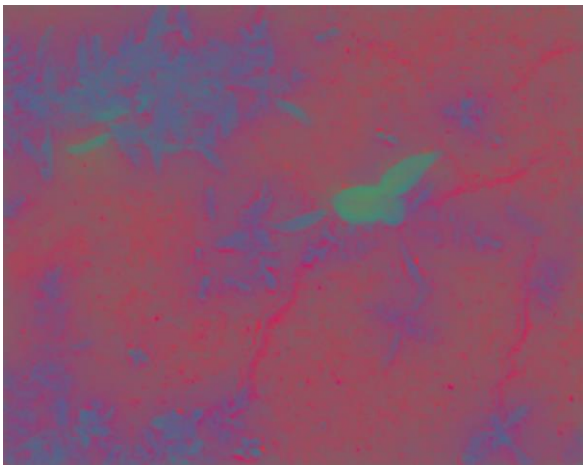

Figure 6a. Output of the network. The amount of red, green and blue indicates the belief of a pixels being soil, maize and weed, respectively.

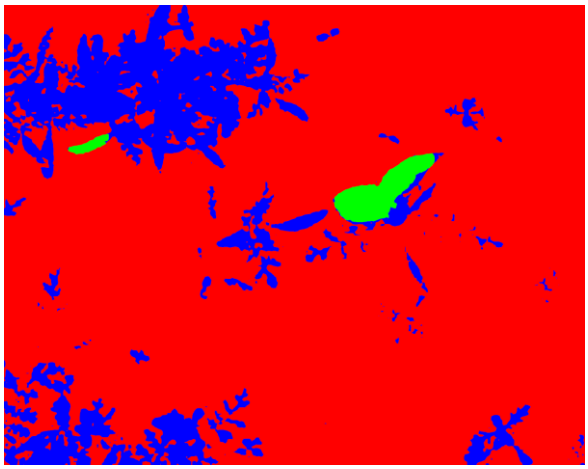

Figure 6b. Non-maximum suppression of network output, where only the dominant color remains.

Figure 6. Result for input image 2

The detection rate of the algorithm is again 100% for both the weeds and the crops. However, a leaf of one of the weeds is recognized as maize. The overall accuracy of the classification is 94.4%.

The intersection over union for crops is 0.71, for weeds it is 0.70 and for the soil the intersection over union is 0.93. The performance is thereby a little lower than for the previous image. However, when looking at the image, it can be seen that mainly the red parts of the weeds are recognized as soil. The reason for this is probably due to the training, in which no non-green weeds has been used and that striding in the network limits the resolution of the output.

3. Real time evaluation

The system can handle 2.43 images per second for images with a resolution of 1MPix, when using an Nvidia Titan X GPU. This performance is achieved without optimization of the network architecture. An improved performance is therefore expected if the neural network is stripped from neurons that only contribute little to the overall loss, whereby real time analysis will be feasible.

4. Conclusions

Some precision farming techniques depends on recognition of weeds and crops in fields. We addressed this problem of detecting weeds and maize by using a fully convolutional neural network that is based on a modified version of the VGG16 architecture. This network produces semantic segmented images as output.

The network has been trained, solely on modelled images of maize and weeds, which has enabled the network to distinguish maize, weeds and soil in real images with accuracies greater than 94%. Furthermore, the network is capable of distinguish weeds from maize when the plants are overlapping each other. The system will therefore also be useful for determine weed and crop coverage.

Acknowledgements

The work was founded by Green Development and Demonstration Programme (GUDP) under the Danish Ministry for Food, Agriculture and Fisheries.

References

Dyrmann, M., & Christiansen, P. (2014). Automated Classification of Seedlings Using Computer Vision. Aarhus. Retrieved from http://plant_recognition.sdu.dk/files/AutomatedClassificationOfSeedlingsUsingComputerVision.pdf

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431–3440. http://doi.org/10.1109/CVPR.2015.7298965

Mottaghi, R., Chen, X., Liu, X., Cho, N., Lee, S., Urtasun, R., & Yuille, A. (2010). The Role of Context for Object Detection and Semantic Segmentation in the Wild. http://doi.org/10.13140/2.1.2577.6000

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 1717–1724. http://doi.org/10.1109/CVPR.2014.222

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. ImageNet Challenge, 1–10. http://doi.org/10.1016/j.infsof.2008.09.005