Semantic Segmentation for the Estimation of Plant and Soil Parameters on Agricultural Machines

Peter Riegler-Nurscher
Josephinum Research
Wieselburg, Austria
Email: p.riegler-nurscher@josephinum.at

Johann Prankl
Josephinum Research
Wieselburg, Austria
Email: johann.prankl@josephinum.at

Markus Vincze
Automation and Control Institute
Vienna University of Technology
Vienna, Austria
Email: vincze@acin.tuwien.ac.at

Abstract*—Many machine vision problems in agriculture, like plant classification, soil cover estimation or agronomic process evaluation in general, can be solved with semantic segmentation approaches. Naturally growing non-rigid organic and inorganic materials and plants are often characterized by blurred class transitions and high intra-class variance. Especially outdoor uncontrolled plant growth and plant decomposition lead to strong occlusions, cluttered scenes and strong illumination variances in images. An agricultural vision system has to cope with these challenges. This work presents four different applications for semantic segmentation in agriculture: (1) soil cover estimation, (2) estimation of grass-legumes ratio, (3) grassland swath detection and (4) grassland cut segmentation. For training, TensorFlow and a convolutional neural network are used. We investigate the influence of different pre-training methods to improve the overall classification performance with a limited number of training samples. The best test accuracy was achieved by initializing the weights from a model based on a semi-artificial clover and grass data set. The use cases with images from closer perspectives, (1) and (2), resulted in less accuracy compared to use cases (3) and (4). In general, all use cases can be solved with sufficient accuracy.

Keywords*–*Semantic Segmentation; Agriculture.

I. Introduction

Most research in the field of computer vision for agriculture focuses on plant and weed detection, pest detection and plant health. However, with increasing autonomy of agricultural machines, the need for process monitoring and evaluation, especially for seeding and harvesting, increases. Many of these applications use semantic segmentation to classify non rigid objects, like plants and soil, or at a higher level, to detect field areas worked of different processing stages.

Development in recent years in semantic segmentation focuses mostly on Convolutional Neural Network (CNN) based methods [1]. CNNs for semantic segmentation consist of an encoder followed by a decoder network. Current works focus on solving the degradation problem, where detailed shape information is discarded by the encoder. Circumventing the degradation problem increases accuracy of the output mask.

We applied semantic segmentation in four use cases on arable fields and grassland. During tillage and seeding, soil cover (1) is an important parameter for soil conservation. The ability to distinguish grasses and legumes (2) during harvesting of grassland is the basis for site specific application of fertilizer and targeted feeding. In harvesting of grassland, detection of swaths (3) and areas of cut grass (4) are the basis for automation of machines and yield estimation. This is work in progress and we want to present preliminary findings in this short paper. The main contribution of this work is the investigation of the influence of pre-training in these use cases in different perspectives and resolutions. Additionally, we adapted the ERFNet CNN [2] for the use cases and tested the inference speed with different hardware. The following Section II gives an overview of the four agronomic use cases for semantic segmentation and its challenges. Section III presents the method applied for the segmentation task. In Section IV, the accuracy and Intersection over Union (IoU) of the different trained models and the inference speeds are shown and discussed.

II. Use Cases

We investigated four different use cases for semantic segmentation on agricultural fields. The images were captured with color cameras mounted on different agricultural implements and annotated manually.

The first use case is soil cover estimation. Soil cover is an important parameter to measure the danger of soil erosion. To objectively quantify the amount of soil cover on a field, camera images are classified into the classes soil, living organic matter, dead organic matter and stones. Studies like [3] and [4] have investigated the problem of segmenting soil cover in images, but often fail because of environmental influences, such as direct sunlight or motion blur. The work in [5] uses CNNs for soil cover estimation, but on a very limited test data set. The image in Figure 1 depicts all four classes. Higher amount of soil cover increases the ability to protect against erosion, where soil cover includes all classes except soil. The percentage of soil cover calculated from the segmentation mask can directly be used to quantify erosion protection.

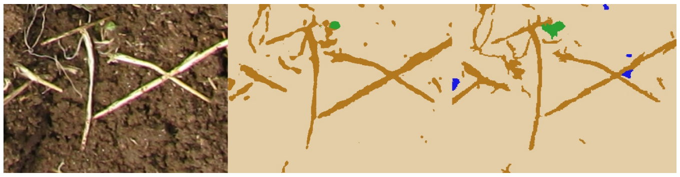

Figure 1. Soil image for soil cover segmentation (left), test mask (middle), ground truth map (right). Living organic matter ●, dead organic matter ●, soil ● and stone ●.

A related problem to soil cover classification is distinguishing plant species. Our special use case is to distinguish between soil, grasses and legumes in grassland. An example image is shown in Figure 2. This segmentation can be the basis for optimized cow feeding and it can serve as an additional parameter for grassland yield estimation.

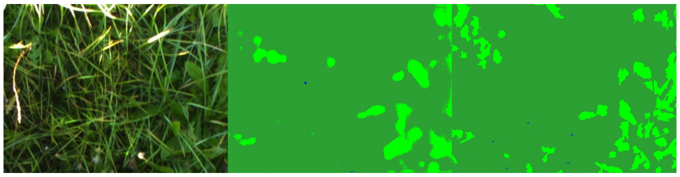

Figure 2. Meadow image for grass/legumes segmentation (left), test mask (middle), ground truth map (right). Grass ● and legumes ●.

In contrast to use cases (1) and (2), the last two have a more global perspective where the classification is not on plant level, but on field area level. The method presented in [6] shows the potential for segmentation of swaths based on stereo depth data and texture information. Our attempt to segment the swath purely on color images in a natural environment makes this problem more difficult, but allows to rely on simpler hardware setups with a single camera. Figure 3 shows an example image for the detection of grassland swaths. This approach can be used for navigation within the field or for yield estimation. The pixels of the images are binary classified into swath or no swath.

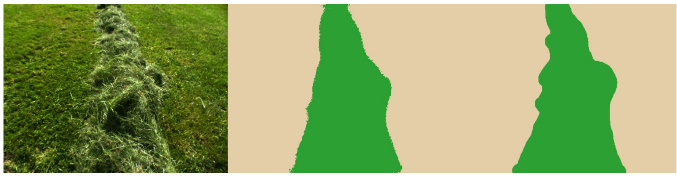

Figure 3. Grassland swath image (left), test mask (middle), ground truth map (right). Swath ● and no swath ●.

The last use case, segmentation of areas of cut grass, is a very similar task to swath detection. The segmentation information can be used for machine control or yield estimation. The image is segmented into the different states of grass during mowing: standing grass, grass turf and mown grass. Due to the camera mounting position, an additional class is introduced to mask the machine. Figure 4 shows an example image from the test data set. Cropped parts of areas with standing grass are included in the grasses and legumes use case.

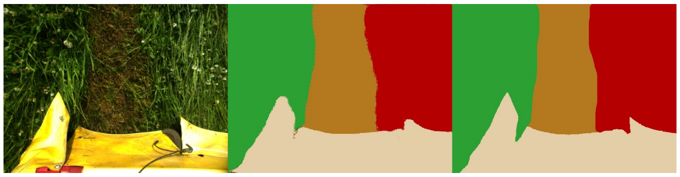

Figure 4. Image for segmentation of areas of cut grass (left), test mask (middle), ground truth map (right). Standing meadow ●, grass turf ●, machine ● and mown grass ●.

CNNs for semantic segmentation are trained in a supervised way. Basis for the training are labelled training samples. Table I shows the number of samples for each use case.

TABLE I. DATASETS FOR EACH USE CASE

Use case	*Number of images*
(1) soil cover estimation	3621
(2) grass/legumes ratio estimation	1030
(3) swath detection	189
(4) cut segmentation	382

To increase the variance of the training data set, image augmentations were added. Usually, images are taken in any orientation, therefore we added horizontally and vertically flipped images. To accommodate for distance changes between soil and camera, with fixed focus cameras, blurring was added randomly. The application on mobile machines with fast optical flow requires short exposure times. At higher speeds, the image brightness decreases. Strong lighting variations in outdoor operations are simulated with linear and non-linear (gamma) brightness changes.

III. SEMANTIC SEGMENTATION WITH CNN

The task of semantic segmentation is to classify each pixel of an image into predefined classes. A CNN for semantic segmentation consists of an encoder block followed by a decoder block. During training, the weights of the network are incrementally adapted to fit the labelled training data. Afterwards, new test images are fed into the network to generate the corresponding classification mask as output of the decoder. The encoder extracts discriminative features from the image to get semantic information for classifying objects. The decoder network reconstructs a class label map, where information from high dimensional encoder layers bypasses the bottle-net in skip connections. This allows to sustain detailed contour information. These CNNs are called U-nets. One variant of a U-net is the ERFNet [2]. The work in [7] compared different state of the art semantic segmentation network architectures. The authors showed that ERFNet provides a good compromise between speed and accuracy and is further used in the proposed work. ERFNet introduces non-bottleneck-1D (non-bt-1D) layers, which combine benefits of bottleneck and non-bottleneck layers. Table II shows the layer architecture of the implemented ERFNet. The implementation is based on an adapted version of the bonnet framework [8].

A major problem in agricultural image processing is that it is quite difficult to generate data, so most data sets are quite small. Different approaches have been introduced, which are able to deal with small data sets. One option is to use data augmentation, where image processing steps, like blurring, affine transformations etc., are performed randomly on the training images. This enriches the training data set. Another possibility is to artificially render a large number of training images. Artificial training images often result in a certain bias; they do not cover the high variance of natural images. Hence, the performance is often weak. Pre-training, on the other hand, uses model weights from similar problems as initial weights for training [9]. We tested combinations of all these approaches and investigated the aspect of pre-training in detail.

Pre-training allows for transferring model parameters from a similar problem. We use model weights of the encoder as initial parameters in the new model. This allows for reusing encoder features and transferring semantic information. The decoder weights are trained from scratch, due to different final classes. The last column of Table II shows which layers are initialized with weights from the pre-trained models.

The publicly available sugar beet data set [10] consists of

TABLE II. LAYERS OF THE ERFNET [2] AS IMPLEMENTED IN THE BONNET FRAMEWORK [8]. THE NUMBER OF OUTPUT FEATURES AND RESOLUTION ARE FOR INPUT IMAGES OF 512X384.

	Layer	Layer Type	# output features	output resolution	initia-lization
Encoder	1	Downsampler block	8	256x192	pre-trained model
	2-3	2 x Non-bt-1D	8	256x192	
	4	Downsampler block	16	128x96	
	5-8	4 x Non-bt-1D	16	128x96	
	9	Downsampler block	64	64x48	
	10-13	4 x Non-bt-1D	64	64x48	
	14	Downsampler block	64	64x48	
	15-18	4 x Non-bt-1D	64	64x48	
Decoder	19	Deconvolution (upsampling)	32	128x96	random
	20-23	4 x Non-bt-1D	32	128x96	
	24	Deconvolution (upsampling)	16	256x192	
	25-28	4 x Non-bt-1D	16	256x192	
	29	Deconvolution (upsampling)	8	512x384	
	30-31	2 x Non-bt-1D	8	512x384	

images of sugar beets and different weeds on various soils. The data set consists of 12,714 images with a resolution of 1296 x 966 pixel, which are resized to 512 x 384 pixel for training. Figure 5 (left) shows an example image of the sugar beet data set. Another publicly available data set is the semi-artificial GrassClover data set from Aarhus University [11]. It consists of artificially generated collages of real cut out clover and grass images. We used 2,600 of the 8,000 images in the data set. The semi-artificial images have a higher resolution than the images in our data set, therefore we resized them by 60% and cut out parts. This resulted in 33,000 image patches with a size of 512 x 384 pixel. Figure 5 (right) shows an example image of the data set.

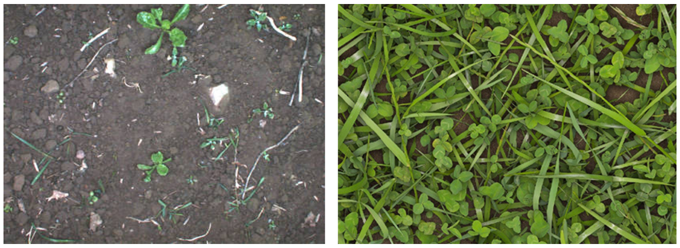

Figure 5. Example image from the sugar beet data set [10] (left), example image of the semi-artificial GrassClover data set [11] (right).

Both data sets contain images, or are based on images, captured under controlled lighting with no direct sunlight. The images are not blurry and contain no impurities. However, the images in our data sets are captured on moving agricultural implements without parasol or additional lighting and, therefore, contain all these environmental influences. Figure 6 shows two examples from the grass/legumes ratio use case with motion blur (top), saturated parts caused by the limited dynamic range under direct sunlight (bottom) and with strong shadows. However, these issues can be overcome to some extent by varying the exposure time of the camera dependent on lighting conditions and driving speed of the machine.

IV. EVALUATION

Each use case was trained in three variants. The weights were either initialized with the sugar beet data set, with the clover grass data set or randomly (without pre-training).

After training convergence, the models were evaluated with a separate test data set. The metrics, IoU and accuracy, were estimated, as presented in Table III. The mean IoU was calculated as shown in (1). Each class, of all C classes contributes equally to the overall IoU, therefore, scarce classes equally influence the IoU.

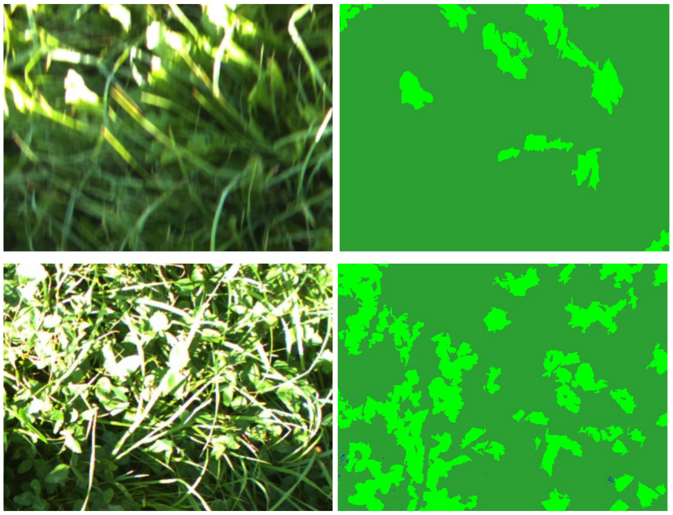

Figure 6. Challenging example images from the grass/legumes use case. The corresponding manually annotated label maps are shown on the right.

$$IoU = \frac{1}{C} \sum_{i=1}^{C} \frac{TP}{TP + FP + FN} \qquad (1)$$

The best model was selected by the best accuracy in the validation data set.

TABLE III. MODEL ACCURACY AND IOU, WITH AND WITHOUT PRE-TRAINING ON THE TEST DATA SETS

Application	*Pre-training*	*Accuracy*	*IoU*
(1) soil cover estimation	sugar beet dataset	0.7993	0.4974
	clover grass dataset	**0.8746**	**0.6640**
	none	0.8546	0.6172
(2) grass/legumes ratio estimation	sugar beet dataset	0.8522	**0.5374**
	clover grass dataset	**0.8859**	0.4480
	none	0.8462	0.5288
(3) swath detection	sugar beet dataset	0.9653	0.9313
	clover grass dataset	**0.9734**	**0.9470**
	none	0.9604	0.9221
(4) cut segmentation	sugar beet dataset	0.9106	0.7903
	clover grass dataset	**0.9340**	**0.8312**
	none	0.9286	0.8241

The results show that all use cases can be solved with satisfactory accuracy. In general, pre-training improves model accuracy. Especially, pre-training with the semi-artificial clover grass data set is beneficial for all use cases. This can be attributed to several factors. In general, the scenes in the sugar beet data set have less soil cover. All four use cases, especially the grassland use cases (2-4), have more soil cover, up to 100%. Additionally, the linear and circular structures within the image data are more similar to the clover grass data set, than to the sugar beet data set. Especially grass and clover are very common in our grassland use cases. The improved performance can be explained by encoder features, taken from the pre-training, which more accurately describe our scenes. The initial weights from the sugar beet data set worsen the accuracy for the cut segmentation even more than no pre-training. This might be attributed to strongly differing requirements and ill-fitting features.

TABLE IV. INFERENCE SPEEDS OF ERFNET ON DIFFERENT DEVICES WITH AN IMAGE RESOLUTION OF 512X384 PX.

Device	Inference time
UP AI Core X Myriad$^{\text{TM}}$ X 2485	268 ms
Intel® Core$^{\text{TM}}$ i7-3630QM CPU	190 ms
NVIDIA® Jetson Nano$^{\text{TM}}$	166 ms
NVIDIA® GeForce RTX 2080 Ti	6.1 ms

As expected, the best accuracy gain was accomplished with the clover grass data set for the grass/legumes ratio estimation problem. Our data set differs mainly in higher naturalness of the images to the semi-artificial clover grass data set.

The use cases swath detection and cut segmentation resulted in better segmentation accuracy and IoU. This can be explained by lower variance within the samples and less conflicting annotations in the data set because of the simpler, less cluttered, scenery. In addition, the effect of motion blurring is more apparent on the soil and plant images, compared to images from more global perspectives. There are more annotation errors, and in general poorer quality, in the data sets with fine grained resolution.

For application on mobile agricultural machines, edge hardware for inference of the models is needed. Depending on the use case, inference times must be guaranteed, especially for real time machine control, and on the other hand, connection to cloud computing is often not an option in rural areas. We investigated the inference speeds on four different devices, as shown in Table IV. The inference on the NVIDIA® GeForce graphics card is shown for reference, but is not eligible for the use on agricultural machines due to active cooling and high power consumption. The Jetson Nano$^{\text{TM}}$ is the most promising edge device for our application, based on inference speed and power consumption, and will be integrated into a vision system.

Results from previous works are published for the common use case of soil cover estimation (e.g., in [4]). In order to show the improvements of CNN methods compared to classic methods, we compared the soil cover estimation results using the established grid method to the results presented in [4]. In the grid method, points are selected in a regular grid pattern from the image and the share for each class is calculated in percent. A regression line between the manual annotation values and the computed results shows the quality of the estimation. The random forest method used in [4] had a regression of $y_{RF} = 0.7573x + 0.233$ ($R^2 = 0.7627$) to the manually annotated test samples x for the class soil and $y_{RF} = 0.5095x + 0.0363$ ($R^2 = 0.7221$) for class dead organic matter. The proposed method in this paper accomplishes a relation of $y_{CNN} = 0.944x + 0.0878$ ($R^2 = 0.8085$) for soil and $y_{CNN} = 0.7687x + 0.0002$ ($R^2 = 0.7467$) for dead organic matter. This shows a significant improvement, especially for the challenging task of distinguishing soil from dead organic matter.

V. CONCLUSION

Semantic segmentation is an important task for many applications in agronomic image analysis. Especially for soil and plant segmentation, CNN based approaches look very promising.

In order to get good results with a low number of training samples, we investigated the influence of pre-training on four different use cases, soil cover estimation, estimation of grass-legumes ratio, grassland swath detection and grassland cut segmentation.

In general, pre-training improves model accuracy. Especially, pre-training with the semi-artificial clover grass data set [11] is beneficial for all use cases. This can be attributed to the similarity of the textures and the consequently well-fitting of the encoder features emerged from the pre-training.

In the further course of the project, we will integrate the use cases into applications and record and annotate additional labelled data. This will allow for validation of the presented models integrated on agricultural machines based on high level agronomic metrics.

ACKNOWLEDGMENT

Parts of this study were financially funded by the Austrian Research Promotion Agency (FFG) under the program BRIDGE 1 between March 2018 and February 2021.

REFERENCES

[1] I. Ulku and E. Akagunduz, "A survey on deep learning-based architectures for semantic segmentation on 2d images," arXiv preprint arXiv:1912.10230, 2019.

[2] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 1, 2017, pp. 263–272.

[3] T. Bauer and P. Strauss, "A rule-based image analysis approach for calculating residues and vegetation cover under field conditions," Catena, vol. 113, 2014, pp. 363–369.

[4] P. Riegler-Nurscher, J. Prankl, T. Bauer, P. Strauss, and H. Prankl, "A machine learning approach for pixel wise classification of residue and vegetation cover under field conditions." Biosystems Engineering, vol. 169, 2018, pp. 188 – 198.

[5] A. K. Mortensen et al., "Semantic segmentation of mixed crops using deep convolutional neural network." in CIGR-AgEng Conference, 26-29 June 2016, Aarhus, Denmark. Abstracts and Full papers. Organising Committee, CIGR 2016, 2016, pp. 1–6.

[6] M. R. Blas and M. Blanke, "Stereo vision with texture learning for fault-tolerant automatic baling," Computers and Electronics in Agriculture, vol. 75, no. 1, 2011, pp. 159 – 168.

[7] S. Mehta, M. Rastegari, A. Caspi, L. G. Shapiro, and H. Hajishirzi, "Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," CoRR, vol. abs/1803.06815, 2018. [Online]. Available: http://arxiv.org/abs/1803.06815

[8] A. Milioto and C. Stachniss, "Bonnet: An open-source training and deployment framework for semantic segmentation in robotics using cnns," in 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 7094–7100.

[9] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" arXiv, no. 1411.1792, 2014.

[10] N. Chebrolu et al., "Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields," The International Journal of Robotics Research, vol. 36, no. 10, 2017, pp. 1045–1052.

[11] S. Skovsen et al., "The grassclover image dataset for semantic and hierarchical species understanding in agriculture," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 2676–2684.