Deep semantic segmentation of 3D plant point clouds

Karoline Heiwolt[1], Tom Duckett[1], and Grzegorz Cielniak[1]

Lincoln Centre for Autonomous Systems, University of Lincoln, UK
{kheiwolt,gcielniak}@lincoln.ac.uk, tom.d.duckett@gmail.com

Abstract. Plant phenotyping is an essential step in the plant breeding cycle, necessary to ensure food safety for a growing world population. Standard procedures for evaluating three-dimensional plant morphology and extracting relevant phenotypic characteristics are slow, costly, and in need of automation. Previous work towards automatic semantic segmentation of plants relies on explicit prior knowledge about the species and sensor set-up, as well as manually tuned parameters. In this work, we propose to use a supervised machine learning algorithm to predict per-point semantic annotations directly from point cloud data of whole plants and minimise the necessary user input. We train a PointNet++ variant on a fully annotated procedurally generated data set of partial point clouds of tomato plants, and show that the network is capable of distinguishing between the semantic classes of leaves, stems, and soil based on structural data only. We present both quantitative and qualitative evaluation results, and establish a proof of concept, indicating that deep learning is a promising approach towards replacing the current complex, laborious, species-specific, state-of-the-art plant segmentation procedures.

Keywords: 3D perception · semantic segmentation · plant phenotyping.

1 Introduction

The global agriculture industry currently faces the challenges of adapting to new climates and reducing its environmental impact, while also feeding a fast-growing world population. One essential effort needed to overcome these challenges is the breeding of new high-yielding plant varieties with various resistances to environmental stresses. While recent advances in plant genome research enable quick development of new plant genotypes, *plant phenotyping*, i.e. evaluation of the plant's structure, performance, and physiological and biochemical traits, is a slow and laborious process, which is considered as a bottleneck in the plant breeding cycle. Thus, there is great demand for fully automated high-throughput in-field phenotyping [16]. Importantly, many essential measurements can be extracted directly from the morphology of the plant. The introduction of new 3D sensing technologies and mobile agricultural robots opens up possibilities for in-field data collection and automation of the morphological analysis. Thus, in recent years there have been a number of scientific contributions towards capturing and automatically interpreting three-dimensional (3D) structural models

of plants. To extract relevant phenotypic measurements, such as leaf area and inclination angle, semantic segmentation of these representations into individual plant organs is needed. The existing algorithms for plant segmentation in 3D space rely heavily on controlled environments, elaborate calibration procedures, hand-picked features, and manually tuned thresholds. As a result, they do not generalise well to new or changing environmental circumstances or new species.

Instead, we propose to train a supervised deep learning algorithm to predict the point-wise segmentation directly from point cloud data. Deep neural networks are heavily data-driven and typically do not require much explicit prior knowledge about the task, other than a suitably large annotated data set. Besides reducing the need for manual tuning, we anticipate that using a supervised learning approach will also maximise the generalisation potential of the same algorithm for a wide range of crop species and environments, by adjusting the training data accordingly. In this paper we use the PointNet++ network architecture for semantic segmentation [14] to discriminate between three semantic categories (soil, leaves, and stems) based on structural data only. Due to the lack of publicly available labelled agricultural data sets, we train and test the network on a data set collected in simulation from synthesised plants. We evaluate the network's segmentation performance in simulation and provide an indicative qualitative validation of the trained network on a smaller selection of real-world depth data taken from the 4D Plant Registration Dataset [12].

The contributions of this work include (i) a novel fully annotated synthetic data set for 3D plant segmentation, (ii) application of the PointNet++ network architecture to the plant segmentation domain, and (iii) a quantitative and qualitative evaluation of semantic segmentation and generalisation performance.

2 Related Work

2.1 3D plant segmentation

Until recently plant segmentation has been focused on classical vision algorithms applied to 2D images. For the purpose of evaluating plant morphology, however, considerable information about 3D configurations and areas obscured by occlusion is lost in 2D projections. There are relatively fewer approaches for semantic segmentation of plants from 3D data.

Several approaches combine clustering techniques with a series of heuristic filters. In [13], a Euclidean clustering procedure is used along with colour and leaf shape heuristics. [18] suggests to apply mean-shift clustering on the depth information of RGB-D images, then classify candidate clusters as vegetation versus background by colour, followed by further instance segmentation based on an active contour model. In [2], individual leaves are segmented in top-down RGB-D images of single plants in a greenhouse environment via blob detection, such that vertically close points are assigned to the same image segments. [10] introduces an elaborate series of filters including removal of statistical outliers, removal of ground points by fixed distance and colour thresholds, followed by

3D equivalents of morphological erosion and subsequent expanding operations. This approach achieves impressive results for a number of greenhouse ornamental plant species, especially in dealing with a complicated occlusion. However, all listed algorithms rely on carefully tuned species-specific parameters and assumptions about plant height, orientation, or colour and lighting conditions. Such tailor-made approaches usually do not generalise well to different species or changes in the testing set-up. More recently, in [1, 12] binary segmentation into stems and leaves is achieved by using a support vector machine (SVM) classifier with feature vectors containing point coordinates and fast point feature histograms. The SVM classification is later refined by density-based clustering into coherent leave and stem areas, discarding of small clusters, and re-assigning points via k-nearest neighbour classification. This approach requires comparably little annotated training data or manual intervention, but addresses the simpler problem of binary segmentation on very high-resolution point clouds, which were recorded with a precision laser scanner from multiple views, resulting in minimal occlusion and no background interference.

In summary, state-of-the-art methods for semantic segmentation of plants rely on hand-crafted filters or controlled environments. While very effective in specific lab settings, their weaknesses lie in their poor generalisability and need for prior knowledge and manual tuning. Their assumptions do not hold in the field and are often violated due to natural variations in plant morphology.

2.2 Deep learning for 3D plant segmentation

Deep learning algorithms, especially convolutional neural networks (CNNs) are well-established as a standard approach to semantic segmentation for 2D images. CNNs take advantage of the ordered spatial pattern of images by performing convolution operations on the input and extract information about local structures in overlapping receptive fields at different scales. Naturally, there have been attempts to translate their success into 3D space. In [15], a multi-view approach is proposed, which allows the use of a CNN to perform semantic segmentation on 2D images. The 2D projections are then combined into a 3D point-cloud and semantic information from the different views is integrated via a voting strategy. Although much less reliant on manual tuning, this approach also suits lab environments best, as it requires exact camera parameters and positions to be known and assumes that the points are visible from all camera angles.

2.3 PointNet++

Recently, novel deep neural network architectures have been introduced, which are specifically designed to accommodate for the irregular structure of point cloud inputs, without the need for projection or discretisation. Point clouds are unordered sets of points in 3D space and frequently vary in size and point density. Unlike CNNs, the PointNet++ network architecture does not require regular input shapes and can be applied to point clouds, independently of its order or size. In a series of *set abstraction levels*, PointNet++ extracts local shape features

from nested subsections of a point cloud and repeatedly aggregates features into higher-level features. Through a corresponding set of *feature propagation levels*, the higher-level features are then interpolated and propagated back into smaller subsections and combined with local features. Eventually, each original input point is described by a feature vector that captures local and global information from all levels of abstraction and can be used for per-point semantic segmentation. This hierarchical architecture has since been applied successfully to tasks such as object detection in indoor scenes [5] and autonomous driving [11], typically focused on rigid, man-made objects and structured environments. There are very few applications to plants, not least because of a lack of training data, but it is also unclear how effectively the network can cope with naturally extreme variations in shape of non-rigid structures.

In [8], PointNet++ is used for a binary segmentation task of detecting strawberry fruit in RGB-D images taken in a real farm. Even though the well-studied 2D CNNs currently still outperform PointNet++, the authors report promising results and suggest further research into using PointNet++ in the agricultural domain, in order to make use of the unique shape information lost in 2D projections. In this work, we aim to apply PointNet++ to the 3D plant segmentation problem using end-to-end deep learning as a possible alternative to the rigid state-of-the-art procedures.

3 Methodology

3.1 Data set

To provide a suitably large annotated 3D data set for deep learning, we created an artificial set of point clouds captured in simulation from synthesised plant models. First, we defined a general model of a tomato plant, describing its branching structure and relative dimensions, using the random tree generating software Arbaro, an open-source implementation of an algorithm introduced by Weber and Penn [17]. Individual mesh objects were then procedurally generated by randomly varying all descriptors, such that the resulting plants are between 0.15 m and 0.35 m tall, and vary in the number, distribution, scale, and relative dimensions of their branches and leaves. In this way a total of 500 unique tomato plant models were produced. An indicative selection is shown in Fig. 1. Using the open-source 3D modelling software Blender [4], a simulated depth camera captured three depth images of each plant from different angles. A common configuration in agricultural robots features a sensor array aimed at the space below the robot [6]. To emulate the in-field deployment, we chose to capture one top-down view from 1.2 m height and two views at 20° and 40° viewing angles, as shown in Fig. 1.

The depth information was captured by ray-casting in a frustum shape covering a 40° square field of view at a resolution of 480×480 rays, yielding 1500 point clouds of 230400 labelled points each. Finally, the data set was shuffled and divided into a training set (1052 examples) and a validation set (224 examples)

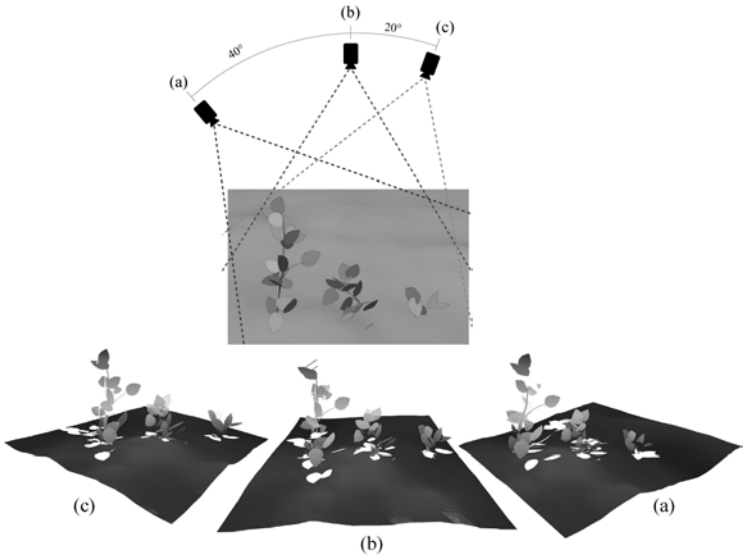

Fig. 1: Schematic diagram of the three partial views captured by simulated depth cameras on a rendered visualisation of three example plant meshes, along with their resulting point clouds.

for training purposes, and a test set (244 examples) which was used for final performance evaluations only.

3.2 Network architecture

We use the PointNet++ architecture adjusted for point-wise segmentation applications introduced by Qi et al. [14]. The network features four set abstraction levels $\mathrm{SA}(K, r, [l_1, ..., l_d])$ with K local regions of ball radius r, and using d fully connected layers of width $l_i (i = 1, ..., d)$ within the abstraction level, followed by four corresponding feature propagation levels $\mathrm{FP}(l_1, ..., l_d)$ with d fully connected layers. We also apply a random dropout with a ratio of 0.5 before the final fully connected layer. Following the original paper's notation conventions, the full parameters are: SA(1024,0.1,[32,32,64])→SA(256,0.2,[64,64,128]) →SA(64,0.4,[128,128,256]) →SA(16,0.8,[256,256,512]) →FP(256,256) →FP(256,256) →FP(256,128) →FP(128,128,128,128,K).

3.3 Performance metrics

In the following evaluation, we report two standard metrics for multi-class segmentation: Categorical Accuracy and mean Intersection over Union (mIoU). The categorical accuracy, however, is susceptible to distortions by imbalanced class sizes. Our synthetic data set is naturally highly imbalanced, containing larger regions of soil than plant matter. Thus, we monitored the network's learning

Table 1: Confusion matrix for all point-wise semantic class labels in the test set, along with the performance metrics for individual classes

		Predicted labels				
		Soil	Leaf	Stem	κ	IoU
True labels	Soil	50545423	3632	276	0.986	0.999
	Leaf	22758	979967	15497	0.966	0.935
	Stem	2759	26419	12869	0.364	0.223
						MIoU: 0.719

progress throughout training by the mIoU, which is the average across all four semantic classes' individual Intersection over Union (IoU) ratios, computed from the number of true positive (TP), false positive (FP), and false negative (FN) classifications as $IoU = \frac{TP}{TP+FP+FN}$. The mIoU places equal importance on all three semantic classes, and is therefore more appropriate for the data considered here. Finally, we also report Cohen's Kappa (κ) [3] for each semantic class, which also takes into account the scale of imbalance and the expected probability of random correct classifications for each class.

3.4 Network training

The network was trained on the designated training set for 120 epochs using the categorical cross-entropy loss function, Adam optimiser, and a learning rate of 0.0001. We selected the trained model after 82 epochs, at which point a rolling average of the validation mIoU across 10 epochs reaches its highest value, as the final model used for the remainder of this work.

4 Evaluation

In the following evaluation we report quantitative and qualitative assessments of the per-point semantic annotations produced by the chosen trained network on a test sample of synthetic data, a sample of a more complex growing scenario, and a real plant data sample.

4.1 Quantitative performance evaluation

On the remaining test sample (see Sec. 3.1), the network achieved an overall categorical accuracy of 0.999 and a mIoU of 0.712. The sources of error are further broken down in a confusion matrix in Table 1. The matrix displays counts for all point-wise classification cases across the test set. We also report individual IoU and κ metrics for each semantic class. This breakdown suggests that the network produces excellent segmentation results for the soil and leaf classes, and the main sources of error are confusions between leaves and stems. In particular, the network has a tendency to be too conservative in assigning stem labels. The stem class is assigned with high specificity, meaning that 0.999 points from other

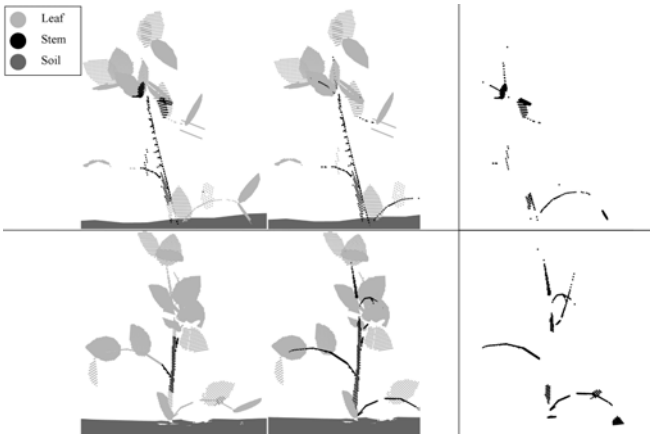

Fig. 2: The network's predicted segmentation (left), corresponding ground truth (centre), and highlighted differences (right) for two example test point clouds.

classes are correctly not labelled as stems, and low sensitivity, with only 0.306 true stem points correctly identified as stems. This failing may well be due to insufficient examples depicting the stem class being presented during training. To counteract the effect of class imbalance, we introduced sample weights to the loss function, such that higher importance was placed on examples of the under-represented classes during training. However, the weighted loss did not significantly influence the network's performance.

4.2 Qualitative performance evaluation

To contextualise the quantitative results, two segmentation outputs from the test set are pictured in Fig. 2, along with the original ground-truth annotation, and a point cloud visualisation highlighting only misclassified points. As expected, the network prediction appears very similar to the ground truth, and the most common segmentation errors occur where stems are confused with leaf regions. Especially thin stems towards the crown of the plant are often not spotted among the leaves. Occasionally, where only small proportions of a leaf are captured in the point cloud, or stems and leaves overlap closely due to perspective and occlusion, the leaf points are misclassified as stems. In summary, however, we can confirm visually that the segmentation of leaves against soil in particular is sensible, and in most cases the base of the stem can be located too.

Transfer to complex growing scenario. One major weakness of existing plant segmentation algorithms is their poor generalisability. Many procedures discussed in Section 2.1 are designed for laboratory settings and can not easily be applied to realistic in-field growing environments with multiple plants, dense foliage, and heavy occlusion patterns. To test how well the concepts our network

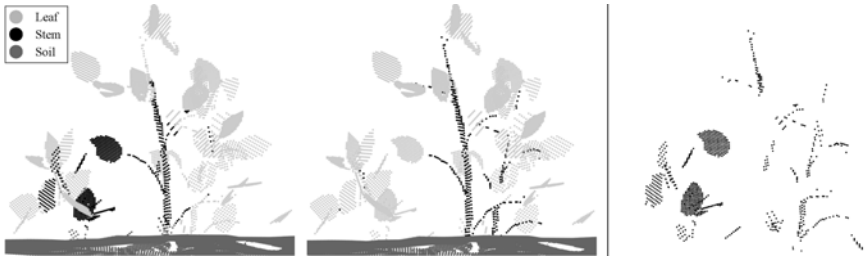

Fig. 3: The network's predicted segmentation (left), corresponding ground truth (centre), and highlighted differences (right) for a complex scene of 4 overlapping plants.

learned from single plants translate to more complex scenarios, we generated one example scene with four plants taken from the test set. The result is pictured in Fig 3. For this point cloud, the network produced slightly more misclassifications, but largely within the same pattern of errors we observed before. The network fails to locate especially thin stems but was able to locate the base and some broader stem regions. We can also observe a few more instances of leaves falsely classified as stems. On the whole, the segmentation works, even though the global shape of this scene was significantly different from any example point cloud presented during training. We conclude that the network did indeed learn to discriminate local shape characteristics and did not over-fit to regularities of the global shape of single plants presented in the training data.

Transfer to real-world data. Finally, we tested whether the concepts learned from synthetic data transfer to real-world plants. For a qualitative validation, we selected two scans of tomato plants and one scan of a maize plant from the 4D Plant Registration Dataset [12]. We down-sampled the high-resolution point clouds by ray-casting to reproduce the perspective and resolution of the simulated depth camera as described in Section 3.1. Fig. 4 shows the segmentation labels predicted by our network for the three resulting point clouds.

The network produces sensible segmentation masks for the two tomato plants, distinguishing well between regions representing soil and plant matter. This experiment serves to demonstrate that knowledge transfer about local shapes to real plants differing from the training data in their acquisition procedure, scale, and natural shape variations, is possible. Finally, the maize plant offers the additional challenge of knowledge transfer to a different species. While similar, the overall morphology and leaf shape differ significantly from our synthetic tomato plants. Still, large regions of all three semantic classes were detected successfully, however, the boundaries between the regions are less precise. Arguably, the semantic separation between stem and leave regions is more ambiguous in this species and previously unseen by the network. Considering these challenging factors, the achieved segmentation demonstrates that the network learned meaningful shape characteristics. It remains to be investigated how well the seg-

Fig. 4: The network's predicted segmentation for real-world depth data of two tomato plants (left and centre) and one Maize plant (right).

mentation algorithm could perform on real data when trained on real data too. The observed knowledge transfer between the synthetic and real-world data and also across species raises the possibility of pre-training in simulation, reducing the amount of fully annotated real-world training data needed.

5 Conclusions and future work

In this work, we successfully applied a supervised machine learning algorithm to the challenge of semantically segmenting plants based on structural data only. On the example of a fully annotated synthetic data set, we demonstrate that the PointNet++ neural network can successfully predict per-point semantic annotations for soil, leaves, and stems directly from point cloud data, and that the learned concepts transfer to new environments and real-world data. Our results serve as a proof of concept, supporting the idea that an understanding of the semantic sub-regions of plants can be learned from data, instead of relying on manually crafted pipelines of classical vision techniques, and that this approach carries potential for increased generalisability compared to current state-of-the-art algorithms. To judge the algorithm's applicability to fully automated in-field phenotyping, more experiments with real-world data are necessary. Our synthetic data set can also be improved by introducing realistic sensor noise and adding models of different crop species to further investigate knowledge transfer and possibilities to learn cross-species concepts for plant organs. To address the network's weakness in segmenting stems, we will trial higher sensor resolutions to improve the sampling density on fine structures and data pre-processing procedures, which augment the training data in such a way as to counter class imbalance without distorting its geometric information (e.g. [7]). Future work should also explore alternative network architectures. Instead of using PointNet++, we are interested in one promising architecture, presented in the context of in-field broccoli head detection [9]. The authors take advantage of the fact that, as a result of the data acquisition technique, point clouds produced by some RGB-D sensors provide an organised structure and allow for the use of a CNN without need for projection.

References

1. Chebrolu, N., Magistri, F., Läbe, T., Stachniss, C.: Registration of spatio-temporal point clouds of plants for phenotyping. PloS one **16**(2), e0247243 (2021)
2. Chéné, Y., Rousseau, D., Lucidarme, P., Bertheloot, J., Caffier, V., Morel, P., Belin, É., Chapeau-Blondeau, F.: On the use of depth camera for 3d phenotyping of entire plants. Computers and Electronics in Agriculture **82**, 122–127 (2012)
3. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
4. Community, B.O.: Blender - a 3D modelling and rendering package. Blender Foundation (2018), http://www.blender.org
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5828–5839 (2017)
6. Emmi, L., Gonzalez-De-Santos, P.: Mobile robotics in arable lands: Current state and future trends. 2017 European Conference on Mobile Robots, ECMR 2017 (2017). https://doi.org/10.1109/ECMR.2017.8098694
7. Griffiths, D., Boehm, J.: Weighted point cloud augmentation for neural network training data class-imbalance. arXiv preprint arXiv:1904.04094 (2019)
8. Le Louedec, J., Li, B., Cielniak, G., et al.: Evaluation of 3d vision systems for detection of small objects in agricultural environments. In: Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (2020)
9. Le Louedec, J., Montes, H.A., Duckett, T., Cielniak, G.: Segmentation and detection from organised 3d point clouds: A case study in broccoli head detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 64–65 (2020)
10. Li, Cao, Y., Shi, G., Cai, X., Chen, Y., Wang, S., Yan, S.: An overlapping-free leaf segmentation method for plant point clouds. IEEE Access **7**, 129054–129070 (2019)
11. Ma, X., Wang, Z., Li, H., Zhang, P., Ouyang, W., Fan, X.: Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6851–6860 (2019)
12. Magistri, F., Chebrolu, N., Stachniss, C.: Segmentation-based 4d registration of plants point clouds for phenotyping. IROS (2020)
13. Nguyen, T.T., Slaughter, D.C., Max, N., Maloof, J.N., Sinha, N.: Structured light-based 3d reconstruction system for plants. Sensors **15**(8), 18587–18612 (2015)
14. Qi, Yi, L., Su, H., Guibas: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in neural information processing systems. pp. 5099–5108 (2017)
15. Shi, W., van de Zedde, R., Jiang, H., Kootstra, G.: Plant-part segmentation using deep learning and multi-view vision. Biosystems Engineering **187**, 81–95 (2019)
16. Tardieu, F., Cabrera-Bosquet, L., Pridmore, T., Bennett, M.: Plant phenomics, from sensors to knowledge. Current Biology **27**(15), R770–R783 (2017)
17. Weber, J., Penn, J.: Creation and rendering of realistic trees. Proceedings of the 22nd annual conference on Computer graphics and interactive techniques - SIGGRAPH '95 (1995). https://doi.org/10.1145/218380.218427
18. Xia, C., Wang, L., Chung, B.K., Lee, J.M.: In situ 3d segmentation of individual plant leaves using a rgb-d camera for agricultural automation. Sensors **15**(8), 20463–20479 (2015)