StalkNet: A Deep Learning Pipeline for High-Throughput Measurement of Plant Stalk Count and Stalk Width

Harjatin Singh Baweja, Tanvir Parhar, Omeed Mirbod and Stephen Nuske

Abstract Recently, a body of computer vision research has studied the task of high-throughput plant phenotyping (measurement of plant attributes). The goal is to more rapidly and more accurately estimate plant properties as compared to conventional manual methods. In this work, we develop a method to measure two primary yield attributes of interest; stalk count and stalk width that are important for many broad-acre annual crops (sorghum, sugarcane, corn, maize for example). Prior work of using convolutional deep neural networks for plant analysis has either focused on object detection or dense image segmentation. In our work, we develop a novel pipeline that accurately extracts both detected object regions and dense semantic segmentation for extracting both stalk counts and stalk width. A ground-robot called the Robotanist is used to deploy a high-resolution stereo imager to capture dense image data of experimental plots of Sorghum plants. We ground-truth validate data extracted using two humans who assess the traits independently and we compare both accuracy and efficiency of human versus robotic measurements. Our method yields R-squared correlation of 0.88 for stalk count and a mean absolute error of 2.77 mm where average stalk width is 14.354 mm. Our approach is 30 times faster for stalk count and 270 times faster for stalk width measurement.

1 Introduction

With a growing population and increasing pressure on agricultural land to produce more per acre there is a desire to develop technologies that increase agricultural output [1]. Work in plant breeding and plant genomics has advanced many varieties of crop through crossing many varieties and selecting the highest performing. This process however has bottlenecks and limitations in terms of how many plant varieties can be accurately assessed in a given timeframe. In particular plant traits of interest; such as stalk count and stalk width are tedious and error prone when performed manually.

H. S. Baweja (✉) · T. Parhar · O. Mirbod · S. Nuske
The Robotics Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh 15213, USA
e-mail: harjatis@andrew.cmu.edu

M. Hutter and R. Siegwart (eds.), *Field and Service Robotics*, Springer Proceedings in Advanced Robotics 5, https://doi.org/10.1007/978-3-319-67361-5_18

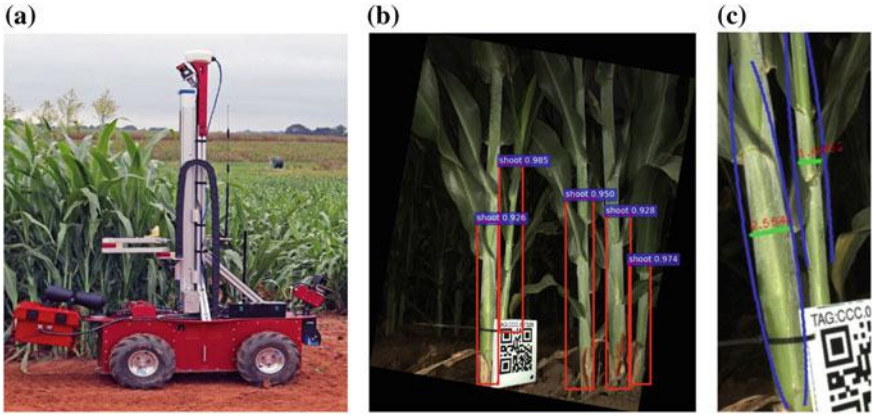

Fig. 1 **a** The Robotanist [2] mobile ground robot. Stereo imager mounted at back of vehicle (far right of image), **b** example image of stalk detection **c** example image of stalk width estimation

Robotics, computer vision and machine learning promise to expedite the process of assessing plant metrics and more precisely identify high performing varieties. This "high-throughput" plant phenotyping has the potential to extract measurements of plants 30 times faster than humans.

We are developing a ground robot called the Robotanist [2] that is capable of navigating tightly spaced rows of plants with cameras and sensors that can assess plant performance. The robot is equipped with machine vision stereo cameras with high-powered industrial flashes to produce quality imagery suitable for extracting robust and precise plant measurements. Figure 1a shows the camera setup where the camera is mounted at bottom right of the robot, b shows sample stalk detections stalk count and c shows stalk width measurement results respectively.

In this paper we present a new image processing pipeline the leverages both recent deep convolutional neural networks for object detection and also semantic segmentation networks together to output both stalk count and stalk width. The combination of networks together provides more precise and accurate extraction of stalk contours and therefore more reliable measurement of stalk width.

We collect data at two separate plant breeding sites one in South Carolina, USA and the other in Puerto Vallarta, Mexico and use one dataset for training our networks and the other dataset we ground truth using manually extracted measurements. We study both the efficacy and efficiency of manual measurements by using two humans to independently measure sets of plants and comparing accuracy and also total time taken per plot to measure attributes. We then compare the human measurements to the robot measurements for validation.

2 Related Work

Machine learning researchers have been making efforts to oust traditional plant phenotyping techniques by adapting neoteric algorithms to work on ground data. A number of such attempts have produced promising results. Singh et al. [3] provide a comprehensive study on how various contemporary ML algorithms can be used as building blocks for high throughput stress phenotyping. Drawing inspiration from traditionally used visual cues to estimate plant health, crop yield etc., Computer Vision has been the mainstay of most Artificial Intelligence based phenotyping initiatives [4]. Our group amongst several research groups is investigating the applications of computer vision to push the boundaries of crop yield estimation [5, 6] and phenotyping. Jimenez et al. [7] provide an overview of many such studies.

Recent advances in deep leaning have induced a paradigm shift in several areas of Machine Learning, especially Computer Vision. With deep learning architectures producing state-of-the-art performances in almost all major computer vision tasks such as image recognition [8], object detection [9] and semantic segmentation [10]; it was only a matter of time for researchers to use these for image based plant phenotyping. A broad variety of studies ranging from yield estimation [11, 12], to plant classification [13], to plant disease detection have been conducted [14].

Pound et al. [15] propose using vanilla CNNs (Convolutional Neural Networks) to detect plant features such as root tip, leaf base, leaf tip etc. Though the results achieved are impressive, the images used are taken in indoor environments thus escaping the challenges of field environments such as occlusion, varying lighting amongst several others. Considering work on phenotype metric of interest- stalks, traditional image processing based approaches tailored for a particular task do well on specific data set but fail to generalize [16, 17]. 3D reconstruction based approaches show promising results, however are almost impossible to reproduce in cluttered field environments [18]. Bargoti et al. [19] provide a pipeline for trunk detection in apple orchards. The pipeline uses Hough Transforms on LIDAR data to initialize pixel wise dense segmentation into a hidden-semi Markov model (HSMM). Hough transform proves to be a coarse initializer for intertwined growth environments with no apparent gaps between adjacent plants.

There has not been much work if any, combining deep learning based state-of-the-art object detection and semantic segmentation for high throughput plant phenotyping. Our work uniquely combines the linchpins in object detection (Faster-RCNN) and semantic segmentation (FCN) for plant phenotyping to get an accuracy that is close to that of human at staggeringly high speeds.

3 Overview of Processing Pipeline

The motivation behind the work was to come up with a high throughput plant phenotyping computer vision based approach that is agnostic to changes in the field

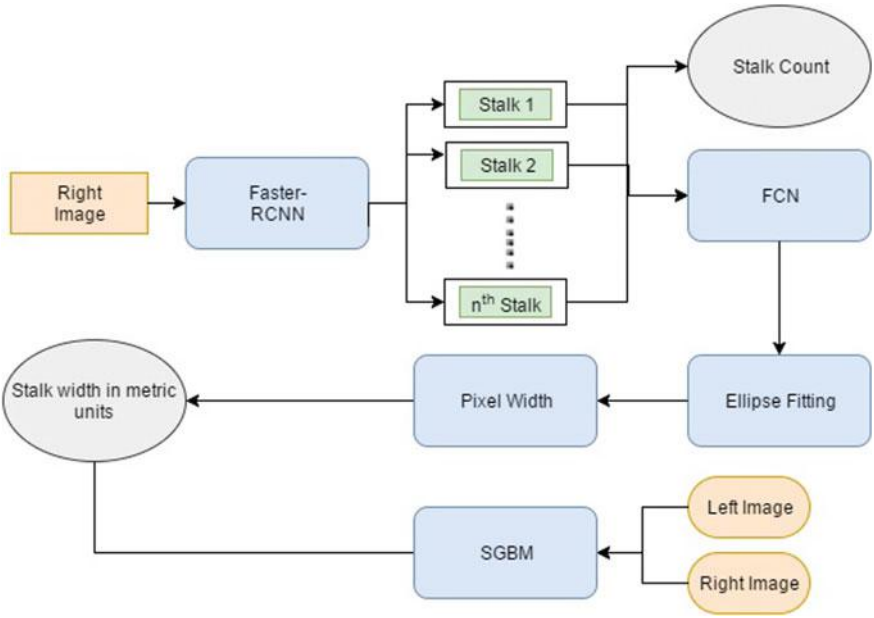

Fig. 2 Overview of the stalk count and width calculation pipeline

conditions and settings such as varying lighting conditions, occlusions etc. Figure 2 shows the overview of the data-processing pipeline, used by our approach. The faster RCNN takes one of the stereo pair images as its input and produces bounding boxes, each for one stalk. These bounding boxes are extracted from the input image (also called as snips) and fed to the FCN, one at a time. The FCN outputs a binary mask, classifying each pixel as either belonging to stalk or the background. To this mask ellipse are fitted, to the blobs in the binary mask, by minimizing the least-square loss of the pixels in the blob [20]. One snip may have multiple ellipses in case of multiple blobs. The ellipse with the largest minor axis is used for width calculation. The minor axis of this ellipse gives us the pixel width of the shoot in the current snip. The corresponding pixels in the disparity map are used to convert this pixel width into metric units.

The whole pipeline takes on an average 0.43 s to process one image, on a GTX 970 GPU. This can make the data-processing on the fly for systems that collect data at 2 Hz.

3.1 SGBM

The stereo pair was used to generate to a disparity map, using SGBM [21] in OpenCV. This was used to get metric measurements from the pixel dimension. It was also used

to calculate the average distance of the plant canopy from the sensor, it was converted into field of view in metric units, so that the estimated stalk count and stalk diameter can be converted into estimated stalk count per meter and stalk diameter per meter respectively.

3.2 Faster RCNN

Fast-RCNN (Fig. 3) by Girshick uses a VGG-16 convnet architecture as feature detector. The network takes pre-computed proposals from images and classifies them into object categories and regresses a box around them. Because the proposals are not computed over the GPU, there is a bottleneck at computing proposals. Faster-RCNN by Girshick et al. is an improvement over the Fast-RCNN, where there is a separate convolution layer that predict object proposals based on the features form the activation of the last layer of the VGG-16 network, called Region Proposal network (RPN). Since the region proposal network is a convolution layer, followed by fully connected layers, it is implemented over GPU, making it almost an order of magnitude faster than Fast-RCNN.

One drawback of the Faster RCNN is the use of non-maximal suppression (NMS) over the proposed bounding boxes. Thus, highly overlapping instances of objects might not be detected, due to NMS rejection. This problem is even severe in highly occluding field images. It was overcome by simply rotating the images by 90° so that the erectness of the stalks may be used to draw tightest possible bounding boxes. We finetuned a pre-trained Faster-RCNN with 2000 bounding boxes. Figure 4 shows sample detections.

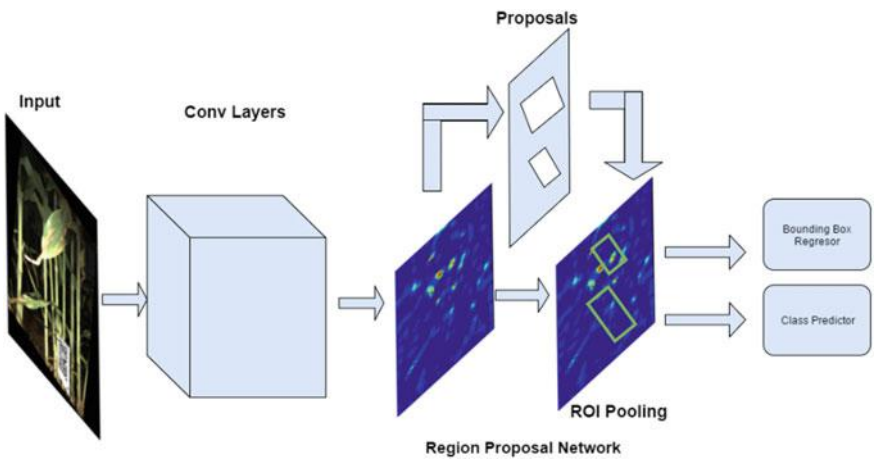

Fig. 3 Faster-RCNN architecture used for stalk detection

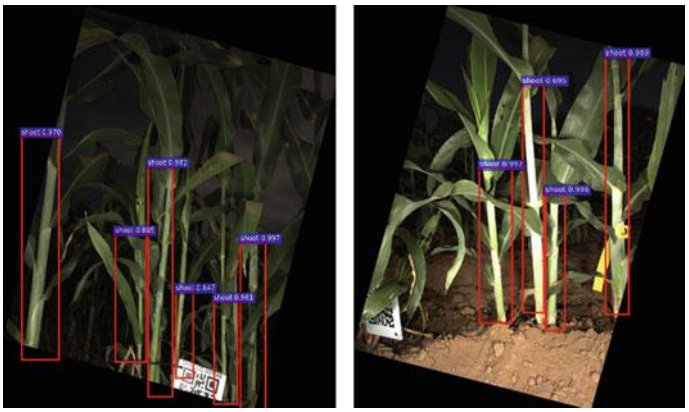

Fig. 4 Example stalk detections using Faster-RCNN

3.3 Fully Convolutional Network (FCN)

FCN (Fig. 5) is a CNN based end to end architecture that uses down-sampling (Convolutional Network) followed by up-sampling (Deconvolutional Network) to take image as input and produce a semantic mask as output.

The snips of stalks detected by Faster-RCNN are sent to FCN for semantic segmentation which by virtue of its fully convolutional architecture can account for different sized incoming image snips. We chose to send Faster-RCNN's output to FCN as input instead of raw image. Output bounding boxes always contain only one stalk and thus FCN is only required to do a binary classification into two classes, namely: stalk and background without having to do instance segmentation also. Our hypothesis was that this would make FCNs job a lot easier and would thus require lesser data to finetune a pretrained version. This hypothesis is validated by results presented in a latter section. We finetuned a pre-trained FCN with just 100 dense

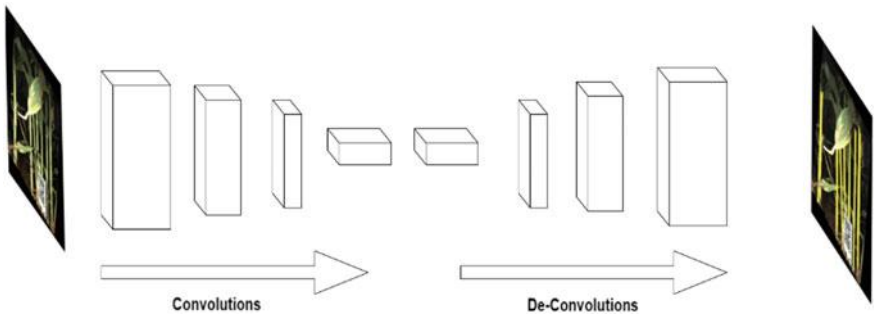

Fig. 5 Fully convolutional network architecture used for dense stalk segmentation

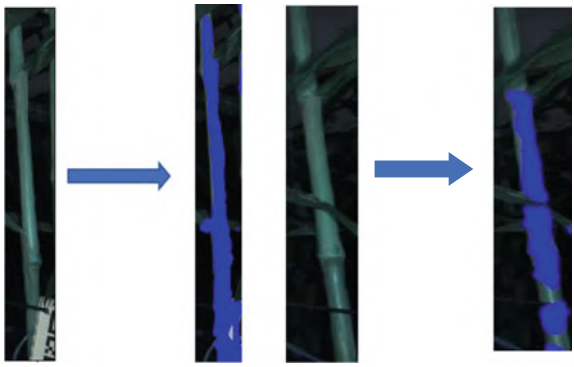

Fig. 6 Sample snipped bounding box input to segmented stalk output

labeled detected stalk outputs from Faster-RCNN. Sample input to output of FCN is shown in Fig. 6.

3.4 Stalk Width Estimation

Once the masks have been obtained, for each of the snippets, ellipses are fitted to each blob of the connected contours of the mask. The ellipses are fitted to minimize the following objective:

$$\varepsilon^2(\theta) = \sum_{i=1}^{n} F(\theta_i, x_i)^2$$

where, $F(\boldsymbol{\theta}; \boldsymbol{x}) = \theta_{xx}x^2 + \theta_{yy}y^2 + \theta_{xy}xy + \theta_x x + \theta_y y + \theta_0$, is the general equation of conics in 2 dimensions. The objective is to find the optimal value of $\boldsymbol{\theta}$ such that we get the best fitting conic over a given set of points. We use OpenCV's inbuilt optimizers to find best fitting ellipses. Figure 7 shows the ellipses fitted to the output mask of the FCN.

Ellipse is fitted to the contours of the blob, so that the minor axis can serve as a starting point for width estimation of the stalk. For the same reason, a simple convex hull fitting was not performed. The minor axes of all the ellipses are then trimmed to

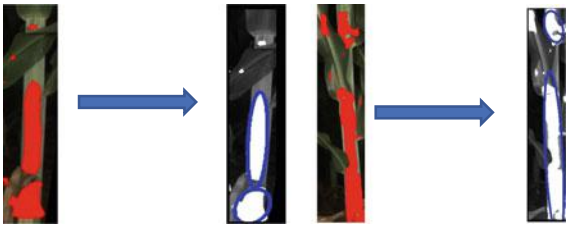

Fig. 7 Result of ellipse fitting on mask output of FCN used for estimating stalk width

Fig. 8 Stalk width estimation pipeline

make sure they lie over the FCN mask. From the trimmed line segments, any segment that night have a slope of greater than 30° is rejected. The remaining line segments are projected on to the disparity map, so that the pixel width can be converted to the width in metric units, as per Algorithm 1. The line segment with the greatest metric width is selected as the width for the stalk in the current snip. The reason behind choosing max width over others is to get rid of the segments proposals that might have leaf occlusions.

Algorithm 1: Stalk width calculation in metric units

- *For each line segment l*
  - *For each end points (x, y) and (x', y') on l*
    - $d = Disparity(x, y)$
    - $d' = Disparity(x', y')$
    - $Z - Z' = (f * b) * \left(\frac{1}{d} - \frac{1}{d'}\right)$
    - $X - X' = (x - x') * Z/f$
    - $Y - Y' = (y - y') * Z/f$
    - width = $\sqrt{(X - X')^2 + (Y - Y')^2 + (Z - Z')^2}$

Hence the overall procedure for width estimation, Fig. 8, can be summarized in the following steps

Algorithm 2: Steps for stalk width estimation

- $widths \coloneqq empty\ list$
- *For all fitted ellipses*
  - *If* $-30^o < minor\ axis\ slope < 30^o$
    - *trim minor axis to fit mask*
    - *Find the metric length of the line segment as per algorithm 1.*
    - *Append the metric length to widths list*
- *Width = max(widths)*

Figure 9 shows the final results for stalk width estimation pipeline.

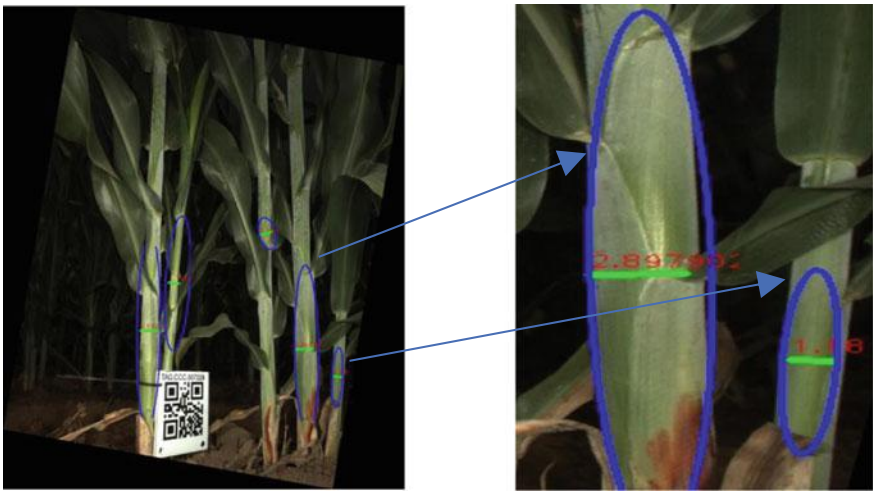

Fig. 9 Stalk width estimation results

4 Results

4.1 Data Collection

Image data was collected in July 2016, in Pendleton, South Carolina using the Robotanist platform. The algorithms were developed on this data. To test the algorithm impartially, another round of data collection with extensive ground truthing was done in February, 2017 in Cruz Farm, Mexico. The images were collected using a 9 MP stereo-camera pair with 8 mm focal length, high power flashes triggered at 3 Hz by Robot Operating System (ROS). The sensor was driven at approximately 0.05 m/s. Distance of approximately 0.8 m was maintained from the plant growth. Figure 10a the Robotanist collecting data in Pendleton, South Carolina, Fig. 10b shows the custom image sensor mounted on the robot.

Each row of plant growth at Cruz Farm is divided into several 7 ft ranges separated by 5 ft alleys. To ground truth stalk count data, all stalks were counted in 29 ranges by two individuals separately. The mean of these counts was taken as the actual ground truth. Similarly, for width calculations, QR tags were attached to randomly chosen stalks for ground truth registration in images. The width of these stalks at height of 12 in. (30.48 cm) and 24 in. (60.96 cm) from the ground was also measured by two individuals separately using Vernier Calipers of 0.01 mm precision. Humans at an average took 210 s to count the stalks in each range and an average of 55 s to measure width of each stalk. Each range at an average has 33 stalks, so on an average it takes 33 min to measure stalk widths of entire range.

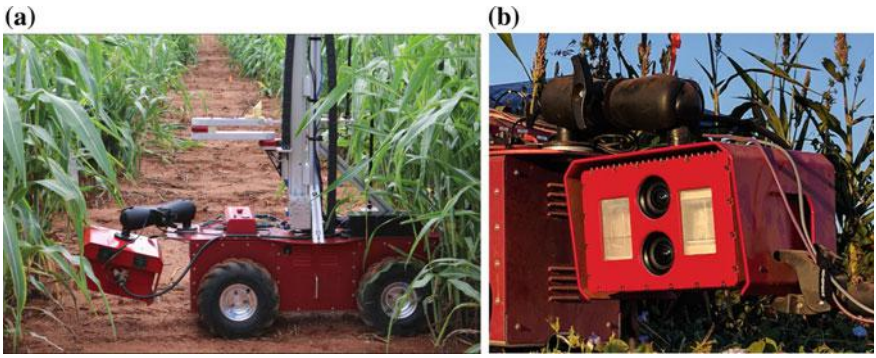

Fig. 10 **a** The Robotanist collecting data in Pendleton, South Carolina **b** Custom stereo imaging sensor used for imaging stalks

4.2 Results for Stalk Count

Faster-RCNN was trained with 400 images with approximately 2000 bounding boxes using alternate optimization strategy. RPN and regressor are trained for 80000 and 40000 iterations respectively in first stage and 40000 and 20000 iterations in the second stage using base learning rate of 0.001 and a step decay of 0.1 every 60000 iterations for RPN and 30000 iterations for regressor. Best test accuracies were achieved by increasing the number of proposals to 2000 and the number of anchor boxes to 21 using different scaling ratios with NMS threshold of 0.2. Due to inability to get accurate homography for data collected at 3 Hz, we resorted to calculating stalk-count/meter using stereo data and do the same for ground truth stalk counts which were collected from ranges, each of constant length 7 ft (2.134 m). Figure 11 shows the R-squared correlation for results of 0.88.

To put the results into perspective, attempting to normalize counts using image widths from stereo data may induce some error as this data is sometimes biased towards stalk count towards the start and end of each range where the mount vehicle is slowed down. Also, there is a little inherent uncertainty in the count data. There are tillers (stems produced by grass plant) growing at the side of some stalks which are hard to discern from stalks with stunted growth. To better understand this, we observe in Fig. 12 that there is a small variation in ground truth stalk counting between two humans as well. The R-squared of Human1's count versus Human2's count should be 1 in an ideal scenario but that is not the case.

4.3 Results for Stalk Width

We plot the stalk width values as measured by Human1, Human2 and our algorithm at approximately 12 in. (30.48 cm) from the ground. At the time of data collection,

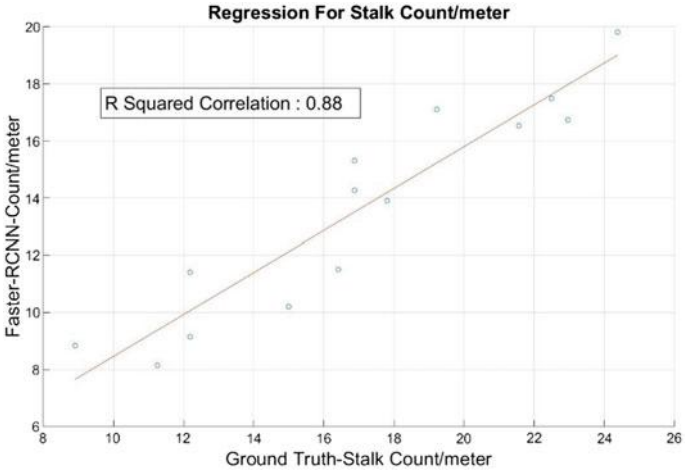

Fig. 11 Linear regression for human stalk count/meter versus Faster-RCNN's count/meter

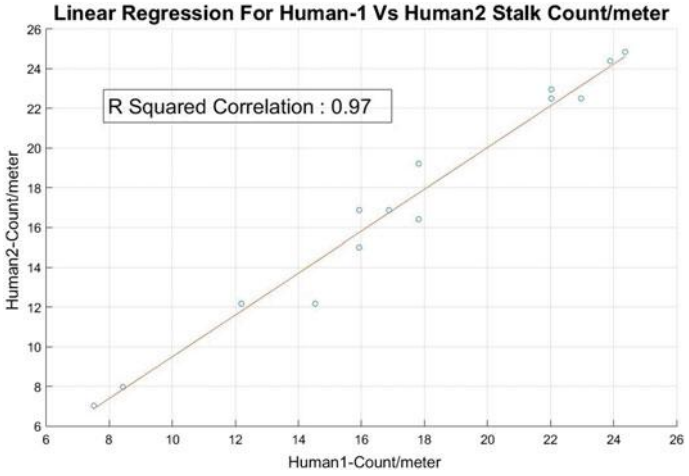

Fig. 12 Linear regression for Human1's stalk count/meter versus Human2's stalk count/meter

it was made sure that a part of ground is visible in every image. This allows us get stalk width at the desired height from the image data. This step is important as there is prevalent tapering in stalk widths as we go higher up from the ground. Figure 13 shows the widths of each ground trothed stalk as per both humans and algorithm.

Since there is a discernible difference in measurements of the two humans, we considered the mean of the two readings as actual ground truth. The mean width of stalks as per this ground truth is 14.354 mm. The mean absolute error between readings of Human1 and Human2 is 1.639 mm and the mean absolute error between readings from human ground truth and algorithm is 2.76 mm. The error can be attributed to

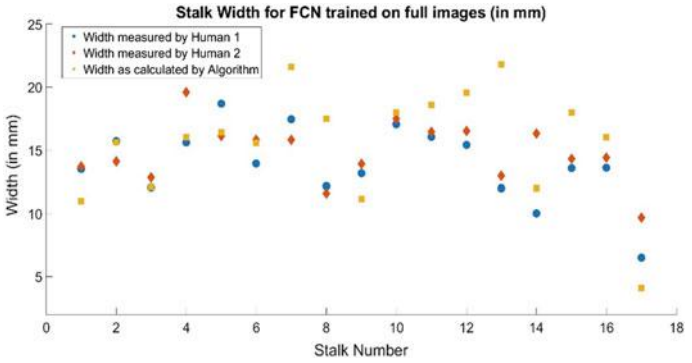

Fig. 13 Width measured by Human1, Human2 and algorithm

rare occlusions that force algorithm to calculate height at a location other than 12 in. (30.48 cm) from the ground. We suspected this as a possibility and thus measure stalk widths at 2 locations during the ground truthing process: at 12 in. (30.48 cm) and 24 in. (60.96 cm) above the ground. Calculations from this data tell us that there was 0.405 mm/in. mean tapering on the measured stalks as we went up from 12 in. (30.48 cm) to 24 in. (60.96 cm).

To validate our hypothesis that providing faster-RCNN's output bounding boxes as inputs to FCN would require lesser dense labeled data to train it. We trained another FCN on densely labeled complete images. This FCN was trained with more than twice the number of densely labeled stalk data (finetuned with approximately 250 densely labeled stalks) than the previous FCN (finetuned with on approximately 100 densely labeled stalks). Even after assuming perfect bounding boxes around it for instance segmentation, the mean absolute error of this FCN was 3.868 mm for width calculation, which is higher than its predecessor having a mean absolute error of 2.76 mm.

4.4 Time Analysis

Table 1 shows the time comparisons of Humans versus algorithm for an average plot. Each plot has approximately 33 stalks and is about 2.133 m in length. We observe that Algorithm is 30 times faster as compared to humans for stalk counting and 270 times faster than human for stalk width calculation.

Table 1 Time analysis for measuring one experimental plot

	Human1 (min)	Human2 (min)	Robot (s)
Stalk count	3.33	3.66	6.5
Stalk width	29	30	
Total	32.33	33.66	6.5

5 Conclusion

We have shown the strength of coupling deep convolutional neural networks together to achieve a high quality pipeline for both object detection and semantic segmentation. With our novel pipeline we have demonstrated accurate measurement of multiple plant attributes.

We find the automated measurements are accurate to within 10% of human validation measurements for stalk count and measure stalk width with 2.76 mm on average. Ultimately though, we identify that the human measurements are 30 times slower than the robotic measurements for count and 270 times slower for measuring stalk width over an experimental plot. Moreover, when translating the work to large scale deployments, that instead of 30 experimental plots are 100's or 1000's of plots in size, it is expected that the human measurements become less accurate and logistically tough to measure in timely fashion during tight growth stage time windows.

In future work we plan to integrate more accurate positioning to merge multiple views of the stalks into more accurate measurements of stalk-count and stalk-width.

References

1. United Nations Department of Economic and Social Affairs Population Division.: http://www.unpopulation.org. Accessed 10 Oct 2014
2. Mueller-Sim, T., et al.: The Robotanist: a ground-based agricultural robot for high-throughput crop phenotyping. In: IEEE International Conference on Robotics and Automation (ICRA), Singapore, Singapore, May 29–June 3 2017
3. Singh, A., et al.: Machine learning for high-throughput stress phenotyping in plants. Trends Plant Sci. **21**(2), 110–124 (2016)
4. Tsaftaris, S.A., Minervini, M., Scharr, H.: Machine learning for plant phenotyping needs image processing. Trends Plant Sci. **21**(12), 989–991 (2016)
5. Sugiura, R., et al.: Field phenotyping system for the assessment of potato late blight resistance using RGB imagery from an unmanned aerial vehicle. Biosyst. Eng. **148**, 1–10 (2016)
6. Pothen, Z., Nuske, S.: Automated assessment and mapping of grape quality through image-based color analysis. IFAC-PapersOnLine **49**(16), 72–78 (2016)
7. Jimenez, A.R., Ceres, R., Pons, J.L.: A survey of computer vision methods for locating fruit on trees. Trans. ASAE-Am. Soc. Agric. Eng. **43**(6), 1911–1920 (2000)
8. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
9. Ren, S., et al.: Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (2015)

10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
11. Sa, I., et al.: On visual detection of highly-occluded objects for harvesting automation in horticulture. In: ICRA (2015)
12. Hung, C., et al.: Orchard fruit segmentation using multi-spectral feature learning. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE (2013)
13. McCool, C., Ge, Z., Corke, P.: Feature learning via mixtures of dcnns for finegrained plant classification. In: Working Notes of CLEF 2016 Conference (2016)
14. Mohanty, S.P., Hughes, D.P., Salathé, M.: Using deep learning for image-based plant disease detection. Front. Plant Sci. **7** (2016)
15. Pound, M.P., et al.: Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. bioRxiv, 053033 (2016)
16. Mohammed Amean, Z., et al.: Automatic plant branch segmentation and classification using vesselness measure. In: Proceedings of the Australasian Conference on Robotics and Automation (ACRA 2013). Australasian Robotics and Automation Association (2013)
17. Baweja, H., Parhar, T., Nuske, S.: Early-season vineyard shoot and leaf estimation using computer vision techniques. ASABE (2017) (accepted)
18. Paproki, A., et al.: Automated 3D segmentation and analysis of cotton plants. In: 2011 International Conference on Digital Image Computing Techniques and Applications (DICTA). IEEE (2011)
19. Bargoti, S., et al.: A pipeline for trunk detection in trellis structured apple orchards. J. Field Robot. **32**(8), 1075–1094 (2015)
20. Fitzgibbon, A.W., Fisher, R.B.: A buyer's guide to conic fitting. DAI Research Paper (1996)
21. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 328–341 (2008)