Length Phenotyping with Interest Point Detection

Adar Vit Guy Shani Aharon Bar-Hillel
Ben Gurion University
Beer-Sheva, Israel
adarv@post.bgu.ac.il, shanigu,barhille@bgu.ac.il

Abstract

Plant phenotyping is the task of measuring plant attributes. We term 'length phenotyping' the task of measuring the length of a part of interest of a plant. The recent rise of low cost RGB-D sensors, and accurate deep networks, provides new opportunities for length phenotyping. In this paper we present a general technique for measuring length, based on three stages: object detection, point of interest identification, and a 3D measurement phase. We address object detection and interest point identification by training network models for each task, and use robust de-projection for the 3D measurement stage. We apply our method to two real world tasks: measuring the height of a banana tree, and measuring the length, width, and aspect ratio of banana leaves in potted plants. Our results indicate satisfactory measurement accuracy, with less than 10% *deviation in all measurements. The two tasks were solved using the same pipeline with minor adaptations, indicating the general potential of the method.*

1. Introduction

In plant phenotyping one measures and assesses complex plant traits related to growth, yield, and other significant agricultural properties [5]. As manual phenotyping is extremely costly, there is a need to develop automated analysis algorithms that are accurate and robust, which are able to measure phenotypic traits in field conditions on real crops [21]. Automated algorithms are required for accelerating cycles of genetic engineering [29], and for automating agriculture processes [5]. Field and greenhouse phenotyping is a difficult challenge, as field conditions are notoriously heterogeneous, and the inability to control environmental factors, such as illumination and occlusion, makes results difficult to interpret [2]. Image analysis algorithms are crucial for advancing large scale and accurate plant phenotyping [18]. A second recent advancement is the abundance of low-cost sensors, from RGB to depth and thermal sensors, useful for capturing plant traits. For example, depth sensors can capture the plant shape in three dimensions, containing useful information about its developmental stage[24].

Figure 1. **Length based phenotyping. Left:** A banana tree with two interest points: basal (red) and upper (blue). The distance between them is the tree height. **Right:** A banana leaf with 4 interest points: basal (red), apex (blue), left (purple) and right (yellow). Only measurable objects are annotated.The line between the two former points is the leaf center line, and its length is the leaf length. The distance between the latter points is the width. Note that the position of the latter points (left and right) is somewhat ambiguous in the direction of the leaf center line.

This paper focuses on the problem of measuring 3D physical lengths of plant parts in field conditions, using a low cost RGBD sensor and a deep network architecture. Measuring the size of plant's parts can provide important cues about the plant state[32] and expected utility. For example, measuring the aspect ratio (the ratio between the length and the width) of young banana leaves in a potted plant prior to planting in the plantation, can determine whether the plant has undergone a mutation that results in undesirable fruits. Specifically, such mutations are characterized by a ratio smaller than 1.8, while normal plant typically have a ratio approaching 2.2 [8]. Another example is estimating the height of a banana tree. It is important since one goal of variety developers is to lower the banana tree height, enabling easier tree treatment for farmers. An example from a different crop is cucumber length measurement. The histogram of cucumber fruit lengths in a given plot provides strong indication regarding the cucumber growth

rate and expected quality [14]. We term such tasks 'length phenotyping'. Given the abundance and repetitive nature of length phenotyping tasks, it is desirable to develop a generic process that can be applied, with minimal adjustments to measuring lengths of different plant parts in various plants.

We focus here on two problems related to banana crops, considered to be the most widely consumed fruit [12]. First, we estimate the height of a banana tree (actually, a banana plant is not a tree, but we use the term tree here for simplicity). This height is determined by measuring the distance between the tree's basal point, and tree's upper interest point, defined by experts as the highest point of the tree's peduncle (top of the arch). These points can be seen in Figure 1 on the left. Second, we estimate the length, width, and aspect ratio of banana leaves. The leaf length is defined as the distance between the leaf's apex and the leaf's connection point with the petiole. The width is the length of the longest line perpendicular to the leaf center line, connecting the two most distant points on the left and right side of the leaf, as can be seen in Figure 1 on the right. The left and right key points are defined with respect to the vector oriented from the leaf basal point to the apex point. While we specifically focus on these two problems, we believe that the algorithmic pipeline we propose is sufficiently general for length measurement tasks, as its components are for the most part not tailored to the specific problems addressed.

Our proposed algorithmic pipeline consists of three stages: (1) detecting the objects of interest, (2) identifying interest points on the detected objects, and (3) de-projection of the interest points to world coordinates to compute distances. We use separate stages of detection and interest point identification rather than a direct identification of the interest points due to two main reasons. First, the interest points are often not visually distinctive on their own, i.e. they do not have a sufficiently unique appearance which will enable their detection without the object context. For example, the left and right interest points of a leaf are only locally characterized by a curved edge, a structure which is abundant in many irrelevant plant parts in the image. Direct detection of such points would hence lead to a proliferation of false positives. Once the leaf is detected, these points can be identified in well defined locations with respect to the leaf, enabling robust finding and accurate localization. A second reason is the need for correspondence determination between pairs of points (and in the leaf case, quadruplets) of the same object, rather than points in different objects, for computing distances between corresponding points.

For the first two stages we use Convolutional Neural Networks (CNNs) trained and applied on RGB images. As object detection in RGB images is well studied [31], we use task transfer from well trained RGB backbones. In the third stage, to perform 3D measurements, we use an RGB-D sensor, providing a depth channel in addition to the RGB measurements. Depth information enables inference of the world coordinates for given image points, using de-projection algorithms [7]. Specifically, we use an Intel D435 (Intel Corporation, Santa Clara, CA, USA) sensor for image acquisition. This camera is based on infra-red active stereoscopy technology, where the target is observed from two different viewpoints and a triangulation method is used to estimate the depth. It has a global shutter sensor, enabling good performance in highly dynamic conditions that exist in fields [7]. It was shown that the Intel D435 sensor perform well in outdoor conditions for phenotyping tasks [28].

For object detection we use a two-staged detector, based on the Faster R-CNN and Mask R-CNN [26, 10] architectures. The first stage in the detector is a Region Proposal Network (RPN), whose goal is to provide image regions which are good candidates for containing objects of interest. The second stage network classifies the object type (or reject it) and refines its bounding box. For the interest point identification architecture we use the interest point subnet proposed in Mask-RCNN [10], with modified loss to cope with the higher difficulty in our specific problems. For the last stage, we develop robust de-projection to cope with the challenging environment, and depth camera instability.

We collected three data sets for each problem. The first is used for network training over the RGB channels information only. The other two data sets are used for testing. Test set A contains RGB images with 2D marked ground truth. Test set B consists of RGB-D images with corresponding 3D ground truth measurements collected manually using a ruler. The accuracy of the 3D point detection (including de-projection) and final 3D measurements is reported on test set B. The accuracy of object detection and interest point identification stages is reported on the union of A and B. The average deviation for tree height estimations is 9.24 centimeters, less than 5% of the true height. For the more challenging leaf aspect ratio task, our average deviations are 2.01 and 0.96 for leaf length and width, respectively, which are less then 10% of the true measurements.

The main contributions of this work are threefold. First, we present a general method for measuring plant part lengths under challenging conditions for agricultural phenotyping. Second, we demonstrate the success of the method on two real world tasks of banana tree height estimation and banana leaf measurements. Third, we provide detailed analysis of the deviations and errors in the various stages of the algorithmic pipeline, supplying important cues regarding further improvement opportunities.

2. Related Work

Plant Phenotyping using Imaging Techniques: Traditional phenotyping is based on extensive human labor, where only a few samples are collected for thorough visual or destructive inspection. These methods are time consum-

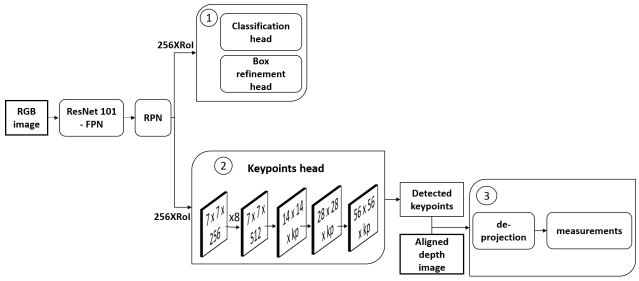

Figure 2. **High level view of our processing pipeline.** Left: the backbone network, extracting a reach feature map in 5 different octaves using the FPN. Middle: Extracted representations of 256 RoIs are processed by classification and box refinement branches (1) and the interest point finder (2), whose structure is detailed. kp denotes the object's number of keypoints. Right: 2D locations of the interest points are extracted from the output heatmap of (2). They are deprojected into world coordinates, and distances among them provide the length measurements (3).

ing, subjective and prone to human error, leading to what has been termed as 'the phenotypic bottleneck'[6]. Computer vision in agriculture has been studied intensively for a few decades, with a primary goal of enabling large scale, automatic visual phenotyping of many phenotyping tasks. Li et al. [18] provided in their work extensive survey of imaging methodologies and their applications in plant phenotyping. Since recently deep learning techniques have emerged as the primary tool in computer vision, they have a growing impact on agricultural applications [17].

RGB-D sensors in phenotyping tasks: In recent years, the use of RGB-D sensors has been expanding due to their increasing reliability and decreasing costs. Well registered depth and color data from such sensors provide a colored 3D point cloud [9] structure, which is useful in many applications. For example, Chéné et al. [4] showed that depth information can resolve individual leaves, allowing automated measurement of leaf orientation in indoor environments. Vit et al. [28] recently compared several depth sensors for a plant phenotyping task, in field condition and in various illumination conditions. It was shown that Intel D435 outperforms the other competitors. In [30] the size of mango fruits in field conditions was estimated. Jiang et al. [15] presented an algorithm for accurately quantifying cotton canopy size in field conditions. They showed that the multi-dimensional traits and multivariate traits were better yield predictors than traditional univariate traits, confirming the advantage of using 3D imaging modalities. Miella et al [20] proposed an in-field high throughput grapevine phenotyping platform for canopy volume estimation and grape bunch detection, using a RealSense R200 depth camera.

Measuring plant height and leaf size: In [22] the area, perimeter, length, and thickness of bananas were measured using a combination of computer vision techniques on gray-scale images. In [3, 16] algorithms for estimation of sorghum height were suggested. Estimation was done in field conditions from autonomously captured stereo images. The methods are based on classic techniques, with Hough transform used for detection and tracking applied to obtain robust measurements. An et al. [1] presented a technique for measuring rosette leaf length by detecting the leaf center and tips in a leaf-segmented binary image. The center was estimated as the centroid of all white (leaf) pixels. Leaf tips were detected as peaks of the rosette-outline curvature. They didn't use depth information in their method.

keypoints detection: Keypoints detection is used extensively in human pose estimation and face recognition. In [10] human joints keypoints are detected by predicting a one-hot spatial mask for each key point type. In [23] another architecture for human joint detection is suggested, based on progressive pooling followed by progressive upsampling. In [25] a CNN is used for localizing face landmarks. Similar to our work, landmark detection is preceded by face detection. In contrast for our work, their proposed architecture explicitly infer the visibility of key points, to account for points invisible at test time. Since our goal is to perform 3D measurements, we have no interest in invisible key points, which do not enable such measurements. Instead, we define as 'measurable object' only objects in which all the relevant key points are visible, and train our detector to only detect measurable object instances.

keypoints detection have not been extensively used in plant phenotyping [27]. Hu et al. [11] developed an algorithm for measuring three size indicators of banana fruit, namely length, ventral straight length, and arc height without using 3D information. They locate five points at the edge of banana and calculated euclidean distances between point pairs for determining these indicators. [13] solves a leaf counting task with an intermediate stage of interest point finding. A model for finding leaf centers is trained with keypoints treated as Gaussian heat map similarly to our work. The heatmap is then used for leaf count regression.

3. Architecture and Algorithms

3.1. Mask R-CNN

The first two stages in our proposed pipeline are based on the Mask R-CNN architecture [10], with some adaptations made for our goal. Mask R-CNN extended a previous architecture termed Faster R-CNN [26] by adding network modules for object segmentation or alternatively, interest point detection. As in Faster R-CNN, Mask R-CNN also consist of two stages. The first stage is the RPN, which generates a set of rectangular object candidates, each accompanied by an 'objectness' score stating the confidence in object existence. The object candidates are chosen from an initial

large set of candidates termed 'anchors'. The anchors set is based on a grid of image locations, and for each location nine anchors are considered with three different sizes and three aspect ratios.

The RPN is deep fully convolutional network whose input is a set of feature maps extracted from a backbone network. The backbone we use is ResNet-101, followed by a Feature Pyramid Network (FPN) [19]. FPN creates multiple-resolution replicas of the high level feature maps computed by the backbone. It hence enables detection at multiple octaves, i.e. scales differing by a factor of two. Among all anchors in all octaves, 256 top object candidate rectangles are chosen for further processing. During training these are labeled as positive if they contain a measurebale object, and negative otherwise. A ratio of 0.33:0.66 is kept for positive versus negative examples during training.

In the second stage, which is a separate network (i.e. no gradient flows between the stages), the model spatially samples $7 \times 7 \times 256$ tensors from each candidate region suggested by the RPN. These region representations are processed by three distinct branches for classification, bounding box refinement and point of interest finding. In the classification branch a softmax layer is used to predict the class of the proposed region, and in the refinement layer offset values are regressed for refining the object's bounding box.

Following [10], the point of interest branch consists of eight 3×3 convolutional layers with 512 maps each, followed by a deconvolution layers scaling up the representation to an output resolution of 56×56. In [10] a single location in the output was designated as the true location, and a spatial softmax classification loss was used to enforce it during training. However, this configuration could not cope with the difficulty of our tasks, and we have adjusted it as described next.

3.2. Adjustments to the phenotyping problem

Unlike in standard detection, in our detection stage we wish to discriminate and detect only measurable objects, i.e. objects with all the relevant interest points visible. Specifically, candidates suggested by the RPN are considered positive during training iff they contain a measurable object including all its key points. This creates a difficult detection problem, since non-measurebale object candidates, which are often very similar to measurable ones, have to be rejected by the object classifier and function as very hard negatives. Furthermore, the measurability constrain adds difficulty since it decreases the number of positively labeled anchors. To allow for a sufficient number of positive anchors we set the non maximum suppression threshold, used for pruning overlapping candidates, to 0.85 during training. Hence, region candidates are only suppressed if they intersect with a higher scoring candidate with a intersection-over-union score larger than 0.85 (0.5 in [10]), so that more positive candidates survive. We also filter the anchoring system based on object size and aspect ratio, e.g. in tree height estimation we do not use small or wide anchors.

The keypoints detection sub-network in [10], described above, was used for a human pose estimation task, which is less ambiguous than our tasks, and hence easier. The locations of human key points like knee, elbow, neck or eye, are spatially well defined, and have unique appearance disambiguating them. In opposed to that, key points in the agriculture tasks considered here are not always clearly defined spatially. For example, the basal point of a tree is defines as the contact point of the tree stem and the ground. However, the contact structure is not a point but a line. We define the point as the middle of the line in our annotation, but this point does not have a unique appearance discriminating it locally from other near points on the contact line. The problem is further complicated since the contact line is often hidden by low vegetation and dead leaves. The exact point position is hence somewhat arbitrary, with nearby points providing identically good candidates. A similar problem exists for the 'left' and 'right' leaf interest points.

The increased key point ambiguity does not allow for a strict single-pixel ground truth location, and enforcing it using the spatial softmax loss of [10] leads to learning failures. Hence instead of using one-hot binary masks, we generate for each annotated keypoint a Gaussian ground truth heat map. In addition, we replace the spatial classification loss with a simple L_2 regression loss, i.e. minimize the square distance, across all pixels, between the Gaussian heat map and the output of the interest point branch.

For the tree height task we used in the ground truth heat map isotropic Gaussians centered on the annotated point location, with fixed variance of $\sigma = 2$ in each axis. For the leaf side points we used a non-isotropic covariance to reflect the fact that these points are well defined in one direction (the direction perpendicular to the leaf center line), but highly ambiguous in the other (see figure 1). Specifically, let $w = (w[1], w[2]) = \frac{x_1 - x_2}{||x_1 - x_2||}$ be the leaf center line direction, with x_1, x_2 the locations of the apex and basal keypoints respectively. We would like this direction to be the large principal direction of the introduced covariance, i.e. its first eigenvector. Given this direction and a hyper parameter S stating the requested ratio between the variance in first and second principal vectors, the covariance matrix is given by $\Sigma = A^t A$ with

$$A = \begin{bmatrix} w[1]S & -w[2] \\ w[2]S & w[1] \end{bmatrix}$$

Figure 3.Middle presents examples of the heat maps generated in the two tasks. At inference time, we take the (x, y) positions of the maximal value in the predicted heat map as the 2D detection for further processing.

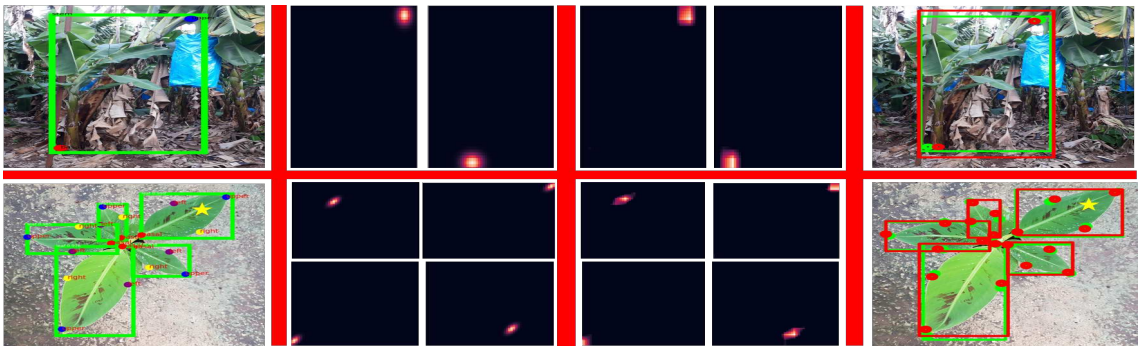

Figure 3. **Visualization of the pipeline operation.** The first row shows progress of banana tree height estimation on a typical example, and the second row shows a case of banana leaves measurement. **Left:** Ground truth bounding boxes and keypoints annotations. **Middle-left:** Ground truth Gausian heat maps constructed for interest point detection. for the leaf measurement example, the maps of the starred leaf are shown. **Middle-right:** Heatmaps inferred by the network. **Left:** Object detections and interest point location as inferred by the network.

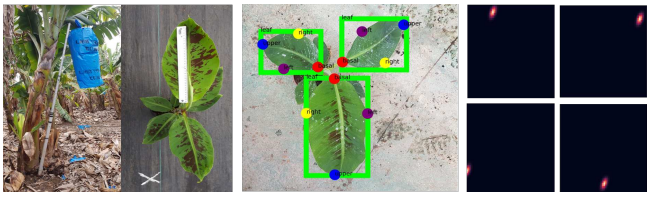

Figure 4. **Ground truth collection. Left:** Measuring the banana tree and banana leaves with a ruler. **Middle:** Annotated ground truth bounding box and keypoints of leaves. **Right:** Ground truth Gaussian heatmaps derived for the interest points of the upper-right leaf.

Task	Train	2D test	3D test
Tree height	577 (757)	87 (105)	33 (33)
Leaf measurements	454 (4409)	74 (236)	30 (101)

Table 1. Number of images and objects (in parentheses) used for training, 2D and 3D testing.

3.3. Obtaining 3D Measurements

In the last stage of the pipeline detected keypoint are back-projected from 2D into 3D world coordinate, and then distances between them are computed. Given the distance D of a point from the sensor imaging plane, measured at 2D coordinates (x, y), one can compute the 3D coordinates (X, Y, Z) by [7]:

$$X = \frac{D \cdot (c_x - x)}{f_x}, Y = \frac{D \cdot (c_y - y)}{f_y}, Z = D \quad (1)$$

where (c_x, c_y) is the sensor's principal point and f_x, f_y are the focal lengths expressed in pixel units. While ideally back projection is simple, in practice there are problems related to lack of depth measurements in some pixels (obtaining hence the value 0), and to depth measurement noise. We hence designed robust depth estimation procedures.

For the banana tree problem, we compute D with the following three stage procedure: 1) Collect the depth values in the ball centered around the detected key point, with 5 pixels radius. 2) drop all the zero (depth failure) measurements, and 3) compute the average of the lowest 10% values. Specifically the last step makes the measurement robust with respect to neighborhood pixels which do not lay on the object, but on far background instead.

For the leaves problem we use a different method. The reason is that the 4 keypoints are located on the margin of the leaf, hence large portion of their neighboring pixels have background depth values. In addition, sometimes the tip of the leaf is not well captured in the depth channel, leaving only background depth measurements in the immediate vicinity of the point. Hence, we use the following two stage procedure. First, we remove background pixels. This is done by computing the minimum depth value in the image M_n, the maximum value M_x and setting to zero all the depth values above $(M_n + M_x)/2$. Then, for each key point we only check the depth values of pixels on the line in the direction of the opposite point (i.e. apex to basal, and left to right). Let x_1, x_2 be the point and its opposite, and define the unit vector between them $w = \frac{x_2 - x_1}{||x_2 - x_1||}$. We then check the depth values of positions $x_1 + l \cdot w$, for the 5 values $l = 0, 1, ..., 4$, using nearest neighbor interpolation to handle float pixel index values. The minimal among the non-zero values is declared as the depth D, and if all are zeros we continue to compute the depth in $x_1 + l \cdot w$ for increasing l values until the first non-zero value is encountered. Following back projection, we use standard Euclidean distance computations to compute the required measurements.

4. Empirical Study

We report here experiments conducted on data collected by experts at an Agritech company Rahan Meristem.

Figure 5. **: confusion between right and left leaf points. Left:**. Cropped image of banana leaf with right keypoint annotated. **Middle:** The ground truth Gaussian heat map for the right keypoint. **Right:** The predicted Gaussian heat map. Confusion between the two points is not surprising, as they have the same local visual characteristics. Only using a large spatial context, containing the whole leaf, it is possible to distinguish between them (according to the locations of the apex and basal points).

4.1. Dataset and Annotation

For both problems, images were captured using two types of sensors. RGB images for training were taken with a Canon HD camera with 4032×3024 resolution. Aligned RGB-D images with resolution 1280×720 were taken using the Intel D435 sensor. For Banana trees, images were taken in several plantations, located in the same area, of banana trees at the fruit harvest stage. The plantations differ w.r.t the banana varieties, which include Gal, Grand-Naine, Adi and Valerie. For the banana leafs, all images were taken in greenhouses conditions. The potted plants were approximately three months old, a stage which enables to determine if the plant has a mutation based on its leaves aspect ratio. Train and test set sizes are summarized in Table 1. For evaluation we used two types of test sets. The first consists of RGB images not used in the training process, including RGB channels of the images taken with the D435 camera. This test set was used for $2D$ evaluations of our methods. The second test set is a subset of the first one, including images taken by Intel D435 for which we collected manual ground truth lengths of the objects - banana tree height and banana leaf length and width measured by a ruler. This test set is used for $3D$ evaluations.

RGB images were annotated with bounding boxes and interest points locations. A total of 757 banana trees, and 1468 banana pots (4409 leaves) were annotated. All annotated objects were measurable, i.e., all their keypoints were visible. Based on the 2D keypoints annotations, Gaussian ground truth heat maps were constructed, as showed in Figure 4. For $3D$ evaluations, the height of 33 banana trees and the length and width of 101 banana leaves in 30 pots were manually measured. Figure 4 (Left) shows the rulers and the measuring procedure.

4.2. 2D Metrics and Results

Detection Rates: The detection threshold was set to 0.5 for both problems as a default compromise between the two types of error. Given a specific application one may chose higher thresholds to achieve higher confidence in the reported estimations, or lower thresholds to be able to detect additional objects, at the cost of more false positives. For the banana tree problem, the mean average precision (mAP) was 0.865. The detector successfully identified 95 of 105 trees, with only two false positives. In each of the detected trees, the model was able to find the two keypoints. The AP for banana leaves detection was 0.885, successfully identifying 221 leaves of 236, with 36 false positive. All interest points of type basal, apex, and left were detected, but key points of the right keypoint were not found in two detected leaves. Observing the predicted heat map (Figure 5), it can be seen that in this case, the key points finder struggled to distinguish between the left and the right keypoints. To overcome this difficulty, we search for the most likely location only in the half space relevant to the keypoint, as determined by the apex and basal point positions.

	Pixel error			Relative error		
keypoint	μ	σ	SE	μ	σ	SE
tree upper	91.37	155.9	16.17	0.03	0.04	0.004
tree basal	121.98	82.27	8.53	0.04	0.02	0.001
leaf apex	34.89	30.27	2.036	0.05	0.03	0.002
leaf basal	32.85	30.52	2.05	0.05	0.03	0.002
leaf left	36.67	42.79	2.87	0.06	0.05	0.003
leaf right	41.16	42.9	2.88	0.06	0.05	0.003

Table 2. Keypoint localization error on the first test set – Euclidean distance between ground truth and detected keypoints. SE denotes standard error.

Point localization accuracy: We compute the Euclidean distance between the detected 2D point location and the ground truth location (deviations in pixels). Since this measure depends on image resolution, its units are somewhat arbitrary. We hence report also the deviations normalized by the length (in pixels) of the relevant distance measured (leaf width, height or tree height). The results can be seen in Table 2. As can be seen, the average point deviation is between 2% and 6% of the length measured. Estimation Deviations in the tree problem are lower than deviations in the leaf problem. The largest deviations were measured for the left and right leaf keypoints, which are clearly more difficult to detect, as they are positioned somewhat arbitrarily along the leaf curve.

4.3. 3D Metrics and Results

For 3D evaluation we examine the absolute deviation of the 3D measurements for our phenotyping tasks, i.e. tree height, leaf length, leaf width (in centimeters) and leaf length-width ratio (unit less). As we performed manual physical measurements of the objects, we are able to evaluate our methods with respect to the true 3D lengths.

Table 3 shows the estimation deviations of our complete pipeline for each of the tasks. As can be seen, the mean rel-

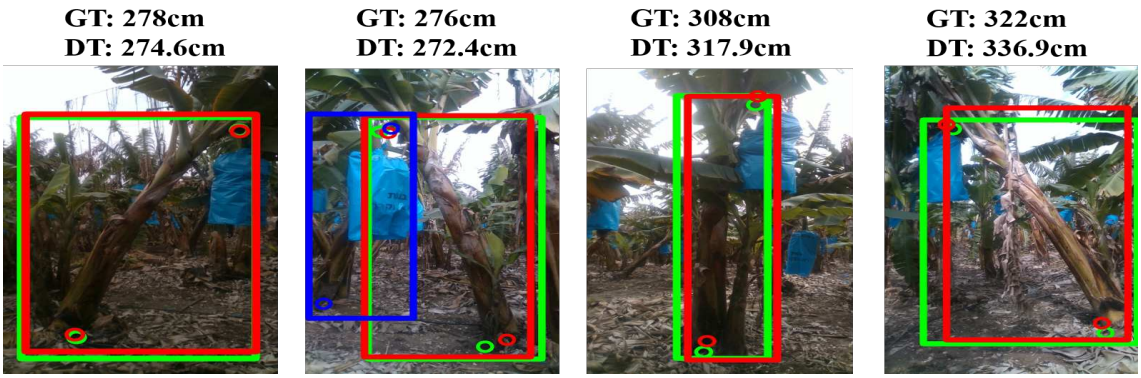

Figure 6. Detection examples in the banana trees problem. Green denotes ground truth, red denotes successful detection, and blue denotes false positives. Above each image we show the height measured by ruler (GT) and the estimated height by our model (DT)

	Algorithmically detected keypoints							Manually marked keypoints						
	Error in cm			Relative error			R^2	Error in cm			Relative error			R^2
size	μ	σ	SE	μ	σ	SE		μ	σ	SE	μ	σ	SE	
tree height	9.24	7.7	1.36	0.03	0.03	0.004	0.74	10.55	6.45	1.14	0.03	0.03	0.003	0.71
leaf length	2.01	1.3	0.13	0.08	0.05	0.005	0.88	0.86	0.76	0.08	0.03	0.03	0.003	0.95
leaf width	0.96	0.73	0.07	0.07	0.06	0.007	0.89	0.94	1.11	0.12	0.07	0.09	0.009	0.77
leaf ratio	0.15	0.17	0.018	-	-	-	-	0.15	0.16	0.017	-	-	-	-

Table 3. Length estimation errors for manually marked and automatically detected keypoints on second test set. SE denotes standard error.

ative deviation for all tasks is under 8% of the true length. The error is smallest for tree height, and largest for leaf length. The reason is that for smaller objects, obtaining accurate estimations is more difficult. This is compensated to some degree, but not completely, by the smaller distance of the leafs to the camera.

To further understand whether the deviations stem from the keypoint detection algorithm, or rather from the inaccuracy introduced by RGB-D sensor, we also estimate the 3D length based on the manually marked ground truth keypoints in the annotated images. In this measurement we skip the first two phases of the pipeline and use only the last de-projection phase (applied to the manually annotated 2D points) to compute the 3D lengths of interest. As Table 3.right shows, for tree height and leaf width there is no statistically significant difference between the algorithmically computed and manually stated points. For leaf length, the error introduced by the automated detection is about twice larger than the error from the manual markings. Since there is no such difference between points related to leaf width and height in 2D accuracy (table 2), the larger deviation is clearly related to larger difficulty of 3D estimation for the leaf apex and/or basal points.

In order to estimate the statistical strength of our method, we compute the fraction of explained variance R^2

$$R^2 = 1 - \frac{Var_{err}}{Var_{total}} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \qquad (2)$$

This essentially compares our estimation method to the trivial estimation method of predicting the mean length for every instance. The results are summarized in the R^2 column in table 3. It can be seen that our estimation explains (i.e. successfully infers) a significant portion $(0.74 - 0.89)$ of the length variance. Interestingly, for leaf width the algorithmic pipeline provides better estimations than relaying on manually marked key-points. The reason is probably the ill-defined nature of the 'left' and 'right' points used for this estimation, which are more consistently found by the algorithm then by the human annotator.

4.4. Detection Examples and Error Analysis

Figures 6 and 7 present successful detection examples alongside various errors of our full pipeline for banana trees and banana leaves problems, respectively. In the detection examples of the trees (Figure 6), the left example shows a case of highly accurate detection and estimation. In the second image an example of false positive detection is shown. It happened because the model associated the peduncle of the ground truth tree to another tree, and since the basal of the other tree is visible the model considers it as a measurable object. In this image, the basal point detected has a significant deviation from the marked ground truth, but since it also lies on the line connecting the stem to ground, tree height estimation is not affected. In the third image 2D deviations in the y axis in both keypoints lead to large deviation in the height estimation. In the right image 2D errors in keypoints position are rather small, but nevertheless the height estimation deviation is relatively large. In addition

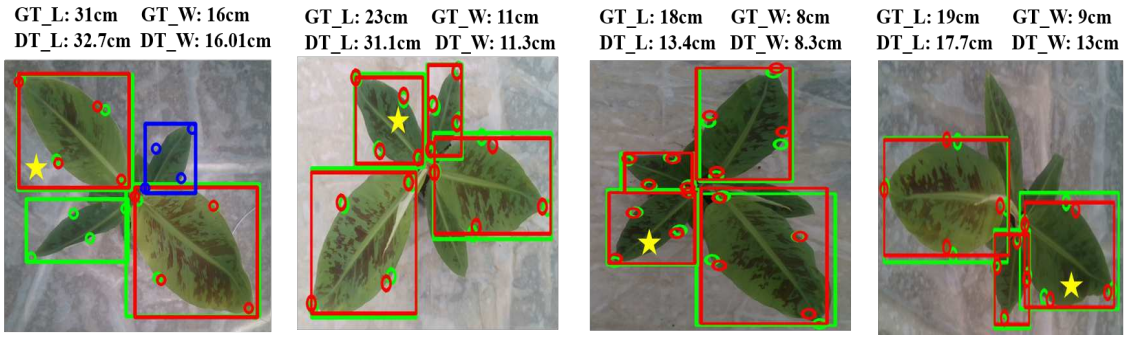

Figure 7. Detection examples in banana leaves problem. The ground truth objects are marked green, and red marks the detected objects. Above each image there are the length and width manually measured using a ruler (GT), and the estimated length and width by our model (DT), for the leaf marked by a yellow star.

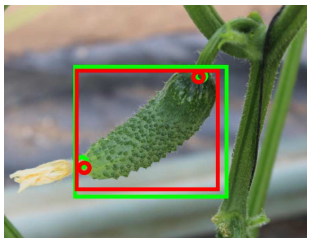

Figure 8. Detected cucumber and its extremities in green house. In green is the ground truth annotations, in red is our model detection

we can see an example of sub-optimal annotation: the annotated basal is marked lower than its correct position.

Figure 7 presents examples for banana leaves. Above each image we show the ground truth and estimated length and width only for the leaf marked with a yellow star, and this is the leaf we refer to in our discussion. In the left image keypoint detections and length measurement are accurate for the marked leaf. In addition there is one relatively small leaf that was not detected. One can argue about the correctness of labeling this leaf as measurable, as it is almost vertical with respect to the sensor. There is also a false positive detection of a leaf above the marked leaf, with a basal hidden by the basal of the marked leaf, and hence, its length cannot be measured. In the second and third images accurate 2D keypoint localization is obtained, but the length estimation error is relatively high. The width error is low for both images. We conjecture that the problem in these images is related to errors in depth estimation by the sensor for the basal or the leaf apex keypoints. The right image seems to suffer from a similar problem, but with the leaf width measurement.

5. Conclusion and Future work

We presented a general technique for length-based plant phenotyping, based on keypoint detection in 2D and depth information from a low-cost 3D sensor. The technique was tested on two specific tasks: banana height estimation and leaf length, width and aspect ratio measurements. It obtained average deviation of 3% (of the total length) for tree height estimation, and $7-8\%$ deviation or for leaf width and length estimation. Statistically, the method was able to explain (infer) $0.74-0.89$ of the total length variance. It is not clear yet whether the results obtained for the aspect ratio estimation are good enough for identifying mutant plants, as current measurements were done on normal banana plants. A set of annotated mutant plant images is required for further testing, and we currently work to obtain such normal/mutant labeling of plants from an expert.

There are a number of possible avenues to take this work forward. First, more data can be used in the problems we considered here both in training, for improving keypoint localization accuracy, and in testing, for verifying our findings. While lengths are currently computed using plain Euclidean distance, we may consider Geodesic distances in the future, taking into account the curvature of the measured surfaces. To demonstrate the generality of the method beyond the two tested problems, we are currently pursuing additional tasks like measuring cucumber length by detecting its extremities. Figure 8 presents an example of our initial results for finding cucumber's extremities in greenhouse conditions. Beyond length measurements, it would be of high interest to develop algorithms extending to measurements of areas and volume based phenotypes. Looking forward, our method can be embedded in a flexible, general phenotyping system as a length estimation module.

6. Acknowledgements

This research is supported by the Israel Innovation Authority through the Phenomics MAGNET Consortium, and by the ISF fund, under grant number 1210/18. We thank Ortal Bakhshian from Rahan Meristem for many helpful discussions and for providing the images.

References

[1] N. An, C. M. Palmer, R. L. Baker, R. C. Markelz, J. Ta, M. F. Covington, J. N. Maloof, S. M. Welch, and C. Weinig. Plant high-throughput phenotyping using photogrammetry and imaging techniques to measure leaf length and rosette area. *Computers and Electronics in Agriculture*, 127:376–394, 2016.

[2] J. L. Araus and J. E. Cairns. Field high-throughput phenotyping: the new crop breeding frontier. *Trends in plant science*, 19(1):52–61, 2014.

[3] T. Baharav, M. Bariya, and A. Zakhor. In situ height and width estimation of sorghum plants from 2.5 d infrared images. *Electronic Imaging*, 2017(17):122–135, 2017.

[4] Y. Chéné, D. Rousseau, P. Lucidarme, J. Bertheloot, V. Caffier, P. Morel, É. Belin, and F. Chapeau-Blondeau. On the use of depth camera for 3d phenotyping of entire plants. *Computers and Electronics in Agriculture*, 82:122–127, 2012.

[5] F. Fiorani and U. Schurr. Future scenarios for plant phenotyping. *Annual review of plant biology*, 64:267–291, 2013.

[6] R. T. Furbank and M. Tester. Phenomics–technologies to relieve the phenotyping bottleneck. *Trends in plant science*, 16(12):635–644, 2011.

[7] S. Giancola, M. Valenti, and R. Sala. *A Survey on 3D Cameras: Metrological Comparison of Time-of-Flight, Structured-Light and Active Stereoscopy Technologies*. Springer, 2018.

[8] G. Grillo, M. José Grajal Martín, and A. Domínguez. Morphological methods for the detection of banana off-types during the hardening phase. In *II International Symposium on Banana: I International Symposium on Banana in the Subtropics 490*, pages 239–246, 1997.

[9] J. Han, L. Shao, D. Xu, and J. Shotton. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE transactions on cybernetics*, 43(5):1318–1334, 2013.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[11] M.-H. Hu, Q.-L. Dong, P. K. Malakar, B.-L. Liu, and G. K. Jaganathan. Determining banana size based on computer vision. *International journal of food properties*, 18(3):508–520, 2015.

[12] Y. H. Hui. *Handbook of food products manufacturing*. Wiley-interscience, 2007.

[13] Y. Itzhaky, G. Farjon, F. Khoroshevsky, A. Shpigler, and A. B. Hillel. Leaf counting: Multiple scale regression and detection using deep cnns, 2018.

[14] L. Jiang, S. Yan, W. Yang, Y. Li, M. Xia, Z. Chen, Q. Wang, L. Yan, X. Song, R. Liu, et al. Transcriptomic analysis reveals the roles of microtubule-related genes and transcription factors in fruit length regulation in cucumber (cucumis sativus l.). *Scientific reports*, 5:8031, 2015.

[15] Y. Jiang, C. Li, A. H. Paterson, S. Sun, R. Xu, and J. Robertson. Quantitative analysis of cotton canopy size in field conditions using a consumer-grade rgb-d camera. *Frontiers in plant science*, 8:2233, 2018.

[16] J. Jin, G. Kohavi, Z. Ji, and A. Zakhor. Top down approach to height histogram estimation of biomass sorghum in the field. *Electronic Imaging*, 2018(15):288–1, 2018.

[17] A. Kamilaris and F. X. Prenafeta-Boldu. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018.

[18] L. Li, Q. Zhang, and D. Huang. A review of imaging techniques for plant phenotyping. *Sensors*, 14(11):20078–20111, 2014.

[19] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[20] A. Milella, R. Marani, A. Petitti, and G. Reina. In-field high throughput grapevine phenotyping with a consumer-grade depth camera. *Computers and Electronics in Agriculture*, 156:293–306, 2019.

[21] M. Minervini, H. Scharr, and S. A. Tsaftaris. Image analysis: the new bottleneck in plant phenotyping [applications corner]. *IEEE signal processing magazine*, 32(4):126–131, 2015.

[22] N. B. A. Mustafa, N. A. Fuad, S. K. Ahmed, A. A. Z. Abidin, Z. Ali, W. B. Yit, and Z. A. M. Sharrif. Image processing of an agriculture produce: Determination of size and ripeness of a banana. In *2008 International Symposium on Information Technology*, volume 1, pages 1–7. IEEE, 2008.

[23] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[24] S. Paulus, J. Behmann, A.-K. Mahlein, L. Plümer, and H. Kuhlmann. Low-cost 3d systems: suitable tools for plant phenotyping. *Sensors*, 14(2):3001–3018, 2014.

[25] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, 2019.

[26] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[27] D. Rousseau, H. Dee, and T. Pridmore. Imaging methods for phenotyping of plant traits. In *Phenomics in Crop Plants: Trends, Options and Limitations*, pages 61–74. Springer, 2015.

[28] A. Vit and G. Shani. Comparing rgb-d sensors for close range outdoor agricultural phenotyping. *Sensors*, 18(12):4413, 2018.

[29] W. Wang, B. Vinocur, and A. Altman. Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. *Planta*, 218(1):1–14, 2003.

[30] Z. Wang, K. B. Walsh, and B. Verma. On-tree mango fruit size estimation using rgb-d images. *Sensors*, 17(12):2738, 2017.

[31] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 2019.

[32] G. Zotz, P. Hietz, and G. Schmidt. Small plants, large plants: the importance of plant size for the physiological ecology of vascular epiphytes. *Journal of Experimental Botany*, 52(363):2051–2056, 2001.