Length Phenotyping with Interest Point Detection

Adar Vit[a], Guy Shani[a] and Aharon Bar-Hillel[b]

[a]Department of Software and Information System Engineering, Ben-Gurion University of the Negev, Beer Sheva, Israel
[b]Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer Sheva, Israel

ARTICLE INFO

Keywords:
Plant Phenotyping; Length Estimation;
RGB-D Sensor; Key-Points detection;

ABSTRACT

Plant phenotyping is the task of measuring plant attributes mainly for agricultural purposes. We term *length phenotyping* the task of measuring the length of a plant part of interest. The recent rise of low cost RGB-D sensors and accurate deep artificial neural networks provides new opportunities for length phenotyping. We present a general technique for length phenotyping 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 develop a robust de-projection procedure for the 3D measurement stage. We apply our method to three real world tasks: measuring the height of a banana tree, the length and width of banana leaves in potted plants, and the length of cucumbers fruits in field conditions. The three tasks were solved using the same pipeline with minor adaptations, indicating the method's general potential. The method is stagewise analyzed and shown to be preferable to alternative algorithms, obtaining error of less than 10% deviation in all tasks. For leaves' length and width, the measurements are shown to be useful for further phenotyping of plant treatment and mutant classification.

1. Introduction

As the world's population is growing the global demand for food and energy is growing as well, requiring innovative advances in the agriculture industry [4, 36]. Plant phenotyping is vital tool for improving growth and yield formation, which are the basis for nursing the world. In plant phenotyping we measure and assess complex plant traits related to growth, yield, and other significant agricultural properties [6]. In some cases plant phenotyping is done in a lab, using advanced costly infrastructure [28], but in many cases phenotyping should be done in field conditions to best capture the true plant traits. In field phenotyping, most measurements are done manually, by a team of workers.

As manual phenotyping is extremely costly, there is a need to develop automated analysis methods that are sufficiently accurate and robust enough for measuring phenotypic traits in field conditions on real crops [28]. Automated algorithms are required for accelerating cycles of genetic engineering [38], and for automating agriculture processes

adarv@post.bgu.ac.il (A. Vit); shanigu@post.bgu.ac.il (G. Shani); barhille@bgu.ac.il (A. Bar-Hillel)
ORCID(s):

(a) Banana trees (b) Banana leaves (c) Cucumber fruits

Figure 1: Length based phenotyping. a: A banana tree with two interest points: basal (red) and upper (blue). The upper is defined by experts as the highest point of the tree's peduncle (top of the arch). The distance between them is the tree height. **b:** 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. **c:** A cucumber with two interest points: basal (red) and upper (blue). The distance between them is cucumber length.

[6]. Field and greenhouse phenotyping are 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 [22], and the recent success of deep neural networks opens new directions [33, 12]. A second recent advancement is the abundance of low-cost sensors, from RGB to depth and thermal sensors, useful for capturing plant traits. While such low cost sensors are typically not as accurate as high end sensors, they are well suited for wide field application in the agricultural industry, which has low profit margins [31].

This paper focuses on the problem of measuring 3D physical lengths of plant parts in field conditions, using a low cost RGB-D sensor and a deep network architecture. Measuring the length of plant's parts can provide important cues about the plant state [41] and expected utility. For example, measuring the width and 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, some mutations are characterized by narrow leaves, and others by an aspect ratio smaller than normal (lower than 1.8, where the normal ratio is 2.2) [10]. Another example is estimating the height of a banana tree. An important 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 [16]. Another important problem is measuring the width of a stem - it is a key characteristic used for biomass potential evaluation [26].

We term such tasks *length phenotyping*. While this paper focuses on three of the above examples, there is an abundance of similar tasks. Given the repetitive nature of length phenotyping tasks, it is desirable to develop a generic process that can be applied, with minimal adjustments, to measure lengths of different plant parts in various plants. The paper focuses on two tasks related to banana crops, considered to be the most widely consumed fruit [14], and a third one relating to cucumbers fruits. The problems are estimation tasks of a banana tree height (actually a banana plant is not a tree, but the term tree is used for simplicity), width and length of banana leaves, and cucumber fruit length. The related interest points and length definitions can be seen in Figure 1.

While focusing on these three problems, we suggest a general algorithmic pipeline, consisting of three stages:

1. Detecting the objects of interest that needs to be measured.
2. Identifying interest points on the detected objects, that we are interested in the distance between them.
3. De-projection of the interest points to world coordinates to compute distances.

For the first two stages we use Convolutional Neural Networks (CNNs) [12] trained and applied on RGB images. RGB images (without depth information) are used for these stages since object detection in RGB images is well studied [40, 33, 12] and it is possible to transfer knowledge (reuse network weights for initialization) from well trained RGB backbones. Such transfer learning is important because, as we later demonstrate, it enables to identify objects successfully by training with only a few hundreds of examples. We use networks based on known architectures [12], but adapt their training procedures to the requirements of the agricultural tasks in several ways.

In the third stage, for measuring distances between interest points, an RGB-D sensor (Intel D435) is used, providing a depth channel registered to the RGB channels. Using the depth information for de-projecting points from the 2D image to their 3D coordinates is possible [8]. However, given the low cost sensors that we use, straightforward de-projection faces problems due to noisy depth values, failing pixels of depth value 0, and *intermediate depth* pixels. An important problem for length measurements are the latter, *intermediate depth* pixels, positioned on the object edge. The depth value of such pixels is interpolated between the depth of the object and its background, providing unreliable measurements. A robust length estimation algorithm was developed in order to cope with those challenges, with two main stages. First, it re-chooses edge points with more reliable depth for length computation, by fitting a fixed-slope interval model. Second, a linear regression is performed over the final depth estimate, to compensate for the shortening of the measured interval.

More direct alternatives to the three-stage pipeline can be proposed. For example, direct detection of the interest points (skipping stage 1) is possible, or even direct inference of the length from the RGBD image without finding the interest points (skipping stage 2). Our choice to separate the stages of detection and interest point identification is based on two arguments. 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 that correspondence determination between pairs of points (and in the leaf case, quadruplets) of the *same* object is required for distance computation. Such correspondence is naturally provided by the detection stage, which finds an object bounding box containing the related points.

Our choice to include an interest point detection stage, rather than directly estimating the 3D length, is based on the assumption that such estimation, while possible, requires the network to bridge a large inference gap on its own, i.e.: it has to learn which cues and image areas are relevant for length regression. While such learning may be possible, it may require a very large annotated training set due to the task difficulty. In addition, our model can be trained with 2D annotation which is easier to obtain, while acquiring 3D annotation (actual length measurements) is more laborious and difficult to collect.

Our choices and assumptions regarding the three-stage pipeline were tested by experiments with alternative more direct algorithms. Specifically, we compare the point detection within an identified object (point detection after object detection), to direct interest points detection. We also compared our length estimation using detected interest points (the proposed three-stages pipe) to two architectures designed for direct length measurements.

The suggested method is tested on collected image datasets which include 3D ground truth measurements, obtained manually using a ruler. The accuracy of the three stages is analyzed and compared to alternatives. The empirical results show measurement deviations lower than 10% for all tasks, and a significant advantage of the proposed pipeline over the alternatives. For leaves' length and width, it is shown that the length measurements enable distinction between certain kinds of plant treatments and mutations with significant accuracy.

The main contributions of this work are hence threefold. First, a general method for plant part length measurement is presented, based on a 3 stage pipeline. Second, the pipeline performance on three important real world problems is measured and shown to be useful. Third, comparison to alternatives and a detailed stage-wise error analysis is conducted, providing valuable information regarding possible further improvements.

2. Related Work

Traditional plant phenotyping is based on extensive human labor, where only a few samples are collected for thorough visual or destructive inspection. These methods are time consuming, subjective and prone to human error, leading to what has been termed as 'the phenotypic bottleneck' [7]. 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. In [22], an extensive survey of imaging methodologies and their application in plant phenotyping is provided. Since recently deep learning techniques have emerged as the primary tool in computer vision, they have a growing impact on agricultural applications [19].

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 [11] structure, which is useful in many applications. For example, Chene et al. [5] showed that depth information can help identifying individual leaves, allowing automated measurement of leaf orientation in indoor environments. Vit et al. [37] recently compared several depth sensors for a plant phenotyping task, in field condition and in various illumination conditions. It was shown that the Intel D435 outperforms other alternatives. In [39] the size of mango fruits in field conditions was estimated. Jiang et al.[17] 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. [27] proposed an in-field high throughput grapevine phenotyping platform for canopy volume estimation and grape bunch detection, using a RealSense R200 depth camera.

Plant's size is an important trait, which can be used as indicator of growth, yield and health of the plant. In [29] the area, perimeter, length, and thickness of bananas were measured using a combination of computer vision techniques on gray-scale images. In [3, 18] 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, typically 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. Lati et al. [21] used 3D-reconstruction algorithm for extracting growth parameters such as plant height and the internode distances in order to detect initial parasitism in potted plants of sunflower broomrape in indoor conditions. They found out that Plant height and the first internode length provided a significant early morphological indication of infection.

Gongal et al. [9] used time-of-flight 3D camera for estimating apple size in tree canopies. They defined the apple size as the length of its longest axis, and used two techniques for measuring it. The first technique measures the

length as the maximal $3D$ euclidean distance between any two individual pixels in the apple region. The second technique used checkerboard images taken in RGBD to estimate pixel size in the real world and from it infer apple size based on the number of pixels. In [23] a pipeline was developed for detecting and localizing citrus fruits in outdoor orchard environment by analyzing RGB-D images. They used Bayes-classifier based image segmentation and density clustering method for localizing each of the fruits. To determine the citrus size, they measured the distance between 3D coordinates of points in the cluster in the x-axis direction.

Deep convolutional neural networks (CNNs) have significantly improved the state-of-the-art in computer vision tasks and are widely used in several classic vision tasks as object classifcation [20], detection [33], and segmentation [12]). These networks are better able to cope with realworld vision problems than previous technologies, hence providing new automation opportunities.

Interest points detection is used extensively in human pose estimation and face recognition. In [12] human joints interest points are detected by predicting a one-hot spatial mask for each interest point type. In [30] another architecture for human joint detection is suggested, based on progressive pooling followed by progressive up-sampling. In [32] 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 infers the visibility of interest points, to account for points invisible at test time. As our goal is to perform 3D measurements, we have no interest in invisible interest points, which do not enable such measurements. Instead, we define as 'measurable object' only objects in which all the relevant interest points are visible and train our detector to only detect measurable object instances. Interest point detection have not been extensively used in plant phenotyping [35]. Hu et al. [13] 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. In [15] a leaf counting task with an intermediate stage of interest point finding was solved. A model for finding leaf centers is trained with interest points treated as Gaussian heat map similarly to our work. The heat map is then used for leaf count regression.

3. Materials and Methods

We here discuss our suggested method - a three stage pipeline for length measurement. A High level architecture of the pipeline can be seen in Figure 2. The three stages are next described in Sections 3.1,3.2 and 3.3. Alternative algorithms compared to the three-staged pipeline are described in 3.4.

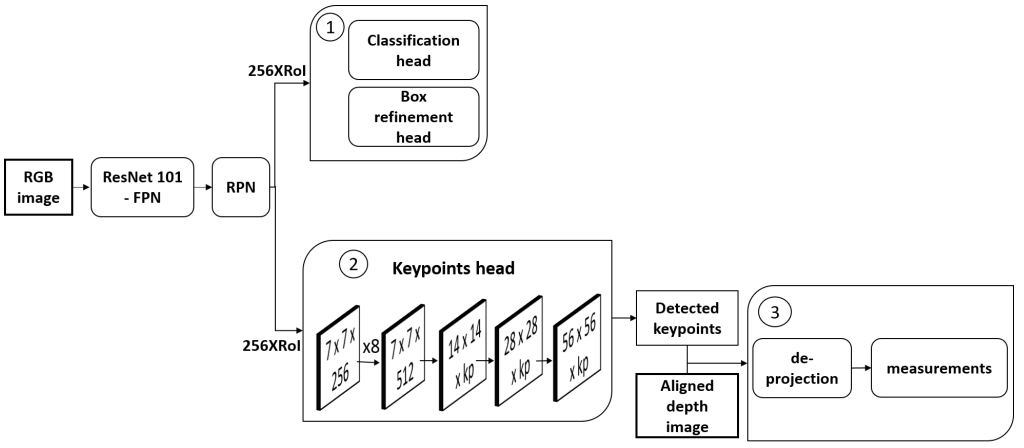


Figure 2: High level view of our processing pipeline. Left: the backbone network, extracting a rich feature map in 5 different octaves using the Feature Pyramid Network (FPN). Middle: Extracted representations of 256 Regions of Interest (RoI) are processed by classification and box refinement branches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

3.1. Detection using Mask R-CNN

The first two stages in our proposed pipeline are based on the Mask R-CNN [12] architecture with adaptations made for our goal. Mask R-CNN extended a previous architecture for object detection termed Faster R-CNN [33], adding network modules for object segmentation or alternatively, interest point finding. As in Faster R-CNN, Mask R-CNN also consists of two stages. The first stage is a Region Proposal Network (RPN)[33], 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*, containing hundreds of thousands of rectangle candidates. The anchor 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 a 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) [24]. The 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 measurable object, and negative otherwise. A ratio of 1:2 is kept for positive versus negative examples during training.

In the second stage, which is a separate network with no gradient flows between the stages, the model spatially

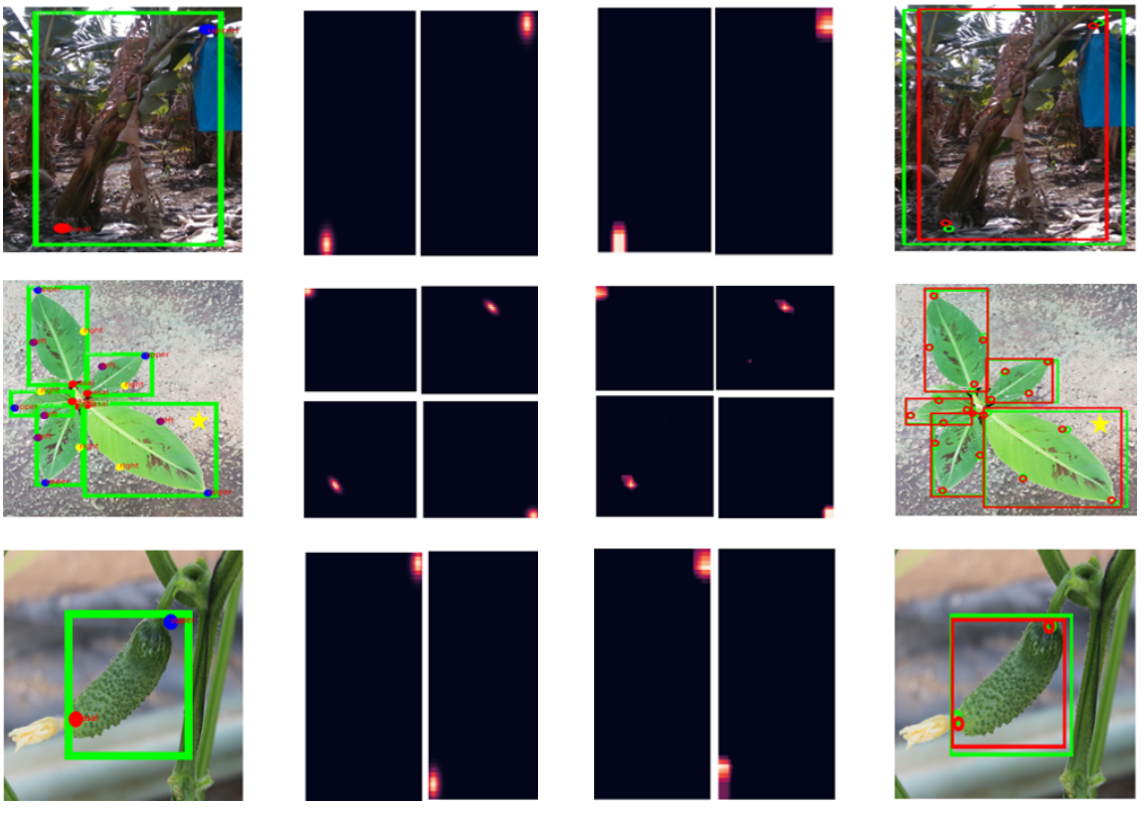
samples a 7 x 7 x 256 tensor from each candidate region suggested by the RPN. These region representations are processed by three distinct branches for classification, bounding box refinement and point of interest finding. In the classification branch a softmax layer is used to predict the class of the proposed region, and in the refinement, layer offset values are regressed for refining the object's bounding box.

Unlike standard detection, in our detection stage we need 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 interest points. This creates a difficult detection problem, since non-measurable object candidates, which are often very similar to measurable ones, have to be rejected by the object classifier and function as hard negatives. Furthermore, the measurability constraint 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 (instead of 0.5 in [12]) so 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.

3.2. Interest Point Detection

Following [12], the interest point finding branch consists of eight 3 x 3 convolutional layers with 512 maps each, followed by deconvolution layers scaling up the representation to an output resolution of 56 x 56. In [12] a single location in the output map was designated as the true location, and a spatial softmax classification loss was used to enforce it during training. However, this configuration was used for a human pose estimation task, which is less ambiguous than our tasks, and hence easier. The locations of human interest points like knee, elbow, neck or eye, are spatially well defined, and have unique appearance disambiguating them. As opposed to that, interest 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 because 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 spatial softmax loss assumes that there is only one "right" point which is clearly distinguished from the other "non right" ones. Hence in our preliminary experiments, the spatial softmax could not cope with the tasks difficulty, and we had to adjusted it as described next.

Instead of using one-hot binary masks, we generate for each annotated interest point a Gaussian ground truth heat map, stating multiple locations as successful hits. In addition, we replace the spatial classification loss with a simple



(a) (b) (c) (d)

Figure 3: Visualization of the pipeline operation. The first row shows progress of banana tree height estimation on a typical example, the second row shows a case of banana leaves measurement and the last shows a case of cucumber measurement. **a:** Ground truth bounding boxes and interest points annotations. **b:** Ground truth Gaussian heat maps constructed for interest point detection. For the leaf measurement example, the maps of the starred leaf are shown. **c:** Heat maps inferred by the network. **d:** Object detections and interest point location as inferred by the network. Red rectangles and points are network outputs, green are the ground truth.

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 all points except leaf 'right' and 'left' points isotropic Gaussians centered on the true point were used, with $\sigma = 2$ variance.

For the leaf side points we used a non-isotropic covariance, reflecting that these points are well defined in the direction perpendicular to the leaf center line, but highly ambiguous in the other direction (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 interest points 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

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

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

3.3. Obtaining 3D Measurements

In the final stage of the pipeline the distance between the detected interest points is estimated. This is done by back-projecting the detected interest point from 2D to 3D world coordinates. 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 [8]:

$$X = \frac{D \cdot (x - c_x)}{f_x} \tag{2}$$

$$Y = \frac{D \cdot (y - c_y)}{f_y} \tag{3}$$

$$Z = D \tag{4}$$

where c_x,c_y are the sensor's principal point coordinates 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 absence of depth measurements in some pixels (represented by a value of 0), and to problematic depth measurement with large errors in intermediate pixels on the edge of an object. Unfortunately, the interest point pixels are usually on an object edge, and hence often are intermediate pixels with corrupted depth value. In order to cope with those challenges, we consider the 2D line between the two identified interest points and try to find two points on it for which a 3D interval model is a good fit. This is done using the procedure described next.

Given two 2D end points suggested by the network, a 2D line of K_1 points between them is sampled, and is also extended K_2 pixels beyond them at each side. Hence a straight line of $K_1 + 2K_2$ points is sampled, indexed as $x_1, .., x_{K_1+2K_2}$, with $x_i \in R^2$. The original points are located at indices $a = K_2 + 1$ and $b = K_2 + K_1$. The depth $(d_1, ..., d_{K_1+2K_2})$ at $x_1, .., x_{K_1+2K_2}$ is sampled with bi-linear interpolation, a common technique for calculating image values in a fractional index location based on its nearby pixels values. A straight line in 3D between the two end points should ideally manifest as a linear function in the d values. Let us denote the number of points by $N = K_1 + 2K_2$. We look for a pair of points (x_i, x_j) on the 2D line $(i, j \in 1, .., N)$ which are close to the original points (x_a, x_b), yet their depth values fit well to a $3D$ interval model. Formally, the criteria optimized are as follows:

- **End points proximity:** The (normalized) square distance of (x_i, x_j) from their original anchors is minimized.

$$S_1(i,j) = \frac{1}{2d_{max}}\left(\|x_i - x_a\|^2 + \|x_j - x_b\|^2\right) \tag{5}$$

with $d_{max} = ||x_N - x_1||$ is the length of the 2D interval considered. d_{max} is used as normalization constant to have $S_1 \in [0,1]$, so it can be combine with other terms without scaling issues.

- **Fixed 3D slope:** The 3D gradient at point i is defined as $g_i = d_{i+1} - d_i$. We assume that the object surface, restricted to the sampled line, is approximately planar. Hence the function expressed by the depth measurements $\{d_i\}$ is expected to be linear, with a fixed slope. The slope in an interval $[x_i, x_j]$, which is also the average gradient is given by

$$E_g = \frac{d_j - d_i}{j - i} = \frac{1}{j-i}\sum_{k=i}^{j-1} g_k \tag{6}$$

Since an approximately fixed slope is expected, we ask for small deviation between the gradient values in the interval $[i, j]$ and the slope, or equivalently, for a low gradient variance

$$S_2(i,j) = \frac{1}{S_T(j-i)}\sum_{k=i}^{j-1}\left(g_k - E_g\right)^2 \tag{7}$$

Where $S_T = S_2(1, N)$ is a normalizing constant. We note that the assumption that the object is planar is restrictive, because in agricultural objects, such as leaves, may have a significant curvature. In such cases, this specific step may be adapted to the specific object of interest. For example, one could use the obtained depth measurements to fit a curve, e.g., using splines.

- **End points gradient:** As x_i, x_j are the end points of an interval in 3D, the squared gradients at these points should be maximized. We hence minimize:

$$S_3(i,j) = -\frac{1}{2g_{max}}\left(\left(g_{i-1}\right)^2 + \left(g_j\right)^2\right) \tag{8}$$

with $g_{max} = \max_{i \in 1,\ldots,N} g_i$

Our final score to optimize is $S = \alpha_1 S_1' + \alpha_2 S_2' + \alpha_3 S_3'$, where α_1, α_2 and α_3 are weighting the relative contribution of each criterion. Since normalization constants were introduced to each term, bringing it to the scale of $[0,1]$, we could use $\alpha_1 = \alpha_2 = \alpha_3 = 1$ without further tuning. For the optimization, a search over all pairs (i, j) with $i < j$ is conducted. Figure 4 illustrates the length estimation procedure with an example.

This function suits cases in which we measure objects from one end to another, like leaves and cucumbers, where

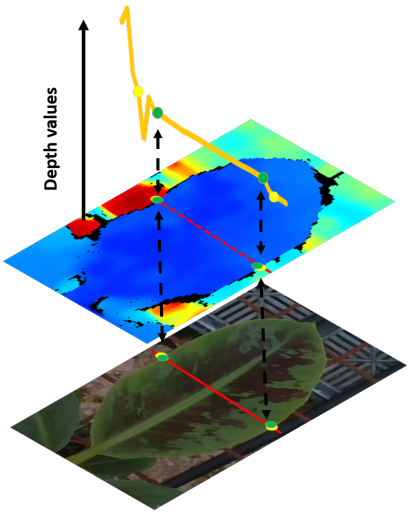


Figure 4: Robust depth estimation illustration: An example of 2D surface leaf and its depth surface. The two points detected by the interest point detector are drawn on them in yellow. The 2D line sampled between these points (and extending a bit beyond them) is shown in red. The depth values sampled on the red line are shown above as the yellow curve. Based on the 3D linear line model, the green points are chosen to replace the original yellow ones, since they have stable depth measurements with a linear line between them. The length between these points is the initial length estimate, but it is slightly biased toward lower lengths. This tendency is latter corrected using a linear regressor.

end points are on a depth edge. For tree height measurements the end points are not positioned exactly on an object edge, though they are not far from the edges. There is a significant change of the depth gradient behavior near the end points, and hence, in this case a three-step alternative is suggested and compared to the algorithm above:

- Collect the depth values in the ball centered around the detected interest point, with a fixed pixel radius (5 pixels in our experiments).
- Ignore all the zero values (depth measurement failure).
- 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, belonging to farther background objects.

The length estimation procedure suggested finds stable estimates, but is has a bias towards short estimations. The reason is that the true end points often do not have stable depth values, and other points which are more interior are used instead. Hence a a simple linear regression model $y = ax + b$ was applied for correction, with the parameters (a, b) estimated on a validation set.

3.4. Alternative Algorithms

Several alternative algorithms were developed and compared to the three staged pipeline. In section 3.4.1 direct detection of interest points is discussed, and in 3.4.2 two architectures are suggested for direct length regression.

3.4.1. Direct Detection of Interest Points

we examined a method for direct detection of interest points, without initial object detection. The 2D interest points annotations were placed in bounding boxes and a Faster R-CNN [33] network was trained for their detection as a multi class problem, where there is a class for each key-point type. For example, in the banana tree example, there is a class for the 'upper' keypoint and another class for the 'basal' keypoint.

We used two different methods of marking the interest points using bounding boxes, introduced due to differences between the tasks. Assume an image of size $W \times H$, an object of size l, where l is the maximum between the object width and height, and an interest point with coordinates (x, y). For cucumber and leaf interest points, the bounding box is a square centered at (x, y) with edge length l (if the point is near the image edge the square side length is $min(l, 2x, 2y, 2(W - x), 2(H - y)))$ to keep the square within the image). For banana tree interest points a different method was used because the points are typically close to the image edge, and $min(l, 2x, 2y, 2(W - x), 2(H - y))$ is small leading to rectangles not containing enough tree context. Instead, a rectangle is used of size $0.5l \times l$ with the interest point located $0.1l$ above the lowest edge for 'basal' points, and $0.1l$ below the upper edge for 'upper' points.

3.4.2. Direct Regression Models

A direct regression model learns to predict the length directly from the input RGBD images. Two direct regression models were developed, both using depth information in the training process instead of 2D interest points annotations. The algorithms' architecture is similar to our model, but the interest points finding branch is replaced with a direct regression module. The module accepts additional maps containing the 3D information, sampled from the object candidate rectangle. As a preliminary step, each pixel of the depth images is de-projected into 3D coordinates. For each candidate RoI, the depth channels are sampled in a 56×56 grid (using an RoI-align layer [12]) into a $56 \times 56 \times 3$ tensor. The coordinates are then normalized by mean subtraction - length measurements are invariant to such subtraction, and it reduces data variability. The resulting tensor, denoted by L is then used in the following architectures:

Plain direct regression: This architecture uses standard CNN machinery for the inference. the RoI tensor representation of size $7 \times 7 \times 256$ runs through 5 convolutions of 3×3 with 512 maps each and then up-sampled to $56 \times 56 \times 8$. The tensor L is added to get a $56 \times 56 \times 11$ tensor. Following this 3 convolutions of $3 \times 3 \times 11 \times 8$ are applied, with L re-added each time to enable repeated depth reasoning. Finally, two global layers of 1024 and 128 neurons are applied followed by a single output neuron. This architecture is comparable in its number of neurons and computational budget to the interest point finding module. It is trained to minimize the l_2 loss w.r.t the true 3D length

Task	Train	2D test	3D test	Measured objects
Tree height	577 (757)	157 (175)	103 (103)	103
Leaf measurements	454 (4409)	320 (1303)	276 (1032)	1032
Cucumber length	446 (2261)	152 (1049)	89 (204)	204

Table 1

Number of images and objects (in parentheses) used for training, 2D testing and 3D testing. The right column presents the number of objects measured by a ruler in each task

measurements.

Heat-map based direct regression: As in the three staged pipeline, the following algorithm offers length computation which is reduced to an interest point finding problem (in contrast to the algorithm above), but with two main differences: there is no supervision for interest points, and length is computed directly by the network instead of a post-processing procedure.

If the locations of the two edge points (i_1, j_1) and (i_2, j_2) in the tensor L are known, the distance d between the points is given by:

$$d^2 = \|L(i_1, j_1, :) - L(i_2, j_2, :)\| = \sum_{d=1}^{3} \Big(L(i_1, j_1, d) - L(i_2, j_2, d) \Big) \tag{9}$$

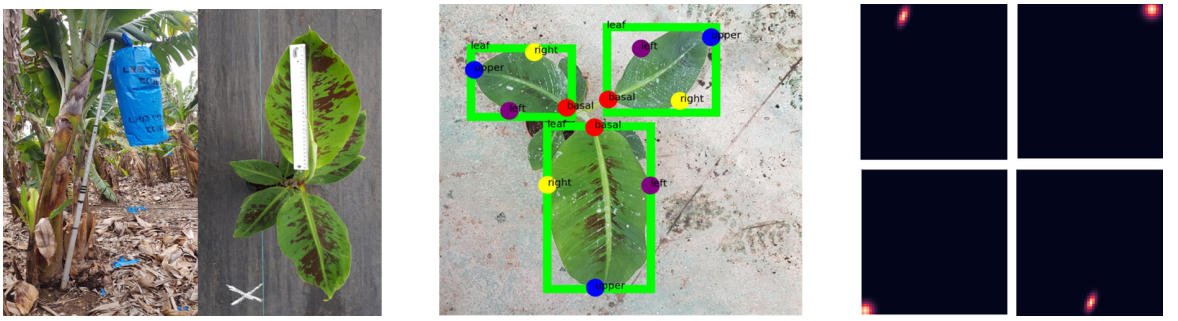
The problem can hence be posed as inferring these locations of the two edge points by using two $56x56$ interest point maps P_1, P_2, where the value P_K in location (i_k, j_k) is 1 and 0 otherwise for $K = 1, 2$. We have

$$d = \sqrt{\sum_{d=1}^{3} \Big(P_1 \cdot L(\ :, :, d) - P_2 \cdot L(\ :, :, d) \Big)^2} \tag{10}$$

Where $A \cdot B$ here is an inner product between vecotorized A and B. This computation can be done in a fixed parameter-less layer. The length output is supervised during training, with its l_2 distance to ground truth length measurements minimized. The keypoint finding head architecture (see Figure 2) is used to infer the heat maps P_1, P_2. A spatial softmax operation is then applied over them, followed by a layer implementing Eq. 10. This architecture is more informed that the plain architecture, with the distance computation encoded into its structure.

4. Results and Discussions

The data sets collected are described in section 4.1, followed by 2D accuracy results in 4.2 and 3D results in 4.3. In section 4.4 algorithmic alternatives and visual error analysis are discussed. In section 4.5 leaf measurements are used for treatment and mutant classification phenotyping.



(a) (b) (c)

Figure 5: Ground truth collection. a: Measuring the banana tree and banana leaves with a ruler (cucumbers were measured in a similar manner). **b:** Annotated ground truth bounding box and interest points of leaves. **c:** Ground truth Gaussian heat maps derived for the interest points of the upper-right leaf presented in b.

4.1. Dataset and Annotations

For all problems, images were captured using two types of sensors. RGB images for training were taken with a Canon HD camera with 4032×3024 resolution. RGB-D images with resolution 1280×720 were taken using the Intel RealSense D435 sensor, based on infra-red active stereoscopy technology. This sensor has a global shutter enabling good performance in highly dynamic conditions [8], and was shown to perform well outdoors for phenotyping tasks [37]. For Banana trees, images of banana trees at the fruit harvest stage were taken in several plantations in the north of Israel. The plantations differ w.r.t the banana varieties, including Gal, Grand-Naine, Adi and Valerie. Cucumber plant images were taken in the south of Israel in greenhouse conditions. For the banana leaves, images were taken in greenhouses in the North of Israel. The potted plants were approximately three months old, a stage which enables to determine if the plant has a mutation based on its leaf's 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 the D435 for which we took manual ground truth length measurements using a ruler of the objects: banana tree height, cucumber length, and banana leaf length and width. Figure 5 (Left) shows the rulers and the measuring procedure. This test set is used for 3D evaluations. RGB images were annotated with bounding boxes and interest point locations. All annotated objects were measurable, i.e., all their interest points were visible. Based on the 2D interest points annotations, Gaussian ground truth heat maps were constructed, as showed in Figure (right) 5. The amount of measured objects by a ruler can be seen in the right column of Table 1.

Leaf measurements based phenotyping: 204 of the potted plants were part of a fertilizer stress experiment in which the pots received four different treatments (see Figure 6, top row), 51 pots per treatments. The difference between

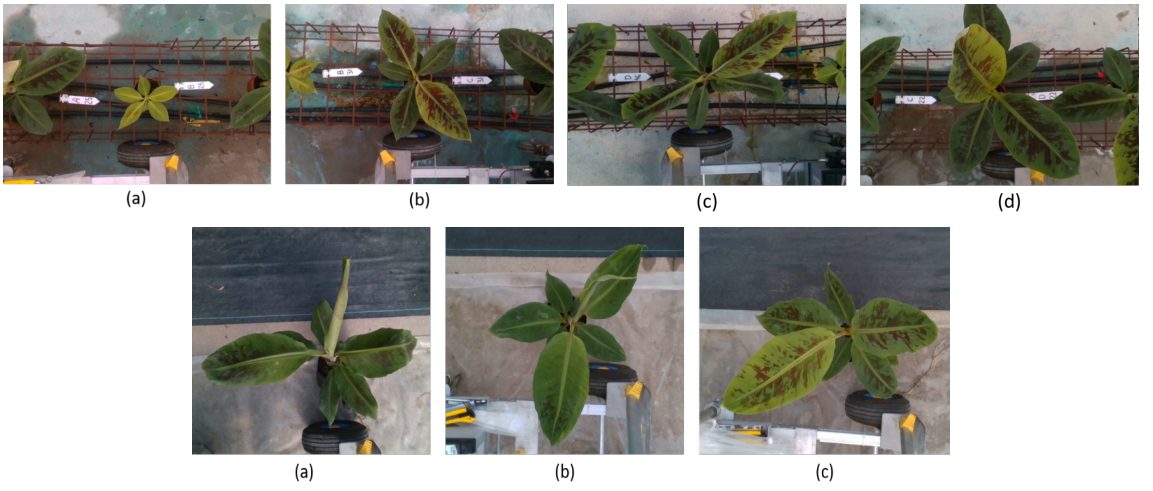


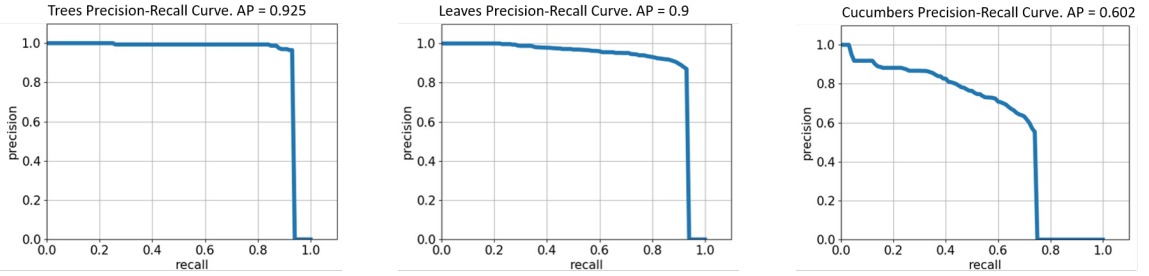
Figure 6: Examples of banana pot treatments and mutations. **Top:** four banana pots, each of which received a different treatment. The letters below designate the treatment, ranging from complete lack of fertilizer in (a) to high fertilizer concentration in (d). See the text for full treatments description. **Bottom:** Examples of two mutants pots and one normal pot. **a:** pot of *masada* mutant. **b:** pot with *dwarf* mutant. **c:** normal pot.

the treatments is in the ratio of water and fertilizer. During the experiment, each pot had been watered for 10 minutes every day. In addition to the watering, there were four three liter tanks connected with a pump to the irrigation system, and their content was added with constant flow. The treatments differ with respect to the content of the tanks, which were: a) no fertilizer at all (3 liters of water) , b) two liters of water and one liter of fertilizer, c) 1.5 liter of water and 1.5 liter of fertilizer, and d) no water (three liters of fertilizer). From a different greenhouse, 44 pots were examined by an expert and classified according to a mutation type: *dwarf* type, *masada* or normal [34]. 15 pots were found as *dwarf*, 14 as *masada* and 15 as normal (see Figure 6, bottom row).

4.2. Metrics and Results in 2D

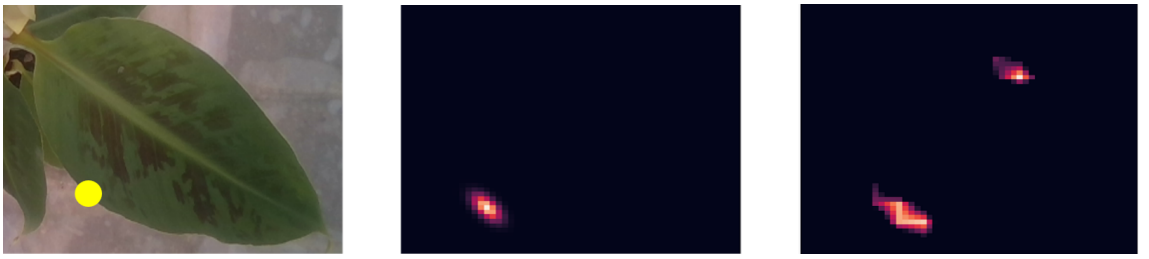
Detection Rates: The network detector provides a probability-like confidence score for each detection, and the two types of errors, miss detection and false positives, can be traded using a detection threshold. Detection performance is measured by drawing the recall-precision trade off graph and measuring the area below it, termed Average Precision (AP) [25]. The recall-precision graphs for the three tasks can be seen in Figure 7. We used a default threshold of 0.5 over the confidence score in our pipeline.

Detection probability: For the banana tree problem, the average precision (AP) was 0.925. The detector successfully identified 165 of 175 trees, with only 5 false positives. In each of the detected trees the two interest points were found successfully. The AP for banana leaf detection was 0.9, successfully identifying 1227 leaves of 1303, with 36 false positive. All interest points of type 'basal' and 'apex' were detected, but interest points of the type 'right' and



(a) Banana trees (b) Banana leaves (c) Cucumber fruit

Figure 7: Recall-Precision graphs for object detection in the three tasks. Average Precison (AP) scores are stated above the graphs.



(a) Right interest point (b) Gt heat map (c) Inferred heat map

Figure 8: : Confusion between right and left leaf points. a:. A cropped image of banana leaf with right interest point annotated. **b:** The ground truth Gaussian heat map for the right interest point. **c:** The Gaussian heat map inferred by the network. 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).

'left' were not found in 4 detected leaves. Observing the predicted heat map in Figure 8, it can be seen that in this case, the interest points finder struggled to distinguish between the left and the right interest points. To overcome this difficulty, we search for the most likely location only in the half space relevant to the interest point, as determined by the apex and basal point positions. The AP for cucumbers detection was 0.602, successfully identifying 822 cucumbers out of 1049, with 348 false positives. Interest points were detected successfully in almost all detected cucumbers, with five cases of undetected 'upper' point and ones case of undetected 'basal' point.

Localization accuracy: The Euclidean distance between the detected 2D point location and the ground truth location (deviations in pixels) was computed. As this measure depends on image resolution, its units are arbitrary. Hence we report also the deviations normalized by the length (in pixels) of the relevant distance measured. The results can be seen in Table 2. As can be seen, the average point deviation is between 3% and 6% of the length measured. Estimation deviations in the tree length are lower than deviations in the leaf and cucumber length. The largest deviations

	Pixel error		Relative error	
interest point	μ	σ	μ	σ
tree upper	60.33	70.79	0.03	0.02
tree basal	91.58	137.18	0.04	0.04
leaf apex	43.13	35.25	0.04	0.02
leaf basal	46.96	33.32	0.04	0.02
leaf left	59.03	98.68	0.06	0.08
leaf right	57.37	74.1	0.06	0.06
cucumber basal	37	37.3	0.05	0.05
cucumber upper	28.75	49.87	0.05	0.05

Table 2

Interest point localization error on the 2D test set. Average and standard deviation of the Euclidean distance between ground truth and detected interest points are reported, in units of pixels and fraction of the object length. It can be seen that deviations are less than 6% of the object length for all interest points.

	Algorithmically detected interest points							Manually marked interest points						
	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	12.35	10.44	1.03	0.043	0.03	0.004	0.7	10.23	8.17	0.81	0.036	0.03	0.003	0.8
leaf length	1.26	1.28	0.042	0.058	0.051	0.002	0.92	0.88	1.04	0.034	0.0431	0.051	0.002	0.95
leaf width	0.93	0.98	0.032	0.089	0.082	0.003	0.88	0.72	1.07	0.035	0.073	0.097	0.003	0.87
cucumber length	1.21	1.34	0.1	0.078	0.074	0.005	0.95	0.66	1.17	0.089	0.038	0.05	0.003	0.97

Table 3

Left: 3D Length estimation errors obtained with the proposed pipeline. **Right:** Length estimation errors obtained by using the manually annotated interest points, shown for comparison. SE denotes standard error.

were measured for the left and right leaf interest points, which are clearly more difficult to detect, as they are positioned somewhat arbitrarily along the leaf curve.

4.3. Metrics and Results in 3D

Object lengths estimated by the suggested pipeline were compared to the true 3D lengths as measured by a ruler. Table 3 shows the estimation deviations for each of the tasks. As can be seen, the mean relative deviation for all tasks is under 9% of the true length. The error is smallest for tree height, and largest for leaf width. 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 between the leaves and the camera. In addition, the left and right interest points are more challenging for accurate detection.

To further understand whether the deviations stem from the interest point detection algorithm, or rather from the inaccuracy introduced by the specific RGB-D sensor, the 3D length based on the manually marked ground truth interest points in the annotated images is also estimated. In this measurement the first two phases of the pipeline are skipped and only the last de-projection phase (applied to the manually annotated 2D points) is used to compute the 3D lengths of interest. Clearly, measurements based on ground truth points are more accurate. However, based on the standard

	Banana's tree height	Banana's leaf length	Banana's leaf width	Cucumber's fruit length
mean prediction (baseline)	0.09	0.35	0.3	0.42
Plain direct regression	0.076	0.11	0.16	0.23
Heat map based regression	0.085	0.23	0.25	0.41
KD + RDE	**0.048**	0.131	0.142	0.13
KD + LR	0.061	0.07	0.095	0.13
KD + RDE + LR	**0.048**	**0.059**	**0.085**	**0.09**

Table 4

Average relative deviation (percent of the total object length) of several alternative pipelines. The deviations of constant prediction of the mean length are given in the first line as a baseline. Results of the two direct length regression algorithms from section 3.4.2 are shown in lines 2,3. Line 4 presents the results of the full pipeline including Keypoint Detection (KD), Robust Depth Estimation (RDE), and Linear Regression (LR). Lines 5 and 6 present ablated versions of the full pipeline with LR or RDE omitted. When RDE is not used (line 6), the direct depth of the detected interest point is used unless it is zero. If it is zero, the first non-zero value in the direction of the opposite interest points is used. The results are of a 10-fold cross validation experiment, conducted only over the limited dataset for which 3D ground truth measurements are available (3D test in Table 1).

errors (column SE in Table 3), for tree height the difference between algorithmic estimates and ground-truth based estimates is not statistically significant. For leaf width and length, the difference is small, but statistically significant. For cucumber length, the error introduced by automated detection is about twice larger than the error from the manual markings. As cucumber interest points do not seem significantly harder to detect than tree or leaf points in 2D (see Table 2), the larger 3D deviation is most likely related to lower robustness of the depth measurements in the cucumber's interest points.

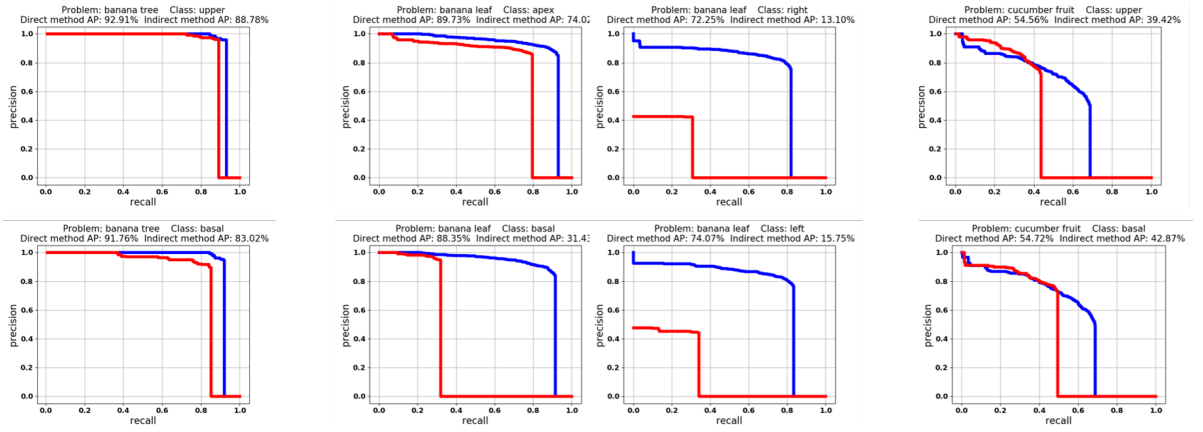
In order to estimate the statistical strength of our method, we compute the fraction of explained variance R^2:

$$R^2 = 1 - \frac{Var_{err}}{Var_{total}} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{11}$$

with $\{y_i\}_{i=1}^{N}$ the ground truth measurements, $\{\hat{y}_i\}_{i=1}^{N}$ the algorithm estimations and $\bar{y}$ the mean of $\{y_i\}_{i=1}^{N}$. This essentially compares our estimation method to the trivial estimation method of predicting the mean length for every instance. The results are summarized in Table 3. Our estimation explains (i.e. successfully infers) a significant portion (0.7 - 0.95) of the length variance.

4.4. Alternatives Algorithms and Error Analysis

Direct Length measurements: Average deviations of alternative length estimation algorithms and ablations of the three-stage pipeline appear in Table 4. As the direct depth regression algorithms use 3D ground truth measurements in their training process, all the algorithms in this comparison were trained using the limited dataset for which 3D annotations were available (see Table 1). It can be seen that the three-staged pipeline outperforms the direct length



(a) trees problem (b) leaves problem (c) cucumbers problem

Figure 9: Precision-Recall curves. Curves are shown in blue for the two stage method, in which the object is first found and than interest points, and in red for the direct method, where interest points are directly detected without first determining the object context ((section 3.4.1)). **a:**. Curves for banana tree points: upper (top) and basal (bottom). **b:** Curves for banana leaf points: apex (top left), right (top right), basal (bottom left) and left (bottom right). **c:** Curves for cucumber fruit points: upper (top) and basal (bottom).

regression methods. While this result may change for significantly larger datasets (differing by one or more orders of magnitude), it shows that in the practical range of several hundred annotated objects, indirect training with interest points is preferable. Both direct methods are better than the baseline prediction, indicating that they were able to learn actual predictors from the data. The heat map based method is worse than the plain direct regression, which is somewhat surprising, as the former has a architecture more tailored to the task. It seems that finding the interest points without guiding annotation is more difficult than direct length regression based on holistic features. The robust depth estimation procedure and the final linear regression contribute both to accuracy improvements.

Interest point finding: Interest point detection rates of the three-staged pipeline are compared to direct interest point detection as described in Section 3.4.1. An inferred interest point is considered a hit if its distance from the ground truth interest points is less than 10% from the object size. Figure 9 presents precision-recall curves for the various interest point types used in our problems. Considering Average Precision as the summary statistic, the two stage method of object detection followed by interest point estimation is preferable to direct interest point detection for all interest point types. For leaves left and right keypoints, the direct detector fails to discriminate between them, and its precision is therefore bounded by 0.5. For the cucumber interest points, there is a range of recall levels for which the direct detector is preferable (low recall levels, specifically 0.1-0.3). For applications emphasizing the most easy-to-find objects (cucumbers/ cucumber points), it is hence easier to find cucumber points directly than to identify the entire cucumber.

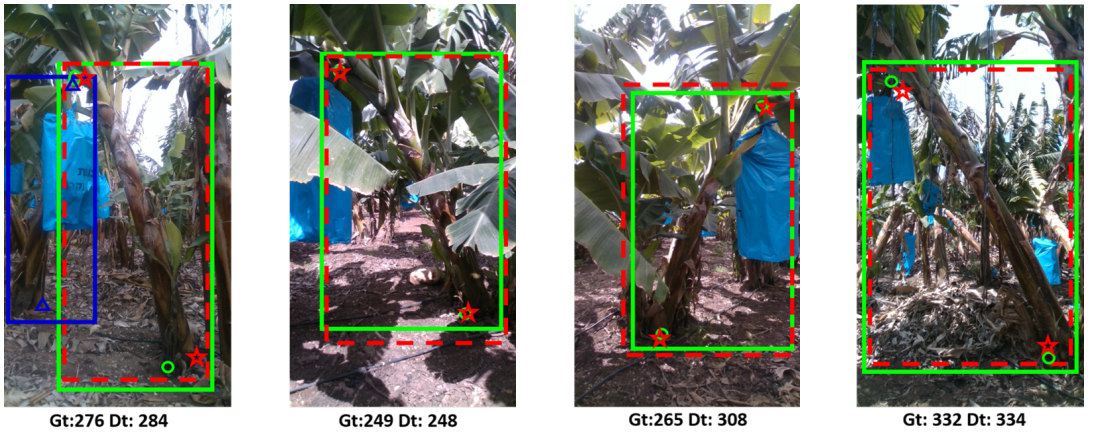


Figure 10: Detection examples in the banana trees problem: Green denotes ground truth, red denotes successful detection, and blue denotes false positives. Below each image we show the height measured by ruler (Gt) and the estimated height by our model (Dt). **a:** a false positive example, caused by a wrong association of a peduncle from one tree to the trunk of another. **b:** An accurate estimation example. **c:** interest points are detected with small 2D deviations, yet height deviation is high due to depth estimation problems. **d:** A case with significant deviations in 2D, yet good height estimation.

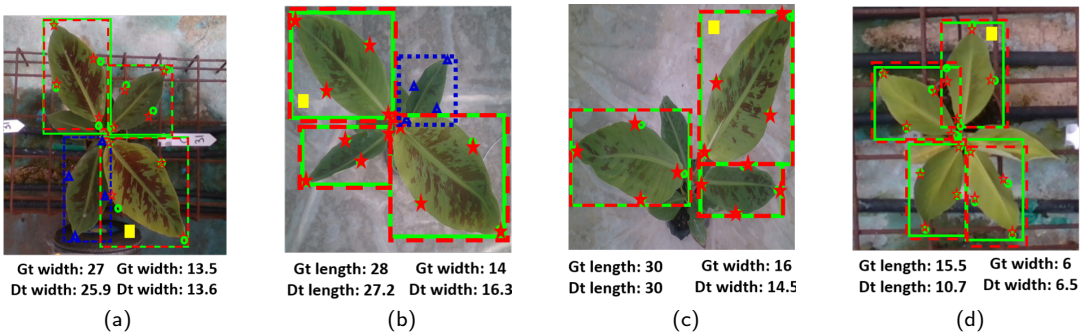


Figure 11: Detection examples in banana leaves problem: Ground truth objects are marked in green, detected objects in red, false detection in blue. Ground truth (Gt) and estimated (Dt) lengths are stated below each image for the leaf marked by a yellow rectangle. **a:** An accurate estimation example. **b:** A relatively large deviation in leaf width despite accurate 2D interest point localization. In addition, the network falsely detects a leaf as a 'measurable leaf' (the left upper one) but his interest points are partially occluded. This is hence a false positive. **c:** A large 2D deviation of the apex point, yet the length estimation is perfect. **d:** Large 2D deviations in all points, leading to corrupted length estimation but fine width estimation.

Figures 10, 11 and 12 present successful detection examples alongside various errors of our full pipeline for banana trees, banana leaves and cucumbers fruits problems, respectively. For the detection task, typical errors include confusion between measurable and non-measurable objects, which is a fine distinction. For length measurements,

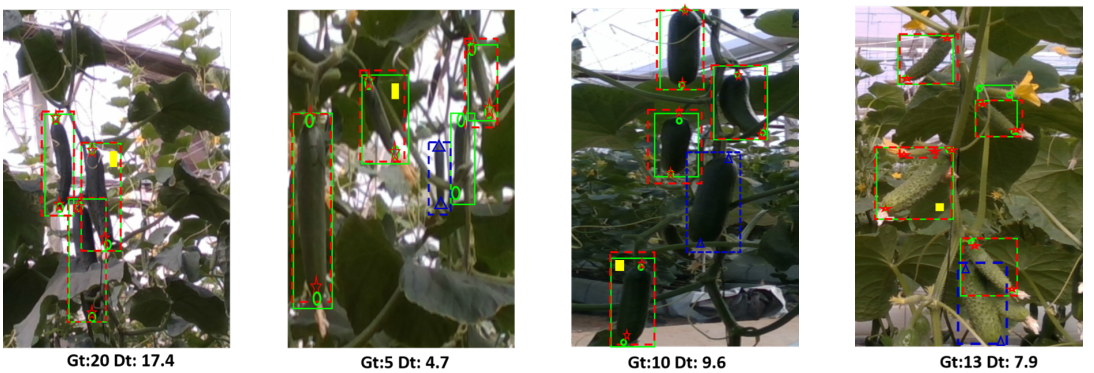


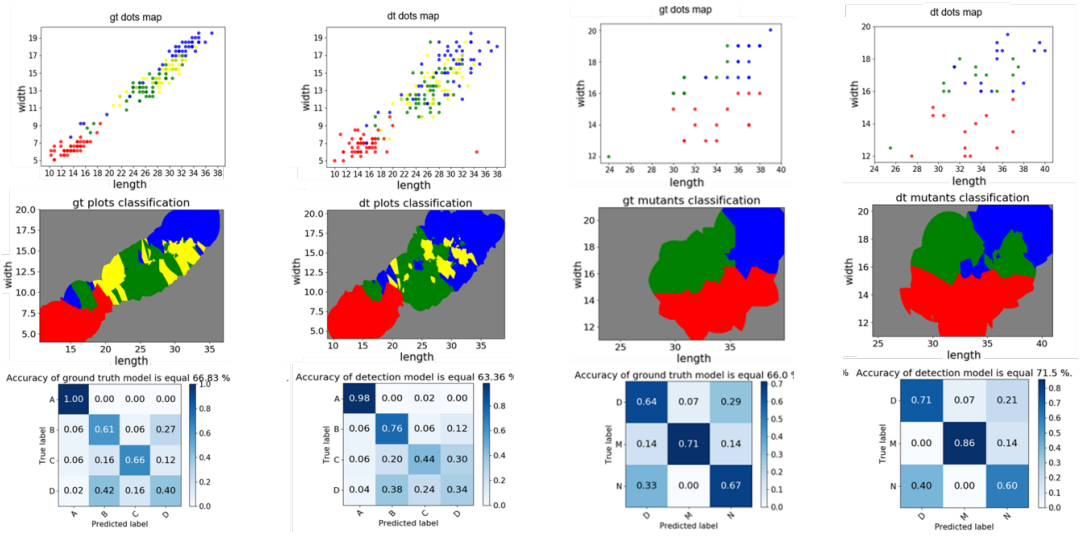
Figure 12: Detection examples in the cucumber fruits problem: Ground truth objects are marked in green, detected objects in red, false detection in blue. Ground truth (Gt) and estimated (Dt) lengths are stated below each image for the cucumber marked by a yellow rectangle. **a:** A 2D deviation of the basal interest point leads to reduced length estimation. **b:** The marked cucumber is accurate estimated. Yet this image contains a cucumber missed by the network, and a cucumber detected by the network which was not annotated (an annotation error). **c:** This cucumber was successfully estimated despite significant deviation in detection of the basal point. the cucumber in blue was not annotated as 'measurable' since its basal point is slightly hidden. **d:** Large 3D length estimation despite good 2D interest points localization.

sometimes there are significant 3D length deviations despite obtaining accurate 2D points localization.

4.5. Length-Based Leaf Phenotyping

We tested the utility of leaf width and height measurements for discrimination among plant treatment and mutation. Only the leaf which was first to grow, and hence the biggest, of each potted plant was considered. We mapped each plant to the two features of leaf length and width, and tested whether a distinction between treatment and mutant classes can be made based on them. The distinction was approached as a classification task, using a KNN classifier in the two dimensional measurement space. K, the number of neighbors, was chosen using a 5-fold cross validation scheme. We examined both classifications using the ground truth measurements (measured by a ruler) and classification using the feature estimates provided by the proposed algorithm. The results are shown in Figure 13, with the main findings summarized below.

Stress phenotyping: The ground-truth KNN model shows that treatment A and D are easier to discriminate, because treatment A results in small leaves and treatment D with the largest ones. Treatments B and C are difficult to identify based on leaf measurements alone. The model based on the algorithmic pipeline is able to perform the same distinctions, though with lower accuracy for class D. Specifically, it characterizes class A as plants with short width (less than 11 cm) and length (less than 16 cm). Plants with width longer than 17 cm and length longer than 30 are inferred to get treatment D. The overall accuracy of the algorithmic model is 63.36%, rather close to the accuracy of



(a) Stress - ground truth model

(b) Stress - estimation model

(c) Mutant - ground truth model

(d) Mutant - estimation model

Figure 13: Length-Based leaf classification: **top row:** Scatter plot of the sample in the two dimensional space of (leaf length, leaf width) for ground truth leaf measurements (a,c) and algorithmic estimations (b,d). **middle row:** Identified clusters by the KNN classifier for ground truth leaf measurements (a,c) and algorithmic estimations (b,d). **bottom row:** confusion matrix for KNN classification using ground truth leaf measurements (a,c) and algorithmic estimations (b,d). For the treatment classification red denotes treatment A (no fertilizer), green denotes B (2:1 ratio of water:fertilizer), yellow denotes C (1:1 ratio), and blue denotes D (full fertilizer). For mutant classes red is the *masada* mutant, green is the *dwarf* mutant, and blue is 'normal'. The gray area in the classification maps includes locations for which at least one of the K closest neighbors is more than 2.5 units away.

the ground-truth model- 66.83%

Mutation phenotyping: The model based on ground truth measurements show that *masada* mutant is easy to identify and is characterized by marrow leave (width lower than 16 cm), with 86% of accuracy in identifying this mutant. The distinction between 'normal' and the *dwarf* mutant is harder. The model based on the algorithmic measurements is very similar to the model obtained from manually measured ground truth, and its total accuracy is even slightly higher (71.5% against 66%).

5. Concluding Remarks

We presented a general technique for length-based plant phenotyping, based on interest point detection in 2D and depth information from a low-cost 3D sensor. The technique was tested on three specific estimation tasks: banana tree height, banana leaf length and width, and cucumber length. Our method obtained an average deviation of 4.3% (of the total height) for tree height estimation, 8.9% and 5.8% for leaf width and length estimation, and 7.8% for cucumber

length. Statistically, the method was able to explain (infer) 70%-95% of the total length variance. The technique was compared to several alternatives and shown to provide better accuracy. While demonstrated for three specific problems, the problem variety suggests that our method can be used in additional phenotyping tasks, requiring only minor task-dependent adaptations. In addition, the measurements potential value was demonstrated for the leaf phenotyping tasks of mutant and stress classification.

There are several possible paths to advancing this work. First, more data can be used in the problems we considered here to obtain improved detection and interest point localization, leading to length estimation accuracy improvements. The benefits of this direction are bounded, tough, as a main source of error is the depth sensor accuracy. The latter can be improved by using a more accurate depth camera, or by fusion with a hi-resolution stereo, for getting more accurate depth information. To demonstrate the generality of the method beyond the three tested problems, we are currently pursuing additional and more challenging tasks such as stem inter-branch length measurements. Beyond length measurements, it would be of high interest to develop algorithms extending to measurements of areas and volume-based phenotypes. Looking forward, the proposed method can be embedded in a flexible, general phenotyping system as a length estimation module.

Acknowledgments

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 and Lena Karol from Hazera Genetics for many helpful discussions and for providing the images and annotations.

References

[1] Nan An, Christine M Palmer, Robert L Baker, RJ Cody Markelz, James Ta, Michael F Covington, Julin N Maloof, Stephen M Welch, and Cynthia 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] José Luis Araus and Jill E Cairns. Field high-throughput phenotyping: the new crop breeding frontier. *Trends in plant science*, 19(1):52–61, 2014.

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

[4] Andrew J Challinor, J Watson, David B Lobell, SM Howden, DR Smith, and Netra Chhetri. A meta-analysis of crop yield under climate change and adaptation. *Nature Climate Change*, 4(4):287, 2014.

[5] Yann Chéné, David Rousseau, Philippe Lucidarme, Jessica Bertheloot, Valérie Caffier, Philippe Morel, Étienne Belin, and François Chapeau-Blondeau. On the use of depth camera for 3d phenotyping of entire plants. *Computers and Electronics in Agriculture*, 82:122–127, 2012.

[6] Fabio Fiorani and Ulrich Schurr. Future scenarios for plant phenotyping. *Annual review of plant biology*, 64:267–291, 2013.

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

[8] Silvio Giancola, Matteo Valenti, and Remo Sala. *A Survey on 3D Cameras: Metrological Comparison of Time-of-Flight, Structured-Light and Active Stereoscopy Technologies*. Springer, 2018.

[9] A Gongal, M Karkee, and S Amatya. Apple fruit size estimation using a 3d machine vision system. *Information Processing in Agriculture*, 5(4):498–503, 2018.

[10] 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.

[11] Jungong Han, Ling Shao, Dong Xu, and Jamie Shotton. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE transactions on cybernetics*, 43(5):1318–1334, 2013.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[13] Meng-Han Hu, Qing-Li Dong, Pradeep K Malakar, Bao-Lin Liu, and Ganesh K Jaganathan. Determining banana size based on computer vision. *International journal of food properties*, 18(3):508–520, 2015.

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

[15] Yotam Itzhaky, Guy Farjon, Faina Khoroshevsky, Alon Shpigler, and Aharon Bar Hillel. Leaf counting: Multiple scale regression and detection using deep cnns, 2018.

[16] Li Jiang, Shuangshuang Yan, Wencai Yang, Yanqiang Li, Mengxue Xia, Zijing Chen, Qian Wang, Liying Yan, Xiaofei Song, Renyi 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.

[17] Yu Jiang, Changying Li, Andrew H Paterson, Shangpeng Sun, Rui Xu, and Jon Robertson. Quantitative analysis of cotton canopy size in field conditions using a consumer-grade rgb-d camera. *Frontiers in plant science*, 8:2233, 2018.

[18] Jihui Jin, Gefen Kohavi, Zhi Ji, and Avideh Zakhor. Top down approach to height histogram estimation of biomass sorghum in the field. *Electronic Imaging*, 2018(15):288–1, 2018.

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

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[21] Ran Nisim Lati, Sagi Filin, Bashar Elnashef, and Hanan Eizenberg. 3-d image-driven morphological crop analysis: A novel method for detection of sunflower broomrape initial subsoil parasitism. *Sensors*, 19(7):1569, 2019.

[22] Lei Li, Qin Zhang, and Danfeng Huang. A review of imaging techniques for plant phenotyping. *Sensors*, 14(11):20078–20111, 2014.

[23] Guichao Lin, Yunchao Tang, Xiangjun Zou, Jinhui Li, and Juntao Xiong. In-field citrus detection and localisation based on rgb-d image analysis. *Biosystems Engineering*, 186:34–44, 2019.

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[25] L. M. Everingham, Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303âĂŞ338, 2010.

[26] Xiaodan Ma, Kexin Zhu, Haiou Guan, Jiarui Feng, Song Yu, and Gang Liu. Calculation method for phenotypic traits based on the 3d

reconstruction of maize canopies. *Sensors*, 19(5):1201, 2019.

[27] Annalisa Milella, Roberto Marani, Antonio Petitti, and Giulio Reina. In-field high throughput grapevine phenotyping with a consumer-grade depth camera. *Computers and Electronics in Agriculture*, 156:293–306, 2019.

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

[29] Nur Badariah Ahmad Mustafa, Nurashikin Ahmad Fuad, Syed Khaleel Ahmed, Aidil Azwin Zainul Abidin, Zaipatimah Ali, Wong Bing Yit, and Zainul Abidin Md 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.

[30] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[31] Stefan Paulus, Jan Behmann, Anne-Katrin Mahlein, Lutz Plümer, and Heiner Kuhlmann. Low-cost 3d systems: suitable tools for plant phenotyping. *Sensors*, 14(2):3001–3018, 2014.

[32] Rajeev Ranjan, Vishal M Patel, and Rama 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.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[34] John Charles Robinson and Victor Galán Sáuco. Nursery hardening of in vitro-produced banana plants. *Fruits*, 64(6):383–392, 2009.

[35] David Rousseau, Hannah Dee, and Tony Pridmore. Imaging methods for phenotyping of plant traits. In *Phenomics in Crop Plants: Trends, Options and Limitations*, pages 61–74. Springer, 2015.

[36] David Tilman, Christian Balzer, Jason Hill, and Belinda L Befort. Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences*, 108(50):20260–20264, 2011.

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

[38] Wangxia Wang, Basia Vinocur, and Arie Altman. Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. *Planta*, 218(1):1–14, 2003.

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

[40] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 2019.

[41] Gerhard Zotz, Peter Hietz, and Gerold 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.