In-field Segmentation and Identification of Plant Structures using 3D Imaging

Paloma Sodhi, Srinivasan Vijayarangan and David Wettergreen

***Abstract*— Automatically correlating plant observable characteristics to their underlying genetics will streamline selection methods in plant breeding. Measurement of plant observable characteristics is called phenotyping, and knowing plant phenotypes accurately and throughout a plant's growth is central to making breeding decisions. In-field plant phenotyping in an automated and noninvasive manner is hence crucial to accelerating plant breeding methods. However, most of the existing methods on plant phenotyping using visual imaging are confined to controlled greenhouse environments.**

This paper presents an automated method of mapping 2D images collected in an outdoor sorghum field to segmented 3D plant units that are of interest for phenotyping. This method leverages multiple horizontal and vertical viewpoints while capturing 2D images from a robotic platform so as to generate in-field 3D reconstructions of the sorghum plant. We develop and quantitatively evaluate segmentation methods on these 3D reconstructions and also compare against reconstructions obtained from a controlled greenhouse environment. We present analysis that contrasts the role of purely local geometric features and the effect of addition of global context in both datasets. This work furthers capabilities of in-field phenotyping which paves the way forward for plant biologists to study the coupled effect of genetics and environment on improving crop yields.

I. INTRODUCTION

Plant phenotyping produces quantitative measurements of observable plant traits. Examples of phenotypic traits include plant height, stem diameter, leaf area, and leaf angle. The ability to correlate phenotypic traits with their genotypes plays a crucial role in improving plant breeding techniques.

Pheontyping in field environments is a *bottleneck* in the plant breeding pipeline and high throughput automated methods are crucial to improved production [1]. The key ability of high throughput phenotyping lies in non-destructively capturing plant traits in an automated manner so as to achieve imaging rates of a minimal hundreds of plants per day [1]. There is an increasing interest in deploying robotic platforms and computer vision methods to achieve these phenotyping rates and hence relieve the bottleneck [1]. The TERRA program by the U.S Department of Energy seeks to develop high-throughput phenotyping methods for accelerated breeding of advanced biofuel crops like *sorghum*.

This paper addresses the problem of automated segmentation and extraction of plant subunits called *phytomers* for the sorghum plant. A plant phytomer unit consists of a leaf, its sheath and the stem segment on which the leaf resides and can be thought of as a functional building block of the plant. It is of special significance for phenotyping, since one can estimate phenotypic traits like leaf angle, stem diameter, leaf length, leaf width, and leaf area from this structure.

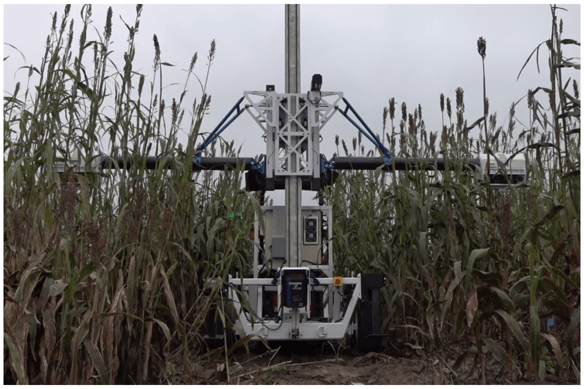


Fig. 1. High throughput phenotyping platform deployed in a sorghum field outdoors. The platform has two arms on either side that extend outwards into crop rows and scan plants vertically at a resolution of 10 cm. Each arm is fitted with two multi-camera sensorpods facing front and back.

A high throughput phenotyping platform developed by National Robotics Engineering Center is shown in Fig. 1. The uniqueness of the platform lies in being able to collect 2D images of a plant at multiple vertical heights from different horizontal viewpoints. Consider now the problem of extracting from a stream of 2D images, a set of 3D segmented phytomer units. The problem becomes especially challenging for field environments, where getting high quality images is difficult due to lighting variations at different heights of the plant canopy. Occlusions, lack of texture and non-rigid body motion of plant structures due to wind further add to the challenges of reliably extracting and matching 2D image features for the purpose of 3D reconstruction.

In this paper, we present an approach for robust extraction of 3D phytomers from raw 2D images of sorghum plants. The key components of the approach are as follows: 3D reconstruction of the plant using multi-view imaging (Section III-B), robust plant segmentation using a mixture of local and global features (Section III-C, III-D) and finally plant phytomer extraction using 3D geometric algorithms (Section III-E). We demonstrate the efficacy of our approach in both indoor greenhouse environments and unstructured outdoor field environments. The main contributions of this paper are:

1) In-field 3D reconstruction and segmentation from a unique and challenging dataset leveraging multiple horizontal and vertical views of the sorghum plant.
2) Qualitative and quantitative comparison of the 3D plant phytomer extraction pipeline for greenhouse and field.

Paloma Sodhi, Srinivasan Vijayarangan, David Wettergreen are with The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA {psodhi,svijaya1,dsw}@ri.cmu.edu

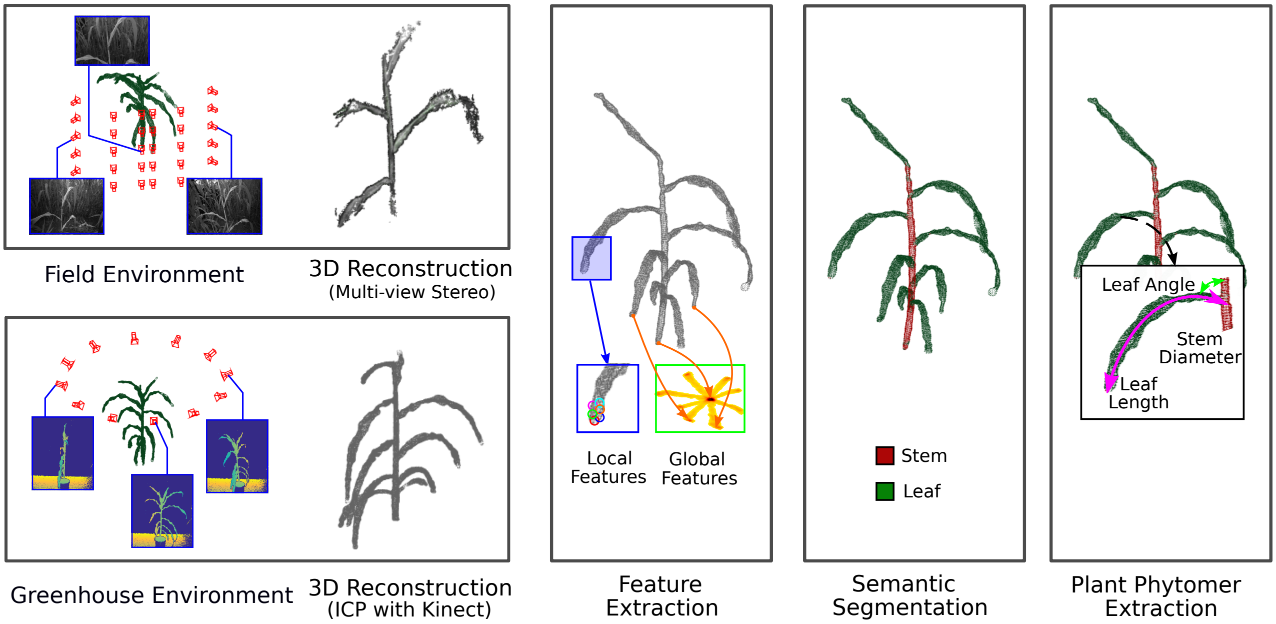


Fig. 2. Approach mapping input plant images to segmented 3D plant phytomer units. The first two stages take plant images captured from multiple viewpoints and generate a 3D point cloud reconstruction of the plant. These two steps differ for greenhouse and field environments. Once the 3D reconstruction is obtained, local and global 3D point features are extracted. The next stage uses a machine learning classifier to assign a semantic class label to each 3D point. Finally, the plant phytomer is extracted from the segmented point cloud.

3) Analysis of the role of purely local geometric features and the effect of addition of global context in both greenhouse and field environments.

II. RELATED WORK

High-throughput phenotyping platforms deploy a variety of imaging modalities like 2D visible imaging, 3D imaging, multispectral imaging, thermal infrared imaging and flourescence imaging [2]. Given its low cost and ease of operation, 2D/3D visible imaging has been commonly used for applications like plant mapping, weed control, fruit counting and yield estimation. For the purpose of phenotyping, the use of 3D visible imaging is important in order to be able to make ground truth metric measurements purely from imaging. However, most of the current state-of-the-art in 3D plant reconstruction, segmentation and phenotyping is in controlled greenhouse environments [3], [4], [5], [6], [7].

Chéné, et al. utilize depth images from a Kinect to segment and extract parameters describing leaves of rosebush, yucca and apple plants in an indoor environment [7]. However, their parametric approach is tailored to top-down views of leaves of these plants in particular. Chaivivatrakul, et al. develop a comprehensive 3D reconstruction and phenotyping system but test it on only five corn plants in a controlled greenhouse setting using a Kinect [5]. Moreover, their segmentation methods utilize global shape fitting methods more suitable for uncluttered, indoor scenes.

Several 3D segmentation methods leverage local shape information by means of local geometric features [3], [6], [8]. Paulus, et al. use point feature histograms to distinguish between grapevine leaves/stems and between wheat stems/ears and provide detailed tabulated results [6]. They, however, collect all their data from a single plant imaged using a hand-held LiDAR in an uncluttered indoor environment.

Sa, et al. utilize 3D point feature histograms for peduncle detection in sweet peppers and validate their results across a variety of sweet pepper plants under clutter [3]. However, they too collect all their data in an indoor greenhouse environment using a Kinect, as a result of which the 3D reconstructions they work with are of high fidelity. Dey, et al. work with field 2D images collected from a grape orchard and perform 3D reconstruction and segmentation of plant organs based on local geometric features [8]. However, unlike sorghum, their application involves segmentation between fairly distinct geometric structures, i.e. grapes (spherical), grape vines (cylindrical) and grape leaves (planar). They are hence able to utilize a low dimensional 3D shape feature constructed using only surface curvature estimates.

Unlike, corn, wheat or grapes, sorghum is a fairly new crop of interest for developing automated segmentation and phenotyping methods. To the best of our knowledge, literature on sorghum phenotyping is sparse and fairly recent [4], [9]. Ribera, et al. utilize UAV imagery to estimate macro-phenotypic traits like plant location and densities [9]. The UAV imagery provides very coarse reconstructions that can not be utilized for reliably extracting individual plant traits. McCormick, et al. generate 3D reconstructions of greenhouse grown sorghum and correlate phenotypes like leaf angle to underlying plant genetics [4]. As detailed later, greenhouse data collected by McCormick, et al. has been used for analysis and comparison to field data in this paper. The focus of their work, however, is correlation of phenotypes to their underlying genotypes, rather than developing automated methods of phenotyping. In this paper, we focus on an automated approach for extracting segmented 3D phytomer units from field as well as greenhouse images.

III. APPROACH

A. Overview

This paper takes the approach of performing multi-view 3D reconstruction from plant images followed by classifying individual 3D points as a stem or a leaf. The overall approach mapping input plant images to 3D phytomers is illustrated in Fig. 2. The following subsections elaborate on the different stages of the overall approach. The first stage involves reconstructing a 3D point cloud of a plant using images captured from multiple viewpoints. The reconstruction step differs for plants in the greenhouse and in the field, as field environments place a constraint on the imaging modalities and the degree of control that can be placed on the environment. Having obtained the 3D reconstructions, the next step computes point-level 3D features using local geometries and a global distance metric to density modes. Each point is then classified as a stem or a leaf by learning a Support Vector Machine (SVM) decision boundary, followed by spatial smoothing using a Conditional Random Field (CRF). Finally, the semantically segmented 3D point cloud is processed to extract 3D plant phytomer units.

Performing segmentation on a 3D representation was chosen over segmenting in 2D due to the following reasons. Firstly, a 3D reconstruction using multiple views reduces the effect of occlusions and background clutter. Secondly, for the sorghum plant in particular, there exists greater distinction between stems and leaves in terms of their 3D geometries rather than just color. Lastly, the final objective of extracting phytomers is to perform metric measurements of various phenotypes. A metrically rectified 3D representation is much more suitable than a 2D image for such an objective.

B. 3D Reconstruction using multi-view imaging

The 3D reconstruction stage takes images of the sorghum plant captured from multiple viewpoints as input and produces a reconstructed 3D point cloud as output. We make a distinction between data collected in an *indoor greenhouse environment* and an *outdoor field environment* since the collected data poses different complexity levels for algorithms involving 3D reconstruction and segmentation. This is because field environments place many additional challenges over controlled greenhouse settings. Firstly, there is a constraint in the imaging modalities that can deployed in a field setting. Majority state-of-the-art 3D plant segmentation and phenotyping algorithms utilize Kinect like sensors or LiDARs, both of which are infeasible for our field settings since Kinect doesn't work outdoors and LiDAR units do not provide high spatial resolution or close range (<10cm). Secondly, getting high quality well-exposed images is difficult due to lighting variations at different heights of the plant canopy. Thirdly, occlusions and non-rigid body motion of plant structures due to wind in the field further add to the challenge of reliably extracting and matching 2D image features for doing 3D reconstruction.

1) In greenhouse environments: Ideally, to obtain a geometrically consistent representation of a plant we would like to leverage 360 degree views of the plant. This is possible to setup for a controlled greenhouse environment. McCormick, et al. [4] place sorghum plants on a turntable and capture 360 degree view depth images at 30° increments using a Kinect camera. The relative poses between the camera and the plant can be seen in Fig. 2. The multiple depth images obtained are then fused together into a single 3D point cloud using the iterative closest point (ICP) algorithm.

2) In field environments: A close-up of one of the multi-camera sensor pods deployed between sorghum crop rows is shown in Fig. 7. The sensor pod contains eight forward facing cameras arranged in two rows along with two additional cameras verged on either ends at an angle of 30°. The sensor pod is mounted on a robotic arm that moves vertically to collect 2D images at multiple plant heights inside the canopy.

We utilize the Multi View Environment (MVE) proposed in [10] for generating 3D point clouds of the sorghum plant from 2D grayscale image sequences captured in the field. The multi-view environment framework begins by taking as input 2D images and applying structure-from-motion technique to reconstruct camera parameters (motion) and a sparse set of 3D scene points (structure). It then computes a depth map for each input image using multi-view stereo [11]. Finally, all

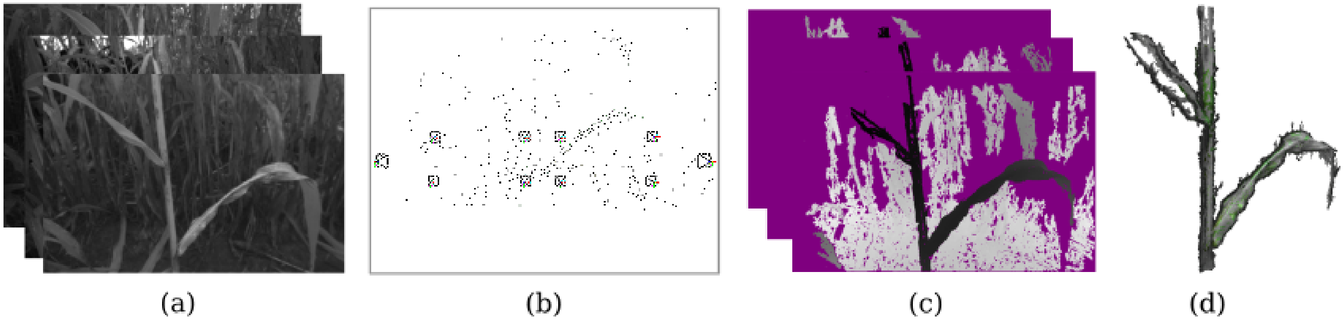


Fig. 3. Multi-view Environment Framework : (a) shows input images from the 10 cameras on the sensorpod. (b) shows output from the structure-from-motion stage that takes in the 2D images and computes camera poses and a sparse set of triangulated 3D points. (c) shows depth maps computed using multi-view stereo for each of the input images. Finally, (d) shows the 3D point cloud reconstructed by combining multiple depth maps.

depth maps are combined to obtain a dense 3D reconstruction of the scene. Fig. 3 illustrates different stages of the MVE framework applied to sorghum plant images from the field.

C. Semantic Segmentation of 3D Point Clouds

The semantic segmentation stage takes as input a 3D point cloud of a plant and uses a support vector machine (SVM) classifier to assigns a stem/leaf class label to each 3D point. Such a semantic representation is important so as to be able to compute phenotypic traits specific to the stem or the leaf.

1) Extraction of Local Features: Point feature representations like surface normals and curvature estimates as used in [8] are somewhat basic in their representations of the geometry around a specific point. To formulate a feature space beyond such representations, like the Point Feature Histogram introduced later, the concept of *dual-ring neighborhood* needs to be defined [12]. Let $\mathcal{P}$ be a set of 3D points with $\{x_i, y_i, z_i\}$ geometric coordinates. A point $\mathbf{p}_i \ \epsilon \ \mathcal{P}$ has a dual-ring neighborhood, $\mathcal{P}^{k_N}$, $\mathcal{P}^{k_H}$, that are defined as,

$$\begin{aligned} &(\exists)\ r_N, r_H \ \epsilon\ \mathbb{R},\ r_N < r_H, \ \text{ such that,} \\ &\mathbf{p}_j \ \epsilon \begin{cases} \mathcal{P}^{k_N} & \text{if } \left|\left|\mathbf{p}_j - \mathbf{p}_i\right|\right|_2 < r_N \\ \mathcal{P}^{k_H} & \text{if } \left|\left|\mathbf{p}_j - \mathbf{p}_i\right|\right|_2 < r_H \end{cases} \\ &\text{with, } 0 < k_N < k_H \end{aligned} \tag{1}$$

The radii r_N, r_H represent two different layers of feature representation for point $\mathbf{p_i}$. The first layer, $\mathcal{P}^{k_N}$, encodes surface normal and curvature estimates, obtained by performing principal component analysis on neighborhood patch $\mathcal{P}^{k_N}$. The second layer, $\mathcal{P}^{k_H}$, can encode Point Feature Histogram (PFH) and Fast Point Feature Histogram (FPFH) representations [13], that are based on relationships between points in $\mathcal{P}^{k_H}$ as well as their normals. Since the second layer constitutes relationships between points and their normals, which are in turn computed using the first layer, it is able to capture more intricate local surface variations.

Details on the PFH and FPFH formulation for a point $\mathbf{p_i}$ are given in [13]. The first step involves estimating surface normals $\overrightarrow{\mathbf{n}}_\mathbf{i}$ using $\mathcal{P}^{k_N}$ neighborhood for all points belonging to $\mathcal{P}^{k_H}$. To compute relative difference between two points $\mathbf{p_i}$, $\mathbf{p_j}$ and corresponding normals $\vec{\mathbf{n}}_\mathbf{i}$, $\vec{\mathbf{n}}_\mathbf{j}$, a Darboux uvw coordinate frame is defined at one of the points as shown graphically in Fig. 4. Using the Darboux frame, the relationship between points and normals are captured as three angular features defined as follows,

$$\begin{aligned} \alpha &= v.\mathbf{n_j} \\ \phi &= u.\frac{(\mathbf{p_j} - \mathbf{p_i})}{||\mathbf{p_j} - \mathbf{p_i}||_2} \\ \theta &= \arctan(w.\mathbf{n_j}, u.\mathbf{n_j}) \end{aligned} \tag{2}$$

These three angular features (α, ϕ, θ) are computed for pairs of points (p_i, p_j) belonging to the $\mathcal{P}^{k_H}$ neighborhood. The method in which these point pairs are chosen depends on the influence region defined for a particular feature representation. Fig. 5 shows the influence region for both PFH and FPFH feature representations. For the PFH representation, the query point p_q and its k_H neighborhood represent a fully interconnected mesh leading to an $O(nk^2)$ computational complexity, where n is the number of points. For the FPFH representation, the query point p_q is connected only to its direct k_H neighbors, which in turn is connected to its own neighbors and the resulting histograms are weighted together with the histogram of the p_q. This has an effect of reducing computational complexity to $O(nk)$. We choose to work with the FPFH representation as a trade-off between running time and accuracy. The FPFH vector is a 33 dimensional vector obtained by placing the combination of the three angular features (α, ϕ, θ) values into 33 bins.

2) Extraction of Global Feature: While purely local features have the advantage of being pose invariant and robust to occlusions, they can be limiting if the 3D reconstructions are noisy enough to prevent local features from being sufficiently discriminative. This is more of a concern for field environments where generating greenhouse like 3D reconstructions with high geometric fidelity is challenging. Segmentation under such environments would benefit notably from addition of some global context as that would increase the discriminative capacity of the point features.

We factor in this global context by using the intuition that even though each 3D model of the plant is sufficiently different, there is still uniformity in the way stem and leaves are connected to each other. In order to exploit this structural uniformity, we collapse the 3D point cloud onto a 2D plane and compute the probability density of this collapsed representation. Each 3D reconstruction of the plant is first transformed such that the stem axis is aligned along the z axis. Since we know the starting camera pose with respect to the plant and the subsequent camera motion, the stem axis

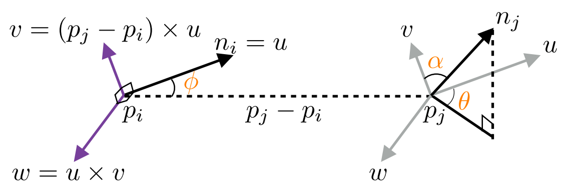


Fig. 4. Darboux uvw frame and angular features (α, ϕ, θ) for a pair of points p_i, p_j along with their corresponding normals $\vec{\mathbf{n}}_\mathbf{i}$, $\vec{\mathbf{n}}_\mathbf{j}$

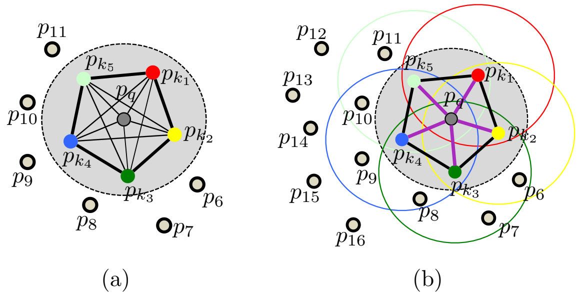


Fig. 5. Influence region for (a) PFH and (b) FPFH for a query point p_q. Unlike FPFH that has $O(nk)$ computational complexity, PFH computes the histogram over a fully interconnected mesh leading to $O(nk^2)$ computational complexity

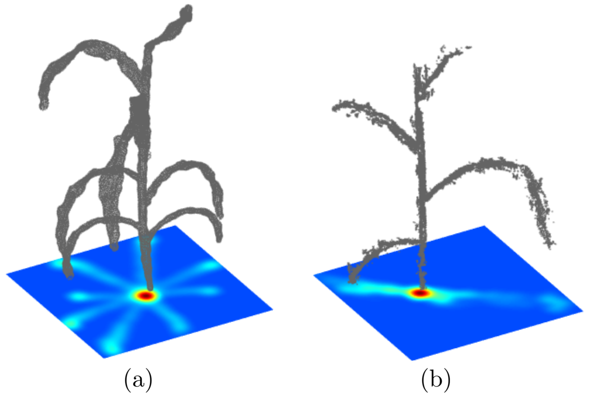


Fig. 6. 3D point cloud reconstructions from (a) greenhouse and (b) field shown along with a heatmap representing probability distributions of their collapsed 2D representations. The higher (red) probability density regions represent the mode which can be seen to lie close to stem position.

determination is an automated step. Post alignment, the 3D point cloud is collapsed onto the (x, y) plane by making all z coordinate values 0. The structural uniformity would cause the mode of the resulting 2D data distribution to be close to the stem location for most cases.

To compute the mode, a kernel density estimator (KDE) [14] is used on the collapsed 2D point distribution. KDE is a non-parametric method to compute the probability density function of a random variable. Let the 2D point coordinates, $(q_1, q_2, \cdots q_n)$, be n two-dimensional samples belonging to an unknown probability distribution f. The kernel density estimator, $\hat{f}_h(q)$, that approximates f is given as,

$$\begin{aligned} \hat{f}_h(q) &= \frac{1}{n}\sum_{i=1}^{n} K_h(q - q_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{q - q_i}{h}\right) \\ q^* &= \underset{q}{\operatorname{argmax}}\left(\hat{f}_h(q)\right) \end{aligned} \tag{3}$$

where, K is the kernel, $h > 0$ represents the bandwidth and q^* the estimated mode. The bandwidth acts as a smoothing parameter, controlling the bias-variance tradeoff of the resulting density distribution. A small bandwidth leads to an unsmooth distribution having high variance, while a large bandwidth leads to a smooth distribution having high bias. We use a gaussian kernel function with standard deviation $\hat{\sigma}$ computed using the n samples. The bandwidth value is taken as $h = (4\hat{\sigma}/3n)^{1/5}$ based on Silverman's rule-of-thumb.

Fig. 6 illustrates the 3D point clouds and probability distribution of their collapsed 2D representations as a heatmap. It can be seen that even with a slightly bent stem, the mode of the collapsed distribution lies close to the (x, y) position of the stem. The global feature for each 3D point $\mathbf{p}_i = \{x_i, y_i, z_i\}$ is then computed as $||[x_i \; y_i] - q^*||_2$, that is its euclidean distance from mode q^* in the xy-plane.

3) Classification using Support Vector Machine: We can now construct a feature vector for an individual 3D point with 34 elements obtained from FPFH (33) and global (1) feature estimation. The SVM classifier is provided a concatenated feature vector (n x 34) as input, where n is the number of 3D points. We make use of kernel SVMs that can perform nonlinear classification by implicitly mapping their inputs into high-dimensional feature spaces. SVMs are a popular and commonly used choice for binary classification problem. The results section provides further details on choosing the kernel function and its parameters as well as the train, validation, test split of the concatenated feature vector.

D. Spatial Smoothing of 3D Point Clouds

The spatial smoothing stage takes as input the segmented output from the SVM and smoothens out the assigned class labels. The SVM predicts a class label for a single 3D point without regard to the label assignment of the neighboring points. A Conditional Random Field (CRF) is hence applied as a post-processing step so as to take context into account. The segmented 3D point cloud produced by the SVM is expressed as a graph, $G = (V, E)$, where the vertices V are formed by the 3D points and edges E represent pairwise connections from a point to every other point. An efficient and tractable inference algorithm for such a fully connected pairwise CRF is detailed in [15]. It begins by formulating a Gibbs energy term $E(\mathbf{x})$ that needs to be minimized,

$$E(\mathbf{x}) = \Sigma_i \psi_u(x_i) + \Sigma_{i<j} \psi_p(x_i, x_j) \tag{4}$$

where, $\psi_u(x_i)$ is the unary potential initialized independently for each point by the SVM classifier. $\psi_p(x_i, x_j)$ is the pairwise potential taking into account pairwise relationships between point classifications, and is of the form,

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \underbrace{\Sigma_{m=1}^{K} w^{(m)} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j)}_{\kappa(\mathbf{f}_i, \mathbf{f}_j)} \tag{5}$$

where, $\kappa(\mathbf{f}_i, \mathbf{f}_j)$ is weighted sum of kernels expressed using positions p_i, p_j and surface normal vectors n_i, n_j as,

$$\begin{aligned} \kappa(\mathbf{f}_i, \mathbf{f}_j) = &\underbrace{w^{(1)} exp\left(-\frac{|p_i - p_j|^2}{2\theta_p^2}\right)}_{\text{smoothness kernel}} \\ &+ \underbrace{w^{(2)} exp\left(-\frac{|p_i - p_j|^2}{2\theta_{pn}^2} - \frac{|n_i - n_j|^2}{2\theta_n^2}\right)}_{\text{surface kernel}} \end{aligned} \tag{6}$$

In eqn (6), the *smoothness kernel* minimizes label differences between neighboring points while the *surface kernel* minimizes label differences between neighboring points with differing surface normal directions. Note that the original 2D image based segmentation kernel in [15] uses an *appearance kernel* instead of the surface kernel so as to minimizes label difference across nearby pixels with similar color values.

E. Extraction of Plant Phytomers

The phytomer extraction stage takes as input the segmented and smoothed point cloud and extracts 3D plant phytomer units. A 3D cylinder model is fit to the points labeled as stem using the random sample consensus (RANSAC) algorithm [16]. The fitted cylinder is then expanded by 25%

Algorithm 1 Region growing for extracting a single leaf

```
Initialize region set $\mathcal{R} = \emptyset$, seed point set $\mathcal{S}$ with 3D
centroid of the input plant node.
for $s \in \mathcal{S}$ do
    Find nearest neighbor set $\mathcal{K}$ of point $s$
    Compute surface normal $n_s$ of point $s$
    for $k \in \mathcal{K}$ do
        Compute surface normal $n_k$ of point $k$
        if $\texttt{angle}(n_k, n_s) < \epsilon_{th1}$ then
            $\mathcal{R} \leftarrow \mathcal{R} \cup k$
            Compute curvature $\lambda$ of point $k$
            if $\lambda < \epsilon_{th2}$ then
                $\mathcal{S} \leftarrow \mathcal{S} \cup k$
return $\mathcal{R}$
```

to intersect leaves branching out from the stem. The intersection points give node positions, where a node is defined as points on the stem from which leaves grow. Individual leaves are then extracted from each node by applying the region growing algorithm [17], with the node serving as a seed point for the algorithm. The region growing algorithm taking a plant node as input and returning the leaf region connected to that node as output is elaborated in Algorithm 1. Once individual leaves at each node are extracted, these are then merged with a section of stem around the node to obtain a phytomer unit corresponding to that node.

IV. RESULTS

In this section, we present results for different stages in the overall approach along with details on parameter selection.

A. System Setup and Data Collection

The greenhouse 3D data used in the paper is the one collected by McCormick, et al. in [4]. The field data used in the paper was collected using the robotic phenotyping platform shown in Fig. 1. A closeup of the multi-camera sensor pod deployed on the robotic platform is shown in Fig. 7. The sensor pod contains eight forward facing cameras arranged in two horizontal rows along with two additional cameras verged on either ends at an angle of 30°. In order to collect different horizontal viewpoints captured by these cameras at multiple plant heights, the sensor pod is attached to a robotic arm that moves it vertically along the plant. All 10 cameras have a synchronized trigger and have been fitted with wide field-of-view lenses since the distance between one crop row to another is only about 0.75m. More details on the hardware system and data acquisition processes can be found in [18]. Field trials for data collection were conducted in an outdoor sorghum field in Weslaco, TX over a period of 5 days in Dec 2016. The sorghum crop rows were pruned so as to approximately have a single plant in foreground view.

B. Experimental Evaluation

Comparison of Two-View and Multi-view Stereo Reconstruction: Fig. 8 shows foreground plant point clouds generated using (a) standard two-view stereo semi-global block matching (SGBM) algorithm and using (b) the multi-view environment framework described earlier. The two cameras chosen for performing two-view stereo SGBM are the ones in the center with a baseline ($\approx$ 0.04m) suitable for stereo imaging at such close range (<0.5m). The images chosen for the multi-view stereo are those from all 10 cameras at a particular robot arm height. It can be seen from Fig. 8 that the dense 3D reconstruction obtained using the multi-view stereo approach is much more geometrically consistent.

Note that when doing multi-view reconstruction using images from the 10 cameras at a particular arm height, we can circumvent the effect of wind motion since all 10 cameras trigger synchronously. However, when doing reconstruction using images across multiple arm heights, we use datasets with low to moderate level effect of wind. This is because high winds cause non-rigid body motion of plants that breaks standard structure-from-motion assumptions, making it a challenging problem beyond the scope of this paper.

Feature Vector Details: The Fast Point Feature Histogram (FPFH) feature representation is chosen over the Point Feature Histogram (PFH) for encoding local geometries as a trade off between runtime and accuracy. Fig. 9 compares the computation times for 10 3D point cloud reconstructions obtained from field data. The neighborhood sizes chosen for computing PFH, FPFH are $k_N = 15$, $k_H = 125$,

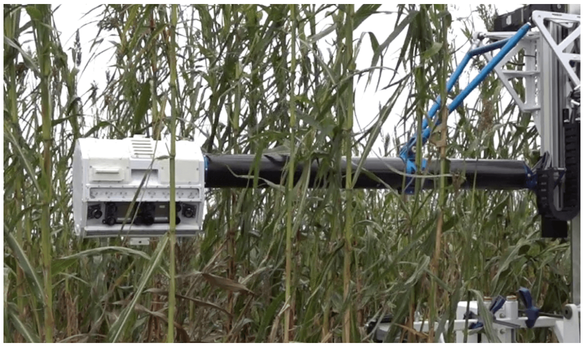
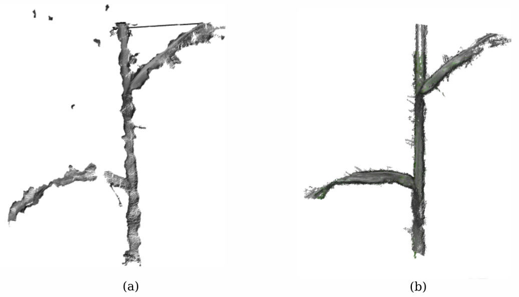


Fig. 7. A closeup of the multi-camera sensor pod deployed on the robotic platform. The sensor pod contains eight forward facing and two additional cameras verged on either ends at an angle of 30°. The sensor pod is connected to a robotic arm and can collect images at multiple plant heights.



(a)

(b)

Fig. 8. Reconstructed foreground plant point clouds using (a) standard two-view stereo semi-global block matching (SGBM) algorithm and using (b) the multi-view environment (MVE) framework. The dense 3D reconstruction using MVE can be seen to be less noisy and more geometrically consistent.

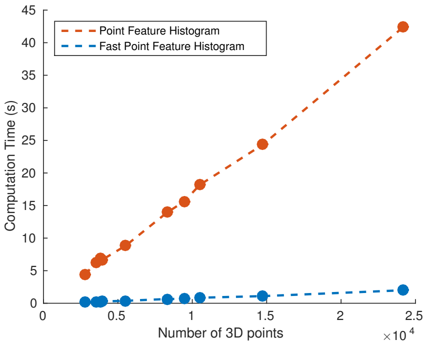


Fig. 9. Comparison of computation times for Point Feature Histogram (PFH) and Fast Point Feature Histogram (FPFH) feature vectors.

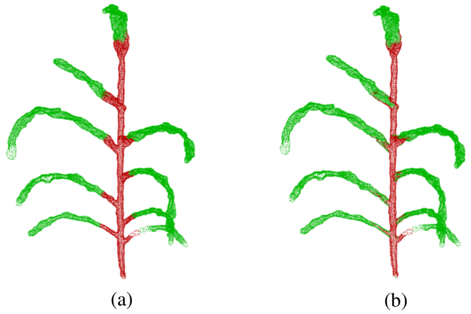


Fig. 12. SVM segmented output in (a) is post processed with CRF to give (b). The CRF corrects leaf false negatives near stem/leaf intersections so as to minimize label difference across neighbors with similar surface normals.

where k_N represents first layer of neighbors, $\mathcal{P}^{k_N}$, used to compute surface normals and k_H represents second layer of neighbors, $\mathcal{P}^{k_H}$, used to encode relationships between points and normals within $\mathcal{P}^{k_H}$. It is important to select a suitable value of k_N so that surface normals capture the underlying geometry of the point cloud at the desired resolution. The value of k_H controls how locally discriminative the point feature is, higher values yielding more discriminative features at the cost of increased computation times.

SVM Training Details: To construct the concatenated feature vector (n x 34) that is input to the SVM, 1000 stem and 1000 leaf 3D points were randomly sampled across 10 different plants, making $n = 2 \times 10^4$. This was done separately for the field data and the greenhouse data. Of these n feature vectors, 70% were chosen for training and 30% for testing. A 3-fold cross-validation was performed on the 70% training data to select optimal parameters for the SVM. Radial basis function (RBF) kernel SVM was found to perform consistently better than the linear SVM. The two parameters that need to be set for a SVM with RBF kernel are γ and C, where γ represents the variance of the (gaussian) RBF kernel and C the misclassification cost. The optimal values for (γ, C) were selected by doing a grid search on a range of values for (γ, C) and using the Area under Curve (AUC) of the precision-recall curve as the comparison metric.

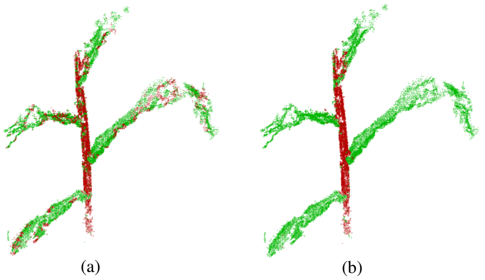


Fig. 10. Segmented SVM output for field data using (a) local and (b) local + global features. Stem false positives on leaf surfaces reduce considerably in (b) due to addition of distance from mode global information.

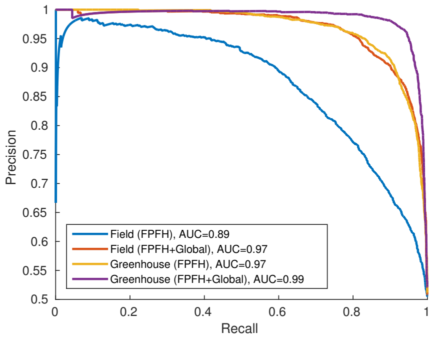


Fig. 11. Precision-Recall curves and their area under curve (AUC) values over 30% test set. The curves bring out the comparison between four scenarios: local, local+global features in field data and local, local+global features in greenhouse data. Addition of global feature improves the AUC value more significantly for field data compared to greenhouse data.

TABLE I

MEAN ACCURACIES FOR THE SEMANTIC SEGMENTATION

	Greenhouse Data		**Field Data**	
	SVM	**SVM+CRF**	**SVM**	**SVM+CRF**
Accuracy	85.5%	90.7%	79.4%	80.2%

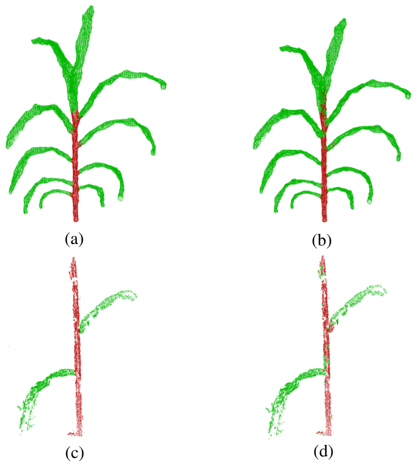


Fig. 13. (b), (d) show qualitative segmentation outputs from SVM followed by CRF smoothing for greenhouse and field environments respectively. (a),(c) show the ground truth segmentations for comparison.

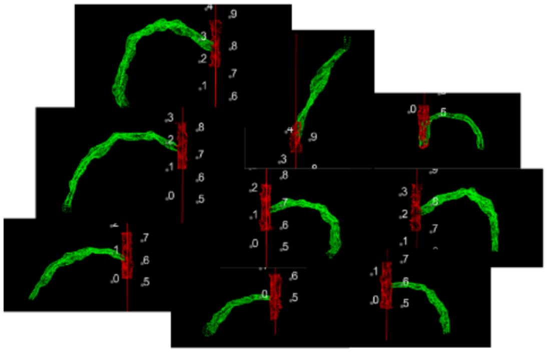


Fig. 14. Segmented 3D phytomers extracted from the 3D point cloud of the plant. The numbers 0-9 represent the detected stem-leaf intersection points (also known as nodes). A plant phytomer is extracted at each plant node.

Effect of local and global features: Having obtained the optimal kernel parameters using the AUC metric, Fig. 11 shows the precision-recall curves for $\gamma = 0.001$, $C = 1$. The curve is obtained over the 30% test samples for two different feature representations for both greenhouse and field data. This helps quantitatively examine the impact of local and global features on semantic segmentation in both field and greenhouse environments. It can be seen from Fig. 11 that the AUC value using purely local features is greater for greenhouse data than the field data. Addition of the global feature to purely local FPFH features causes the AUC values to increase noticeably for field data. This effect is also captured qualitatively in Fig. 10, where addition of the global feature reduces the stem false positives significantly.

Effect of CRF spatial smoothing: Table I shows the quantitative effect of spatially smoothing the SVM output using a CRF. It can be seen that there is an increase in average accuracy after the CRF post-processing step. This effect is also qualitatively visualized for the greenhouse data in Fig. 12. The CRF corrects leaf false negatives near the stem/leaf intersection points. This is primarily due to the surface kernel term in the CRF that penalizes label difference across neighbors with similar surface normal orientations.

Results across varying plant anatomies: Fig. 13 shows qualitative results for the semantic segmentation followed by CRF smoothing. The quantitative accuracies computed across 10 representative plants in greenhouse environments and 10 representative plants in field environments is tabulated in Table I. The final segmented 3D phytomer units extracted from the point cloud of a plant are visualized in Fig. 14.

V. CONCLUSION

This paper presents an approach for mapping 2D plant images to segmented 3D plant phytomer units that are of interest for phenotyping. It begins by performing multi-view 3D reconstruction, followed by segmentation using local and global features, and finally using the segmented cloud to extract phytomers. We show results for different stages of the approach for both field and greenhouse environments.

ACKNOWLEDGEMENTS

This research is supported by the U.S. Department of Energy grant DE-AR0000596. This work was supported through collaboration with Texas A&M University Agrilife Research, principal investigator William Rooney. We also gratefully acknowledge advice and assistance of John Mullet, Alex Tomlinson, Sandra Truong and Ryan McCormmick.

REFERENCES

[1] Noah Fahlgren, Malia A Gehan, and Ivan Baxter. Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. *Current opinion in plant biology*, 24:93–99, 2015.

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

[3] Inkyu Sa, Chris Lehnert, Andrew English, Chris McCool, Feras Dayoub, Ben Upcroft, and Tristan Perez. Peduncle detection of sweet pepper for autonomous crop harvestingcombined color and 3-d information. *IEEE Robotics and Automation Letters*, 2(2):765–772, 2017.

[4] Ryan F McCormick, Sandra K Truong, and John E Mullet. 3d sorghum reconstructions from depth images identify qtl regulating shoot architecture. *Plant Physiology*, pages pp–00948, 2016.

[5] Supawadee Chaivivatrakul, Lie Tang, Matthew N Dailey, and Akash D Nakarmi. Automatic morphological trait characterization for corn plants via 3d holographic reconstruction. *Computers and Electronics in Agriculture*, 109:109–123, 2014.

[6] Stefan Paulus, Jan Dupuis, Anne-Katrin Mahlein, and Heiner Kuhlmann. Surface feature based classification of plant organs from 3d laserscanned point clouds for plant phenotyping. *BMC bioinformatics*, 14(1):1, 2013.

[7] 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.

[8] Debadeepta Dey, Lily Mummert, and Rahul Sukthankar. Classification of plant structures from uncalibrated image sequences. In *Applications of Computer Vision (WACV), 2012 IEEE Workshop on*, pages 329–336. IEEE, 2012.

[9] Javier Ribera, Fangning He, Yuhao Chen, Ayman F Habib, and Edward J Delp. Estimating phenotypic traits from uav based rgb imagery. 2016.

[10] Simon Fuhrmann, Fabian Langguth, and Michael Goesele. Mve-a multi-view reconstruction environment. In *GCH*, pages 11–18, 2014.

[11] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M Seitz. Multi-view stereo for community photo collections. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[12] Radu Bogdan Rusu. Semantic 3d object maps for everyday manipulation in human living environments. *KI-Künstliche Intelligenz*, 24(4):345–348, 2010.

[13] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 3212–3217. IEEE, 2009.

[14] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[15] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

[16] Ruwen Schnabel, Roland Wahl, and Reinhard Klein. Efficient ransac for point-cloud shape detection. In *Computer graphics forum*, volume 26, pages 214–226. Wiley Online Library, 2007.

[17] Radu Bogdan Rusu and Steve Cousins. 3d is here: Point cloud library (pcl). In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1–4. IEEE, 2011.

[18] Srinivasan Vijayarangan, Paloma Sodhi, Prathamesh Kini, James Bourne, Simon Du, Hanqi Sun, Barnabas Poczos, Dimitrios Apostolopoulos, and David Wettergreen. High-throughput robotic phenotyping of energy sorghum crops. In *Field and Service Robotics*. Springer, 2017.