Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 Keller Hall
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 18-008

Registering Reconstructions of the Two Sides of Fruit Tree Rows

Pravakar Roy, Wenbo Dong, Volkan Isler

April 10, 2018

Registering Reconstructions of the Two Sides of Fruit Tree Rows

Pravakar Roy, Wenbo Dong and Volkan Isler

Abstract—We consider the problem of building accurate three dimensional (3D) reconstructions of orchard rows. This problem arises in many applications including yield mapping and measuring traits (e.g. trunk diameters) for phenotyping. While 3D reconstructions of side views can be obtained using standard methods, merging the two side-views is difficult due to the lack of overlap between the two partial reconstructions.

We present a novel method that utilizes global features to constrain the solution. Specifically, we use information from the silhouettes and the ground plane for alignment. The method is evaluated using multiple simulated and real datasets.

I. INTRODUCTION

The problem of building accurate 3D reconstructions of orchard rows arises in a number of agricultural automation tasks. 3D models can be used for automated pruning of trees [1]. Geometric traits such as tree volume and trunk diameter used for phenotyping can be extracted from such models. It is also important for yield mapping: while image-based methods can be used to count fruit in individual images, 3D models can be used for tracking them across images and to avoid double counting fruit visible from both sides of the row. For example in the case of the tree shown in Fig. 1, most of the apples are visible from both sides and can be counted twice in independent single side scans.

There are many techniques such as Structure from Motion (SfM) or RGB-D SLAM [2], [3] which can generate reconstructions of individual sides of the rows. However, existing methods can not merge these two reconstructions: Even with the manual selection of correspondences, Iterative Closest Point (ICP) techniques fail (Fig. 3). Large-scale SfM techniques can produce consistent reconstructions with the presence of overlapping side views or with loop closure. Obtaining such views in orchard settings is hard because the rows can be extremely long (sometimes spanning a thousand meters or more). The use of very precise Real-Time Kinematic (RTK) GPS can be used to solve the registration problem, but it is costly and not always available.

In this paper, we present a novel method to merge reconstructions from both sides of a row without the need for overlapping views or GPS coordinates (Fig. 2). We utilize a key observation: “The orthographic projection of the occlusion boundary of the trees in a row in the fronto-parallel plane from opposite sides of the row are symmetric.” Coupling this fact with the assumption of the existence of a common ground plane, we solve the problem of merging the reconstructions from both sides first by solving the problem

Fig. 1: The front view of an apple tree. Many apples in this tree are visible from both sides. If we add counts from individual sides, these apples will be counted twice.

of finding a rigid body transformation between the occlusion boundaries of a fronto-parallel view. Next, we address the problem of estimating a single overlapping depth distance. We solve this problem using existing 2D shape matching methods and semantic constraints (e.g. the tree trunks are well approximated by cylinders, their projection in a side view and front view have the same width). The method is evaluated using multiple simulated and real datasets.

II. RELATED WORK

There has been a lot of recent work on yield estimation for specialty crops [4]–[8]. Most of the existing systems rely on external sensors to register the fruits from single side/ both sides of the row. Wang et al. [4] used stereo cameras coupled with GPS and odometry sensors to avoid double counting. They aligned the apples globally in 3D space and removed the ones which are within .05 meter of a previously registered apple. Hung et al. [9] and Bargouti et al. [5] used sampling at certain intervals to remove overlap between images. Das et al. [7] used optical flow and navigational sensors to avoid duplicate apples. In our previous work [8], we presented a method for registering apples from the single side of a row based on affine tracking and incremental Structure from Motion (SfM). None of these previous methods build a consistent 3D model of the entire row. In contrast, in this

Fig. 2: Given two up to scale reconstruction of both sides of a row (first and second image from left), our goal is to merge the two reconstructions into a single coherent model (rightmost image).

Fig. 3: Simulation results. 1st: Input reconstruction of both sides. 2nd: Applying the ICP algorithm fails to merge the trunks. 3rd: Merging from proposed method. 4th: Refinement using semantic information.

work, we focus on aligning SfM reconstructions from both sides of a row to create a single consistent model.

Recent studies present methods for merging visually disconnected SfM models [10], [11] for urban environments. These methods estimate the scale and relative height for all submodels based on the Manhattan world assumption and find possible connection points between the reconstructions utilizing semantic information (facade edges, windows, doors etc). Afterwards, they generate all possible fully connected models and find the most likely model according to loop closure and symmetry alignment-alignment. In contrast, for aligning SfM reconstructions from both sides of a row, we utilize the symmetry of their occlusion boundaries from a fronto-parallel view. Afterwards, we adjust the overlap in depth direction using semantic information.

III. PROBLEM FORMULATION AND OVERVIEW OF TECHNICAL APPROACH

Consider a row of trees in an orchard. Suppose an imaging device moves along one side of the row (which we arbitrarily call the “front side”) and captures images. Then it moves to the “back” side and captures a second set of images. The images can be standard RGB images or they may also include depth information. The images in each set are used to obtain two independent reconstructions represented as point clouds. The main problem we address is to merge these two reconstructions by computing the scale, rotation and translation to align them. It is formalized as follows:

Problem Definition: Given two sets of input images $\{\mathcal{F}\mathcal{I}_i, \mathcal{B}\mathcal{I}_j\}$ from the front and back sides of a row, two reconstructions $\{\mathcal{F}, \mathcal{B}\}$ of $\{\mathcal{F}\mathcal{I}_i, \mathcal{B}\mathcal{I}_j\}$ along with extrinsic camera poses $\{\mathcal{F}\mathbf{T}_i, \mathcal{B}\mathbf{T}_j\}$ where $i = 1, \dots, m$ and $j = 1, \dots, n$ of each image, the goal is to merge these two reconstructions into a single coherent model ${}^3\mathcal{P}$ with combined set of camera poses $\{\mathbf{T}\}$ by finding a transformation

${}_{\mathcal{B}}^{\mathcal{F}}\mathbf{T} = s[{}_{\mathcal{B}}^{\mathcal{F}}\mathbf{R}|{}_{\mathcal{B}}^{\mathcal{F}}\mathbf{t}]$ that merges the back side reconstruction with the front (see Fig. 2).

Next, we introduce the objective function used for optimization.

Objective Function: The front and back side reconstructions $\mathcal{F}, \mathcal{B}$ generally do not share any local feature matches (point correspondences). To constrain the system, we propose to use global features based on the following observations: (1) the occlusion boundary of an object from the front and back orthographic views are same (see Fig. 4) (2) tree trunk segments at the same height from two sides can be treated approximately as cylinders. When projected to the ground plane they share the same center of the elliptical shape. If we align the median plane of the detected trunks, the maximum depth alignment error is bounded by the trunk widths.

We use the following geometric concepts to formulate the optimization problem. An orthographic front view is the parallel projection of all the points in the XY plane. The occlusion boundaries of this view can be approximated by the well-known concept of alpha shapes [12]. An alpha hull is the generalized version of the convex hull. The boundaries of an alpha hull α_{bnd} are point pairs that can be touched by an empty disc of radius alpha.

In most orchard settings, there are no leaves/branches attached to the trunks near the ground plane and the number of 3D points in this region are very sparse. We can use this observation to cluster the 3D points belonging to tree trunks. Let $\mathcal{F}_{tr}, \mathcal{B}_{tr}$ denote the detected trunk points close to the median trunk plane. $\mathbf{P}_{xy}, \mathbf{P}_{xz}$ denotes the orthogonal projection matrix to the front and top plane. With this we can define our problem as the following minimization problem:

$$\begin{aligned} s^*, {}_{\mathcal{B}}^{\mathcal{F}}\mathbf{R}^*, {}_{\mathcal{B}}^{\mathcal{F}}\mathbf{t}^* = & \\ \operatorname{argmin}_{s, {}_{\mathcal{B}}^{\mathcal{F}}\mathbf{R}, {}_{\mathcal{B}}^{\mathcal{F}}\mathbf{t}} & \operatorname{dist}(\alpha_{bnd}(\mathbf{P}_{xy} \cdot \mathcal{F}), \alpha_{bnd}(\mathbf{P}_{xy} \cdot {}_{\mathcal{B}}^{\mathcal{F}}\mathbf{T} \cdot \mathcal{B})) \\ & + \operatorname{dist}(\mathbf{P}_{zx} \cdot \mathcal{F}_{tr}, \mathbf{P}_{zx} \cdot {}_{\mathcal{B}}^{\mathcal{F}}\mathbf{T} \cdot \mathcal{B}_{tr}), \end{aligned} \quad (1)$$

where α_{bnd} computes the alpha shape boundary points. To find similarity between two point sets $\mathcal{P}, \mathcal{Q}$ in $\mathbb{R}^D$, we use the following metric:

$$dist(\mathcal{P}, \mathcal{Q}) = \sum_{p \in \mathcal{P}} \min_{q \in \mathcal{Q}} (p - q)^2. \quad (2)$$

Technical Approach: We can solve Eq. (1) using trusted region methods such as Levenberg-Marquardt (LM) algorithm [13] for which we need to find a good initial solution. Trivial initial values such as $\mathbf{I}_{3 \times 3}$ and zero translation $\mathbf{0}_{3 \times 1}$ do not work. In this paper, we develop a method to find a good initial solution. Essentially, we solve the two parts on the right-hand side of Eq. (1) sequentially along with some preliminary steps. The steps we follow are the following:

- 1) For both $\mathcal{F}, \mathcal{B}$, we detect the ground plane.
- 2) We perform a Principle Component Analysis (PCA) on both the reconstructions. Utilizing the ground plane normal, and camera poses we figure out the depth and up direction of the PCA components and align the point clouds roughly.
- 3) To get rid of the rotational component left after PCA, we perform alignment of the ground plane normals.
- 4) We fix the scale of the reconstructions using median scene height and fix the height of the ground plane using the median height of ground plane inliers.
- 5) Now, we only have to compute the translations. We use alpha volume analysis to compute the occlusion boundary and 2D shape matching techniques [14] to compute the translation in XY plane.
- 6) Next, we align the trunk points close to the median trunk plane

After these steps, the point clouds are roughly aligned and the trivial initial solution $s = 1$, ${}^{\mathcal{F}}\mathbf{R} = \mathbf{I}_{3 \times 3}$, ${}^{\mathcal{F}}\mathbf{t} = \mathbf{0}_{3 \times 1}$ leads to convergence. This method does not compute the trunk overlap distance precisely. However, it provides correspondence between the trunks from both side. Using the Random Sample Consensus (RANSAC) [15] scheme, the trunk and the ground area around each tree are modeled as a cylinder and a plane, respectively. This semantic information, i.e., trunks and ground areas, can be exploited into the bundle adjustment to further eliminate misalignment of two-sides reconstructions by adjusting camera poses and 3D information of semantic objects and feature points.

We describe each of these steps in details. We start with the initial step of PCA and ground plane alignment.

IV. GROUND PLANE ESTIMATION AND ALIGNMENT USING PCA

The main goal of this step is to eliminate most of the rotational difference required to align the two reconstructions. As is well known, this is normally solved by Principle Component Analysis (PCA). We assume that the length of the portion of the row covered by the input reconstruction is always longer than the height of the trees and the depth captured. Therefore, the first principle component always denotes the length of the row covered. The other two principle components though vary from reconstruction to

Fig. 4: The silhouettes of an apple tree from both sides. As the silhouettes are from the orthographic front and back view of the tree, they should align.

reconstruction. Therefore, while aligning the principal components we need to be aware of which principle component denotes scene depth and which one denotes height.

To automatically figure out the scene “up” and “depth” directions we estimate the ground plane. We perform a simple three-point RANSAC method for plane estimation [16]. Afterward, we align the corresponding principal components. If necessary, we flip the depth direction of one of the reconstructions to ensure that the frontal depth planes are opposing each other. Afterwards, we align the ground plane normals and rotate the point clouds to a canonical frame of reference (X, Y, Z) where $X = [1, 0, 0]^T$, $Y = [0, 1, 0]^T$ (up direction) and $Z = [0, 0, 1]^T$ (depth direction). Next, we fix the scale of the reconstructions using the maximum and minimum height of the trees and fix the height of the ground plane using inliers. The reconstructions are now roughly aligned in terms of rotation and translation in Y direction. Fig. 5 shows the result after these steps on a sample input reconstruction.

Fig. 5: Input and PCA Standardization. Left image shows the input to our algorithm. Right image shows the reconstructions after PCA based alignment, ground plane alignment, and scale and height adjustment. In the input we have only six trees but in the right image we see more than six trees. This shows that though most of the rotational differences are gone, the alignment is wrong in terms of translation in the X direction.

V. ALIGNMENT OF ORTHOGRAPHIC PROJECTION BOUNDARIES

In the last section, we roughly aligned the two reconstructions in terms of rotation, scale and, translation with respect

to the ground plane. Now we have to estimate translation in the canonical direction $X = [1, 0, 0]^\top$ and $Z = [0, 0, 1]^\top$. We start with solving for the translation in X direction. In practice, reconstructions are not perfect, and the ground plane is not perfectly planar. Consequently, our estimation in the previous step contains some error in terms of rotation, scale and translation in X and Y -directions. We use a method that computes this residual rotation, translation and scaling along with the translation in the X direction.

As outlined in Section III, to solve this we utilize the occlusion boundary of the reconstructions from orthographic front views. We use alpha volume analysis to compute the occlusion boundaries. The alpha hull boundaries are basically a set of 2D points. Thus, essentially we are solving a 2D point set registration problem. This problem is very well studied and many solutions exist in the literature. As our alpha boundaries are noisy we use a well-known shape alignment method, Coherent Point Drift algorithm (CPD) [14]. Myronenko et. al cast the point set registration problem as a probability density estimation problem. They represent one of the input point set as the centroids of a Gaussian Mixture Model(GMM) [17] and the other input as data. For the rigid transformation case, they reparameterize the GMM centroids in terms of rotation transformation and scale. They estimate the parameters by minimizing the negative log likelihood using the well-known Expectation Maximization algorithm [18]. Additionally, they add an extra component in the GMM to account for noise and outliers. At the optimum value of the parameters, two point sets are aligned. We apply the transformation computed by CPD to the entire point cloud to align them in the XY directions. The top figure in Fig. 6 shows the alignment of occlusion boundaries.

VI. ALIGNMENT IN DEPTH DIRECTION USING TRUNK INFORMATION

The principle ambiguity left is the relative depth distance between the two reconstructions. In an orchard row, trees are generally planted in straight lines and tree trunks are perpendicular to the ground. Therefore, we can imagine the existences of a central trunk plane bisecting the trunks. For each individual reconstructions, this bisector plane can be approximated by the median depth-plane of the detected trunks and we can align the reconstructions roughly by aligning the points close to this median plane.

Trunk Detection: In most orchard settings, there are no leaves/branches attached to the trunks near the ground plane. Consequently, the number of 3D points in the trunk region close to the ground is very small. We capture this region by binning the 3D points in terms of their distance from the ground plane (see Fig. 7). The flat region in this curve shown in Fig. 7 is easily captured by the utilizing the derivative. We use the region with the minimum number of points and two knee points around it to find the trunks close to the ground.

We perform a simple median filtering of the trunk points ${}^3B_{tr}$, ${}^3B_{tr}$. Afterward, like the previous section, we can align them using Coherent Point Drift(CPD) method [14].

Fig. 6: Alignment based on occlusion boundary and trunks. The top image shows how we align the occlusion boundaries of the reconstructions. The reconstructions after this step is shown in the middle image which demonstrates some alignment error in the depth direction (trunks and poles do not merge). Finally, the bottom image shows merged reconstruction using trunk information which does not have any duplicate trunk/poles.

Fig. 7: Trunk detection. Figure on the left shows the detected trunks on the simulated trees. Figure on the right shows a plot of the number of 3D points plotted against their distance from the ground plane. The flat region in this plot correspond to the 3D points belonging to the trunk regions.

After this step, both the reconstructions are in the same frame of reference and they have a very small difference in terms of rotation, translation, and scale. The trivial initial solution of $s = 1$, ${}^F\mathbf{R} = \mathbf{I}_{3 \times 3}$, ${}^F\mathbf{t} = \mathbf{0}_{3 \times 1}$ leads to fast convergence. The bottom figure in Fig. 6 shows the merged reconstruction. As a byproduct from this process, we have established trunk to trunk correspondences between the two reconstructions. We can use this to optimize our alignment further. We discuss the details of this procedure in the next section.

VII. SEMANTIC BUNDLE ADJUSTMENT

Give a height interval from the estimated ground plane, the trunk slice can be treated as a cylinder from two-sides reconstructions, which can be parameterized by its axis, center and radius. Without loss of generality, the local ground area of a tree is assumed as a plane defined by its normal and center. To eliminate the misalignment of two-sides reconstructions, the initial transformation is first calculated using such semantic information, i.e., tree trunks and ground areas around each tree. To address the issue of accumulated errors of camera poses, two-sides 3D reconstructions after initial alignment need to be further optimized by minimizing the cost of the semantic bundle adjustment

$$\underset{\mathbf{R}_c, \mathbf{t}_c, \mathbf{R}_s, \mathbf{t}_s, \mathbf{X}_p}{\operatorname{argmin}} J' = J + \sum_s \sum_c \sum_{p \in \mathcal{V}(s,c)} \rho(\lambda_s E_b(s, c, p)) \quad (3)$$

$$E_b(s, c, p) = \phi_l([\mathbf{R}_s | \mathbf{t}_s][\mathbf{R}_c | \mathbf{t}_c]^{-1} \mathbf{c} \bar{\mathbf{X}}_p, \mathbf{b}_s)^2$$

where ϕ_l ($l = 0, 1$) is the loss function of the distance between an s -th object and observed points belongs to it. J is the cost function for the standard bundle adjustment [19] and ${}^c\bar{\mathbf{X}}_p$ is the p -th observed 2D feature and its 3D location in the c -th camera frame. ρ is the robust Huber cost function. The weight λ_s balances between the cost J of feature points and the cost of semantic object points. See our technical report [20] for more details.

VIII. EXPERIMENTS

In this section, we report our experimental setup and results. We evaluate our merging method using both simulated and real datasets. The 3D reconstructions for the real datasets are produced using a commercial photogrammetry software Agisoft [21].

A. Simulated Dataset:

The purpose of the simulated dataset is to evaluate the accuracy of the method, when we have perfect 3D reconstructions as inputs. We generate a tree row using 3D models of apple trees and simulate the ground plane as planar points with Gaussian noise. We split the row into two arbitrary sections and transform the splitted reconstructions using randomly generated rotation, translation and scale. We run our method on these reconstructions and find how well it performs. The metric we use is the percentage of apples correctly merged. During each split we keep track of the number of apples belonging to both reconstructions. This is easily accomplished by computing a bounding box for every

apple and see how many of them lie within the intersection of the split clouds. After merging we simply find out the number of intersecting bounding boxes. Fig. 8 shows our accuracy against the overlap between the reconstructions. The overlap is actually the common area between two split reconstructions. For example, if we take 70% of the point cloud from the front side and 50% of the point cloud from the back side, we will have an overlap of 20%. For every overlap percentage, we performed this simulation twice. It is evident from Fig. 8 that, we merge more than 95% for any amount of overlap.

Fig. 8: Percentage of apples merged correctly in simulation.

It is notable that (see Fig. 3), even with perfect reconstructions and manually provided correspondences, Iterative Closest Point (ICP) based methods fail. We show an example of the apples merging in Fig. 9. For clarity of understanding, we color the apples from two sides differently.

B. Real Dataset:

In real datasets, the input reconstructions often contain missing components and holes. Therefore, we evaluate the effectiveness of the merged reconstruction for yield estimation. Specifically, we focus on apple orchards and merge reconstructions obtained from two sides of apple tree rows. We count the number of apples from the merged reconstructions and match it with the ground truth as well as single side counts. Additionally, we demonstrate a few qualitative results, such as how well trunks and poles are aligned and so on.

Yield Estimation in Apple Orchards: We use two datasets for this purpose:

Dataset1: Dataset1 contains six trees and it has 270 apples in total (Obtained by harvesting). See Fig. 10 (first image from left) for a sample image from this dataset.

Dataset2: Dataset2 contains eight trees and it has 274 apples in total (Obtained by harvesting). See Fig. 10 (last image from left) for a sample image from this dataset.

Both of these datasets were collected from University of Minnesota Horticultural Research Center and for both these datasets the images were acquired using a Samsung Galaxy S4 cell phone. Images from single side of the trees were plugged into the aforementioned photogrammetry software

Fig. 9: The top view and side view of apples from both sides.

Agisoft for single side 3D reconstructions. The reconstruction time was up to three hours for each dataset on a computer with 6th generation Intel Core i7 CPU.

Fig. 10: Sample images from used datasets. Left figure shows a sample image from Dataset1 and the right figure shows a sample image from Dataset2

We estimate the fruit count from each side of a row and merge them. First, we detect the fruits using our previously developed segmentation method [22]. Afterwards, we back-project the detected fruits in the images to obtain the fruit location in the 3D reconstruction. Fig. 11 shows an example. We perform a connected component analysis to detect the apple clusters in 3D. Then we project individual 3D clusters back to the images by utilizing the recovered camera motion. We count the fruits from these reprojected images using our counting method developed in [6]. A 3D cluster can be tracked over many frames. We choose three frames with the highest amount of detected apple pixels (from the 3D cluster) and report the median count of these three frames as the fruit count for the cluster. We follow this procedure for all the detected 3D clusters and aggregate the fruit count from a single side. It is notable that we remove the apples on

the ground for all the single side counts by using our found ground plane.

To merge counts from both sides, we compute the intersection of the connected components from both sides. Afterwards, we compute the total counts by using the inclusion-exclusion principle [23]. Essentially, we sum up the counts from all the connected components, compute the intersections area among them (among $1, 2 \dots$ total number of intersecting clusters) and add/subtract the weighted parts accordingly. Fig. 12 shows our result. Our counting accuracy from both sides for Dataset1 and Dataset2 are 94.81% and 91.98%. Compared to the both side count if we just add the single side counts we overcount significantly -128.8% for Dataset1 and 136.86% for Dataset2. Therefore, it is evident that merging the rows from both sides is essential to obtain accurate fruit counts.

Qualitative Results: Fig. 13 shows our merged reconstruction for Dataset2. Again, we can see that compared to the single side reconstructions shown on the top row (Fig 13), the merged reconstructions are considerably denser and provide more reliable fruit counts.

To further validate the proposed merging algorithm, we visually check if the misalignment of landmarks (e.g. poles and tree trunks) is eliminated. Dataset3 contains six trees collected by a camera attached to a stick in a tilted-top view of tree canopies. Fig. 14 shows that tree trunks and poles are well-aligned using proposed merging algorithm.

IX. CONCLUSION

In this paper, we presented a method for merging reconstructions from two sides of a tree row utilizing the constraints posed by the occlusion boundaries from orthographic views and tree trunks. We verified the method with multiple simulated and datasets. Additionally, we showed that coupled with accurate detection we can estimate the yield close to 95% of the ground truth. In future, we would like to extend our existing counting method [6] so that it utilizes 3D information for counting. Additionally, we want to localize each individual apples in the 3D space and build a complete map which a picker robot can use for path planning.

Fig. 11: Apples in 3D and intersecting clusters from both sides. Left figure shows the apples from the front side of Dataset1 detected in 3D. Figure in the middle shows the computed connected components on this apples. Figure in the right shows how the connected components from the back are intersecting with the front ones.

Fig. 12: Total fruit counts computed from the merged reconstruction and independent single side reconstructions compared to ground truth. The single side fruit counts are computed by summing up independent single side counts. Evidently the counts from merged reconstructions are more consistent.

X. ACKNOWLEDGEMENT

The authors thank Professors Emily Hoover, Cindy Tong and James Luby from Department of Horticultural Science, University of Minnesota, for their expertise and help with the experiments.

REFERENCES

- [1] A. Tabb and H. Medeiros, "A robotic vision system to measure tree traits," *arXiv preprint arXiv:1707.05368*, 2017.
- [2] S. Agarwal, K. Mierle, *et al.*, "Ceres solver," 2012.
- [3] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012, pp. 573–580.
- [4] Q. Wang, S. Nuske, M. Bergerman, and S. Singh, "Automated crop yield estimation for apple orchards," in *Experimental Robotics*, ser. Springer Tracts in Advanced Robotics, J. P. Desai, G. Dudek, O. Khatib, and V. Kumar, Eds. Springer International Publishing, 2013, vol. 88, pp. 745–758.
- [5] S. Bargoti and J. P. Underwood, "Image segmentation for fruit detection and yield estimation in apple orchards," *CoRR*, vol. abs/1610.08120, 2016. [Online]. Available: <http://arxiv.org/abs/1610.08120>
- [6] P. Roy and V. Isler, "Vision-based apple counting and yield estimation," in *Experimental Robotics*. Springer, 2016.
- [7] J. Das, G. Cross, C. Qu, A. Makineni, P. Tokekar, Y. Mulgaonkar, and V. Kumar, "Devices, systems, and methods for automated monitoring enabling precision agriculture," in *Proceedings of IEEE Conference on Automation Science and Engineering*, 2015.
- [8] P. Roy and V. Isler, "Surveying apple orchards with a monocular vision system," in *Automation Science and Engineering (CASE), 2016 IEEE International Conference on*. IEEE, 2016, pp. 916–921.
- [9] C. Hung, J. Underwood, J. Nieto, and S. Sukkarieh, "A feature learning based approach for automated fruit yield estimation," in *Field and Service Robotics*. Springer, 2015, pp. 485–498.
- [10] A. Cohen, T. Sattler, and M. Pollefeys, "Merging the unmatched: Stitching visually disconnected sfm models," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [11] A. Cohen, J. L. Schönberger, P. Speciale, T. Sattler, J.-M. Frahm, and M. Pollefeys, "Indoor-outdoor 3d reconstruction alignment," in *European Conference on Computer Vision*. Springer, 2016, pp. 285–300.
- [12] H. Edelsbrunner and E. P. Mücke, "Three-dimensional alpha shapes," *ACM Transactions on Graphics (TOG)*, vol. 13, no. 1, pp. 43–72, 1994.
- [13] J. J. Moré, "The levenberg-marquardt algorithm: implementation and theory," in *Numerical analysis*. Springer, 1978, pp. 105–116.
- [14] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 12, pp. 2262–2275, 2010.
- [15] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [16] M. Y. Yang and W. Förstner, "Plane detection in point cloud data," in *Proceedings of the 2nd int conf on machine control guidance, Bonn*, vol. 1, 2010, pp. 95–104.
- [17] C. E. Rasmussen, "The infinite gaussian mixture model," in *Advances in neural information processing systems*, 2000, pp. 554–560.
- [18] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm," *IEEE transactions on medical imaging*, vol. 20, no. 1, pp. 45–57, 2001.
- [19] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment modern synthesis," in *International workshop on vision algorithms*. Springer, 1999, pp. 298–372.
- [20] W. Dong and V. Isler, "Tree morphology for phenotyping from semantics-based mapping in orchard environments," Technical Report TR-18-0XX, University of Minnesota, Computer Science & Engineering Department, Tech. Rep., 2018.
- [21] "Agisoft photoscan," <http://www.agisoft.com/>, accessed: 2017-09-15.

Fig. 13: Merged reconstruction for Dataset2. The top two images show the input to our method. The bottom two images show two sides of the merged reconstruction.

Fig. 14: Merged reconstruction for Dataset3. 1st: Input reconstruction of both sides. 2nd: Initial alignment before semantic bundle adjustment. 3rd: Further refinement using semantic bundle adjustment.

- [22] P. Roy, N. Stefas, C. Peng, H. Bayram, P. Tokekar, and V. Isler, "Robotic surveying of apple orchards," University of Minnesota, Department of Computer Science, Tech. Rep., 07 2015.
- [23] T. Andreescu and Z. Feng, "Inclusion-exclusion principle," in *A Path to Combinatorics for Undergraduates*. Springer, 2004, pp. 117–141.