Toward Semantic Scene Understanding for Fine-Grained 3D Modeling of Plants

Mohamad Qadri, Harry Freeman, Eric Schneider, George Kantor

Carnegie Mellon University, Robotics Institute
5000 Forbes Ave, Pittsburgh, PA 15213

Abstract

Agricultural robotics is an active research area due to global population growth and expectations of food and labor shortages. Robots can potentially help with tasks such as pruning, harvesting, phenotyping, and plant modeling. However, agricultural automation is hampered by the difficulty in creating high resolution 3D semantic maps in the field that would allow for safe manipulation and navigation. In this paper, we build toward solutions for this issue and showcase how the use of semantics and environmental priors can help in constructing accurate 3D maps for the target application of sorghum. Specifically, we 1) use sorghum seeds as semantic landmarks to build a visual Simultaneous Localization and Mapping (SLAM) system that enables us to map 78% of a sorghum range on average, compared to 38% with ORB-SLAM2; and 2) use seeds as semantic features to improve 3D reconstruction of a full sorghum panicle from images taken by a robotic in-hand camera.

1 Introduction

Imagine a fully automated mobile manipulator with two cooperative robotic arms tasked to create a full 3D reconstruction of all fruits in a tree canopy, with some fruit initially occluded by branches and leaves. One arm pushes a branch aside while the other arm moves through free space to take images of the exposed area. Our vision is to move towards developing such a system. The first step towards this goal is to develop algorithms that can understand and reason about 3D semantics in the scene to allow for safe and reliable manipulation. This requires accurate high-resolution 3D reconstruction.

Existing 3D reconstruction, visual Simultaneous Localization and Mapping (SLAM), and Structure from Motion (SFM) algorithms fundamentally rely on the accuracy of traditional visual feature matching methods, such as SIFT (Lowe 2004) and ORB (Rublee et al. 2011) (used by popular feature-based SLAM methods such as ORB-SLAM2 and ORB-SLAM3). These features perform poorly in agricultural environments due to the lack of texture in the images, variations in luminosity levels, and the dynamics of the environment (for example, leaves or crops moving due to wind).

In this paper, we demonstrate 1) how the use of semantics and environmental constraints, such as the structure of robotic navigation in agricultural fields, enables the development of robust SLAM systems for 3D mapping in agriculture and 2) how semantics can improve ICP-based registration for high-definition 3D modeling of plants. We focus on two target applications: mapping in sorghum fields and full panicle 3D reconstruction using a robotic arm with an in-hand stereo camera.

2 Related Work

There has been significant progress in visual SLAM for both direct (Engel, Schöps, and Cremers 2014) and indirect (Mur-Artal and Tardós 2017), (Campos et al. 2021) methods. However, these methods fail to extend to agricultural settings due to varying lighting conditions and repeated patterns. (McCormac et al. 2018), (Ok et al. 2019), (Choudhary et al. 2014), (Nicholson, Milford, and Sünderhauf 2018) use learned features and landmarks, sometimes with scene structure or prior model assumptions, but these approaches are either indoor or assume less cluttered scenes. There are several impressive 2D object detection (Bochkovskiy, Wang, and Liao 2020), (Liu et al. 2016), (Ren et al. 2015) and semantic segmentation (Zhao et al. 2017), (He et al. 2017) networks which various works build upon. In agriculture, (Baweja et al. 2018) and (Parhar et al. 2018) use segmentation to measure stalk width. (Nellithimaru and Kantor 2019) build a SLAM system using geometric primitive shapes fitted to grapes as semantic 3D landmarks. (Liu et al. 2018) uses semantic data for point cloud alignment to count apples. (Dong, Roy, and Isler 2020) and (Santos et al. 2020) tackle similar issues for apples and grapes. (Sodhi et al. 2018) uses a robotic system to create in-field 3D reconstructions of sorghum plants. (Sepúlveda et al. 2020) present segmentation, planning, and occlusion algorithms to increase the picking accuracy of a dual-arm aubergine harvesting robot. (Zine-El-Abidine et al. 2021) presents a method to delineate apple trees in a trellis structured orchard and perform fruit count. We build on top of these works and present promising results along with future research directions to promote in-field 3D semantic mapping and safe manipulation in agriculture.

3 Semantic Features as SLAM Landmarks

In this section, we demonstrate how using semantics, leveraging environment specific constraints, and reasoning about the geometry of the scene can alleviate some of the the 3D reconstruction and data association challenges in agriculture. One such prior is robotic navigation in agricultural environments; robots traverse the field one row at a time, generally moving in a straight line. We show how incorporating this prior and assumptions about the geometric relationship between semantic landmarks leads to improvements in data association accuracy and hence increased SLAM robustness. Our focus application is 3D mapping in sorghum fields using sorghum seeds as semantic landmarks.

3.1 Semantic SLAM Leveraging Robotic Navigation Constraints

This section describes the front-end system where semantic and geometric constraints are enforced. The full SLAM system is further described in (Qadri 2021).

Feature Extraction The feature extraction pipeline (Fig. 1) is based on (Parhar et al. 2018). A Faster-RCNN network with a VGG16 backbone is used for detection, and returns a bounding box for each seed in the image. Each bounding box is cropped and passed to a pix2pix (Isola et al. 2017) network, which generates a new image with a segmentation mask for each detected seed. After segmentation, a 2D ellipse is fitted to the segmented areas, and ellipse centers are used as semantic keypoints.

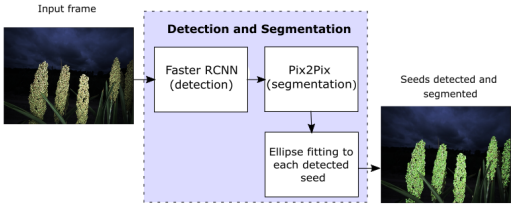


Figure 1: Detection and segmentation pipeline.

Data association algorithm The object-level data association between stereo pairs and successive temporal frames is framed as linear sum assignment problem (LSAP) optimization (Burkard, Dell'Amico, and Martello 2012). We define a bipartite graph $G = (U, V, E)$. Each vertex $s_{ab} \in U$, with coordinates (a, b) in the camera frame, corresponds to the projection of a 3D landmark onto image A. Similarly, each vertex s_{mn} with coordinates $(m, n) \in V$ corresponds to a projection onto image B. An edge $c_{ij} \in E$ between nodes s_{ab} and s_{mn} defines the cost of associating s_{ab} to s_{mn}. By introducing an assignment matrix φ where $\varphi_{ij} \in \{0, 1\}$, LSAP can be framed as the following optimization problem:

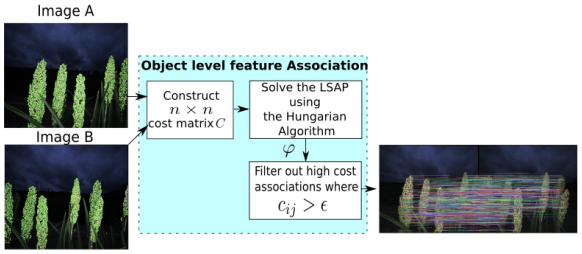


Figure 2: Proposed feature association pipeline.

$$\min_{\varphi \in S} \sum_{i=1}^{N} \sum_{j=1}^{N} c_{ij}\varphi_{ij} \text{ subject to : } \sum_{j=1}^{N} \varphi_{ij} = 1,\ i \in U$$

$$\sum_{i=1}^{N} \varphi_{ij} = 1,\ j \in V$$

Where S is the set of all possible assignments of nodes in U to nodes in V. Since sorghum panicles are rigid bodies, the distance from a particular seed to its neighboring seeds should stay approximately constant as the robot moves in the environment. Hence, we add the constraint that the sum of the Euclidean distances of a node $s_{ab} \in U$ to its surrounding nodes in U should be approximately equal to the sum of the Euclidean distances of $\varphi^*(s_{ab}) \in V$ and its surrounding nodes in V, where φ^* is the optimal assignment. We define a heuristic cost function that captures this geometric structure between the landmarks. For each node $s_{ab} \in U$, we define sets of neighbouring nodes: L_{ab} (left), R_{ab} (right), T_{ab} (top), and B_{ab} (bottom) satisfying the conditions:

$$L_{ab} = \{\forall s' = (c, d) \in U \mid 0 < a - c < \Delta \text{ and } |d - b| < \epsilon\}$$
$$R_{ab} = \{\forall s' = (c, d) \in U \mid 0 < c - a < \Delta \text{ and } |d - b| < \epsilon\}$$
$$T_{ab} = \{\forall s' = (c, d) \in U \mid 0 < b - d < \Delta \text{ and } |c - a| < \epsilon\}$$
$$B_{ab} = \{\forall s' = (c, d) \in U \mid 0 < d - b < \Delta \text{ and } |c - a| < \epsilon\}$$

We define $L_{mn}, R_{mn}, T_{mn}, B_{mn}$ similarly for $s_{mn} \in V$. The cost of associating node s_{ab} to node s_{mn} is defined as:

$$C(s_{ab}, s_{mn}) = r \cdot C'(L_{ab}, L_{mn}) + r \cdot C'(R_{ab}, R_{mn}) + r \cdot C'(B_{ab}, B_{mn}) + r \cdot C'(T_{ab}, T_{mn}) + |b - n|$$

$$\text{where } C'(X, Y) = \frac{\sum_{s' \in X} \sqrt{(s'_x - a)^2 + (s'_y - b)^2}}{\sum_{s' \in Y} \sqrt{(s'_x - m)^2 + (s'_y - n)^2}}$$

r is a constant, and $|b-n|$ is a term added to penalize matching landmarks with high vertical difference due to the horizontal nature of the robot trajectory. s'_x and s'_y are the x and y coordinates of one of the surrounding nodes s'.

Cost as a matching confidence measure The Hungarian algorithm returns the optimal assignment matrix φ^*, which is a bijection from U to V. Each row in φ^* is a one-hot vector, where $\varphi^*_{ij} = 1$ indicates node i has been matched with node j, and has a cost c_{ij}. Removing assignments where c_{ij} is over a threshold keeps only high confidence matches.

3.2 SLAM Results

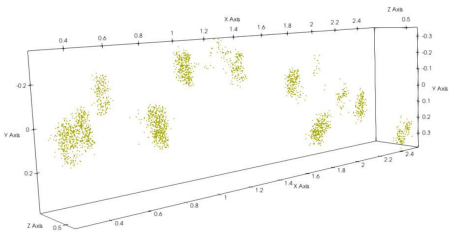


Figure 3: Example of a reconstructed 3D scene. The green dots correspond to 3D sorghum seeds and grouping of points correspond to different sorghum panicles. The first 1m is reconstructed from the sequence of images in Fig. 7 (appendix).

Once features are detected and associated, we frame the back-end optimization as a standard factor-graph problem which returns the optimized 3D landmark locations and camera trajectory. Fig. 3 is an example of a reconstructed sorghum range[1]. We use Maximum Distance Mapped as an indicator metric for the stability of SLAM systems and performance of data association algorithms. This is the distance that SLAM was able to map before a failure occurred in the back-end optimization or the system lost track. Table 1 compares the maximum distance mapped with our semantic features against traditional descriptors with a brute force matcher (BF). Using our proposed matching algorithm, we can map 3 out of 8 sorghum ranges completely and map 65% of the remaining 5 ranges on average (78% on average across the 8 sorghum ranges). SIFT performed the best out of the remaining four algorithms, with which we are able to map around 38% of the 8 sorghum ranges on average. This shows that using geometric relationships between semantic landmarks can improve performance when visual feature descriptors perform poorly.

Range ID (length in m)	**Feature detector + Matcher**				
	SIFT + BF	***SURF + BF***	***ORB + BF***	***AKAZE + BF***	***OURS***
1 (3.56 m)	1.86 m	1.55 m	0.2 m	1.55 m	**3.56 m**
2 (5.00 m)	0.25 m	0.25 m	Failed	0.19 m	**5.00 m**
3 (4.42 m)	0.5 m	0.38 m	Failed	0.38 m	**2.85 m**
4 (4.1 m)	1.47m	0.6 m	Failed	1.14 m	**2.31 m**
5 (4.78 m)	0.57 m	0.74 m	failed	0.74 m	**2.31 m**
6 3.94 m	1.4 m	0.19 m	0.11 m	0.45 m	**3.2 m**
7 5.03 m	3.15 m	0.26 m	Failed	0.46 m	**3.72 m**
8 4.43 m	4.04 m	0.94 m	Failed	0.94 m	**4.43 m**

Table 1: Maximum distance mapped. For fair comparison, we remove all matches that do not adhere to the camera motion assumptions (horizontal travel) for all methods. All ground truth distances are extracted from GPS.

In Table 2, we compare the performance of our SLAM algorithm against ORB-SLAM2, a feature-based SLAM system using DBoW2 for feature matching. We report the maximum distance mapped before the system become "lost".

[1]Sorghum fields are composed of rows, each containing several ranges. A range is ≈4m long and may contain different varieties of sorghum. Empty spaces with no plants separate consecutive ranges.

	ORB-SLAM2	***OURS***
Range 1 (3.56m)	0.35m	**3.56m**
Range 2 (5.00m)	0.25m	**5m**
Range 3 (4.42m)	0.18m	**2.85m**
Range 4 (4.10m)	0.25m	**2.31m**
Range 5 (4.78m)	0.31m	**2.31m**
Range 6 (3.94m)	0.12m	**3.2m**
Range 7 (5.03m)	0.33m	**3.72m**
Range 8 (4.43m)	0.26m	**4.43m**

Table 2: ORB-SLAM2 vs. OURS

These results illustrate the expected performance of feature-based SLAM methods when feature descriptors, a fundamental building block, perform poorly due to lack of texture and variations in luminosity levels, which further motivates the use of semantic features. We also ran ORB-SLAM3 on our dataset which improves on the relocalization capabilities of ORB-SLAM2 by building local maps when the system is lost. Local maps are merged when revisiting already mapped areas. We observed that ORB-SLAM3 performs similarly to ORB-SLAM2; the system repeatedly enters one of the "lost" states every few frames, indicating that ORB-SLAM3 is only able to construct local maps using a few images before losing track again.

4 High Resolution 3D Modeling with Semantic Features

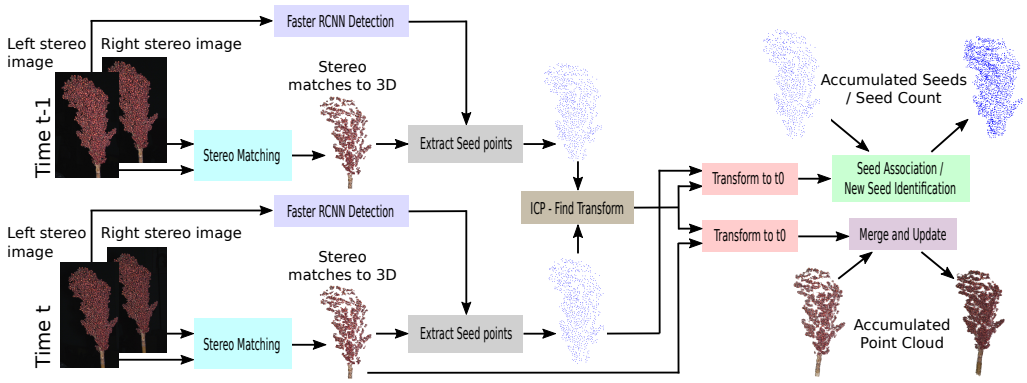
In the previous section, we presented a system with a camera rigidly attached to a mobile robot moving in the environment, commonly used in agricultural robotics. However, a fundamental shortcoming of this approach is its inability to build full 3D reconstructions because it is limited to mapping the visible face of an object. This is a substantial limitation since occlusions are common in agricultural settings. Hence, development of novel methodologies on the system and algorithmic levels should be made to reason and deal with such occlusions. In this section, we propose a process in which a robotic arm with an attached in-hand camera (Fig. 4b) is able to capture a full 360° set of stereo images of a single sorghum panicle and calculate phenotyping data. We qualitatively show how combining forward kinematics (FK) with semantic features can improve 3D reconstruction. We plan to eventually evaluate reconstruction and seed matching accuracy on a surrogate metric, seed count.

4.1 Reconstruction and Seed Matching

The main components for the reconstruction and matching pipeline are shown in Fig. 4a: at each time step, a stereo image pair is used to generate a 3D point cloud and predict bounding boxes as described in 3.1. To mitigate the effect of noise in the robot kinematics, ICP is used to refine the 3D cloud by finding the relative transformation between the current and previous frame. We compare running ICP on the full point clouds vs ICP only on the projected 3D seeds centers. The final 3D reconstructed cloud is then updated.

4.2 Preliminary Results

Captured stereo images using the robotic arm are downsampled to 3cm spacing and then passed to the 3D reconstruction pipeline as described in 4.1, forming the final point



(a) 3D Reconstruction and Seed Matching System

(b) Prototype end-effector for a UR5 allowing a 360° view of the panicle. Camera from (Silwal et al. 2021).

Figure 4

cloud. Preliminary 3D reconstruction results can be seen in Fig. 5. We note that running ICP only on the centers of detected seeds produces results that are less blurred, brighter, and capture more of the seed surface compared to full-cloud ICP, where we register full point clouds. The comparatively greater blurring in full-cloud ICP is a result of mixing between seed surfaces and inter-seed points in the final 3D reconstruction, and occurs because ICP is prone to run into local minima. In contrast, ICP using only seed centers shows comparatively better results since we operate on fewer semantically meaningful points (seed centers). These results are highlighted in Fig. 6. Note that there are black areas in the 3D reconstruction not present in the RGB image. This is a result of invalid disparity values calculated by SGBM (Hirschmuller 2008). An interesting direction is to explore is deep-learning depth generation networks such as FCRN Depth Prediction (Laina et al. 2016).

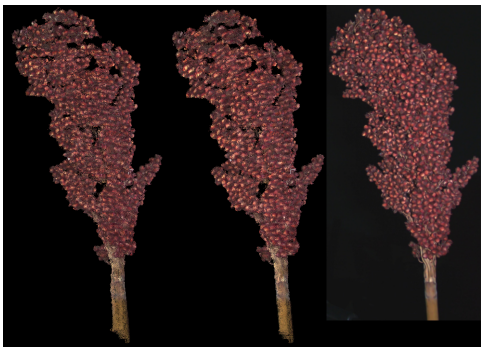


Figure 5: Reconstructions of a sorghum panicle over 90°. Shown are full-cloud ICP (left), ICP on seed centers (middle), and source RGB (right). Enlarged version in appendix.

5 Future Work

The SLAM results and initial 3D reconstructions show promise, but there is still much to do to fully realize the benefits of semantic reasoning. Going forward, both SLAM and

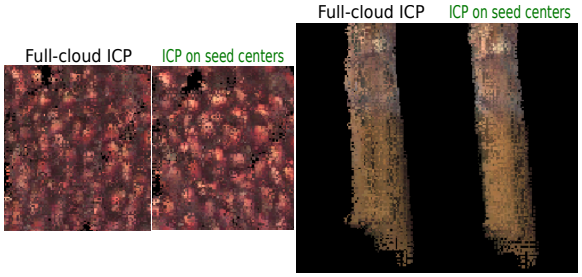


Figure 6: Zoomed view of Fig. 5. We see that there is better stem alignment and the reconstructed surface appears brighter (less blurred) when using ICP on seed centers.

reconstruction efforts could be improved by considering additional relevant information in the feature matching step. For example, integrating wind speed into the optimization formulation and incorporating learned deep visual features such as (Wang et al. 2020) for a more robust data association process on semantic landmarks. In future work, we also plan to explore high-resolution reconstructions with a larger dataset, containing more varied panicles. An important issue which will need to be addressed is how to assess both the 3D reconstruction and seed match steps in a scaleable manner without ground truth. Eventually, we envision that a better understanding of the scene semantics will allow for motion planning of robotic arms over semantic occupancy maps which could lead to safer and efficient manipulation in various agricultural settings. An interesting question is how to combine scene understanding with reasoning about environment dynamics to create robotic systems that can confidently perform complex manipulation tasks.

Acknowledgments

This work was supported by DOE APRAE TERRA 3P and USDA NIFA CPS 2020-67021-31531. Thanks to Clemson University's Pee Dee Research Center for assisting with data capture. Thanks to CMU PhD students John Kim and Mark Lee for designing and building the camera mount.

References

Baweja, H. S.; Parhar, T.; Mirbod, O.; and Nuske, S. 2018. StalkNet: A Deep Learning Pipeline for High-Throughput Measurement of Plant Stalk Count and Stalk Width. In *Field and Service Robotics*, 271–284. Springer.

Bochkovskiy, A.; Wang, C.; and Liao, H. M. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *CoRR*, abs/2004.10934.

Burkard, R.; Dell'Amico, M.; and Martello, S. 2012. *Assignment Problems: revised reprint*. SIAM.

Campos, C.; Elvira, R.; Rodríguez, J. J. G.; Montiel, J. M.; and Tardós, J. D. 2021. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. *IEEE Transactions on Robotics*.

Choudhary, S.; Trevor, A. J.; Christensen, H. I.; and Dellaert, F. 2014. SLAM with object discovery, modeling and mapping. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1018–1025. IEEE.

Dong, W.; Roy, P.; and Isler, V. 2020. Semantic mapping for orchard environments by merging two-sides reconstructions of tree rows. *Journal of Field Robotics*, 37(1): 97–121.

Engel, J.; Schöps, T.; and Cremers, D. 2014. LSD-SLAM: Large-scale direct monocular SLAM. In *European conference on computer vision*, 834–849. Springer.

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

Hirschmuller, H. 2008. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2): 328–341.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; and Navab, N. 2016. Deeper Depth Prediction with Fully Convolutional Residual Networks. *CoRR*, abs/1606.00373.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. SSD: Single Shot MultiBox Detector. In *European conference on computer vision*, 21–37. Springer.

Liu, X.; Chen, S. W.; Liu, C.; Shivakumar, S. S.; Das, J.; Taylor, C. J.; Underwood, J. P.; and Kumar, V. 2018. Monocular Camera Based Fruit Counting and Mapping with Semantic Data Association. *CoRR*, abs/1811.01417.

Lowe, D. G. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision*, 60(2): 91–110.

McCormac, J.; Clark, R.; Bloesch, M.; Davison, A.; and Leutenegger, S. 2018. Fusion++: Volumetric Object-Level SLAM. In *2018 international conference on 3D vision (3DV)*, 32–41. IEEE.

Mur-Artal, R.; and Tardós, J. D. 2017. ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. *IEEE Transactions on Robotics*, 33(5): 1255–1262.

Nellithimaru, A. K.; and Kantor, G. A. 2019. ROLS: Robust Object-level SLAM for Grape Counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Nicholson, L.; Milford, M.; and Sünderhauf, N. 2018. Quadricslam: Dual Quadrics from Object Detections as Landmarks in Object-Oriented SLAM. *IEEE Robotics and Automation Letters*, 4(1): 1–8.

Ok, K.; Liu, K.; Frey, K.; How, J. P.; and Roy, N. 2019. Robust object-based slam for high-speed autonomous navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, 669–675. IEEE.

Parhar, T.; Baweja, H.; Jenkins, M.; and Kantor, G. 2018. A Deep Learning-Based Stalk Grasping Pipeline. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 1–5. IEEE.

Qadri, M. 2021. *Robotic Vision for 3D Modeling and Sizing in Agriculture*. Master's thesis, Carnegie Mellon University, Pittsburgh, PA.

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

Rublee, E.; Rabaud, V.; Konolige, K.; and Bradski, G. R. 2011. ORB: An efficient alternative to SIFT or SURF. In Metaxas, D. N.; Quan, L.; Sanfeliu, A.; and Gool, L. V., eds., *ICCV*, 2564–2571. IEEE Computer Society. ISBN 978-1-4577-1101-5.

Santos, T. T.; de Souza, L. L.; dos Santos, A. A.; and Avila, S. 2020. Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. *Computers and Electronics in Agriculture*, 170: 105247.

Sepúlveda, D.; Fernández, R.; Navas, E.; Armada, M.; and González-De-Santos, P. 2020. Robotic aubergine harvesting using dual-arm manipulation. *IEEE Access*, 8: 121889–121904.

Silwal, A.; Parhar, T.; Yandún, F.; and Kantor, G. A. 2021. A Robust Illumination-Invariant Camera System for Agricultural Applications. *ArXiv*, abs/2101.02190.

Sodhi, P.; Sun, H.; Póczos, B.; and Wettergreen, D. 2018. Robust plant phenotyping via model-based optimization. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7689–7696. IEEE.

Wang, Q.; Zhou, X.; Hariharan, B.; and Snavely, N. 2020. Learning feature descriptors using camera pose supervision. In *European Conference on Computer Vision*, 757–774. Springer.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid Scene Parsing Network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zine-El-Abidine, M.; Dutagaci, H.; Galopin, G.; and Rousseau, D. 2021. Assigning apples to individual trees in dense orchards using 3D colour point clouds. *Biosystems Engineering*, 209: 30–52.

Appendix

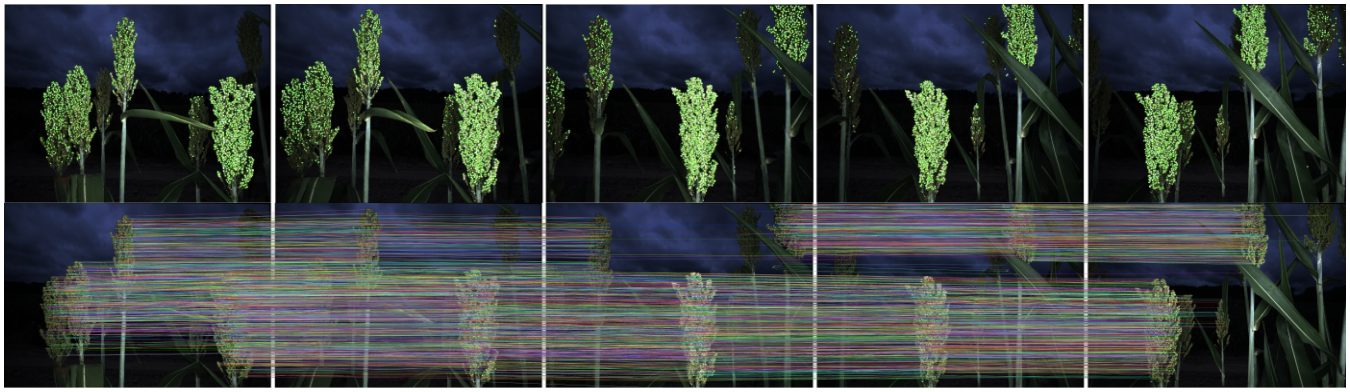


Figure 7: The first row shows the bounding box detections for the SLAM system presented in Section 3, one bounding box per sorghum seed. The second row shows the output of the proposed data association pipeline for five consecutive images.

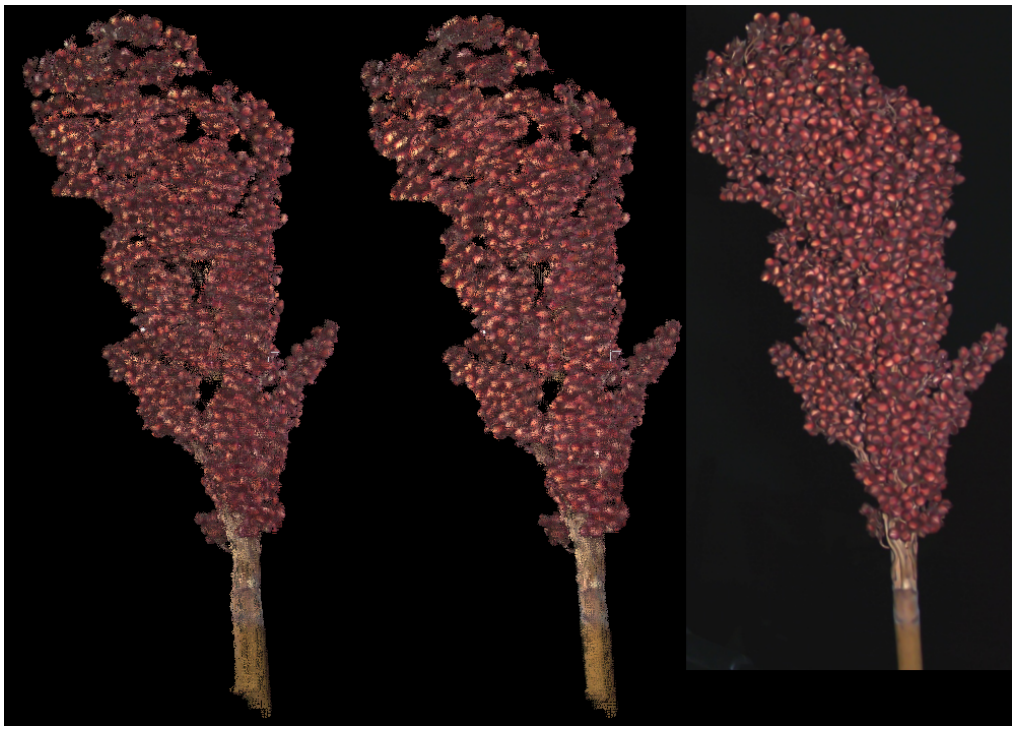


Figure 8: Full-size version Fig. 5. Shown are full-cloud ICP (left), ICP on seed centers (middle), and source RGB image (right).