4D Crop Monitoring: Spatio-Temporal Reconstruction for Agriculture

Jing Dong, John Gary Burnham, Byron Boots, Glen Rains, Frank Dellaert

***Abstract*— Autonomous crop monitoring at high spatial and temporal resolution is a critical problem in precision agriculture. While Structure from Motion and Multi-View Stereo algorithms can finely reconstruct the 3D structure of a field with low-cost image sensors, these algorithms fail to capture the dynamic nature of continuously growing crops. In this paper we propose a *4D reconstruction* approach to crop monitoring, which employs a spatio-temporal model of dynamic scenes that is useful for precision agriculture applications. Additionally, we provide a robust data association algorithm to address the problem of large appearance changes due to scenes being viewed from different angles at different points in time, which is critical to achieving 4D reconstruction. Finally, we collected a high-quality dataset with ground-truth statistics to evaluate the performance of our method. We demonstrate that our 4D reconstruction approach provides models that are qualitatively correct with respect to visual appearance and quantitatively accurate when measured against the ground truth geometric properties of the monitored crops.**

I. Introduction & Related Work

Automated crop monitoring is a key problem in precision agriculture, used to maximize crop yield while minimizing cost and environmental impact. Traditional crop monitoring techniques are based on measurements by human operators, which is both expensive and labor intensive. Early work on automated crop monitoring mainly relies on satellite imagery [1], which is expensive and lacks sufficient resolution in both space and time. Recently, crop monitoring with Unmanned Aerial Vehicles (UAVs) [2], [3], [4], [5], [6] and ground vehicles [7], [8], [9], [10], [11] has garnered interest from both agricultural and robotics communities due to the abilities of these systems to gather large quantities of data with high spatial and temporal resolution.

Computer vision is a powerful tool for monitoring the crops and estimating yields with low-cost image sensors [7], [10], [12], [13]. However, the majority of this work only utilizes 2D information in individual images, failing to recover the 3D geometric information from sequences of images. Structure from Motion (SfM) [14] is a mature discipline within the computer vision community that enables the recovery of 3D geometric information from images. When combined with Multi-View Stereo (MVS) approaches [15], these methods can be used to obtain dense, fine-grained 3D reconstructions. The major barrier to the direct use of these methods for crop monitoring is that traditional SfM and MVS methods only work for *static scenes*, which cannot solve 3D reconstruction problems with *dynamically growing* crops.

J. Dong, B. Boots and F. Dellaert are with College of Computing, Georgia Institute of Technology, USA. jdong@gatech.edu, {bboots,frank.dellaert}@cc.gatech.edu. J. G. Burnham and G. Rains are with the Department of Entomology, University of Georgia, USA. {burnhamj,grains}@uga.edu. Supplementary video is available at: https://youtu.be/BgLlLlsKWzI

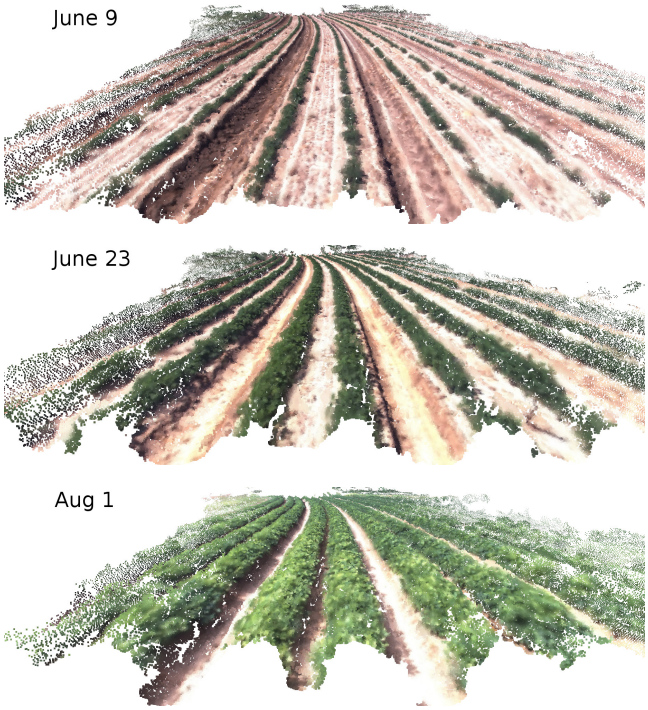


Fig. 1: Reconstructed 4D model of a peanut field by our approach. Each time slice of the 4D model is a dense reconstructed point cloud.

The computer vision community has worked on time-lapse reconstruction of dynamic scenes for years, but most existing approaches do not obviously apply to crop monitoring applications. Change detection in images [16], [17], point clouds [18], and volumes [19], [20] have been studied. Martin et al. [21], [22] synthesizes smooth time-lapse videos from images collected on the Internet, but this work is limited to 2D results, without any 3D geometric information. Early work on 4D reconstruction includes [23], [24], which build city-scale 3D reconstructions via temporal inference from historical images. Further work includes [25], which offers better scalability and granularity. A probabilistic volumetric 4D representation of the environment was proposed by [26], and then used in actual 4D reconstructions enabled by 3D change detection [20]. The major issue with most existing approaches is that they assume each geometric entity keeps nearly-constant appearance for the temporal duration, which is not the case in crop monitoring since crops are changing *continuously*.

Other related work includes data association approaches that build visual correspondences between scenes with appearance variations. Data association is a key element of 3D reconstruction: by identifying geometric correspondences between scenes with appearance changes (e.g. due to changing viewpoint), the scene can be reconstructed while minimizing misalignment. Recently, data association has been applied to even more difficult problems. For example, Griffith et al. studied the alignment of natural scene images with significant seasonal changes [27], [28]. Localization and place recognition are applications that require data association approaches that are highly robust to illumination and seasonal appearance changes [29], [30]. A more comprehensive survey of visual place recognition and localization can be found in [31]. Particularly, Beall et al. build a spatio-temporal map, which helps to improve robustness of localization across different seasons [32]. The difficulty with applying these existing methods to field reconstruction is that they are designed for autonomous driving applications: the camera's angle of view is assumed to vary little while traveling along roads. As a result, these approaches are not designed to solve the large-baseline data association problems that are prevalent in field reconstruction due to the large variations in the angle of view that are typically encountered by vehicles traversing a field.

The precision agriculture community has also developed spatio-temporal reconstruction approaches. For example, [11], [33] developed a LIDAR-based pipeline to retrieve height information over time. Unfortunately these approaches are designed for particular types of crops and scenes, using strong prior information about crop shape, which limits the general usage in other precision agriculture applications.

In this paper we address the problem of time-lapse 3D reconstruction with *dynamic scenes* to model *continuously growing* crops. We call the 3D reconstruction problem with temporal information *4D reconstruction*. The output of 4D reconstruction is a set of 3D entities (point, mesh, etc.), associated with a particular time or range of times. An example is shown in Fig. 1. A 4D model contains all of the information of a 3D model, e.g. canopy size, height, leaf color, etc., but also contains additional temporal information, e.g. growth rate and leaf color transition.

We also collected a field dataset using a ground vehicle equipped with various sensors, which we will make publicly available. To our knowledge, this will be first freely available dataset that contains large quantities of spatio-temporal data for robotics applications targeting precision agriculture.

Our paper contains three main contributions:

- We propose an approach for 4D reconstruction for fields with continuously changing scenes, mainly targeting crop monitoring applications.
- We propose a robust data association algorithm for images with highly duplicated structures and significant appearance changes.
- We collect a ground vehicle field dataset with ground truth crop statistics for evaluating 4D reconstruction and crop monitoring algorithms.

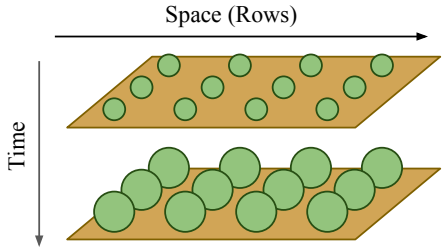


Fig. 2: The field 4D model. The field contains multiple rows, and there are multiple time sessions of the field.

II. Method

We begin by stating several assumptions related to crop monitoring, before specifying the details of our 4D reconstruction algorithm.

- The scene is *static* during each data collection session.
- The field may contain multiple *rows*.

The first assumption is acceptable because we only focus on modelling crops and ignore other dynamic objects like humans, and the crop growth is too slow to be noticeable during a single collection session. The second assumption is based on the geometric structure of most fields. The 4D field model reflecting these two assumptions is illustrated in Fig. 2.

Our proposed system has three parts.

- A multi-sensor Simultaneous Localization and Mapping (SLAM) pipeline, used to compute camera poses and field structure for a *single* row in a *single* session.
- A data association approach to build visual correspondences between different rows and sessions.
- A optimization-based approach to build the full 4D reconstruction across all rows and all sessions.

To generate the 4D reconstruction of the entire field, we first compute 3D reconstruction results for each row at each time session, by running multi-sensor SLAM independently. Next we use the data association approach to match images from different rows and sessions, building a joint factor graph that connects the individual SLAM results. Finally we optimize the resultant joint factor graph to generate the full 4D results.

To clarify notation, we assign the superscript of each symbol to the row index and time session index in the remaining article, unless otherwise mentioned. The superscript $\langle t_i, r_j \rangle$ indicates the variable is associated to row r_j of the field at time session t_i.

A. Multi-Sensor SLAM

The SLAM pipeline used in this work has two parts, illustrated in Fig. 3.

The first part of the SLAM system is a front-end module to process images for visual landmarks. SIFT [34] features are extracted from each image, and SIFT descriptor pairs in nearby image pairs are matched by the approximate nearest neighbor library FLANN [35]. The matches are further filtered by 8-point RANSAC [36] to reject outliers. Finally a single visual landmark is accepted if there are more than

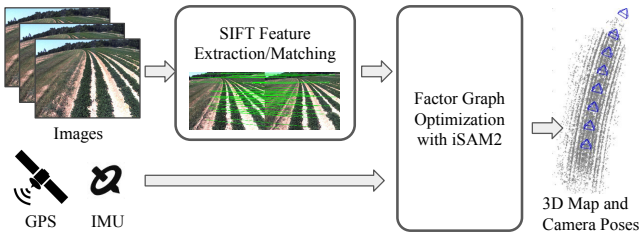


Fig. 3: Overview of multi-sensor SLAM system.

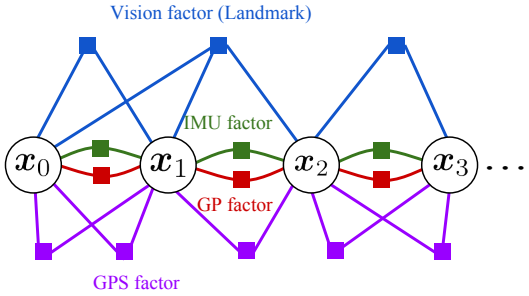


Fig. 4: Factor graph of multi-sensor SLAM.

6 images that have corresponding features matched to the same landmark.

The second part of the SLAM system is a back-end module for estimating camera states and landmarks using visual landmark information from the front-end and other sensor inputs. Since the goal of the multi-sensor SLAM system is to reconstruct a *single* row during a *single* data collection session, the back-end module of the SLAM system estimates a set of N camera states $X^{\langle t_i,r_j\rangle} = \{\boldsymbol{x}_0^{\langle t_i,r_j\rangle}, ..., \boldsymbol{x}_{N-1}^{\langle t_i,r_j\rangle}\}$ at row r_j and time t_i, given visual landmark measurements from the front-end module, and other sensor measurements, including an Inertial Measurement Unit (IMU) and GPS. For each camera state, we estimate $\boldsymbol{x}_j = \{\mathbf{R}_j, \mathbf{t}_j, \boldsymbol{v}_j, \boldsymbol{\omega}_j, \mathbf{b}_j\}$, which includes camera rotation $\mathbf{R}_j$, camera translation $\mathbf{t}_j$, translational velocity $\boldsymbol{v}_j$, angular rotation rate $\boldsymbol{\omega}_j$, and the IMU sensor bias $\mathbf{b}_j$.

The SLAM problem is formulated on a factor graph [37] where the joint probability distribution of estimated variables $X^{\langle t_i,r_j\rangle}$ given measurements $Z^{\langle t_i,r_j\rangle}$ is factorized and represented as the product

$$p(X^{\langle t_i,r_j\rangle}|Z^{\langle t_i,r_j\rangle}) \propto \prod_{k=1}^{K} \phi(X_k^{\langle t_i,r_j\rangle}), \tag{1}$$

where K is the total number of factors, $X_k^{\langle t_i,r_j\rangle}$ is the set of variables the kth factor involved, and ϕ is the factor in the graph which is proportional to measurement likelihood $l(X_k^{\langle t_i,r_j\rangle}; z_k^{\langle t_i,r_j\rangle})$, given kth measurement $z_k^{\langle t_i,r_j\rangle} \in Z^{\langle t_i,r_j\rangle}$. The states can then be computed by Maximum a Posteriori (MAP) estimation

$$\hat{X}^{\langle t_i,r_j\rangle} = \underset{X^{\langle t_i,r_j\rangle}}{\operatorname{argmax}}\, p(X^{\langle t_i,r_j\rangle}|Z^{\langle t_i,r_j\rangle}). \tag{2}$$

The details of the factor graph is shown in Fig. 4. We use smart factors [38] for visual landmarks, to reduce memory storage by avoiding the explicit estimation of landmark variables. Outlier rejection is used to reject landmarks with re-projection error larger than 10 pixels. IMU measurements are incorporated into the factor graph by preintegrated IMU factors [39]. GPS measurements may have different sensor rate than images (see Sec. III-A), so we use a continues-time SLAM approach which formulates the SLAM problem as a Gaussian Process (GP) [40]. The GPS measurements are thus easily incorporated into factor graph as interpolated binary factors (magenta factors in Fig. 4), with GP prior factors (red factors in Fig. 4). Details about continuous-time SLAM as a GP can be found in [41], [40], [42].

We optimize the factor graph by iSAM2 [43]. Once camera states are estimated, M landmarks $L^{\langle t_i,r_j\rangle} = \{l_0^{\langle t_i,r_j\rangle}, ..., l_{M-1}^{\langle t_i,r_j\rangle}\}$ are triangulated by known camera poses.

B. Robust Data Association over Time and Large Baseline

The second key element of our approach is robust data association. Data association is a key technique to get reconstruction results of more than a single row at a single time; however, the data association problem between different rows or times is difficult, since there are significant appearance changes due to illumination, weather or view point changes.

The problem is even more challenging in crop monitoring due to *measurement aliasing* [44]: fields contain highly periodic structures with little visual difference between each plants (see Fig. 1). As a result, data association problems between different rows and times is nearly impossible to solve by image-only approaches.

Rather than trying to build an image-only approach, we use geometry information output by SLAM as a prior for data association across rows and time. The SLAM results provide camera poses and field structures from all of the sensors (not just images), which helps us to improve the robustness of data association.

Specifically, the data association problem involves finding visual correspondences (matches between SIFT feature points) between two images, I_1 and I_2, which are taken by camera C_1 and C_2 respectively, as shown in Fig. 5. Cameras C_1 and C_2 may come from the same or different rows, during the same or different time sessions. Each camera $C_i = \{\mathbf{K_i}, \mathbf{R_i}, \mathbf{t_i}\}$ contains the camera intrinsic calibration $\mathbf{K_i}$ which is calibrated offline, and camera transformation in global frame $\{\mathbf{R_i}, \mathbf{t_i}\}$ estimated by SLAM on a single row.

We combine two methods to build a data association using prior information from SLAM, *back projection bounded search* and *homography image warping*. The two methods are detailed here.

Back Projection Bounded Search: The basic idea of back projection bounded search is to reduce number of possible outliers by limiting the search range while seeking visual correspondences. Assume L_1 is the set of all estimated landmarks visible in C_1, and each landmark in L_1 has corresponding feature points in I_1. For each $l_i \in L_1$, the linked feature point $f_{1,i} \in I_1$ might have a corresponding matched point at $p_{2,i} \in I_2$, which is the back-projected point of l_i on I_2, if C_1, C_2 and l_i are accurately estimated and l_i does not change its appearance. With estimation errors, we

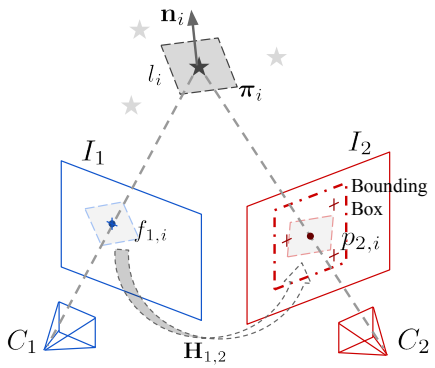


Fig. 5: Diagram of robust data association.

define a relaxed search area as a bounding box centered at $p_{2,i}$ on I_2, as shown in Fig. 5, to search for the corresponding feature point for $f_{1,i}$. This significantly limits the search area to match $f_{1,i}$ and reject many possible outliers, compared with searching the whole of I_2.

Homography Image Warping: Although the back projection bounded search rejects the majority of outliers, data association is still difficult when the viewing angle changes: the object's appearance may change significantly with large baselines, causing the search for a match in object's appearance from I_1 to I_2 based on SIFT descriptors, which also change with appearance. This is the major challenge for data association across images collected in different rows.

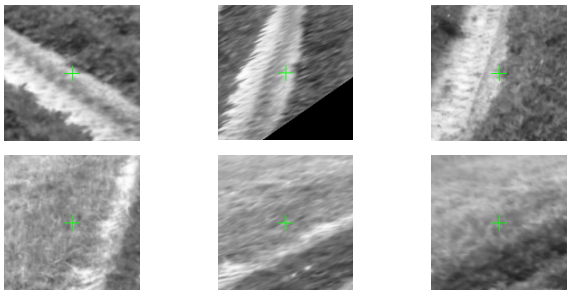
To combat this problem, we use a homography based method to eliminate appearance variations in l_i due to viewpoint changes. This is partially inspired by [45], but with the major difference that our proposed method uses feature points instead of patches. We assume that l_i lies on a local *plane* $\boldsymbol{\pi}_i$. If this assumption is satisfied, $\boldsymbol{\pi}_i$ induces a homography $\mathbf{H}_{1,2}$ from I_1 to I_2 [36, p.327]

$$\mathbf{H}_{1,2} = \mathbf{K}_2\big(\mathbf{R}_{1,2} - \frac{\mathbf{t}_{1,2}\mathbf{n}_i^\top}{d}\big)\mathbf{K}_1^{-1} \tag{3}$$

where $\{\mathbf{R}_{1,2}, \mathbf{t}_{1,2}\}$ define the relative pose from C_1 to C_2, $\mathbf{n}_i$ is the normal vector of $\boldsymbol{\pi}_i$, and d is the distance from C_1 to $\boldsymbol{\pi}_i$. We use $\mathbf{H}_{1,2}$ warp I_1 to get I_1', which has same view point with I_2, and thus similar appearance. We next extract a SIFT descriptor $f_{1,i}'$ on I_1' for bounded search rather than using the original $f_{1,i}$. Two example patches are provided in Fig. 6: although I_1 and I_2 have significant appearance variation, since they are taken from different rows, the warped I_1' has a very similar appearance to I_2, which makes SIFT descriptor matching possible.

The full data association pipeline is summarized in Algo. 1. We estimate the normal vector $\mathbf{n}_i$ by a local landmark point cloud around l_i in L_1. Homography image warping is only enabled when the baseline between C_1 and C_2 is longer than a threshold, in our system this is set to 0.5m. After getting all nearest neighbour feature matches by back projection bounded search, a final outlier rejection is performed by 8-point RANSAC.

Experiments validate the performance and robustness of proposed approach. A cross-row (1st vs. 3rd row) data association result is shown in Fig. 7. The naive FLANN+RANSAC approach can only recover feature matches on the top, where far away objects do not change their appearance, and it fails to register any crops correctly in the field. However, the proposed approach can register feature points in the field, with significant changes of appearance. A successful cross session matching result is also shown in Fig. 8.



(a) I_1 (b) I_1' (c) I_2

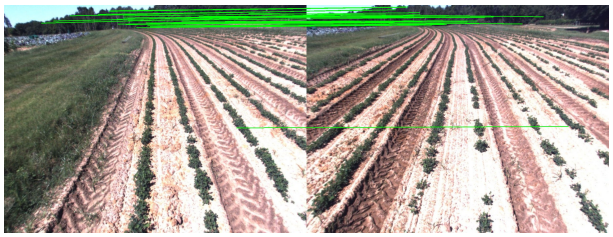
Fig. 6: Homography image warping on two random images patches. (a) original I_1, (b) warped image I_1', and (c) original I_2. Patch center with green cross is feature point $f_{1,i}$ on (a), $f_{1,i}'$ on (b), and back project point $p_{2,i}$ on (c).

Algorithm 1: Robust Data Association

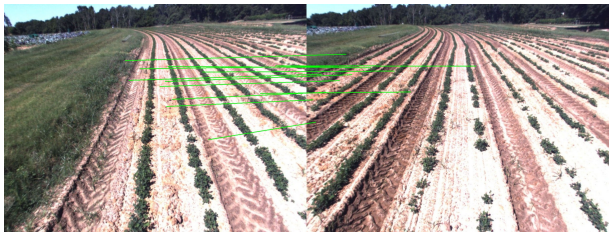
Input : Image I_1, I_2, Camera C_1, C_2, Landmarks L_1
Output: Set of matched feature point pairs $P_{1,2}$
set match set $P_{1,2} = \emptyset$
foreach $l_i \in L_1$ **do**
  back project l_i to $C_2 \rightarrow p_{2,i}$
  if C_1 *and* C_2 *baseline length* $<$ *threshold* **then**
    $f_{1,i}' = f_{1,i}$
  else
    calculate homography $\mathbf{H}_{1,2}$ use Eq. 3
    use $\mathbf{H}_{1,2}$ warp $I_1 \rightarrow I_1'$
    calculate SIFT descriptor at $p_{2,i}$ on $I_1' \rightarrow f_{1,i}'$
  end
  set l_i's match set $P_i = \emptyset$
  foreach *feature point* $f_{2,j} \in I_2$ **do**
    if $f_{2,j}$ *in bounding box of* $p_{2,i}$ **then**
      insert $[f_{1,i}', f_{2,j}] \rightarrow P_i$
    end
  end
  find min $L2$ of SIFT descriptor in $P_i \rightarrow [f_{1,i}', f_{2,k}]$
  insert $[f_{1,i}, f_{2,k}]$ into $P_{1,2}$
end
RANSAC_8pt_reject_outlier$(P_{1,2})$

C. 4D Reconstruction

The third of last part of our pipeline is a 4D reconstrcution module. The complete 4D reconstruction pipeline is illustrated in Fig. 9. We define the goal of 4D optimization as jointly estimating all camera states $X = \bigcup_{t_i \in T, r_j \in R} X^{\langle t_i, r_j \rangle}$ and all landmarks $L = \bigcup_{t_i \in T, r_j \in R} L^{\langle t_i, r_j \rangle}$, where R and T are set of rows and sessions, respectively. The measurements



(a) data association by FLANN and 8-point RANSAC



(b) data association by proposed approach

Fig. 7: Data association results of a image pair between 1st and 3rd row of June 9. Best viewed in digital.

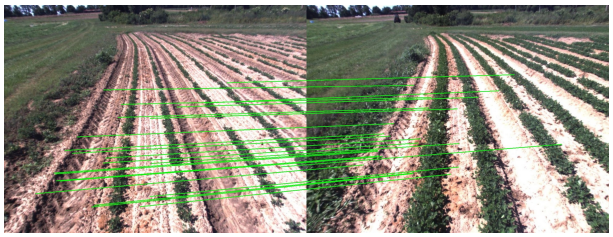


Fig. 8: Data association results of a image pair between 1st row of June 9 and June 20. Best viewed in digital.

$Z = \bigcup_{t_i \in T, r_j \in R} Z^{\langle t_i, r_j \rangle} \cup Z_{cr}$ includes all single row information as well as data association measurements Z_{cr} that connect rows across space and time.

Similar to the multi-sensor SLAM, we also formulate the joint probability of all camera states

$$p(X|Z) \propto \prod_{t_i \in T} \prod_{r_j \in R} \phi(X^{\langle t_i, r_j \rangle}) \prod_{h=1}^{H} \phi(X_{cr,h}), \quad (4)$$

where H here is the size of Z_{cr}, and $X_{cr,h}$ is the set of states hth measurement of Z_{cr} involved. This joint probability can be expressed as a factor graph, shown in Fig. 10(a). The first part of the joint probability consists of the factor graphs from all single rows. And the second part is the cross-row and cross-session measurements Z_{cr}, which are vision factors generated from cross-row and cross-session data association. We call the added factors *shared landmarks*, since they are shared by two (or possibly more) rows, and they have two (or more) sessions associated with them.

Solving the MAP estimation problem results in estimated camera states

$$\hat{X} = \underset{X}{\operatorname{argmax}}\, p(X|Z). \quad (5)$$

We use the Levenberg-Marquardt algorithm to solve the optimization problem, with initialization from the result of multi-sensor SLAM. Outlier rejection of vision factors [38] is also enabled during optimization, to reject possible false positive feature matches from cross-row and cross-session data association. Landmarks $\hat{L}$ are estimated by triangulation given estimated camera poses.

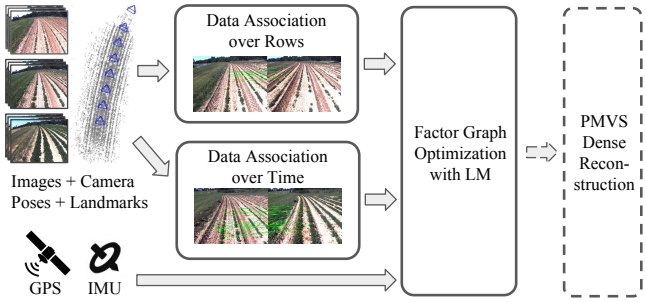
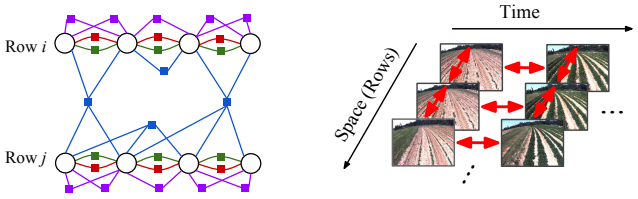


Fig. 9: Overview of 4D reconstruction pipeline. Dash box of PMVS dense reconstruction step means it is optional.



(a) 4D factor graph (b) 4D data association pattern

Fig. 10: (a) Factor graph of two rows with data association, connected vision factors are shared (matched) landmarks in two rows. (b) Data association pattern of 4D reconstruction.

Data association is performed across different rows and times to get Z_{cr}. Exhaustive search between all row pairs is not necessary, since distances or timespans that are too large make data association impossible. In our approach, we only match rows next to each other in either the space domain (near-by rows in the field), or the time domain (near-by date) using the proposed robust data association approach, as shown in Fig. 10(b).

The point cloud $\hat{L}$ is relatively sparse, since it comes from a feature-base SLAM pipeline, where only points with distinct appearance are accepted as landmarks (in our system SIFT key points are accepted). An optional solution is to use PMVS [15], which takes estimated camera states $\hat{X}$ to reconstruct dense point clouds.

III. Evaluation

A. Dataset

To evaluate the performance of our approach with real-world data, we collected a field dataset with large spatial and temporal scales. Existing datasets with both large scale spatial and temporal information include the CMU dataset [46], the MIT dataset [47], and the UMich dataset [48]. However, all of these datasets are collected in urban environments, and are not suitable for precision agriculture applications.

The dataset was collected from a field located in Tifton, GA, USA. The size of the field is about 150m×120m, and

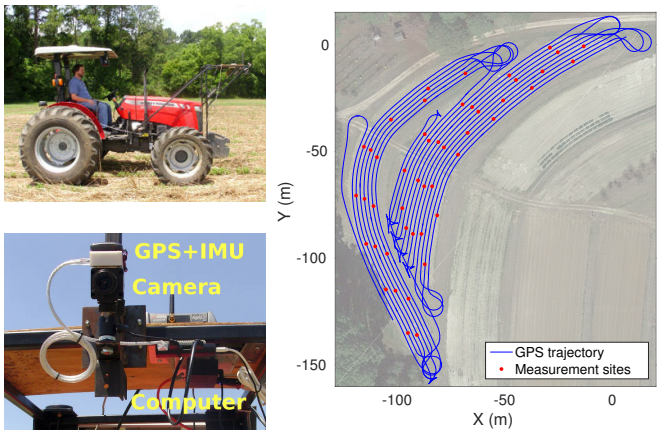


Fig. 11: Top left is the tractor collected the dataset; Down left shows sensors and computer (RTK-GPS is not shown); Right is a sample RTK-GPS trajectory, and sites of manual measurements are taken, overlay on Google Maps.

Fig. 12: Eight sample images taken at approximately same location in the field, dates taken are marked on images.

it contains total 21 rows of peanut plants. The map of the field is shown in Fig. 11.

We use a ground vehicle (tractor) equipped with multiple sensors, shown in Fig. 11, to collect all of the sensor data. The equipped sensors include: (1) a Point Grey monocular global shutter camera, 1280×960 color images are streamed at 7.5Hz, (2) a 9DoF IMU with compass, acceleration and angular rate are streamed at 167Hz, and magnetic field data is streamed at 110Hz, (3) a high accuracy RTK-GPS, and a low accuracy GPS, both of them stream latitude and longitude data at 5Hz. No hardware synchronization is used. All data are stored in a SSD by an on-board computer.

We recorded a complete season of peanut growth which started May 25, 2016 and completed Aug 22, right before harvest. The data collection had a total of 23 sessions over 89 days, approximately two per week, with a few exceptions due to severe weather. Example images of different dates are shown in Fig. 12. Each session lasted about 40 minutes, and consisted of the tractor driving about 3.8km in the field.

In addition to sensor data, ground truth crop properties (height and leaf chlorophyll) at multiple sampling sites in the field were measured weekly by a human operator. There were a total of 47 measuring sites, as shown in Fig. 11.

B. Results

We ran the proposed 4D reconstruction approach on the peanut field dataset. We implemented the proposed approach with the GTSAM C++ library.[1] We used RTK-GPS data from the dataset as GPS input, and ignored lower accuracy GPS data. Since the peanut field contains two sub-fields with little overlap (see Fig. 11), the two sub-fields were reconstructed independently. Since the tractor runs back and forth in the field, we only use rows in which the tractor driving south (odd rows), to avoid misalignment with reconstruction results from even rows. Example densely reconstructed 4D results are shown in Fig. 1.

Although Fig. 1 shows that the 3D reconstruction results for each single session qualitatively appear accurate, to make these results useful to precision agriculture applications, are interested in evaluating the approach quantitatively. In particular we wanted to answer the following questions:

- Are these 3D results correctly aligned in space?
- Are these 3D results useful for measuring geometric properties of plants useful for crop monitoring (height, width, etc.) ?

To answer the first question, we visualize the 4D model by showing all 3D point clouds together. We visualize part of the 4D sparse reconstruction result in Fig. 13. Point clouds from different dates are marked in different colors. We can see from the cross section that the ground surface point clouds from different sessions are aligned well, which shows that all of the 3D point clouds from different dates are registered accurately into a single coordinate frame. This suggests that we are building a true 4D result. We can see the growth of the peanut plants, as the point cloud shows 'Matryoshka doll' like structure, earlier crop point clouds are contained with in point clouds of later sessions.

For comparison, the 3D point clouds aligned by ICP [49] are also shown in Fig. 13. The ICP results are significantly worse than alignments computed by our proposed approach, since ICP can only compute a single rigid relative transformation for each point cloud pair. In contrast, the proposed approach can perform data association in multiple places in the field, equivalent to a 'non-rigid' transformation.

To answer the second question, we show some preliminary crop height results using reconstructed 4D point clouds, compared with ground truth manual measurements. We set up a simple pipeline to estimate the height of peanut plants from sparsely reconstructed 4D point clouds at multiple sites, by: (1) estimating local ground planes by RANSAC from May 25's point cloud for each site (when peanuts are small and ground plane is well reconstructed); (2) separating the peanut canopy's point cloud by color (using RGB values); and finally (3) estimating the distance from the peanut canopy's top to the local ground plane.

Preliminary height estimates during the whole growing season at 12 sampling sites are shown in Fig. 14. With the exception of sites 22 and 25, which have slightly biased height estimates due to poor RANSAC ground plane estimates, the results match the ground truth measurements well. For all of the sites, the root-mean-square(RMS) error of height estimation is 2.93cm. This is better compared to

[1] https://bitbucket.org/gtborg/gtsam

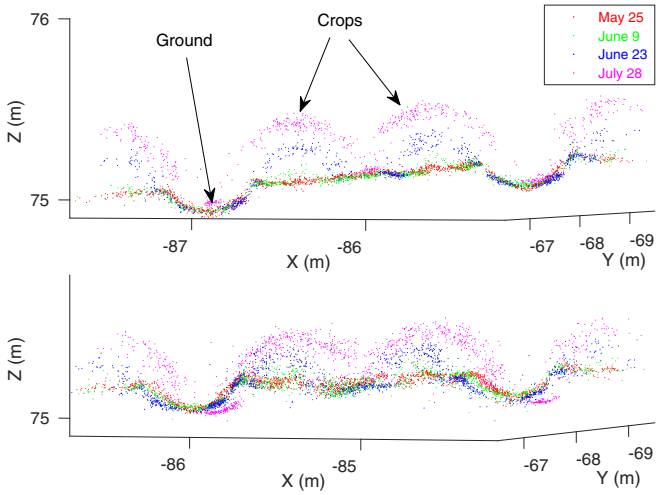


Fig. 13: Cross section of part of the sparse 4D reconstruction results at 3rd row. Upper subfigure is results of proposed approach, lower subfigure is results of ICP. Only 4 sessions are shown to keep figure clear. Best viewed in digital.

reported performances of LIDAR based methods [3], shows that we can compute reasonable height estimates even with a simple method, and proves that the 4D reconstruction results contain correct geometric statistics.

IV. Conclusion and Future Work

In this paper we address the 4D reconstruction problem for crop monitoring applications. The outcome of the proposed 4D approach is a set of 3D point clouds, with pleasing visual appearance and correct geometric properties. A robust data association algorithm is also developed to address the problems inherent in matching image features from difference dates and different view points, while performing 4D reconstruction. A side product of this paper is a high quality field dataset for testing crop monitoring applications, which will be released to the public.

Although we show some preliminary results of crop height analysis, this paper is mainly solving the problem of reconstructing the 4D point clouds, not the point cloud analysis. We leave the use of existing methods for analyzing point clouds [50], [51] or proposing more sophisticated point cloud analysis, as future work.

One possible extension to the proposed approach is further improvement of the data association process. Although our data association method outperforms existing approaches, it still needs assistance from high accuracy RTK-GPS as prior information, and we experience a few failure cases due to thunderstorms that wash out most features on the ground between sessions.

Acknowledgement

This work is supported by National Institute of Food and Agriculture, U.S. Department of Agriculture, under award number 2014-67021-22556. The authors thank Chuchu Zhang's help making supplementary video.

References

[1] F. Rembold, C. Atzberger, I. Savin, and O. Rojas, "Using low resolution satellite imagery for yield prediction and yield anomaly detection," *Remote Sensing*, vol. 5, no. 4, pp. 1704–1733, 2013.

[2] M. Bryson, A. Reid, F. Ramos, and S. Sukkarieh, "Airborne vision-based mapping and classification of large farmland environments," *J. of Field Robotics*, vol. 27, no. 5, pp. 632–655, 2010.

[3] D. Anthony, S. Elbaum, A. Lorenz, and C. Detweiler, "On crop height estimation with UAVs," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2014.

[4] 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 *IEEE Intl. Conf. on Automation Science and Engineering (CASE)*, 2015.

[5] K. Zainuddin, M. Jaffri, M. Zainal, N. Ghazali, and A. Samad, "Verification test on ability to use low-cost UAV for quantifying tree height," in *IEEE Intl. Colloquium on Signal Processing & Its Applications (CSPA)*, 2016.

[6] S. K. Sarkar, J. Das, R. Ehsani, and V. Kumar, "Towards autonomous phytopathology: Outcomes and challenges of citrus greening disease detection through close-range remote sensing," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2016.

[7] T. Hague, N. Tillett, and H. Wheeler, "Automated crop and weed monitoring in widely spaced cereals," *Precision Agriculture*, vol. 7, no. 1, pp. 21–32, 2006.

[8] J.-F. Lalonde, N. Vandapel, and M. Hebert, "Automatic three-dimensional point cloud processing for forest inventory," *Robotics Institute*, p. 334, 2006.

[9] S. Singh, M. Bergerman, J. Cannons, B. Grocholsky, B. Hamner, G. Holguin, L. Hull, V. Jones, G. Kantor, H. Koselka, *et al.*, "Comprehensive automation for specialty crops: Year 1 results and lessons learned," *Intelligent Service Robotics*, vol. 3, no. 4, pp. 245–262, 2010.

[10] S. Nuske, K. Wilshusen, S. Achar, L. Yoder, S. Narasimhan, and S. Singh, "Automated visual yield estimation in vineyards," *J. of Field Robotics*, vol. 31, no. 5, pp. 837–860, 2014.

[11] J. P. Underwood, G. Jagbrant, J. I. Nieto, and S. Sukkarieh, "Lidar-based tree recognition and platform localization in orchards," *J. of Field Robotics*, vol. 32, no. 8, pp. 1056–1074, 2015.

[12] D. Font, M. Tresanchez, D. Martínez, J. Moreno, E. Clotet, and J. Palacín, "Vineyard yield estimation based on the analysis of high resolution images obtained with artificial illumination at night," *Sensors*, vol. 15, no. 4, pp. 8284–8301, 2015.

[13] I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. McCool, "Deepfruits: A fruit detection system using deep neural networks," *Sensors*, vol. 16, no. 8, p. 1222, 2016.

[14] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building Rome in a day," in *Intl. Conf. on Computer Vision (ICCV)*, 2009.

[15] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multi-view stereopsis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 8, pp. 1362–1376, 2010.

[16] K. Sakurada, T. Okatani, and K. Deguchi, "Detecting changes in 3D structure of a scene from multi-view images captured by a vehicle-mounted camera," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[17] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," in *Robotics: Science and Systems (RSS)*, 2016.

[18] W. Xiao, B. Vallet, M. Brédif, and N. Paparoditis, "Street environment change detection from mobile laser scanning point clouds," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 107, pp. 38–49, 2015.

[19] T. Pollard and J. L. Mundy, "Change detection in a 3-D world," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[20] A. O. Ulusoy and J. L. Mundy, "Image-based 4-D reconstruction using 3-D change detection," in *European Conf. on Computer Vision (ECCV)*, 2014.

[21] R. Martin-Brualla, D. Gallup, and S. M. Seitz, "Time-lapse mining from internet photos," in *ACM SIGGRAPH*, 2015.

[22] R. Martin-Brualla, D. Gallup, and S. M. Seitz, "3D time-lapse reconstruction from internet photos," in *Intl. Conf. on Computer Vision (ICCV)*, 2015.

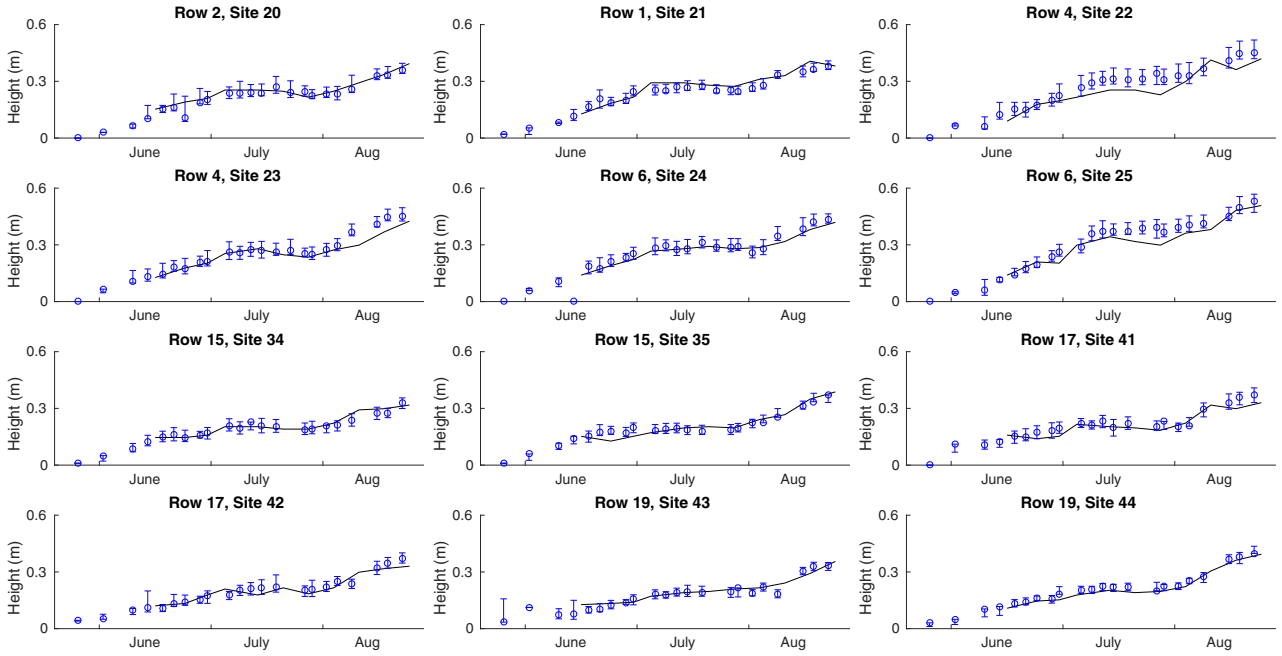


Fig. 14: Estimated peanut heights at 12 sampling sites in blue, with ground truth manual measurements in black lines.

[23] G. Schindler, F. Dellaert, and S. Kang, "Inferring temporal order of images from 3D structure," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[24] G. Schindler and F. Dellaert, "Probabilistic temporal inference on reconstructed 3d scenes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[25] K. Matzen and N. Snavely, "Scene chronology," in *European Conf. on Computer Vision (ECCV)*, 2014.

[26] A. O. Ulusoy, O. Biris, and J. L. Mundy, "Dynamic probabilistic volumetric models," in *Intl. Conf. on Computer Vision (ICCV)*, 2013.

[27] S. Griffith, F. Dellaert, and C. Pradalier, "Robot-enabled lakeshore monitoring using visual SLAM and SIFT flow," in *RSS Workshop on Multi-View Geometry in Robotics*, Citeseer, 2015.

[28] S. Griffith and C. Pradalier, "Reprojection flow for image registration across seasons," in *British Machine Vision Conf. (BMVC)*, 2016.

[29] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2012.

[30] T. Naseer, B. Suger, M. Ruhnke, and W. Burgard, "Vision-based markov localization across large perceptual changes," in *European Conf. on Mobile Robots (ECMR)*, 2015.

[31] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Trans. Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[32] C. Beall and F. Dellaert, "Appearance-based localization across seasons in a metric map," *IROS Workshop on Planning, Perception and Navigation for Intelligent Vehicles (PPNIV)*, 2014.

[33] M. Stein, S. Bargoti, and J. Underwood, "Image based mango fruit detection, localisation and yield estimation using multiple view geometry," *Sensors*, vol. 16, no. 11, p. 1915, 2016.

[34] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. J. of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[35] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 36, no. 11, pp. 2227–2240, 2014.

[36] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, second ed., 2004.

[37] F. Dellaert and M. Kaess, "Square Root SAM: Simultaneous localization and mapping via square root information smoothing," *Intl. J. of Robotics Research*, vol. 25, pp. 1181–1203, Dec 2006.

[38] L. Carlone, Z. Kira, C. Beall, V. Indelman, and F. Dellaert, "Eliminating conditionally independent sets in factor graphs: A unifying perspective based on smart factors," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2014.

[39] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation," *Robotics: Science and Systems (RSS)*, 2015.

[40] S. Anderson and T. D. Barfoot, "Full STEAM ahead: Exactly sparse gaussian process regression for batch continuous-time trajectory estimation on SE (3)," *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2015.

[41] T. Barfoot, C. H. Tong, and S. Sarkka, "Batch continuous-time trajectory estimation as exactly sparse gaussian process regression," *Robotics: Science and Systems (RSS)*, 2014.

[42] X. Yan, V. Indelman, and B. Boots, "Incremental sparse GP regression for continuous-time trajectory estimation & mapping," *Proc. of the Intl. Symp. of Robotics Research (ISRR)*, 2015.

[43] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping using the Bayes tree," *Intl. J. of Robotics Research*, vol. 31, pp. 217–236, Feb 2012.

[44] V. Indelman, E. Nelson, J. Dong, N. Michael, and F. Dellaert, "Incremental distributed inference from arbitrary poses and unknown data association: Using collaborating robots to establish a common reference," *IEEE Control Systems*, vol. 36, no. 2, pp. 41–74, 2016.

[45] C. Wu, B. Clipp, X. Li, J.-M. Frahm, and M. Pollefeys, "3d model matching with viewpoint-invariant patches (VIP)," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.

[46] H. Badino, D. Huber, and T. Kanade, "The CMU visual localization data set." `http://3dvis.ri.cmu.edu/data-sets/localization`, 2011.

[47] M. F. Fallon, H. Johannsson, M. Kaess, D. M. Rosen, E. Muggler, and J. J. Leonard, "Mapping the MIT stata center: Large-scale integrated visual and RGB-D SLAM," in *RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, 2012.

[48] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of Michigan north campus long-term vision and lidar dataset," *Intl. J. of Robotics Research*, vol. 35, no. 9, pp. 1023–1035, 2016.

[49] F. Pomerleau, F. Colas, R. Siegwart, and S. Magnenat, "Comparing ICP variants on real-world data sets," *Autonomous Robots*, vol. 34, no. 3, pp. 133–148, 2013.

[50] Y. Li, X. Fan, N. J. Mitra, D. Chamovitz, D. Cohen-Or, and B. Chen, "Analyzing growing plants from 4D point cloud data," *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia*, vol. 32, no. 6, p. 157, 2013.

[51] L. Carlone, J. Dong, S. Fenu, G. Rains, and F. Dellaert, "Towards 4D crop analysis in precision agriculture: Estimating plant height and crown radius over time via expectation-maximization," in *ICRA Workshop on Robotics in Agriculture*, 2015.