Robot Localization Based on Aerial Images for Precision Agriculture Tasks in Crop Fields

Nived Chebrolu Philipp Lottes Thomas Läbe Cyrill Stachniss

***Abstract*— Localization is a pre-requisite for most autonomous robots. For example, to carry out precision agriculture tasks effectively, a robot must be able to localize itself accurately in crop fields. The crop field environment presents unique challenges such as the highly repetitive structure of the crops leading to visual aliasing as well as the continuously changing appearance of the field, which makes it difficult to localize over time. In this paper, we present a localization system, which uses an aerial map of the field and exploits the semantic information of the crops, weeds, and their stem positions to resolve the visual ambiguity problem and to enable robot localization over extended periods of time. We evaluate our approach on a real field over multiple sessions spanning several weeks. Experiments suggest that our approach provides the necessary accuracy required by precision agriculture applications and works in cases where current techniques using typical visual features tend to fail.**

I. Introduction

An agricultural robot needs to localize itself accurately to navigate and perform treatment actions in the field effectively. Although several approaches for localization exist, the crop field environment poses several unique challenges, which are difficult to cope with. For example, the repetitive structure of the crops in the field gives rise to aliasing. This easily results in multi-modal distributions about the robot's pose that are difficult to resolve. In addition to this, the appearance of the vegetation in the field changes continuously over time, even on the same day. This makes it challenging to localize over multiple sessions, which is, however, a requirement in most precision agriculture applications.

Presently, the solution for operating in such environments is the use of high-precision real-time kinematic (RTK) GPS. Although these sensors are capable of providing the desired accuracy most of the time, they are rather expensive and are still vulnerable to signal outages resulting in degraded estimates. Several other localization approaches based on visual features such as SIFT [13] or similar features fail due to the large difference in the appearance of the field over the crop season, see also [1].

In this paper, we present a solution to the localization problem for ground robots operating in crop fields over long periods. It can also be used independently to provide redundancy to other localization systems such as GPS etc. Our system only requires the ground robot to be equipped with a monocular camera, an odometer, and uses an aerial map of the field as a reference. Such a map can be obtained easily from unmanned aerial vehicles (UAVs) flying over the field once. The key idea of our approach is to take advantage of the salient features in the field that remain invariant over the crop season, even when the visual appearance changes dramatically. We use the locations of the plants and the gaps in the field as features capturing the inherent geometry of the field and exploit the plant semantics to further tackle visual ambiguities. Furthermore, we also capture the changes in the field by explicitly modeling the existence of plants as a probabilistic belief and use this information to curate the map after each session.

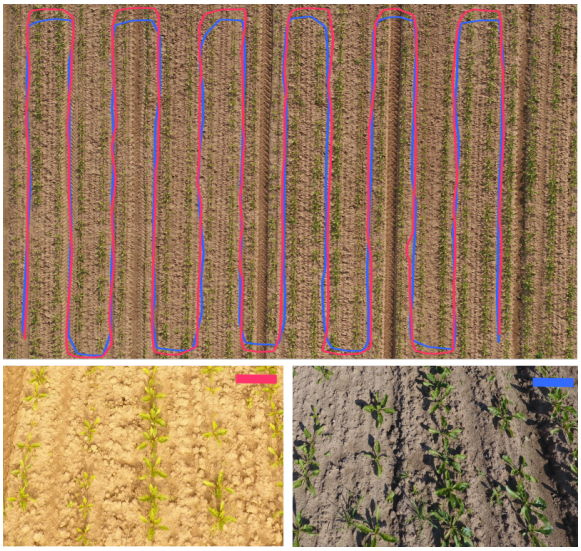

Fig. 1: Top: Robot trajectory estimated by our approach recorded at two different points in time (sessions) visualized on top of the aerial map used as a reference map. Bottom: Images from the ground robot recorded at same location but at the different times of data acquisition.

The main contribution of this paper is a novel localization system for robots operating in crop fields over an extended period of time. Initially, we construct a reference map as a set of sparse features encoding the geometry and semantics of the field using the images taken from a UAV. For localization, we use the feature detections from the ground robot within a Monte-Carlo localization algorithm to estimate the pose of the robot using an observation model targeted to the crop field domain. In addition to that, we update the map of the field at the end of each session based on the belief of the existence of each feature. The map update allows us to reduce the potential for wrong feature associations and thus improve the performance of the feature-based localization over time in changing environments such as agricultural fields.

All authors are with the University of Bonn, Germany.

This work has partially been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 - 390732324 - PhenoRob.

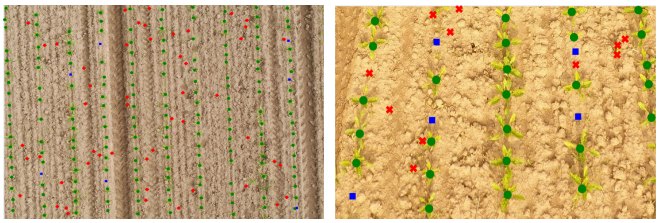

Fig. 2: Feature detections from an UAV (left) and ground robot image (right). We visualize crop (green), weed (red) and gap (blue) features detected from both views.

In sum, our approach is able to (i) localize with sufficient accuracy allowing the robot to navigate within the gap between side-by-side crop rows, (ii) provide better performance than standard GPS approaches and is more robust to environmental changes over season as compared to methods relying on purely visual features, (iii) perform localization over multiple sessions without needing to remap the field each time, and (iv) maintain an updated map of the environment by integrating the current measurements.

II. Related Work

In the past, several works have used aerial imagery as a reference map for localization. This is motivated by the fact that this information about the environment can be exploited to improve the localization quality. For example, Kümmerle *et al.* [8] show that by incorporating aerial imagery into a SLAM system, they are able to both localize better and acquire maps with increased global consistency.

The main challenge in exploiting information from aerial images is to find the data associations between the ground robot sensor data. This is often challenging due to large viewpoint difference between the two sources. To find these associations, several approaches such as [19], [14], [10], [9] propose new features that can be detected and matched against aerial images. These approaches typically employ a robust outlier rejection mechanism to deal with the large number of false correspondences. While other approaches such as [4], [20] use depth information to find vertical structures and other keypoints such as corners and planes from man-made structures which can be observed both from the ground and aerial views. However, most of the methods that we have discussed are tailored for urban environments having planar surfaces and vertical edges which is not the case for crop fields. But in the spirit of identifying features which can be commonly observed, we exploit plant and gap locations as features which are more suitable for agricultural fields as these locations do not change.

Several other approaches exploit semantic information from the environment to find better and robust features for the localization task. For example, Ruchti *et al.* [17] and Christie *et al.* [2] use range information from the laser scanner along with semantics of the environment for matching ground level images against aerial images or OpenStreetMap data. In the agricultural domain, some recent works such as [5], [21], [7] have aimed at exploiting situations specific to crop fields for performing data association. Taking inspiration from these works, we incorporated semantics of the plants in terms of crops and weeds which help both in finding correspondences and tackling the visual ambiguity problem.

III. Features for Localization in Crop Fields

In order to localize the robot using aerial images of fields, we need to find data associations between the UAV and the ground robot images. As these images are taken from two very different viewpoints, we need to extract features which are visible in both the images. In this section, we describe how to compute these features and use them as measurements for estimating the robot's pose.

A. Generating aerial landmark map

To generate the aerial landmark map, we capture images of the field from an UAV with a downward-looking camera over the whole field. We align these UAV images with a standard bundle adjustment procedure and estimate the camera poses as well as the digital elevation model of the field. Using these estimates, we stitch individual images to generate the aerial orthomosaic. This is done by the commercially available software Agisoft PhotoScan.

From this orthomosaic, we compute a landmark-based representation, which consists of stem locations of the plants. The main idea behind using the plant stem locations as landmarks is that they provide a representation of the field that is comparably static. In addition to the stem location, we further use the camera images to classify each plant as a crop or a weed and use this information for avoiding inter-class associations of the features during localization. Thereby, our approach takes advantage of the natural semantics of the field.

Both, the stem locations of individual plants and the semantic label, are estimated using an end-to-end trainable fully convolutional network (FCN) that we developed in [12]. In addition to that, we compute the gap locations between crops in the field, i.e., positions of missing crops on the field surface. We are able to do this because we expect the crops to grow in the row. The gap locations provide a more distinctive pattern, which allows us to tackle the problem of visual aliasing in row crops as described in [1]. From these features, we construct a map $\mathcal{M}$ of the environment as collection of landmark tuples

$$\mathcal{M} \;=\; \left\{ (l^{(1)}, s^{(1)}), \ldots, (l^{(L)}, s^{(L)}) \right\}, \qquad (1)$$

where $l = (l_x, l_y)^T$ is the location of the landmark in the global coordinate frame derived from the aerial reference map and $s \in \{\mathsf{crop}, \mathsf{weed}, \mathsf{gap}\}$ denotes its corresponding semantic label. Fig. 2 (left) show an example with landmarks computed for an UAV image with crop (green), weed (red), and gap (blue) features.

B. Extracting features from ground robot images

For the ground robot, we extract features for every incoming image to find a data association between the current observation and the aerial landmark map. We extract the same features as in the aerial image, i.e., the locations of crop

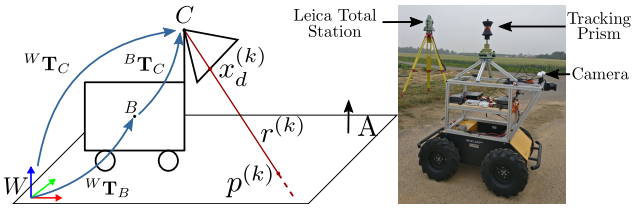

Fig. 3: Left: Projecting a feature $x_d^{(k)}$ from the image plane to the coordinate frame of the map $\mathcal{M}$. Right: Our Clearpath Husky robot and the tracking setup used to record ground truth trajectory in the crop fields.

and weed stems as well as gaps between the crops. However, extracting precise locations of plants from the ground robot images is more challenging as the camera orientation is tilted with respect to the nadir view. This camera setup results in images with varying ground sampling distance, where the farther parts of the field have a lower resolution. Furthermore, during the later stages when the crops are bigger, this view suffers from occlusions induced by large crops in the front of the scene. To deal with these challenges, we deploy a re-trained version of the same FCN [12], which we had used for the aerial data. To allow the FCN to detect stems from a perspective view, we fine-tune it by an additional training on a small portion of labeled image data captured from the ground robot. This yields in a set of feature detections for an image

$$\mathcal{F} \quad = \quad \left\{(x_d^{(1)}, s^{(1)}), \ldots, (x_d^{(K)}, s^{(K)})\right\}, \tag{2}$$

where x_d is the pixel location of a feature in the image coordinate frame and s denotes the semantic label of the detection. Fig. 2 (right) shows the feature detections for a ground robot image extracted using this procedure.

C. Camera projection

To match the features detected in the ground robot image, we project each detection $x_d^{(k)} \in \mathcal{F}$ onto the aerial map $\mathcal{M}$. For making this projection, we need to know the pose of the camera $^W\mathbf{T}_C$ and the parameters of the ground plane A in the world frame (illustrated in Fig. 3). We assume to have a calibrated camera and that the relative transformation from the robot base to the camera $^C\mathbf{T}_B$ is known . The pose of the robot base $^W\mathbf{T}_B$ is estimated by the localization algorithm explained in the next section.

To obtain the projected point $p^{(k)}$ in the map $\mathcal{M}$ corresponding to $x_d^{(k)}$, we first compute the direction of the ray $r^{(k)}$ in world coordinates using the camera calibration matrix $\mathbf{K}$ and the rotation matrix $\mathbf{R}$ from the $^W\mathbf{T}_C$ as

$$r^{(k)} \quad = \quad \mathbf{R}^\mathsf{T}\mathbf{K}^{-1}x_d^{(k)}. \tag{3}$$

Then, we compute the location $p^{(k)}$ of the feature observation on the plane A as the intersection of the ray $r^{(k)}$ and A [6]. This can be obtained rather elegantly by expressing $r^{(k)}$ in Plücker coordinates. We express $r^{(k)}$ as a line $L^{(k)}$ joining the camera projection center C and a point $q = C + r^{(k)}$ along the ray as

$$L^{(k)} = \begin{bmatrix} L_h \\ L_0 \end{bmatrix} = \begin{bmatrix} C - q^{(k)} \\ C \times q^{(k)} \end{bmatrix}. \tag{4}$$

From $L^{(k)}$, we compute the transposed Plücker matrix

$$\mathbf{\Gamma}^\mathsf{T}(L^{(k)}) \quad = \quad \begin{bmatrix} \mathsf{S}(L_0) & L_h \\ -L_h^\mathsf{T} & 0 \end{bmatrix}, \tag{5}$$

where $\mathsf{S}(L_0)$ is the skew symmetric matrix computed through the vector L_0. Finally, we obtain $p^{(k)}$ as

$$p^{(k)} \quad = \quad \mathbf{\Gamma}^\mathsf{T}(L^{(k)})A. \tag{6}$$

Due to the limited field of view of the camera, the number of features detected in a single image frame is often small (≈ 30). Typically, such data is not distinct enough to cope with the visual aliasing in the environment. Therefore, we aggregate features from consecutive images into a small sub-map untill it covers an area of 15 m^2. The accumulated data represents an observation for the particle filter described in the subsequent section.

This allows the accumulated observations to have sufficient features in order to be able to match against the aerial map effectively. These accumulated observations, i.e the set of all points $p^{(k)}$ and their semantics $s^{(k)}$ in the sub-map, form the measurement $\mathcal{Z}$ for our system

$$\mathcal{Z} \quad = \quad \{(p^{(1)}, s^{(1)}), \ldots, (p^{(N)}, s^{(N)})\}, \tag{7}$$

where p is location of feature projected in the global coordinate frame and s denotes corresponding the semantic label.

IV. Global Localization In Crop Fields

Due to repetitive structure of the environment, data association between the ground robot observations $\mathcal{Z}$ and the aerial landmark map $\mathcal{M}$ is potentially ambiguous. Additionally, detecting features from the ground robot images is noisy, and can result in false detections that have no correct associations in the map. This makes the application of EKF-based systems challenging. Therefore, we use the Monte-Carlo localization or MCL [3] to estimate the pose of the ground robot as it provides a natural way to better deal with multiple hypotheses.

MCL estimates a belief over the robot pose using a set of weighted particles where each particle represents a possible pose of the robot. For our implementation, we consider pose as the position and its orientation of the robot on the field surface. The MCL filter performs two main steps to maintain the belief over the pose. It first propagates the particles based on the odometry estimated by the wheel encoder measurements from the robot. We use the odometry motion model based on the wheel encoder readings as described in [18] to implement this step. Whenever a new measurement $\mathcal{Z}$ is available, it updates the weight of each particle based on an observation model. This model provides a measure of how well the observation agrees with the map given the current pose. Finally a new set of particles is re-sampled from the old ones, where the chance of survival for each particle is proportional to its weight in the old particle set.

We propose an observation model that takes into account the semantics of the features in addition to their locations. By considering the semantics, we are able to reduce the number

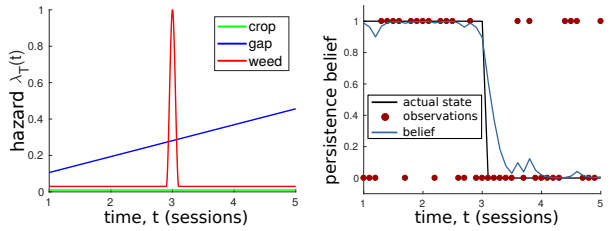

Fig. 4: Left: Hazard function $\lambda_T(t)$ for crop, weed and gap features specifying the prior information regarding their survival. Right: existence belief for a gap feature computed by the persistent filter.

of wrong data associations, which helps us deal with the aliasing in the field. We define an error $\xi^s(z_i)$ for each point of the semantic type s. The error $\xi^s(z_i)$ is computed as the distance to the nearest landmark l of the same semantic type s in the map $\mathcal{M}$. This means that we only associate crop features in the observations to crops in the map. Similarly, weeds and gaps in the observations are associated against their counterparts in the map.

We can compute $\xi^s(z_i)$ efficiently using a distance transform map $\mathcal{D}^s$, which is pre-computed separately for each feature type, i.e. crop, weed, and gap. $\mathcal{D}^s$ is essentially a look-up table providing the distance to the nearest landmark for each location in the map. Therefore, the error for the measurement z_i is obtained simply by looking-up the value in $\mathcal{D}^s$ at $p^{(i)}$, which is the projection of the feature on the map $\mathcal{M}$. We then compute the average error from all points belonging to a semantic type $Q_s = \{(p, \zeta) \in \mathcal{Z} | \zeta = s\}$:

$$\xi^s_{avg}(z, m) \quad = \quad \frac{1}{|Q_s|} \sum_{Q_s} \mathcal{D}^s(p^{(i)}), \tag{8}$$

and update the particle weight under our observation model as

$$w \quad \propto \quad \prod_s \exp\left(-\frac{\xi^s_{avg}}{\sigma_d}\right)^2, \tag{9}$$

where σ_d is the expected measurement noise in the feature detection. The average error ξ^s_{avg} is truncated to a maximum value for robustness against outliers, which is equivalent to using a truncated Gaussian. This turns our observation model into a likelihood field which can be evaluated efficiently.

V. Map Update

Typically, the map $\mathcal{M}$ used for localization is constructed only once at the beginning of the crop season. However, over time as the field appearance changes, new plants will appear and some of the existing ones may no longer be present. For example, new weeds appear in the field, while some of the gaps are no longer present when the crops increase in size. Therefore, to account for these changes in the field, we integrate the ground robot observations into our map $\mathcal{M}$ and curate it over time. If our observations were perfect, we would only need to remove a feature from the map if that feature's location is re-observed and the feature is not detected, and equivalently for the situation of adding a new feature to the map. In reality, however, the detector outputs are noisy, which compounded by the error in estimated pose of the robot itself, makes it difficult to determine unambiguously if a feature is present in the map or not. Therefore, the best we can achieve is to estimate a belief over the presence of the feature given the observations.

To compute this belief, we realize the so-called persistence filter described by Rosen *et al.* [15]. In addition to integrating the observations $\mathcal{Z}$, the persistence filter also provides an elegant way to incorporate prior information about the feature in terms of its expected survival time in the environment. One of the ways to integrate this survival prior is through a hazard function $\lambda_T(t)$, which encodes the information of how the feature disappearance rate varies over time. A hazard function $\lambda_T(t)$ allows us to describe the various changes occurring in the field in an intuitive manner. In our implementation, we design three different hazard functions to model the survival priors for crops, weeds and gaps. These three hazard function are visualized in the left image of Fig. 4 (left). For example, $\lambda_T(t)$ for crops (green) is very small and constant through out as we expect the crops to survive till the end of the season. Instead, $\lambda_T(t)$ for gaps (blue) decays over time as some of the gaps get covered by the nearby crops and are not detectable anymore. Finally, for weeds (red) we see a sharp rise at $t = 3$, this is to account for a weeding treatment performed at $t = 3$ on field after which we expect the majority of weeds to die.

Once the survival priors are defined, the filter fuses the observation at time t to update the feature existence belief. Thus, every feature maintains an existence probability that can be used to add/remove features. As an example, we show the belief computed by the filter for a weed feature in the field (Fig. 4, right). We observe that despite the false detections, the filter maintains a belief close to the ground truth by exploiting the prior information and using successive observations.

At the end of each session, we update the map $\mathcal{M}$ by removing the features whose existence belief is less than a fixed threshold. We also add the newly discovered features from the current session and initialize them with an existence probability of one.

VI. Experimental Evaluation

A. Dataset description

The experiments were performed on a real sugarbeet field, where we recorded data over several weeks. The images for generating the aerial map were taken at the beginning of the season using a DJI Phantom 4 UAV. These images were captured from a height of 10 m covering the whole field. The orthomosaic map generated from the images has a ground resolution of 5 mm per pixel. For recording the ground robot data, we use a Clearpath HuskyA200 equipped with wheel encoders, Ublox EVK-7 GPS and a ZED stereo camera. We only use the RGB images from the left camera for our experiments. The camera was mounted at a height of 1.2 m from the base, tilted at an angle of 45° towards the ground, see Fig. 3. We operated the ground robot by manually joysticking it with an average speed of 0.6 m/s. We collected the data over five different sessions, each roughly separated

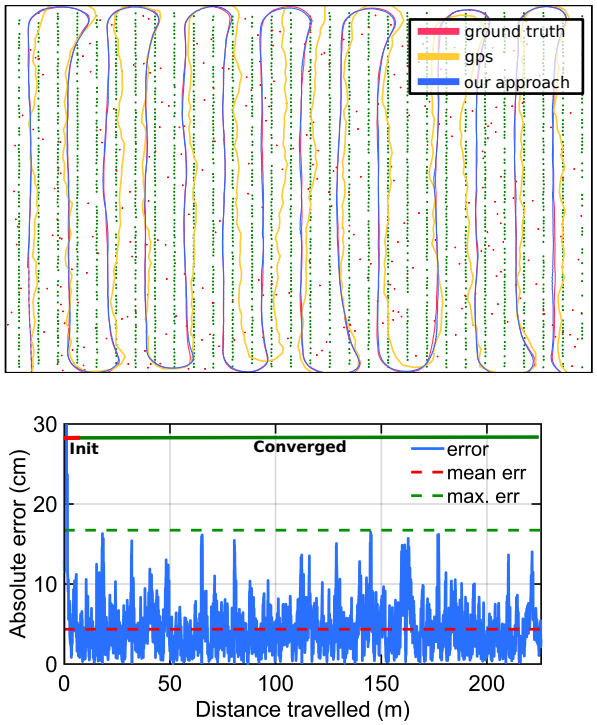

Fig. 5: Top: Comparison of the trajectory estimated by our localization approach against GPS and ground truth. Bottom: Absolute trajectory error over the whole trajectory.

TABLE I: Ablation study on localization performance

	Feature type	μ, σ (cm)	max (cm)
With Semantics	crops + weeds + gaps	(4.3, 2.8)	16.7
	crops + weeds	(5.8, 3.9)	18.3
Without Semantics	crops + weeds + gaps	(5.1, 3.1)	22.7
	crops + weeds	(6.6, 3.5)	28.4
	crops	(54.5, 3.5)	79.3

by a week. During this period, the crop size ranged between 5 cm to 20 cm in diameter. Additionally, a weeding treatment was performed by the farmers just before the third session where most of the weeds in the field were removed.

B. Localization accuracy

The first experiment is designed to support the claim that we are able to localize with sufficient accuracy required to carry out precision agriculture tasks in crop fields. This essentially requires that the robot both localizes in the correct crop row and is accurate enough to navigate within that row. This means a global accuracy of under 25 cm is required which is the inter-crop row distance in our field. This accuracy has to be achieved under changing appearance and strong visual aliasing.

We begin the experiment by initializing our MCL filter with 5,000 particles with an initial variance of 5 m around the estimate provided by the GPS. Here, we show the localization results for the first session using aerial map created on the same day. To visualize the performance of the filter, we plot the estimated mean pose of the robot (blue), GPS (yellow) and ground truth (red) measurements on the

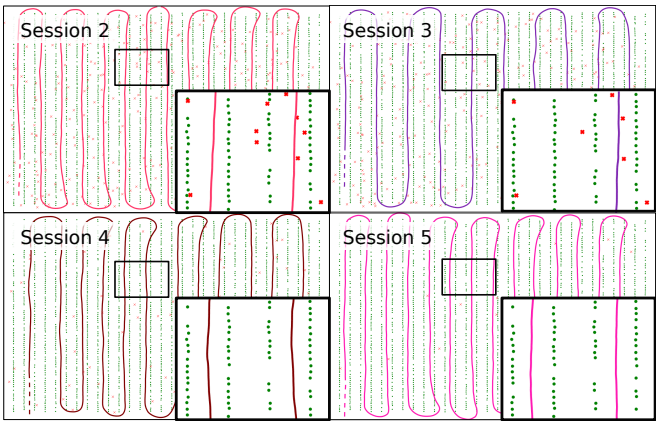

Fig. 6: Robot localizing over multiple sessions overlaid on updated reference map from the previous session. Dashed trajectory corresponds to the initialization phases. Zoomed-in view visualizes the changes in the map due to the update step reflecting the actual changes in the field.

map in Fig. 5, top. The filter converges after the robot travels a distance of about 6 m (dashed-blue). The ground truth measurements were obtained by tracking a prism target placed on top of the robot using a Leica Total Station TS50 with an accuracy under 1 cm.

In Fig. 5, top, we see that our approach (blue) provides a smooth estimate of robot's path along the crop rows whereas the GPS measurements often jump between the crop rows. We evaluate the accuracy of our trajectory estimate in terms of the absolute difference between our solution and the ground truth. We also note that our estimated trajectory is very close to the ground truth, indicated by the blue (ours) follows the red (ground truth) trajectory. Once the filter has converged, we obtain an average error of 4.3 cm, with the maximum error being around 17 cm. This error is less than the inter-crop row distance of 25 cm which is necessary to navigate safely without going over the crops. The localization error over the entire trajectory is plotted in Fig. 5, bottom.

Further, we perform an ablation study, which highlights the effect of the different feature types and semantics on the localization performance. The results are summarized in Tab. I. We observe that using detections from all features types, i.e. crops, weeds and gaps, provides the best performance. Also, we see that by additionally using semantics, localization the performance is better than using the same features but without the semantic information. In particular, the maximum error is lower while using the semantics. In the last row of the Tab. I, we observe that when using just the crop features (and not gaps), the filter estimate converges to wrong row indicated by its mean error of around 50 cm (shifted by two rows). This is caused by the high visual aliasing in the crop fields and indicates that crop locations alone are not sufficient to address this aliasing challenge.

C. Localization performance over multiple sessions

In this experiment, we demonstrate that our system is able to localize the robot successfully over multiple sessions spanning several weeks, whereas state-of-the-art methods

TABLE II: Localization performance over multiple sessions

Session	Initial Map μ (σ)[cm]	Initial Map max [cm]	Updated Map $\mu(\sigma)$ [cm]	Updated Map max [cm]
1	4.3 (2.8)	16.7	–	–
2	5.3 (3.6)	18.6	4.8 (2.9)	16.2
3	6.1 (3.8)	21.2	6.9 (3.5)	16.7
4	7.4 (6.2)	29.2	5.1 (3.3)	12.2
5	∞	∞	4.2 (3.8)	14.9

TABLE III: Number of features after each map update

Feature type	Session 1	Session 2	Session 3	Session 4	Session 5
crops	2472	2422	2406	2384	2353
gaps	417	341	221	211	206
weeds	306	303	76	14	22

relying on visual features are unable work properly. For this experiment, we update the map at the end of each session and use it as the reference for localizing the robot in the next session. Fig. 6 visualizes the estimated trajectory for sessions 2-5. We were able to localize successfully over all the sessions with an average error of about 5 cm and a maximum error of about 17 cm. Note that in Fig. 6, the trajectory from different sessions sometimes visit different crop rows. This is because the robot was actually joysticked through these rows and is not an error in the estimated trajectory. In the same figure, we see the zoomed-in view for a particular location in field, where the landmarks in the map have been updated based on the observations from the previous sessions. This allows the robot to localize accurately despite the changes in the field.

We also analyze the advantage of the map update step by comparing the performance against the setup where the initial map was used as the reference for all the sessions. The results are summarized in Tab. II. We see that in general, using the updated map results in better performance, both in terms of a lower mean and maximum error. In particular, we see that for session 5, the filter fails to localize using the initial map while it is successful while using the update map. This is due to the fact that the field changed substantially since the initial map was acquired.

As a qualitative evaluation of the map update step, we report the number of features in the updated map after each session in Tab. III. We note that the number of crops remain roughly the same over the whole season, and the number of gaps reduce gradually over time as crops grow and gaps are closed by the canopy cover. In particular, we can see that after the map update for session 3, the number of weeds drops from 303 in session 2 to 76 in session 3, reflecting the actual state of the field due the execution of weed control by the farmer. This number further goes down in session 4 when more measurements from the robot are integrated by the persistence filter described in Sec. V.

TABLE IV: Performance of visual matching across sessions

Desc. Type	% pairs matched successfully Session 1 vs. 1	Session 1 vs. 2	Session 1 vs. 3	Session 1 vs. 4	Session 1 vs. 5
SIFT [13]	93.8	25.0	12.5	8.3	4.2
ORB [16]	91.7	22.9	18.7	6.2	0
BRISK [11]	89.6	16.7	0	0	0

Our approach was able to localize over multiple sessions due to the combination of features that can be detected effectively in a crop field, and a map which is curated after each session using robot observations. In contrast, we were unable to localize over multiple sessions using visual features such as SIFT, ORB, BRISK or similar. This is because we are not able find data associations reliably between different sessions. Tab. IV reflects this situation where matched images from each session and the corresponding images UAV images taken from the first session. Here, we observe that while matching images from session 2 to session 1, about 75% percent of the images fail to match against corresponding images from session 1 when using SIFT descriptor for matching. The situation gets worse when matching images from session 5, where 96% of the pairs do not match. These results are consistent with the results obtained in [1].

VII. Discussion

Presently, our approach assumes the field to be locally planar while projecting the feature detections on the map, and in the MCL filter where we estimate the pose of the robot. In principle, to deal with fields with slopes, we can estimate the height by augmenting it to each particle in MCL and defining an appropriate motion update model. However, in our experiments we do not take it into consideration. Also, as our approach relies on the location of crops, weeds and gaps, it is suited for crop fields such as sugarbeet, carrot, maize, strawberry etc, but would not work for example in wheat/rice fields. Also note that based on the type of weed treatment, for example chemical treatment instead of mechanical removal, the prior introduced in the persistent filter needs to changed reflecting the actual physical process of the weeds dying slowly rather than instantly.

VIII. Conclusion

In this paper, we presented a novel approach for localizing robots in crop fields using aerial images. Our method exploits features specific to fields such as crop, weed and gap locations to find data associations. We also keep the map of the environment updated by integrating observations into the map. This allows us to successfully to localize for multiple sessions over the crop season, where typically methods relying on visual descriptors fail. We evaluated our approach on a real sugarbeet field and show that it is accurate enough to navigate along the crop rows safely and localize for multiple sessions over several weeks.

IX. Acknowledgments

We thank the team from Campus Klein-Altendorf near Bonn, Germany, for maintaining the sugar beet fields used for the experimental evaluation.

REFERENCES

[1] N. Chebrolu, T. Läbe, and C. Stachniss. Robust long-term registration of uav images of crop fields for precision agriculture. *IEEE Robotics and Automation Letters*, 3(4):3097–3104, 2018.

[2] G. Christie, G. Warnell, and K. Kochersberger. Semantics for UGV Registration in GPS-denied Environments. *arXiv preprint*, 2016.

[3] F. Dellaert, D. Fox, W. Burgard, and S. Thrun. Monte carlo localization for mobile robots. In *IEEE International Conference on Robotics and Automation (ICRA)*, May 1999.

[4] M. Ding, K. Lyngbaek, and A. Zakhor. Automatic registration of aerial imagery with untextured 3d lidar models. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[5] J. Dong, J.G. Burnham, B. Boots, G. Rains, and F. Dellaert. 4D Crop Monitoring: Spatio-Temporal Reconstruction for Agriculture. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2017.

[6] W. Förstner and B. Wrobel. *Photogrammetric Computer Vision – Statistics, Geometry, Orientation and Reconstruction*. Springer Verlag, 2016.

[7] F. Kraemer, A. Schaefer, A. Eitel, J. Vertens, and W. Burgard. From Plants to Landmarks: Time-invariant Plant Localization that uses Deep Pose Regression in Agricultural Fields. In *IROS Workshop on Agri-Food Robotics*, 2017.

[8] R. Kummerle, B. Steder, C. Dornhege, A. Kleiner, G. Grisetti, and W. Burgard. Large scale graph-based slam using aerial images as prior information. *Autonomous Robots*, 30(1):25–39, Jan 2011.

[9] T.B. Kwon and J.B.Song. A new feature commonly observed from air and ground for outdoor localization with elevation map built by aerial mapping system. *Journal of Field Robotics*, 28(2):227–240, 2010.

[10] K. Y. K. Leung, C. M. Clark, and J. P. Huissoon. Localization in urban environments by matching ground level video images with an aerial image. In *2008 IEEE International Conference on Robotics and Automation*, pages 551–556, May 2008.

[11] S. Leutenegger, M. Chli, and R. Siegwart. BRISK: Binary robust invariant scalable keypoints. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, pages 2548–2555. IEEE, 2011.

[12] P. Lottes, J. Behley, N. Chebrolu, A. Milioto, and C. Stachniss. Joint Stem Detection and Crop-Weed Classification for Plant-specific Treatment in Precision Farming. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2018.

[13] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Intl. Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[14] A. Majdik, D. Verda, Y. Albers-Schoenberg, and D. Scaramuzza. Air-ground matching: Appearance-based gps-denied urban localization of micro aerial vehicles. *Journal of Field Robotics*, 32(7):1015–1039, 2015.

[15] D. Rosen, J. Mason, and J. Leonard. Towards Lifelong Feature-Based Mapping in Semi-Static Environments. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2016.

[16] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: an efficient alternative to sift or surf. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2011.

[17] P. Ruchti, B. Steder, M. Ruhnke, and W. Burgard. Localization on OpenStreetMap Data Using a 3D Laser Scanner. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2015.

[18] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2005.

[19] T. A. Vidal-Calleja, C. Berger, J. Sol, and S. Lacroix. Large scale multiple robot visual mapping with heterogeneous landmarks in semi-structured terrain. *Journal on Robotics and Autonomous Systems (RAS)*, 59(9):654 – 674, 2011.

[20] X. Wang, S. Vozar, and E. Olson. FLAG: Feature-based Localization between Air and Ground. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2017.

[21] W. Winterhalter, F. V. Fleckenstein, C. Dornhege, and W. Burgard. Crop row detection on tiny plants with the pattern hough transform. *IEEE Robotics and Automation Letters*, 3:3394–3401, 2018.