2017 IEEE International Conference on Robotics and Automation (ICRA)
Singapore, May 29 - June 3, 2017

Robust Visual-Inertial Localization with Weak GPS Priors for Repetitive UAV Flights

Julian Surber, Lucas Teixeira and Margarita Chli
Vision for Robotics Lab, ETH Zurich, Switzerland

***Abstract*— Agile robots, such as small Unmanned Aerial Vehicles (UAVs) can have a great impact on the automation of tasks, such as industrial inspection and maintenance or crop monitoring and fertilization in agriculture. Their deployability, however, relies on the UAV's ability to self-localize with precision and exhibit robustness to common sources of uncertainty in real missions. Here, we propose a new system using the UAV's onboard visual-inertial sensor suite to first build a Reference Map of the UAV's workspace during a piloted reconnaissance flight. In subsequent flights over this area, the proposed framework combines keyframe-based visual-inertial odometry with novel geometric image-based localization, to provide a real-time estimate of the UAV's pose with respect to the Reference Map paving the way towards completely automating repeated navigation in this workspace. The stability of the system is ensured by decoupling the local visual-inertial odometry from the global registration to the Reference Map, while GPS feeds are used as a weak prior for suggesting loop closures. The proposed framework is shown to outperform GPS localization significantly and diminishes drift effects via global image-based alignment for consistently robust performance.**

Video— `https://youtu.be/TvxYG4MC7KU`

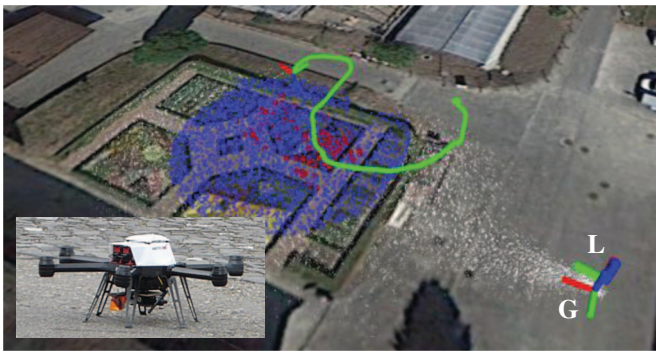

Fig. 1: Visualization of the real-time UAV localization during flight over a garden with respect to the Reference Map (white dots). Out of the blue points selected to register the current view to the Reference Map, the red ones highlight successful matches. The green line denotes the final estimated path of the UAV. The global (G) and the local (L) base frames are visible on the right, while the AscTec Neo UAV used in all our experiments is visible in the inset.

I. INTRODUCTION

Unmanned Aerial Vehicles (UAVs) have recently captured the interest of academia and industry as they can assist in a plethora of cumbersome tasks. Whether this is infrastructure inspection and maintenance, mapping and 3D scene reconstruction or field monitoring and fertilization the agility of small UAVs promises significant impact. One key challenge hindering progress in this direction is accurate robot localization; the robot needs to know the precise six Degree-of-Freedom (6DoF) pose (position and orientation) within space at any time. Exhibiting uncertainty of up to $7m$ with 95% confidence [1], GPS-based localization is not accurate enough to guide a UAV with the precision necessary in these tasks. More accurate localization can be achieved if the robot's environment is artificially modified. Differential GPS, for example, achieves much more precise localization employing a second GPS ground station [2], and so does Ultra Wide Bandwidth (UWB) using anchors on the ground [3]. These approaches, however, require modification of the environment and expensive infrastructure that is not always favourable or available. As a result, the literature has turned to alternatives employing onboard sensing cues only and performing Simultaneous Localization And Mapping (SLAM).

This research was supported by the Swiss National Science Foundation (SNSF, Agreement no. PP00P2_157585) and EC's Horizon 2020 Programme under grant agreement n. 644128 (AEROWORKS).

With one of the first systems to perform SLAM using a single camera in real-time, PTAM [4] pioneered the keyframes approach to SLAM, with a plethora of variants proposed since its emergence. A significant improvement in robustness and accuracy of visual SLAM has been achieved by incorporating an Inertial Measurement Unit (IMU) into the estimation [5] providing information about the gravity's direction and rather accurate estimates of the acceleration of the platform in the short term. Visual-Inertial (VI) SLAM has been approached from two different directions: recursive filtering and batch non-linear optimization. Filtering-based methods (e.g. [5]) tend to be computationally cheaper, while non-linear optimization based methods (e.g. [6]) achieve better accuracy. Following the findings of [7], demonstrating that non-linear optimization is preferable to filtering-based methods, we base our work on OKVIS [6], [8].

VI-SLAM using information within a limited time-horizon in the past (i.e. rendering loop-closures impossible), is often referred to as VI-odometry. It tracks landmarks through consecutive camera frames and estimates the 6 DoF pose of the robot and the 3 DoF position of the tracked landmarks simultaneously. One fundamental problem of VI-odometry is the unobservability of the global position and heading of the robot: Only an estimate of the robots pose with respect to the starting point is supported. A second problem of VI-odometry is drift: Due to the limited memory capacity and processing power onboard a UAV, old information constantly needs to be marginalized out and small discretization and

linearisation errors sum up over time. As a result, the uncertainty of the estimated pose is bound to increase with the duration of the mission. In order to address this, in this work we propose to perform a first reconnaissance flight in the region of interest and process all sensing data (visual, inertial and GPS) offline to compile an accurate, metric 'Reference Map'. Later on, in subsequent flights over the same area, new images can be registered to this Reference Map, providing localization of the current UAV pose to this map, essentially recovering the UAV's global position and heading and correcting for drift.

While there exist relevant works on combining image-based localization to a pre-built map with visual odometry (e.g. [9]), there is only a handful of works combining it with VI-odometry. In [10] for example, the estimation is based on recursive filtering and image-based localization is performed against a pre-built map with an efficient indexed nearest-neighbour search. Using a keyframe-based system subject to non-linear optimization instead, [11] combines it with an image-based localization against the pre-built map. In this work, we employ the same keyframe-based VI-odometry pipeline, but propose a novel image-based localization scheme that exploits the geometric relationship between the UAV and its environment, rendering the alignment to the Reference Map more robust and scalable. We demonstrate that the system successfully corrects for the VI-odometric drift by using GPS information to suggest a first, rough registration to the Reference Map and then refining it via 2D-3D feature matching between the current view of the UAV and the map. A snapshot of the proposed system in action is shown in Fig. 1. Similarly to [11], we keep the local tracking of the VI-odometry and the alignment to the map decoupled, but here we perform the alignment by a computationally more efficient recursive filtering approach. In a nutshell, the main contributions of this work comprise of:

- a novel, geometrically motivated image-based localization scheme against a geo-referenced map, which, together with a keyframe-based VI-odometry, forms an accurate localization system,
- a novel, computationally efficient filtering-based alignment to the Reference Map, which is decoupled from the local tracking and mapping, and
- demonstration of the performance of the system in real outdoor flights and evaluation of the estimation against accurate 3D position ground truth.

II. METHODS

During a first flight over the specific region of interest visual, inertial, and GPS data is captured from the onboard sensors to construct the 'Reference Map', which comprises of a set of 3D landmarks with their associated descriptors. The Reference Map is constructed after the flight in an offline stage using Structure from Motion (SfM) [12]. During a new flight within the same region, a keyframe-based VI-odometry is executed on the UAV to smoothly track the robot's pose extracting and tracking the same type of image features as in the first flight. To complete this local estimation, the visual features used within the VI-odometry are queried against the Reference Map and whenever an image is localized successfully, the local VI-odometry estimation is realigned to the Reference Map.

A. The Reference Map

Following a SfM paradigm, we use the sequence of images captured during the reconnaissance flight to construct the Reference Map of the desired workspace. For every image in this sequence, BRISK keypoints [13] are detected and described. Matching descriptors across frames leads to correspondences between 3D landmarks observed from different viewpoints. The correspondences can be used to simultaneously estimate the 3D position of the landmarks and the 6D pose of the images. This is done via non-linear optimization that minimizes reprojection error terms across all images. To estimate the metric scale of the map, we include IMU measurements into the optimization following the OKVIS tight VI fusion approach [6], [8]. The minimization cost function with reprojection error terms $\mathbf{e}_r^{k,j}$ and IMU error terms $\mathbf{e}_s^k$ is

$$J = \sum_{k=1}^{K} \sum_{j \in \mathcal{J}(k)} \mathbf{e}_r^{k,j\,T} \mathbf{W}_r^{k,j} \mathbf{e}_r^{k,j} + \sum_{k=1}^{K-1} \mathbf{e}_s^{k\,T} \mathbf{W}_s^k \mathbf{e}_s^k \ , \quad (1)$$

where K denotes the number of keyframes considered in the optimization, $\mathcal{J}(k)$ is the set of keypoints within keyframe k and $\mathbf{W}_r^{k,j}$ and $\mathbf{W}_s^k$ are the weights for the reprojection and IMU error terms, respectively. The optimization is performed in a sequential way, where the algorithm loops though the images in the order they have been captured. Comparing every frame with the latest keyframe, a frame is selected to serve as a new keyframe if their overlap is below a certain threshold. The optimization is run on all selected keyframes, while the frames in between consecutive keyframes are marginalized out. Once the optimization is finished, all points with a quality measure below a certain threshold are removed to reduce the memory-usage and decrease the processing time during online localization. As a last step, minimizing the euclidean distance between SfM and GPS trajectory, the Reference Map is geo-referenced. This could also be achieved by directly including GPS error terms in Equation 1.

B. Online localization

We divide the process of online localization into a local VI-odometry step and the alignment of this local estimation to the Reference Map. The odometry tracks the robot's pose based on the information of a camera and the IMU. The alignment is performed by two subsequent modules; the geometric image-based localization module localizes new images within the Reference Map based on a position prior, while the base frame transformation estimation module keeps track of the slowly changing transformation between the base frame of the map and the base frame of the odometry. An overview over the system used for online localization is outlined in Fig. 2.

We work with four different coordinate frames as illustrated in Fig. 3. The global base frame G is fixed in the world

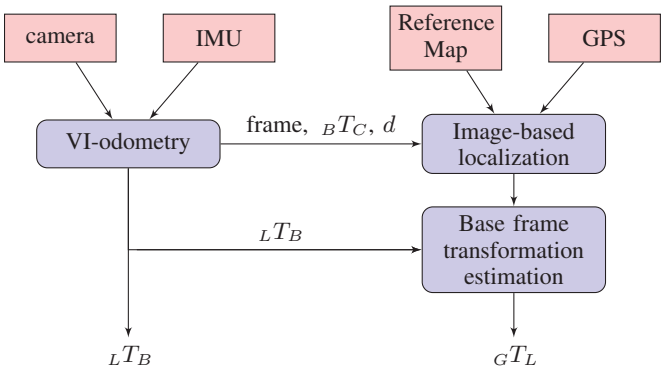

Fig. 2: Overview over the online localization system. Based on the information of a camera and an IMU, the VI-odometry estimates the transformation ${}_LT_B$ between the local base frame L and the body frame B. The image-based localization module localizes a new frame within the Reference Map. Based on successfully located images, the base frame transformation estimation module estimates the transformation ${}_GT_L$ between global and local base frame.

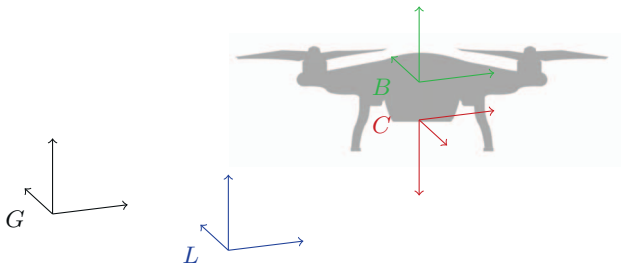

Fig. 3: The different coordinate frames used for the localization of the UAV: the body frame B and the camera frame C are rigidly attached to the UAV. The global base frame G is world-fixed and denotes the origin of all map points and the local base frame L is placed at the starting point of the UAV and forms the origin of the local VI-odometry estimation.

and serves as the origin for all saved Reference Map points. The estimation of the VI-odometry is performed with respect to the local base frame L. The body frame B is rigidly attached to the UAV and the camera frame C is attached to the optical center of the camera. The transformation ${}_BT_C$ between the body frame and the camera frame is calibrated online by the VI-odometry. The final goal of the online localization is to estimate the transformation ${}_GT_B$ between the global base frame and the body frame in real-time. This transformation can also be expressed as the product of the transformation between the global and local base frames, ${}_GT_L$, and the VI-odometry estimate ${}_LT_B$. Below, we describe each stage in this process.

1) VI-odometry: Tracking BRISK keypoints across images, the keyframe-based OKVIS system jointly minimizes the reprojection and IMU error terms according to the same cost function as in the offline stage (Equation 1) to estimate the position of new 3D landmarks and the pose of the robot with respect to the local base frame L, simultaneously. To achieve real-time operation, here, a sliding window of 3 keyframes and 2 recent frames is used, while older frames are marginalized out.

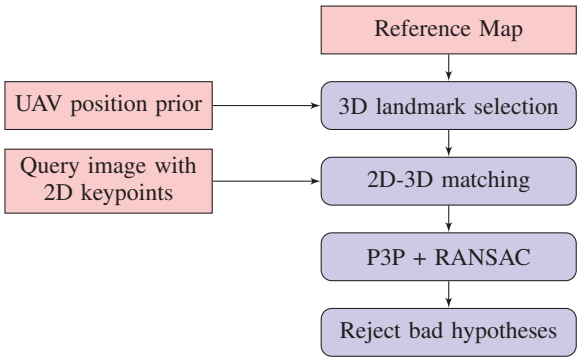

Fig. 4: Processing steps to register a query image (e.g. the current view of the UAV) to the Reference Map.

2) Image-based localization: To register an image to the Reference Map, the described 2D keypoints of this image need to be matched against the 3D landmarks of the map. Typically, the number of landmarks in the Reference Map is some orders of magnitude larger than the number of keypoints in the image. To handle this asymmetry, the map points can be clustered into visual words to speed-up the matching with a nearest neighbours search [14]. Such place recognition algorithms face a tradeoff between precision and recall – one wants to avoid false positives, while keeping the number of false negatives small. The geometric relationship between different 3D landmarks and between the robot and the landmarks is only considered in a later verification step. We propose, instead, a geometric landmark selection in the first step, which bounds the number of 3D points and allows for a brute-force 2D-3D matching. Fig. 4 gives an overview of the processing steps involved to localize a new image against the Reference Map.

We use a position prior to select a part of the Reference Map for the 2D-3D matching. In the beginning or during periods, where the uncertainty of the estimation with respect to the map is higher than the GPS uncertainty, we use the GPS measurement as a prior for the landmark selection. If the uncertainty of the estimation is smaller than the GPS uncertainty, which is the case during normal operation, the estimated pose with respect to the Reference Map is used as a prior. For the landmark selection, we make the assumption of a downward-looking camera and select all points that could potentially lie in the field-of-view of the camera. A landmark is selected if its horizontal world position, i.e. its (x, y)-coordinates, fulfils the following requirements:

$$\begin{aligned} |x - x_{prior}| &< r \\ |y - y_{prior}| &< \sqrt{r^2 - (x - x_{prior})^2}, \end{aligned} \tag{2}$$

with $r = d * \tan(FOV/2) + k * \sigma^2$, where d denotes an estimate for the average depth of the visible points, FOV the camera's field of view, k is a tuning parameter to define the uncertainty bounds of the prior to be searched for (where $k = 12$) and σ is the standard deviation of the prior in x and y direction. x_{prior} and y_{prior} denote the point in the scene at distance d along the optical axis of the camera.

Once a set of landmarks is selected from the Reference Map, we perform a brute-force 2D-3D matching between the BRISK descriptors of the new image and the BRISK descriptors of the selected landmarks. The resulting correspondences are prone to outliers. To remove these outliers and to estimate the most likely 6DoF pose of the image with respect to the map, we perform a Perspective-three-Point (P3P) algorithm in a RANSAC scheme [15]. As a last step, the estimated image pose is rejected if either the number of RANSAC inliers is smaller than 10 or if the (x, y)-position lies outside the 10-σ bounds of the GPS measurement. If an image-based localization passes these checks, the resulting transformation ${}_G T_{B,meas}$ is passed to the base frame transformation module.

3) Base frame transformation estimation: Successful image-based localizations within the Reference Map are noisy, accessed with latency, and appear only occasionally, i.e. whenever the appearance of the new image is similar to the Reference Map. These are the reasons why the estimate of ${}_G T_{B,meas}$ is not sufficient as real-time localization for the robot. We need to perform another important step, which constantly realigns the VI-odometry to the map and forms, together with the VI-odometry, an accurate real-time localization of the robot. We capture the transformation between the VI-odometry and the Reference Map by estimation of a base frame transformation that is modelled with 4 states, see below, performing a random walk. In contrast to [11], we formulate the base frame transformation estimation as a computationally cheaper recursive filtering problem and apply a Kalman filter framework [16]. The transformation between the global map base frame G and the local VI-odometry base frame L can be expressed by:

$$ {}_G T_L = {}_G T_B * {}_L T_B{}^{-1}. \tag{3} $$

For every successful image-based localization, we get an estimate for ${}_L T_B$ from the VI-odometry and an estimate for ${}_G T_B$ from the image-based localization. We initialize the filter with the first successful image-based localization. The filter performs two subsequent steps. It keeps the latest estimate and increases the uncertainty in a propagation step at a constant rate. Whenever a new image is successfully localized within the map, it corrects the belief in an update step.

Thanks to the IMU measurements, two rotational degrees of freedom (roll and pitch) are observed already by the VI-odometry, and we can model the base frame transformation with only four states, the translation (x, y, z) and the rotation around the world-z axis yaw (heading). The rotation around the other two axes is set to zero. In the end, the base frame transformation ${}_G T_L$ captures the unobserved global position and heading and the drift of the VI-odometry. Together with the VI-odometry estimate ${}_L T_B$, it forms an accurate localization system for the robot.

III. EVALUATION

We evaluate our system on real data, captured outdoors over a garden with size of $40m \times 25m$. The visual-inertial data was collected with the VI-Sensor described in [17] that consists of an ADIS16448 MEMS IMU and two WVGA monochrome cameras hardware-wise time synchronized. In this work, we focus on monocular-inertial estimation for a more generally applicable framework, so we only use the feeds of the left camera in all experiments. The images are captured with a frame rate of 20 Hz, have a resolution of 752×480 pixels and the cameras provide a field-of-view of $122°$. The GPS was acquired by the onboard GPS module of the AscTec Neo hexacopter, to which the sensor system was attached, as seen in Fig. 5. For the set of flights conducted for this experiment, the UAV flew around the same area mostly between $3m$ and $10m$ above the ground. Accurate 3D ground truth for the position of the UAV was acquired by a Leica Nova TM50 ground station, which is able to track a prism mounted on the UAV with sub-centimetre accuracy.

Fig. 5: The sensor system, providing visual (here we use only one camera) and inertial measurements, and a Leica prism underneath are attached to the AscTec Neo platform (left image). The garden environment, in which the first set of experiments have been conducted (right image).

A. Accuracy

We create the Reference Map with the data of the first flight of the UAV above the area of interest (i.e. flight no. 0) and evaluate the online localization on eight other flights performed at the same site. Fig. 6 shows the estimated trajectory in comparison to the ground truth trajectory for flight no. 1. Before the UAV localizes itself for the first time within the Reference Map, the absolute translation error is high, for the rest of the flight the median stays below $60cm$ (see Fig. 7).

Analysing the performance of the system on all eight flights (Table I), we can see, that the mean translation error lies between $30cm$ and $60cm$, which is an order of magnitude better than conventional GPS-based localization [1]. Due to overload of the onboard computational power (CPU), the VI-odometry drops frames from time to time (see Sec. III-B), which makes the system non-deterministic on the presented data. The results between different runs on the same data differ up to 20% and the numbers presented in Tables I and IV are avaraged values over three runs.

For all eight flights, the UAV localizes successfully within the Reference Map, even though no special attention has been spent on similar viewpoints or illumination across the flights. Flight no. 3 includes an exploratory sequence, in which the UAV left the boundaries of the Reference Map. During this exploration, the position errors increased, but as soon as the UAV was back in the Reference Map, the algorithm corrected for this by updating the base frame transformation. This sequence demonstrates the possibility of exploring new areas

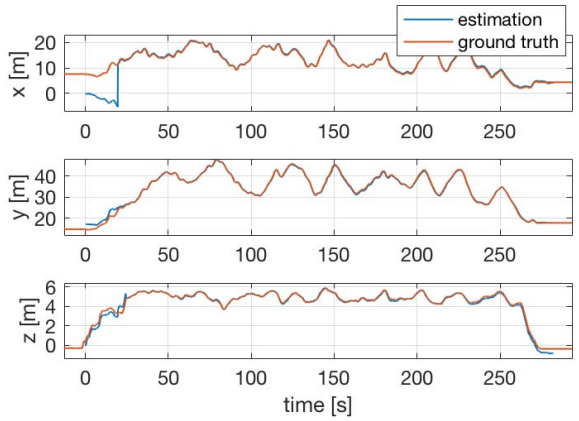

Fig. 6: The estimated position of the UAV during flight no. 1 in comparison to the ground truth trajectory. After the first successful image-based localization at about 20s in the flight, the robot can localize itself in the map and the estimated position jumps to the true value. From then on, the estimated position follows consistently the ground truth.

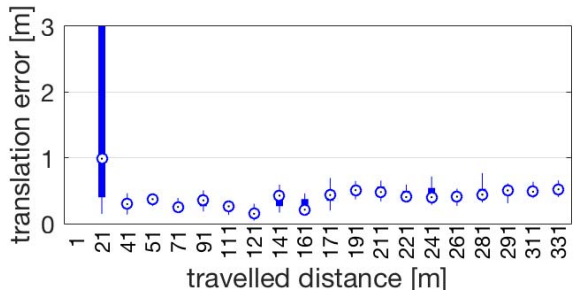

Fig. 7: The absolute translation error, which describes the Euclidean distance between the estimated position and ground truth, for flight no. 1 over the travelled distance. The error statistics for a set of data points that has been collected after a certain travelled distance (between value on x-axis and next value) are given by median and 25th and 75th percentiles.

around the Reference Map to get additional data, which could be potentially used to extend the Reference Map later on.

The errors in the online localization are mainly caused by errors within the Reference Map. To evaluate the Reference Map, we compare the SfM trajectory with the ground truth trajectory for the reference flight. The mean absolute translation error for the SfM trajectory of the map is $0.26m$ with a standard deviation of $0.13m$, which lies in the same order of magnitude as the final localization accuracy. This implies that the online localization could be improved, if a more accurate Reference Map can be provided, e.g. using denser and computationally more demanding approaches [18].

B. Processing time

All the experiments have been performed in real-time on a 3.1GHz Intel Core i7 CPU processor onboard the UAV. The processing times to query a new image against the Reference Map and perform the base frame transformation estimation can be seen in Table II. The matching of keypoints within the new image against the selected 3D landmarks is the computationally most expensive part of the image-based localization. The Kalman filter update achieving the base frame transformation estimation takes a negligible amount of time and is computationally cheaper than similar optimization based approaches (e.g. [11]).

flight	travelled distance [m]	translation error [m]	std [m]
1	325	0.39	0.12
2	97	0.50	0.30
3	305	0.87*	0.91*
4	65	0.42	0.16
5	79	0.41	0.25
6	59	0.56	0.16
7	95	0.30	0.15
8	19	0.57	0.16

TABLE I: The mean translation error lies between $30cm$ and $60cm$ for all eight evaluated flights within the area of the Reference Map. The standard deviation describes the error over the flight sequence. $*$ in flight 3, the UAV left the map and the error increased. Removing this exploratory trajectory from the evaluation, the translation error diminishes to $0.54 \pm 0.27m$.

Step	Time [ms]
Image-based localization	**32**
landmark selection	3
3D-2D matching	26
P3P + RANSAC	3
Base frame trans. estimation	**0.03**

TABLE II: Timings used to query a new image against the Reference Map, averaged over all 24 runs. The matching step is the most expensive part of the algorithm. The filtering step achieving the base frame transformation estimation is computationally very inexpensive.

The system runs on full CPU load with several frames being dropped due to overload. Slightly better results than shown in Sec. III-A are achieved, if the system is slowed down such that the information of all frames can be explored as shown in Table III. To this end, the system was slowed down by a factor between 2 and 3, depending on the distance to structure, which varies on each flight.

flight	translation error [m]	std [m]
1	0.34	0.10
2	0.48	0.18
3	0.83*	0.92*
4	0.48	0.15
5	0.35	0.13
6	0.46	0.33
7	0.21	0.10
8	0.42	0.10

TABLE III: The mean translation error for the slowed-down system that explores the information of all frames (no frame dropping), lies between $20cm$ and $50cm$.
$*$ in flight 3, the UAV left the Reference Map and the error increased. Removing this exploratory sequence from the evaluation, the translation error for flight no. 3 is $0.49 \pm 0.25m$.

C. Base frame transformation estimation

The transformation between the global map base frame G and the local odometry base frame L is described by the translation (x, y, z) and the rotation around the world-z

axis, yaw (heading). Fig. 8 shows these estimates for flight no. 1. The estimates are initialized with the first successful localization within the Reference Map and updated based on every new localization. The estimates perform a random walk capturing the drift of the VI-odometry estimation.

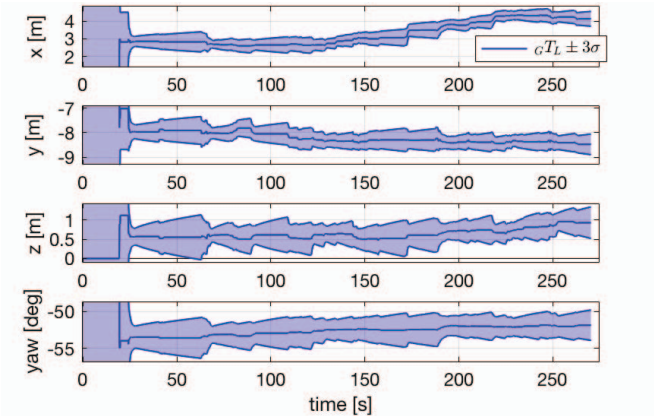

Fig. 8: The transformation between the global map base frame G and the local VI-odometry base frame L for flight no. 1. The transformation is described by the position (x, y, z) and rotation around the world-z axis, yaw. The estimates perform a random walk capturing the drift of the odometry. The uncertainty, described by the 3σ uncertainties, increases during periods without successful localizations.

D. Changing appearance

In outdoor sceneries, appearance is constantly changing depending on the lighting conditions, the weather, the season and environmental changes such as growing plants. To get an insight into the robustness of the system under appearance changes, we performed repeated flights at the same site two weeks later. The appearance of the garden changed in the meantime too much to let the system localize itself successfully within the garden, see Fig. 9. At the same time, the appearance around the garden stayed similar enough for successful localization permitting successful estimates of base frame transformation.

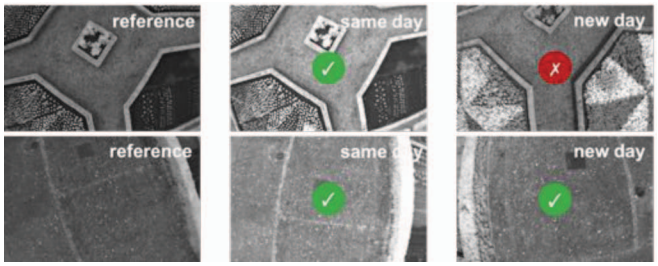

Fig. 9: Changing appearance: On the left side we see two images used to create the Reference Map, in the middle images of another flight at the same day, and on the right side images of a flight two weeks later. Successful image-based localizations are only possible in parts of the map where the appearance is similar enough to the reference (indicated by a green check mark).

E. Comparison to VI-odometry without a Reference Map

As stated in Sec. I, any VI-odometry lets the robot localize itself with respect to a local coordinate frame. Without knowledge about its absolute position and orientation, the robot has to reason about the environment and make sophisticated decisions by itself. Providing the robot with a map in which it can localize itself and, hence, providing it with knowledge about its absolute position and orientation, enables the robot to do path planning or follow a user-defined path. Moreover, via constant realignment of the VI-odometry to the Reference Map, the proposed approach compensates for the drift within the VI-odometry. As illustrated in Fig. 10, performing the VI-odometry alone (i.e. without registering the pose with respect to the Reference Map), the errors in the estimation are constantly increasing, while the errors of the proposed system stay bounded as long as images are successfully registered to the Reference Map from time to time.

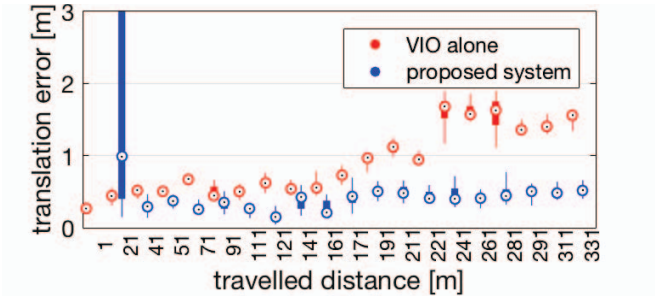

Fig. 10: Comparison of the errors of the proposed system and the VI-odometry (VIO) without registering images to the Reference Map for flight no. 1. While the errors of the odometry alone are increasing over time, the errors of the proposed system stay bounded.

F. Operation in GPS-denied environments

The proposed system uses GPS measurements as a weak prior for 3D landmark selection whenever the estimation accuracy of the system itself is low. This is the case before initialization or after longer periods, where no successful localization within the Reference Map is acquired (see Sec. II-B.2). If GPS is not available and no other position prior is provided, the processing time for the initialization is increased, as the descriptors of a new image are queried against all points within the map. Moreover, the robustness of the initialization is reduced, since the probability of a false localization is increased. This initialization process could be enhanced by a bag of words approach in future work. However, for the data used to evaluate the system, we did not experience any false initialization and only a marginal increase in processing time during initialization if the system is run without GPS. Exploring the GPS priors becomes more important in larger work spaces (see also [19]), which will be demonstrated in future work.

G. Increasing the map size

If the current SfM-pipeline is applied to larger workspaces, the errors within the map scale with the map size. To get

an insight into the effects of an increased workspace, we evaluate the system on additional data that was captured on a vegetable field of $100m \times 60m$, where the UAV was flown in a relative altitude between $3m$ and $20m$. An impression about the environment and a sample image out of the sequence are shown in Fig. 11. We generate the Reference Map from data of a first flight and evaluate the system on three different flights, see Table IV.

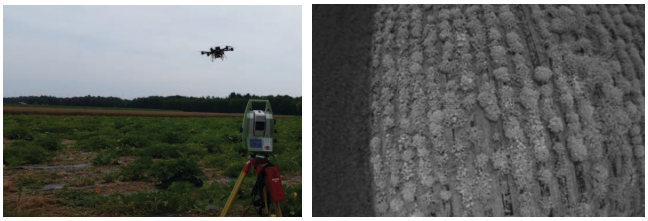

Fig. 11: To evaluate the system in a larger environment, the UAV was flown above a vegetable field and tracked by a Leica ground station to acquire ground truth data (left). The vegetable field shows typically repetitive structure as seen in the sample image (right).

flight	travelled distance [m]	translation error [m]	std [m]
A	378	1.14	0.47
B	440	1.05	0.54
C	506	1.65	0.89

TABLE IV: The mean translation error with standard deviation for the flights over the vegetable field lies between $1m$ and $1.7m$.

The mean translation errors are higher than for the smaller workspace due to higher errors within the Reference Map. The mean translation error of the SfM-trajectory is $0.97m$ with a standard deviation of $0.64m$ and as such in the same order of magnitude as the final localization accuracy.

IV. CONCLUSIONS

Constructing a Reference Map during a first flight in the workspace of interest, this paper presents a system to localize against this map in subsequent flights using primarily visual and inertial cues, as well as weak GPS priors, which are often available (albeit highly unreliable) in outdoor environments. UAV localization is achieved by combining keyframe-based VI-odometry with a novel image-based localization within the Reference Map that exploits the geometric relationship of the features in the current view and the 3D landmarks of the Map. Successful localizations within the Map are used to estimate the transformation between a local odometry base frame and a global map base frame in real-time, decoupling the local tracking and the alignment to the map. Hereby, the estimation of the base frame transformation is achieved by a computationally efficient recursive filtering algorithm.

The proposed system is evaluated on an extensive testbed of real outdoor flights against accurate 3D position ground truth acquired by a Leica tracking station. The framework bounds the otherwise continuously accumulating drift in traditional VI-odometry systems and achieves localization accuracy an order of magnitude better than conventional GPS localization. Moreover, tests on larger environments and varying scenery and illumination conditions reveal consistently robust and accurate performance. With robotic perception constituting the largest hurdle before robots are employed in real missions, this work promises to bring UAV systems a step closer to real deployment.

Future directions will focus on refinement and densification of the Reference Map for increased localization accuracy, on augmentation to larger maps, and on extension of the Reference Map on-the-fly in the case of visiting adjacent areas in a new flight.

REFERENCES

[1] W. H. Center, "Global Positioning System (GPS), Standard Positioning Service (SPS), Performance Analysis Report. Submitted to Federal Aviation Administration, Washington." Report no. 93, 2016.

[2] R. Siegwart, I. Nourbakhsh, and D. Scaramuzza, *Introduction to autonomous mobile robots.* MIT press, 2011.

[3] K. Guo, Z. Qiu, C. Miao, A. Zaini, C. Chen, W.Meng, and L. Xie, "Ultra-wideband-based localization for quadcopter navigation." Unmanned Systems 4.01, pages 23-34, 2016.

[4] G. Klein and D. W. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2007.

[5] A. I. Mourikis and S. I. Roumeliotis, "A multi-state kalman filter for vision-aided inertial navigation," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2007.

[6] S. Leutenegger, P. Furgale, V. Rabaud, M. Chli, and R. Siegwart, "Keyframe-based Visual-Inertial SLAM using Nonlinear Optimization," in *Proceedings of Robotics: Science and Systems (RSS)*, 2013.

[7] H. Strasdat, J. M. M. Montiel, and A. J. Davison, "Real-time monocular SLAM: Why filter?" in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2010.

[8] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual-inertial odometry using nonlinear optimization," *International Journal of Robotics Research (IJRR)*, vol. 34, no. 3, pp. 314–334, 2015.

[9] H. Lim, J. Lim, and H. J. Kim, "Real-time 6-dof monocular visual slam in a large-scale environment," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[10] S. Lynen, T. Sattler, M. Bosse, J. Hesch, M. Pollefeys, and R. Siegwart, "Get out of my lab: Large-scale, real-time visual-inertial localization," in *Proceedings of Robotics: Science and Systems (RSS)*, 2015.

[11] H. Oleynikova, M. Burri, S. Lynen, and R. Siegwart, "Real-time visual-inertial localization for aerial and ground robots," in *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, 2015.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[13] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary Robust Invariant Scalable Keypoints," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011.

[14] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *International Journal of Robotics Research (IJRR)*, vol. 30, no. 9, pp. 1100–1123, 2011.

[15] X. Gao, X. Hou, J. Tang, and H. Cheng, "Complete solution classification for the perspective-three-point problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 25, no. 8, pp. 930–943, 2003.

[16] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, vol. 82, pp. 35–45, 1960.

[17] J. Nikolic, J. Rehder, M. Burri, P. Gohl, S. Leutenegger, P. Furgale, and R. Siegwart, "A synchronized visual-inertial sensor system with FPGA pre-processing for accurate real-time SLAM," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[18] R. A. Newcombe, S. L. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[19] O. Vysotska, T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Efficient and effective matching of image sequences under substantial appearance changes exploiting gps priors," in *International Conference on Robotics and Automation (ICRA)*, 2015, pp. 2774–2779.