GPS-Denied Global Visual-Inertial Ground Vehicle State Estimation via Image Registration

Yehonathan Litman*, Daniel McGann*, Eric Dexheimer, Michael Kaess

***Abstract*— Robotic systems such as unmanned ground vehicles (UGVs) often depend on GPS for navigation in outdoor environments. In GPS-denied environments, one approach to maintain a global state estimate is localizing based on preexisting georeferenced aerial or satellite imagery. However, this is inherently challenged by the significantly differing perspectives between the UGV and reference images. In this paper, we introduce a system for global localization of UGVs in remote, natural environments. We use multi-stereo visual inertial odometry (MSVIO) to provide local tracking. To overcome the challenge of differing viewpoints we use a probabilistic occupancy model to generate synthetic orthographic images from color images taken by the UGV. We then derive global information by scan matching local images to existing reference imagery and then use a pose graph to fuse the measurements to provide uninterrupted global positioning after loss of GPS signal. We show that our system generates visually accurate orthographic images of the environment, provides reliable global measurements, and maintains an accurate global state estimate in GPS-denied conditions.**

I. Introduction

A global state estimate is often crucial to robotic platforms during autonomous navigation. In particular, planning algorithms require a global state estimate whenever their mission objectives are tied to global locations. When available, a GPS receiver is the best source for global state information. These sensors are relatively accurate and good signals are common in most places. However, GPS is not infallible: natural and urban terrain can disrupt GPS signals, GPS can be jammed in adversarial settings, and the global navigation satellite system itself can experience failures. Failing to provide global localization estimates can at best impede a robot's operation and at worst result in a failed mission and the loss of the robot. We therefore concern ourselves with overcoming these GPS failure modes by providing a global state estimate to an unmanned ground vehicle (UGV) after the loss of signal.

We introduce a system for real-time global position estimation in remote, natural environments using preexisting aerial or satellite imagery. Our system consists of four modules. First is a multi-stereo visual inertial odometry (MSVIO) module that provides robust local odometry using multiple stereo-camera pairs. Second is a mapping module that uses a probabilistic 3D occupancy model to generate visually accurate synthetic orthographic images of the UGV's local surroundings. Third is a registration module that derives global state measurements by registering images from the mapping module with georeferenced aerial or satellite imagery via a robust scan matching algorithm. Finally, the fourth module combines MSVIO measurements with registration results in a global pose graph to provide a continuous and consistent global state estimate. Our complete system can operate in real-time. In sum, we make four main contributions:

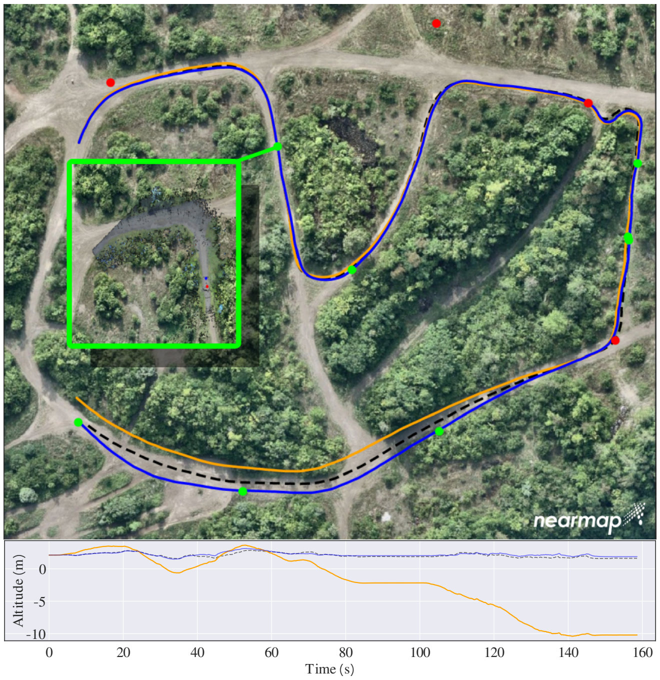

Fig. 1: An example of the localization process where synthetic orthographic images generated by the UGV are matched to corresponding locations in the reference aerial imagery, with an example alignment shown in the green box. Successful registrations, or green dots, are used as global measurements to correct the drift of the orange MSVIO trajectory, resulting in a blue corrected trajectory. After global optimization, the corrected trajectory exhibits less drift from the dashed black ground truth trajectory than MSVIO.

1) We develop a MSVIO formulation for efficient and fault-tolerant local state estimation.
2) We overcome challenges to existing work with an image generation method that provides visually accurate orthographic views of the UGV's local environment.
3) We provide a full localization pipeline to provide real-time global position estimates.
4) We compare our method to the state of the art for GPS-denied visual localization on real world datasets.

* The first two authors contributed equally to this work.

Yehonathan Litman, Daniel McGann, Eric Dexheimer, and Michael Kaess are all with the Robotics Institute, Carnegie Mellon University: {ylitman, dmcgann, edexheim, kaess}@andrew.cmu.edu.

This material is based upon work supported by the U.S. Army Research Office and the U.S. Army Futures Command under Contract No. W911NF-20-D-0002. The content of the information does not necessarily reflect the position or the policy of the government and no official endorsement should be inferred.

II. Related Work

Historically, the standard solutions to GPS-denied localization are dead reckoning and simultaneous localization and mapping (SLAM). These methods are well studied, efficient, and widely used. However, both solutions can drift relative to a global frame even with known initial state. SLAM solutions account for drift with loop closures but such measurements require re-visitations which, in general operation, cannot be assumed. An alternate method to loop closures that can correct drift in GPS-denied environments is to compare sensor measurements to georeferenced information. Such measurements can be combined with dead reckoning or a SLAM solution to provide global state estimation.

A popular source of georeferenced information for urban environments is a high definition (HD) map [4], [22]. HD maps provide rich and highly accurate reconstructions of roadways and enable recovering accurate global state. However, these maps and the algorithms that use them to localize rely on known semantic meaning of urban infrastructure (e.g. lane markings, street signs). Additionally, the creation and maintenance of HD maps is prohibitively expensive even in urban centers [8], [24], [27], [29]. Constructing and maintaining these maps for much larger natural, remote environments would be impractical given current state-of-the-art methods and even if constructed these map's reliance on known semantic meaning would significantly reduce their efficacy in remote environments.

In this paper we focus on GPS-denied localization in natural remote environments where the lack of semantic structures necessitates different approaches to GPS-denied localization than in urban centers. To operate in natural, remote environments we first need a source of georeferenced information collected and maintained at scale in these environments. There are two clear candidates for this role: 1. Aerial and satellite imagery, referred to jointly as "reference imagery", and 2. Digital Elevation Models (DEMs).

One method to recover global state using a DEM is to perform horizon matching [3], [25]. The horizon's profile is extracted from UGV images and matched to a DEM giving a rough location of the UGV. However, all horizon matching methods assume a clear view of the horizon. This assumption is violated when operating in and around vegetation or structures. Another method is to construct a local elevation model that can be matched against the reference DEM [14]. However, matching to the reference DEM requires observation of non-planar terrain features that can well constrain the UGV position. These features exist in the DEM at the scale of hills, mountains, and valleys. The local elevation model would therefore have to be large enough to contain such features. It is very likely that any UGV local state estimate will drift significantly before a sufficiently large model could be constructed. Such drift would result in the construction of a self-inconsistent local model that would not match against the reference DEM.

Unlike DEMs, reference imagery contains features at a scale that is practical for a UGV to observe. However, UGV localization to reference imagery is challenged by the significant view point difference; a UGV sees a scene much differently than a satellite does. [35] addresses this challenge by warping a 360° image onto the ground plane (assumed flat) to approximate the viewpoint of the reference imagery and compared to reference imagery using a whole image SIFT descriptor. However, it is noted in this work that where the flat ground assumption is violated (by vegetation, objects, buildings, etc.), significant artifacts appear in the resulting image which leads to decreased performance.

Another approach to tackle the view-point challenge is to learn a deep model to embed corresponding ground and reference images closely in feature space [18], [21]. These methods are notably inspired by the geolocalization work of [1], [36] with the added complexity of handling "cross view" image pairs. Both of these methods suffer from high variance individual measurements and, conditionally, poor generalization. The later of which is of particular importance given natural remote environments are often underrepresented in training data.

A parallel line of work has studied localization of aerial vehicles using reference imagery, where the view point difference is often negligible. Given a common viewpoint, methods similar to those explored for UGVs are possible including deep feature matching [2], [33], and classical feature matching [32]. In addition, many more measurement methods are possible including visual scan matching [9], visual feature matching [5], semantic feature alignment [6], [23], and pose optimization [13], [28], [37]. Many of these methods provide more accurate, lower variance measurements than those for ground images and enable the use of modern optimization techniques for recovering a global state estimate [7], [20], [26], [30].

Related work has shown promising results for UGV localization relative to reference imagery, yet the issue of viewpoint difference remains. This motivates our work to create a UGV localization pipeline that allows for the construction and registration of synthetic top down images that are visually accurate to existing reference imagery even in the presence of surrounding vegetation. Furthermore, by fusing the registration results with visual inertial sensing we can conduct high accuracy global state estimation in real time.

III. Methodology

Our global localization pipeline consists of four modules that carry out the following consecutive operations: 1) obtain local position estimates from MSVIO, 2) use the MSVIO estimates to build a local map, 3) register the local map to reference aerial imagery with scan matching, and 4) fuse the global registration measurements with the local odometry from MSVIO into a pose graph to produce global position estimates.

A. Multi-Stereo Visual-Inertial Odometry

Our MSVIO module is driven by the design described in [19]. Instead of running multiple independent VIO algorithms across individual cameras, we opt to track features across frames from all camera pairs and gather the features

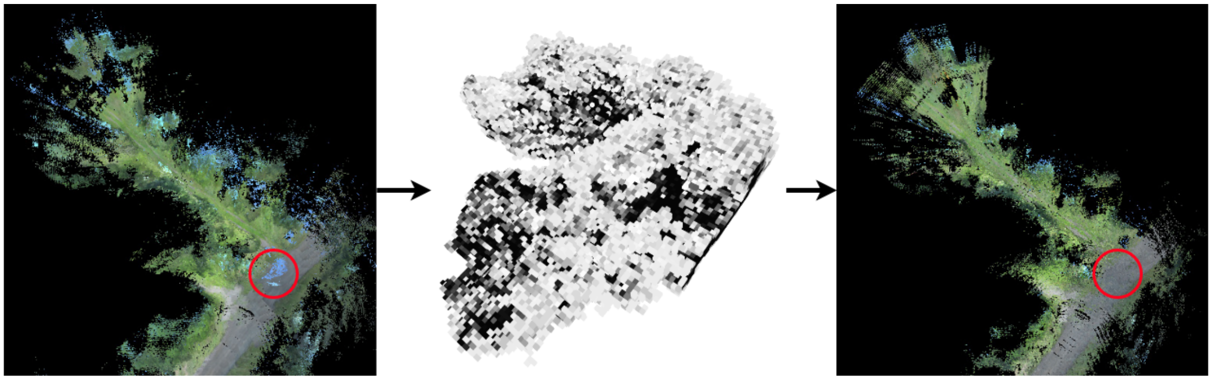

Fig. 2: A local map shown on the left can exhibit artifacts like blue sky pixels due to noise in disparity. By explicitly modeling the occupancy probability in a 3D occupancy grid, we can filter out voxels of low occupancy probability by rendering a local map with only voxels of high occupancy probability, shown in dark in the center image. This removes the most significant artifacts from the final local map shown on the right. For visualization, we ignore voxels with occupancy probability below 0.5.

into a single set. Given the disparity calculated via semi-global block matching (SGBM) [15] from the previous frame, points can be triangulated and matched to existing 2D features from the current frame. Instead of performing RANSAC with a generalized camera model, which may require a large sample size, we opt for a simpler solution. Points from the front camera are selected for P3P, but the inlier check is performed across all cameras. Since the side cameras are easily occluded by vegetation, they may not always provide reliable points, while the front camera does as it faces the direction of motion. Finally we pass the collection of inliers and inertial measurements into a fixed-lag smoother to jointly optimize for the relative motion of the UGV.

MSVIO is substantially more robust than the more common single stereo VIO. The main advantage of using multiple cameras is a wider field of view. In situations where one stereo pair captures an image that may be too challenging for tracking visual features, the system uses the features from other frames that are tracking well. Thus, the system is able to compute accurate odometry in challenging scenarios where traditional single stereo VIO approaches would fail. Alternatively, this can be done with a single fish-eye camera, but at the cost of reduced resolution.

B. *Local Map Construction*

With the position estimation from the MSVIO and color images from each stereo pair, the local map construction module generates the image used for the global registration process. First, we use the computed disparity to project a dense sampling of pixels from each image into 3D space around the robot. This provides us with a 3D point cloud for which each point has an associated RGB value. From this step we could directly generate the orthographic image by spatially binning this point cloud into a 2D image.

However, stereo matching often provides noisy results when applied in real-world scenarios. This causes significant artifacts in the resulting local map, as shown in Fig. 2, and would in turn decrease registration performance. To address this challenge we accumulate points into a 3D probabilistic occupancy grid based on the binary Bayes filter derived in [16]. Since our UGV traverses over long distances, we implement a scrolling occupancy grid which is centered around the vehicle and purges voxels that are outside the grid's bounds. This ensures the local map remains visually consistent with the reference imagery and not affected by drift from MSVIO.

With the stereo depth data $\mathbf{z}_{t_1:t_2}$ and VIO poses $\mathbf{x}_{t_1:t_2}$, the probability that a voxel is occupied or free is denoted as $p(v \mid \mathbf{z}_{t_1:t_2}, \mathbf{x}_{t_1:t_2})$. Note that t_1 to t_2 is the timeframe where the voxel is inside the 3D grid. This can be efficiently computed with a log-odds formula that uses the prior occupancy probability, which we set as $p(v) = 0.5$ as we do not have any occupancy information at the beginning:

$$l(v \mid \mathbf{z}_{t_1:t_2}, \mathbf{x}_{t_1:t_2}) = \log\left(\frac{p(v \mid \mathbf{z}_{t_1:t_2}, \mathbf{x}_{t_1:t_2})}{1 - p(v \mid \mathbf{z}_{t_1:t_2}, \mathbf{x}_{t_1:t_2})}\right) \tag{1}$$

The sensor model we use for determining occupancy raytraces from the vehicle position until it hits the position of the voxel containing the computed 3D point from stereo depth. Using the log-odds equation, the occupancy probability is incremented at the hit voxel and decremented for missed voxels. Our stereo depth model weighs hits higher than misses, and constrains the maximum and minimum occupancy probability for each voxel as in [16]. In addition to occupancy, our model tracks the color for each occupied cell by interpolating the color of all points within it independently for each channel.

Using this tracked occupancy and color information we generate the local map as a synthetic orthographic image. A 2D image is initialized to the exact width and length as the occupancy grid. Each pixel in the image is colored using the color information provided by the top most cell at the corresponding position in the occupancy grid whose occupancy probability is greater than a predefined threshold. An example local map generated with and without our occupancy modeling is shown in Fig. 2.

C. *Registration*

We can derive global state measurements by registering the local maps onto reference imagery. In addition to the local map, our registration algorithm requires a current global state estimate that is provided by the global pose graph module.

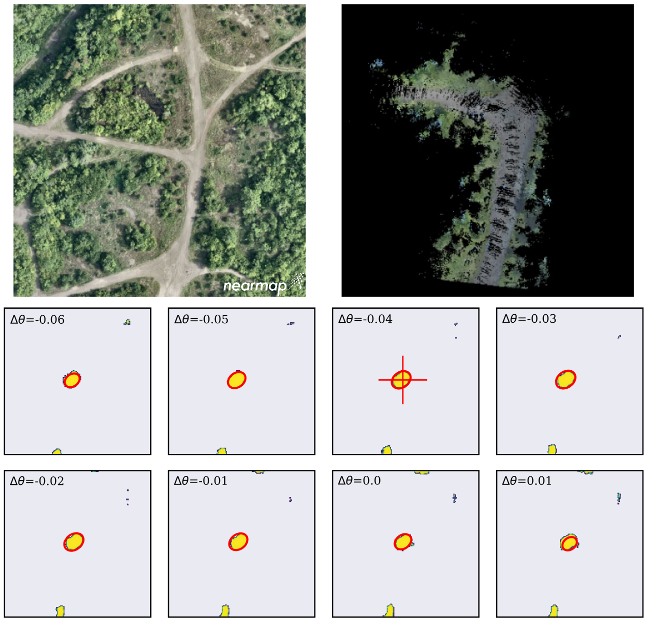

Fig. 3: The search region in the top left is extracted from reference imagery around the current global position estimate. The local map in the top right is matched against this region with 3D scan matching. The 3D cost volume $\mathbf{C}^{th}$ after thresholding is shown on the bottom for a subset of search angles. Overlaid on the cost volume is the optimum's location and covariance denoted by the red "+" and ellipses, respectively.

With the global pose estimate and local map, we perform scan matching over translation and rotation differences $\Delta\mathbf{r} = [\Delta x, \Delta y, \Delta\theta]^\top$ between the local map and a subset of the georeferenced imagery defined around the current global state estimate. We then extract the optimum from the resulting volume, as shown in Fig. 3. This, along with the known location of the reference image and the vehicle's position relative to the local map resolve the vehicle's global location and heading. Finally, to fully constrain the vehicle's translation we perform a lookup on a DEM at the measured global location to determine the UGV's altitude.

The scan matching process can make use of any similarity or difference measure. Our algorithm uses normalized cross correlation (NCC). Due to sparsity in the local map we employ a variant of NCC that uses an image mask $\mathbf{M}$ to calculate the cost volume $\mathbf{C}$ between our reference imagery $\mathbf{R}$ and the local map $\mathbf{T}$. For a single rotation angle, such that $\mathbf{T}$ has been rotated by $\Delta\theta$ around the vehicle's location in the local map to form $\mathbf{T}_{\Delta\theta}$, we compute NCC as

$$\mathbf{C}_{\Delta\mathbf{r}} = \frac{\sum_{i,j}(\mathbf{T}_{\Delta\theta_{i,j}} \cdot \mathbf{R}_{\Delta x+i,\Delta y+j} \cdot \mathbf{M}_{i,j})}{\sqrt{\sum_{i,j}\left(\mathbf{T}_{\Delta\theta_{i,j}} \cdot \mathbf{M}_{i,j}\right)^2 \cdot \sum_{i,j}\left(\mathbf{R}_{\Delta x+i,\Delta y+j} \cdot \mathbf{M}_{i,j}\right)^2}} \tag{2}$$

This is performed for each $\Delta\theta$ in the search space to construct the cost volume.

An alternative to scan matching is to perform a nonlinear optimization over the cost function. However, this cost function is non-convex and therefore optimization is highly susceptible to converging to local optima. Scan matching provides a global (or pseudo-global given we limit our search to a region) view of the cost function. Therefore, at the cost of computation, scan matching ensures that we find the true optimum within the search region. Additionally, the pseudo-global view of the cost function enables us to perform covariance estimation and outlier rejection that would not be possible within an optimization based registration algorithm.

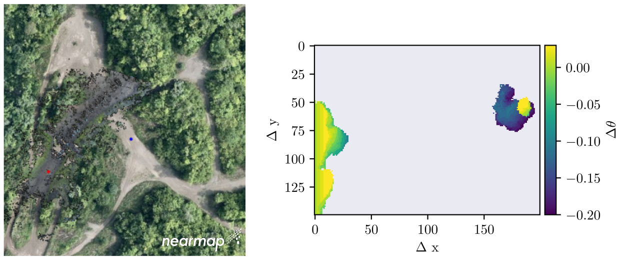

Fig. 4: An example of an outlier registration. On the right is a top down view of the non-zero entries of the cost volume after thresholding. We can see three distinct modes in this volume indicating a poor or ambiguous measurement. On the left is the alignment according to the cost volume optimum. The correct registration would align the red and blue dots at the proper angle.

To calculate the covariance we first threshold $\mathbf{C}$ to retain only weights that are within one standard deviation of the optimum to create $\mathbf{C}^{th}$. The remaining non-zero entries represent weighted samples from the measurement distribution. Next, the weights for these samples are normalized into probabilities $p(\Delta\mathbf{r})$. NCC weighs are strictly positive and normalized according to

$$p(\Delta\mathbf{r}_i) = \frac{\mathbf{C}^{th}_{\Delta\mathbf{r}_i}}{\sum_j \mathbf{C}^{th}_{\Delta\mathbf{r}_j}} \tag{3}$$

and the covariance is calculated with the mean μ as

$$\Sigma_{\Delta\mathbf{r}} = \sum_i p(\Delta\mathbf{r}_i)(\Delta\mathbf{r}_i - \mu)(\Delta\mathbf{r}_i - \mu)^\top \tag{4}$$

The costmap produced by scan matching also allows for robust outlier rejection. We expect a good registration to produce a single peak within the interior of the cost volume. This indicates that the search region contains what is likely the global optimum and that this optimum is well defined and unique. This expected behavior leads to two heuristics used for outlier rejection. First, a measurement is considered an outlier if the optimum lies on the edge of the cost volume. Such positioning suggests that the true optimum is outside the current search region and the registration should be performed again. Second, a measurement is considered an outlier when less than a specified proportion (e.g., 90%) of samples in the $\mathbf{C}^{th}$ are within the same 6-neighbor connected component as the optimum. This condition is violated when there are multiple significant peaks indicating a poor or ambiguous registration. An example of an registration identified as an outlier by these heuristics can be seen in Fig. 4.

D. Global Registration Pose Graph

After we get the global measurements, we pair them with the local odometry estimates from MSVIO into one pose graph optimization scheme, motivated by [30]. We represent

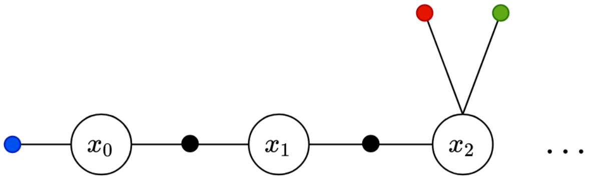

Fig. 5: An illustration of the global pose graph. The white nodes represent the 6 DOF pose of the vehicle in the global frame, the black factors represent the MSVIO relative constraints, the red and green factors represent the registration and elevation constraints, respectively, and the blue factor is a prior on the initial state.

our estimation as a maximum a posteriori (MAP) problem where we estimate the poses of all frames up to a time t

$$\mathcal{X}_t = \{\mathbf{x}_0, \dots, \mathbf{x}_t\} \tag{5}$$

For our scenario, MSVIO measurements are used as the relative constraints between states and registration and elevation measurements, denoted by h, are used as unary factors. The elevation factor is constructed using the elevation value obtained directly from a DEM at the given registration measurement coordinates and a constant Gaussian noise derived from the DEM's resolution. We also impose a 6 DOF prior which comes from the assumption that our localization pipeline starts after loss of GPS signal and therefore that the initial state is known. The complete pose graph scheme, or solution to the MAP, is seen in Fig. 5. This is under the assumption that the measurement noises follow a zero-mean noise Gaussian distribution, and thus the MAP solution simplifies to a nonlinear least-squares problem [11] as

$$\begin{aligned}\mathcal{X}_t^\star = \underset{\mathcal{X}_t}{\operatorname{argmin}} & \underbrace{||\mathbf{x}_0||^2_{\Sigma_0}}_{\bullet\ \text{Prior}} + \sum_{i=1}^{t} \Big(\underbrace{||P(\mathbf{x}_{i-1}, \mathbf{x}_i)||^2_{\Sigma_P}}_{\bullet\ \text{MSVIO Factor}} \Big) \\ & + \sum_{i=1}^{t/N} \Big(\underbrace{||H(\mathbf{x}_{Ni}, h_{Ni})||^2_{\Sigma_H}}_{\bullet\ \text{Elevation Factor}} + \underbrace{||R(\mathbf{x}_{Ni}, \mathbf{r}_{Ni})||^2_{\Sigma_R}}_{\bullet\ \text{Registration Factor}} \Big)\end{aligned} \tag{6}$$

where the measurement covariances for the corresponding factors are $\Sigma_0, \Sigma_P, \Sigma_H, \Sigma_R$, $||v||^2_\Sigma$ is the squared Mahalanobis distance of v, and N is the number of frames between adding registration and elevation factors.

It is important to note that sequential MSVIO odometry measurements are in reality correlated, as features can be tracked between sequential segments. They are, however, assumed independent in the factor graph. We build the graph using the GTSAM framework [10] and incrementally optimize in real time as MSVIO and registration measurements are acquired. Since this is a nonlinear problem, we solve using the Gauss-Newton method.

IV. Experimental Validation

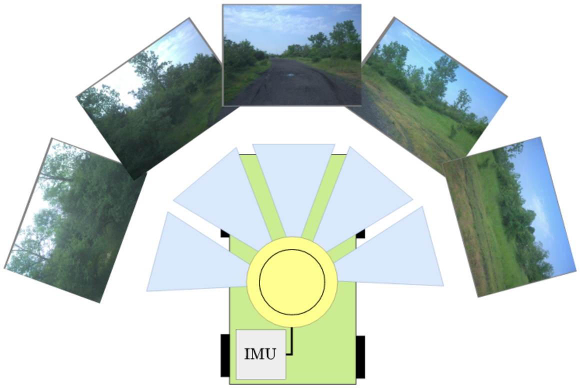

Fig. 6: A diagram of the vehicle used for data collection, with an example of the stereo images (one from each stereo pair). The vehicle was equipped with five stereo pairs as well as an IMU module, all time synchronized with an FPGA.

Data was collected by a UGV platform with 5 stereo pairs which are synchronized with an IMU through an on-board FPGA, shown in Fig. 6. The extrinsic parameters of the cameras were estimated using the method described in [12]. Images are captured at a rate of 4 Hz, and IMU outputs data at 100 Hz. The vehicle was driven around a field testing site in Pittsburgh to collect data for two trajectories, with ground truth provided by real time kinematic GPS.

Reference aerial imagery of the test site was acquired from a third party[1]. The imagery was captured in the same season but a year prior to data collection at a resolution of approximately 0.23 meters per pixel. We use a DEM from the National Elevation Dataset [34]. We generate and register the local map against the reference imagery every 50 frames.

For comparison we implement two alternative methods based on [35] and [18], referred to as “ORB” and “CVM” respectively, owing to the basis of their implementation. For more details, we refer the reader to Sec. II. Our implementations of both methods use the same particle filter and a motion model derived for the data collection platform. We make two modifications to our implementation of [35]. First, we use the open source ORB descriptor [31] in favor of the SIFT descriptor used in the original work. Second, we compute the query descriptor on our synthetic local map images to match the view point achieved by image warping in the original work. For the approach described in [18], we use the publicly available pre-trained weights for the CVM-Net-II model [17]. We compute the CVM query descriptor from a panorama stitched from the UGV's forward facing cameras to match the panoramic image format with which the network was trained. For both we report the trajectory taken by the location of the most probable particle at every timestep. We also compare against an alternate version of our method, referred to as "Ours (Binning)" in which we replace our probabilistic mapping technique with spatial binning of the colored pointcloud generated by the mapping module.

Experiments were run on a machine equipped with an Intel i7-8650 CPU and 16 GB of RAM. The MSVIO, local mapping, and registration processes are all modular and run on separate threads. We first outline our system's performance with respect to the GPS groundtruth on the first sequence of approximately 650 meters in length, and then compare our system's performance to the alternative methods on the second sequence of approximately 2.3 kilometers.

The results of the first experiment can be seen in Fig. 1.

[1] Nearmap: `nearmap.com`

TABLE I: Quantitative localization metrics for all methods, in meters.

	MSVIO	ORB	CVM	**Ours**	Ours (Binning)
Max Error	45.27	20.13	59.09	**8.28**	75.13
RMSE	18.75	9.53	23.33	**2.94**	34.85

Our results are expressed in terms of the absolute trajectory error (ATE). The maximum error for MSVIO and our position estimation was 16.28 meters and 4.87 meters, while the RMSE was 7.72 meters and 2.11 meters, respectively. The final drift of our approach was 3.73 meters, or 0.57% of the total trajectory length, showing that while MSVIO alone can experience significant amounts of drift, our method recovers from drift and converges toward the ground truth. We also observe that our outlier rejection is very effective. All registrations that deviate significantly from the ground truth are correctly rejected, while a majority (8 out of 12) are correctly identified as inliers.

In our second experiment we compare our system to the state-of-the-art methods outlined above. The qualitative and quantitative comparisons can be found in Fig. 7, Table. I respectively. Overall, we see that our method outperformed the state of the art and maintained the most accurate global estimate across the 2.3 kilometers long sequence and that, similarly to the first sequence, it was able to correct the drift that arises from using only MSVIO for estimation. Notably, we observe that our method significantly out preforms the non-probabilistic variant indicating that our probabilistic mapping technique has a significant positive impact on performance. In addition, only our method was able to function in real time. The ORB method's runtime was $8\times$ slower than ours while CVM's was $100\times$ slower.

Both comparison methods produced significantly less accurate estimates than our approach. We hypothesize that the cause of this decreased performance is derived from the fact that the descriptor comparison measurement model has high variance. This can cause the most probable particle to jump around the true vehicle location at every sensor measurement, and in extreme cases cause the entire distribution to diverge from the ground truth trajectory.

It is also necessary to note that both comparison methods had, unfortunately, non-optimal experimental conditions. The ORB descriptor was designed for dense patches, but the ORB method computed its query descriptor on our sparse local map images as it was the only top down image we could provide. Additionally, the CVM-Net used in this experiment was trained using data on roadways. Therefore, it is possible that the model was not able to generalize for the natural environment of our experiments. These conditions, however, are likely representative of those experienced in real-world operation where a dense top down image may be impossible to acquire due to occlusions, and data may not exist for the deployment environment to pre-train a neural model. Our method is able to generalize to never-before-seen environments and perform well even with significant occlusion from environmental features like vegetation.

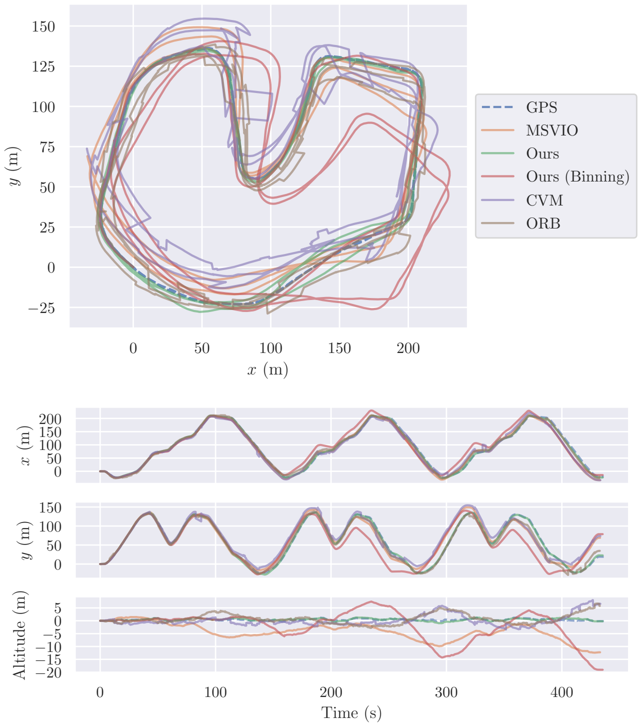

Fig. 7: Trajectories for all methods are shown in the top image while the position on the 3 axes with respect to time is shown in the bottom 3 figures.

V. Conclusion and Future work

In this paper, we outlined the design of a global localization pipeline for GPS-denied scenarios. A multi-stereo VIO module was extended to provide robust odometry for challenging environments. A probabilistic 3D occupancy grid was created to generate accurate synthetic top down images without significant artifacts and thus address the issue of drastically differing perspectives between vehicle and aerial imagery. A registration module was designed to align these images with reference imagery to measure global location. Finally, a pose graph was formulated to fuse odometry and global measurements and provide a continuous global state estimate for robot operation after loss of GPS signal. We show that our system can localize in real time and outperforms existing state-of-the-art methods on real world datasets.

In its current form we have also found that our method is sensitive to visual differences between the local map and reference imagery. Such differences can be induced due to photometric qualities of the captured ground images (e.g. exposure, white balance) or by temporal changes (e.g. reference imagery was captured during a different season). Such visual differences can cause decreased performance of our image registration method and in-turn degraded localization accuracy. In future work, we plan to explore registration techniques that generalize to a wider variety of visual conditions as well as methods to normalize the sensed and reference imagery to mitigate visual differences. Both directions focus on robustifying the method to a variety of different scenarios.

REFERENCES

[1] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5297–5307, June 2016.

[2] M. Bianchi and T. D. Barfoot. UAV localization using autoencoded satellite images. *IEEE Robotics and Automation Letters*, 6(2):1761–1768, 2021.

[3] X. Bouyssounouse, A. V. Nefian, A. Thomas, L. Edwards, M. Deans, and T. Fong. Horizon based orientation estimation for planetary surface navigation. In *IEEE International Conference on Image Processing (ICIP)*, pages 4368–4372, 2016.

[4] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays. Argoverse: 3D tracking and forecasting with rich maps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8740–8749, June 2019.

[5] H.-P. Chiu, A. Das, P. Miller, S. Samarasekera, and R. Kumar. Precise vision-aided aerial navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 688–695, 2014.

[6] J. Choi and H. Myung. BRM localization: UAV localization in GNSS-denied environments based on matching of numerical map and UAV images. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4537–4544, 2020.

[7] G. Cioffi and D. Scaramuzza. Tightly-coupled fusion of global positional measurements in optimization-based visual-inertial odometry. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5089–5095, 2020.

[8] O. Dabeer, W. Ding, R. Gowaiker, S. K. Grzechnik, M. J. Lakshman, S. Lee, G. Reitmayr, A. Sharma, K. Somasundaram, R. T. Sukhavasi, and X. Wu. An end-to-end system for crowdsourced 3D maps for autonomous vehicles: The mapping component. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 634–641, 2017.

[9] G. J. V. Dalen, D. P. Magree, and E. N. Johnson. Absolute localization using image alignment and particle filtering. In *AIAA Guidance, Navigation, and Control Conference*, 2016.

[10] F. Dellaert. Factor graphs and GTSAM: A hands-on introduction. Technical report, Georgia Institute of Technology, 2012.

[11] F. Dellaert and M. Kaess. *Factor Graphs for Robot Perception*. Now Publishers Inc., August 2017.

[12] E. Dexheimer, P. Peluse, J. Chen, J. Pritts, and M. Kaess. Information-theoretic online multi-camera extrinsic calibration. *IEEE Robotics and Automation Letters*, 2022.

[13] H. Goforth and S. Lucey. GPS-denied UAV localization using pre-existing satellite imagery. In *International Conference on Robotics and Automation (ICRA)*, pages 2974–2980, 2019.

[14] G. Hemann, S. Singh, and M. Kaess. Long-range gps-denied aerial inertial navigation with lidar localization. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1659–1666, Oct 2016.

[15] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008.

[16] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Autonomous Robots*, 2013. Software available at `http://octomap.github.com`.

[17] S. Hu, M. Feng, R. M. H. Nguyen, and G. H. Lee. CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7258–7267, 2018.

[18] S. Hu and G. H. Lee. Image-based geo-localization using satellite imagery. *International Journal of Computer Vision*, 128:1205–1219, 2019.

[19] J. Jaekel, J. Mangelson, S. Scherer, and M. Kaess. A robust multi-stereo visual-inertial odometry pipeline. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4623–4630, Oct 2020.

[20] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert. iSAM2: Incremental smoothing and mapping with fluid relinearization and incremental variable reordering. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3281–3288, Shanghai, China, May 2011.

[21] D.-K. Kim and M. R. Walter. Satellite image-based localization via learned embeddings. *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2073–2080, 2017.

[22] R. Liu, J. Wang, and B. Zhang. High definition map for automated driving: Overview and analysis. *Journal of Navigation*, 73:324–341, 2020.

[23] A. Masselli, R. Hanten, and A. Zell. Localization of Unmanned Aerial Vehicles Using Terrain Classification from Aerial Images. In E. Menegatti, N. Michael, K. Berns, and H. Yamaguchi, editors, *Intelligent Autonomous Systems 13*, pages 831–842, Cham, 2016. Springer International Publishing.

[24] G. Mattyus, S. Wang, S. Fidler, and R. Urtasun. HD Maps: Fine-grained road segmentation by parsing ground and aerial images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3611–3619, June 2016.

[25] A. V. Nefian, X. Bouyssounouse, L. Edwards, T. Kim, E. Hand, J. Rhizor, M. Deans, G. Bebis, and T. Fong. Planetary rover localization within orbital maps. In *IEEE International Conference on Image Processing (ICIP)*, pages 1628–1632, 2014.

[26] T. H. Nguyen, T.-M. Nguyen, and L. Xie. Range-focused fusion of camera-imu-uwb for accurate and drift-reduced localization. *IEEE Robotics and Automation Letters*, 6(2):1678–1685, 2021.

[27] D. Pannen, M. Liebner, W. Hempel, and W. Burgard. How to keep HD maps for automated driving up to date. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2288–2294, 2020.

[28] B. Patel, T. D. Barfoot, and A. P. Schoellig. Visual localization with google earth images for robust global pose estimation of UAVs. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6491–6497, 2020.

[29] D. Paz, H. Zhang, Q. Li, H. Xiang, and H. I. Christensen. Probabilistic semantic mapping for urban autonomous driving applications. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2059–2064, Oct. 2020.

[30] T. Qin, S. Cao, J. Pan, and S. Shen. A general optimization-based framework for global pose estimation with multiple sensors, 2019. Preprint arXiv:1901.03642.

[31] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An efficient alternative to SIFT or SURF. In *International Conference on Computer Vision*, pages 2564–2571, 2011.

[32] M. Shan, F. Wang, F. Lin, Z. Gao, Y. Z. Tang, and B. M. Chen. Google map aided visual navigation for UAVs in GPS-denied environment. In *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 114–119, 2015.

[33] A. Shetty and G. X. Gao. UAV pose estimation using cross-view geolocalization with satellite imagery. In *International Conference on Robotics and Automation (ICRA)*, pages 1827–1833, 2019.

[34] U.S. Geological Survey. USGS NED ned19_n40x50_w080x00_pa_southwest_2006 1/9 arc-second 15x15 minute IMG, 2010. `www.sciencebase.gov/catalog/item/581d2b68e4b08da350d63d02`.

[35] A. Viswanathan, B. R. Pires, and D. Huber. Vision-based robot localization by ground to satellite matching in GPS-denied situations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 192–198, Sept 2014.

[36] N. Vo, N. Jacobs, and J. Hays. Revisiting IM2GPS in the deep learning era. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2640–2649, Oct 2017.

[37] A. Yol, B. Delabarre, A. Dame, J. E. Dartois, and E. Marchand. Vision-based absolute localization for unmanned aerial vehicles. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3429–3434, 2014.