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

Yehonathan Litman

CMU-RI-TR-22-46

August 1, 2022

The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, 15213

Thesis Committee:
Prof. Michael Kaess, *chair*
Prof. Sebastian Scherer
Prof. Ji Zhang
Montiel Abello

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Robotics.

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 thesis, 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 the system generates visually accurate orthographic images of the environment, provides reliable global measurements, and maintains an accurate global state estimate in GPS-denied conditions.

Acknowledgments

I am incredibly thankful to my advisor, Prof. Michael Kaess, for his extensive support during my studies. His guidance was critical in helping me develop the skills I need as a researcher. His belief and investment in me meant everything to the development of my knowledge and skills as a roboticist. I literally could not have asked for a better advisor.

I'd also like to extend my appreciation to my lab mates at the Robot Perception Laboratory (RPL), namely Ruoyang, Samiran, Akash, Wei, Akshay, Dan, Tianxiang, Monty, and Suddhu, as well as the rest of the CMU RI community. Everyone always had my back and were willing to help me in my research and answering questions I had whenever they could. They made my experience at CMU mean what it did.

Lastly, I'd like to thank my family for their limitless support and love.

Funding

This work was 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.

Contents

1 Introduction **1**
1.1 Motivation 1
1.2 Contributions 1
1.3 Overview 3

2 Related Work **5**
2.1 GPS-Denied Localization 5
2.2 Georeferenced Information 5
2.2.1 HD Maps 5
2.2.2 Digital Elevation Models 6
2.2.3 Satelite Imagery 7

3 Approach **11**
3.1 Multi-Stereo Visual-Inertial Odometry 11
3.2 Local Map Construction 12
3.3 Registration 13
3.4 Global Registration Pose Graph 16

4 Experiments **19**

5 Conclusion and Future Work **25**
5.1 Conclusion 25
5.2 Limitations 25
5.3 Future Work 26

Bibliography **27**

List of Figures

1.1 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. 2

2.1 A comparison of state of the art work on visual matching of sensed data to georeferenced imagery. The images are from [35] [37], and [38] from left to right, respectively. All three methods were designed to operate in different spaces, from natural to urban environments, without exception, with varying levels of success. 8

3.1 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, voxels with occupancy probability below 0.5 are ignored. 12

3.2 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. . . . 15

3.3 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. 16
3.4 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. 17

4.1 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. 20
4.2 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. 23

List of Tables

4.1 Quantitative localization metrics for all methods, in meters. 21

Chapter 1

Introduction

1.1 Motivation

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.

1.2 Contributions

In this thesis, I introduce a system for real-time global position estimation in remote, natural environments using preexisting aerial or satellite imagery. The 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

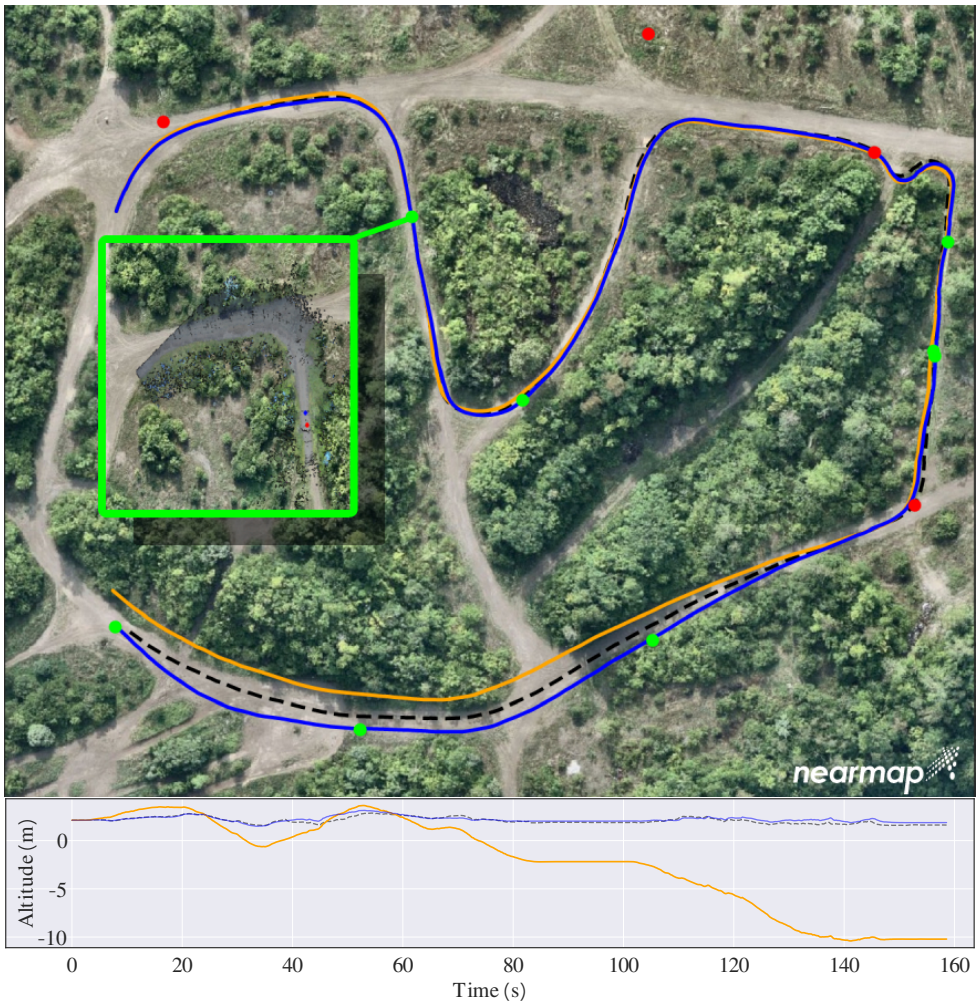

Figure 1.1: An example of the localization process where synthetic orthographic images generated by the UGV are matched to corresponding locations in the reference aerial imagery, with an example alignment shown in the green box. Successful registrations, or green dots, are used as global measurements 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.

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[1]. Finally, the fourth module combines MSVIO

[1]This module, described in section 3.3, describes joint work with Daniel McGann.

measurements with registration results in a global pose graph to provide a continuous and consistent global state estimate. The complete system can operate in real-time and its localization performance is shown in Fig. 1.1.

In sum, this thesis presents the following four main contributions:

1. An MSVIO formulation for efficient and fault-tolerant local state estimation.
2. An image generation method that provides visually accurate orthographic views of the UGV's local environment and overcomes challenges to existing work in visual registration and matching.
3. A full localization pipeline to provide real-time global position estimates.
4. A comparison of the method to the state of the art for GPS-denied visual localization on real world datasets.

1.3 Overview

In Chapter 2, we begin with a discussion of previous related work in GPS-denied localization that utilized different ways to match sensed data to previously acquired data. Then, Chapter 3 introduces the localization pipeline and its individual module components in depth. Chapter 4 compares the pipeline with other state of the art GPS-denied localization methodologies and then qualitatively discusses the benefits of the proposed pipeline. Lastly, Chapter 5 concludes the thesis and discusses some of the limitations of the methodology as well as some proposed directions for future work.

Chapter 2

Related Work

2.1 GPS-Denied Localization

The historically 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 the global frame even with known initial state. SLAM solutions can account for drift via loop closures but such measurements require re-visitations which, in general operation, cannot be assumed.

Yet, because it is not guaranteed that scenes will be seen repeatedly in a sequence, another solution has been to pair the system with a priori georeferenced global information in order to give a global localization estimate in GPS-denied settings. Such comparisons can then be used to correct for drift relative to the global frame in an existing local (dead reckoning or SLAM) solution.

2.2 Georeferenced Information

2.2.1 HD Maps

One source of georeferenced information are high definition (HD) maps which have been popularized by autonomous vehicles in recent years. HD maps have been formulated in many ways. All formulations include at a minimum the geometric

and semantic structure of roadways and often also include the geometric and visual structure of the environment represented by visual features, point clouds, or even dense geometric reconstructions [4, 22]. These maps are created through the aggregation of data from many sources, including satellite imagery, public street maps, and importantly data collection using ground vehicles in the environment [8, 24, 29]. These maps provide rich and highly accurate reconstructions of the environment, and in turn enable high accuracy global localization. However, the creation and maintenance of HD maps is prohibitively expensive even in urban centers [27]. Constructing and maintaining these maps for much larger and less frequently transited remote environments would be impractical given current state-of-the-art methods.

To operate in natural, remote environments we need alternative source of georeferenced information to the HD map that can be collected and maintained at scale. There are two clear candidates for this role: 1. Satellite and aerial imagery, considered as the reference imagery, and 2. Digital elevation models (DEMs). Both provide sources of georeferenced information for global localization and exist for the entire landmass of the Earth. While the appearance of the Earth changes with high frequency aerial imagery is collected across the planet at a daily rate. DEM data collection is less frequent as its content changes at a much slower geologic rate. The data availability and ease of collection for both DEMs and satelite imagery make them ideal for fast robot deployment and localization.

2.2.2 Digital Elevation Models

One method to perform localization using a DEM is to perform horizon matching [3, 25]. The horizon's profile is extracted from UGV images and is matched to a DEM. However, all horizon matching methods assume a clear view of the true horizon. While this assumption holds for extraterrestrial environments and some environments on Earth, it is violated when operating in and around vegetation or man made structures that partially or fully obstruct the horizon. Another method to localize using a DEM is to construct a local DEM that can be matched against the reference DEM [14]. However, matching to the reference DEM requires observation of unique features which exist in the DEM at the scale of hills, mountains, and valleys. The local map would therefore have to be large enough to contain such features. While observing

these features was possible for a large aerial platform, 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 DEM that would not represent the true structure of the terrain.

2.2.3 Satelite Imagery

Unlike DEMs, reference imagery contains unique features at a scale that are practical for a UGV to observe. However, UGV global localization to reference imagery is challenged by the significant view point difference, e.g., a UGV sees a much different scene than that from a satellite. One approach [35] addresses the view point challenge to approximate the viewpoint of an aerial image, where a 360° image is warped onto the ground plane (assumed flat) to create a synthetic top down image. However, it is noted in this work that where the flat ground assumption is violated (by vegetation, objects, buildings), significant artifacts appear in the resulting image which leads to decreased performance. To address this, localization is performed using a particle filter where the probability of the measurement given UGV location is computed from the distance between whole image SIFT descriptors computed on the warped ground image and a sample of the reference imagery taken at each particle's position.

Another approach to tackle the view-point challenge is to learn a deep model to embed matching ground and aerial images closely in feature space [17, 21]. These methods are notably inspired from geolocalization work [1, 36] with the added complexity of handling "cross view" image pairs. In geolocalization applications, one attempts to localize a single image, rather than continuously localize a robot. Thus, like [35], the image descriptors are used in a particle filter to localize the robot. A shortcoming of both of these methods is that the use of embedded descriptors results in noisy individual measurements. It is only through the aggregate of many measurements in the form of a particle filter that a reasonable solution is found. As such these methods benefit from a particle filter's strengths (i.e. no need for initial state estimate), but also suffer from a particle filter's downsides (i.e. non-determinism, and computational cost for tracking high dimensional state). While these approaches were able to avoid issues visual differences better than in classical approaches due to a higher space representation, they required extensive computational resources that

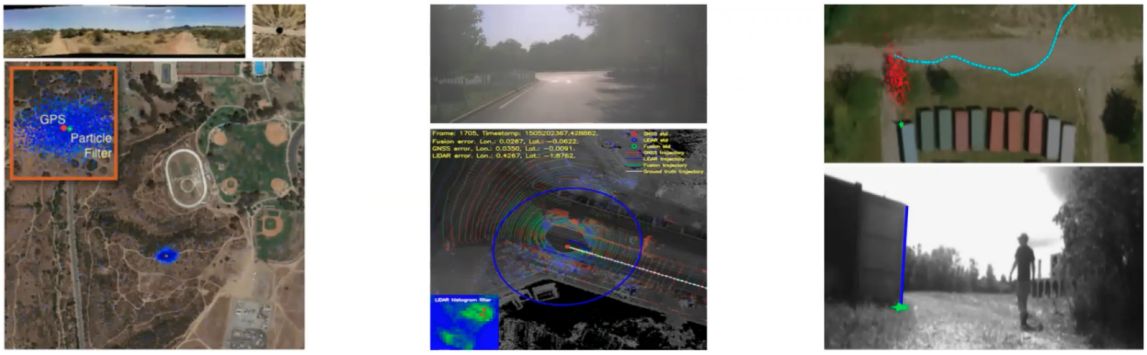

Figure 2.1: A comparison of state of the art work on visual matching of sensed data to georeferenced imagery. The images are from [35] [37], and [38] from left to right, respectively. All three methods were designed to operate in different spaces, from natural to urban environments, without exception, with varying levels of success.

are difficult to place in robots, and, because the models are trained with the same maps used for testing, also suffer from a lack of operational scope, in addition to the challenge of collecting data for learning.

On the other hand, other approaches that fused visual and inertial sensors with global estimates have shown that tracking could be done with very high accuracy and at a high rate [7, 26, 30] by reducing the drift in the global frame of reference. This offers an avenue for solving the issue of poor localization estimates from the particle filter by using additional sensing that is superior and better modelled to counteract the higher uncertainty in registration estimates.

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, 39]. 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 when aerial imagery is used as a georeferenced source, yet the issue of viewpoint differences

remains, especially in the context of conducting accurate localization. Different approaches have been proposed to address this issue, as shown in Fig. 2.1, all with certain benefits and drawbacks, but they all heavily rely on structural cues in the specific environment in which robots operate. This motivates our work to construct a UGV localization pipeline that allows for the construction and registration of synthetic top down images that are visually accurate to existing reference imagery, such that the system does not overly depend on persistent structural cues in its environment. Furthermore, by fusing the registration results with visual inertial sensing and jointly optimizing, we can conduct high accuracy state estimation in real time that is globally consistent.

Chapter 3

Approach

In this section I present a global localization pipeline formulation that consists of four modules which 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.

3.1 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 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.

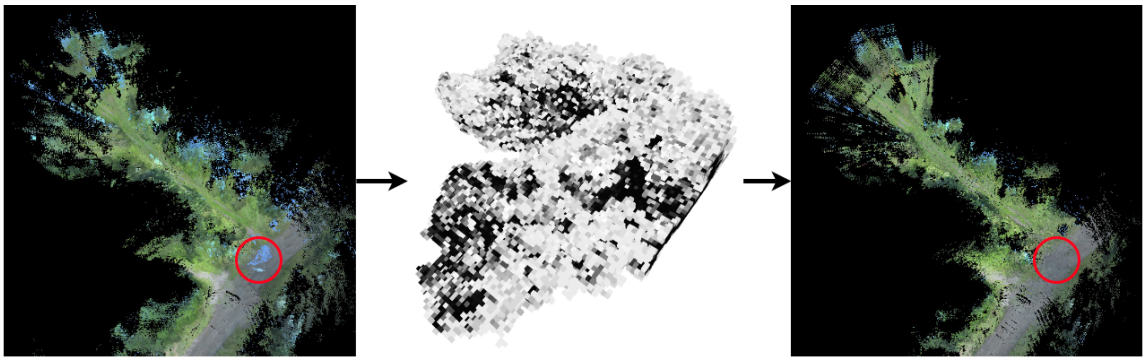

Figure 3.1: A local map shown on the left can exhibit artifacts like blue sky pixels due to noise in disparity. By explicitly modeling the occupancy probability in a 3D occupancy grid, we can filter out voxels of low occupancy probability by rendering a local map with only voxels of high occupancy probability, shown in dark in the center image. This removes the most significant artifacts from the final local map shown on the right. For visualization, voxels with occupancy probability below 0.5 are ignored.

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.

3.2 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. 3.1, 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:

$$l(v \mid \mathbf{z}_{t_1:t_2}, \mathbf{x}_{t_1:t_2}) = \log\left(\frac{p(v \mid \mathbf{z}_{t_1:t_2}, \mathbf{x}_{t_1:t_2})}{1 - p(v \mid \mathbf{z}_{t_1:t_2}, \mathbf{x}_{t_1:t_2})}\right) \tag{3.1}$$

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. 3.1.

3.3 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.

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. 3.2. 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

$$\mathbf{C}_{\Delta\mathbf{r}} = \frac{\sum_{i,j}(\mathbf{T}_{\Delta\theta_{i,j}} \cdot \mathbf{R}_{\Delta x+i,\Delta y+j} \cdot \mathbf{M}_{i,j})}{\sqrt{\sum_{i,j}\left(\mathbf{T}_{\Delta\theta_{i,j}} \cdot \mathbf{M}_{i,j}\right)^2 \cdot \sum_{i,j}\left(\mathbf{R}_{\Delta x+i,\Delta y+j} \cdot \mathbf{M}_{i,j}\right)^2}} \tag{3.2}$$

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 non-linear 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.

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

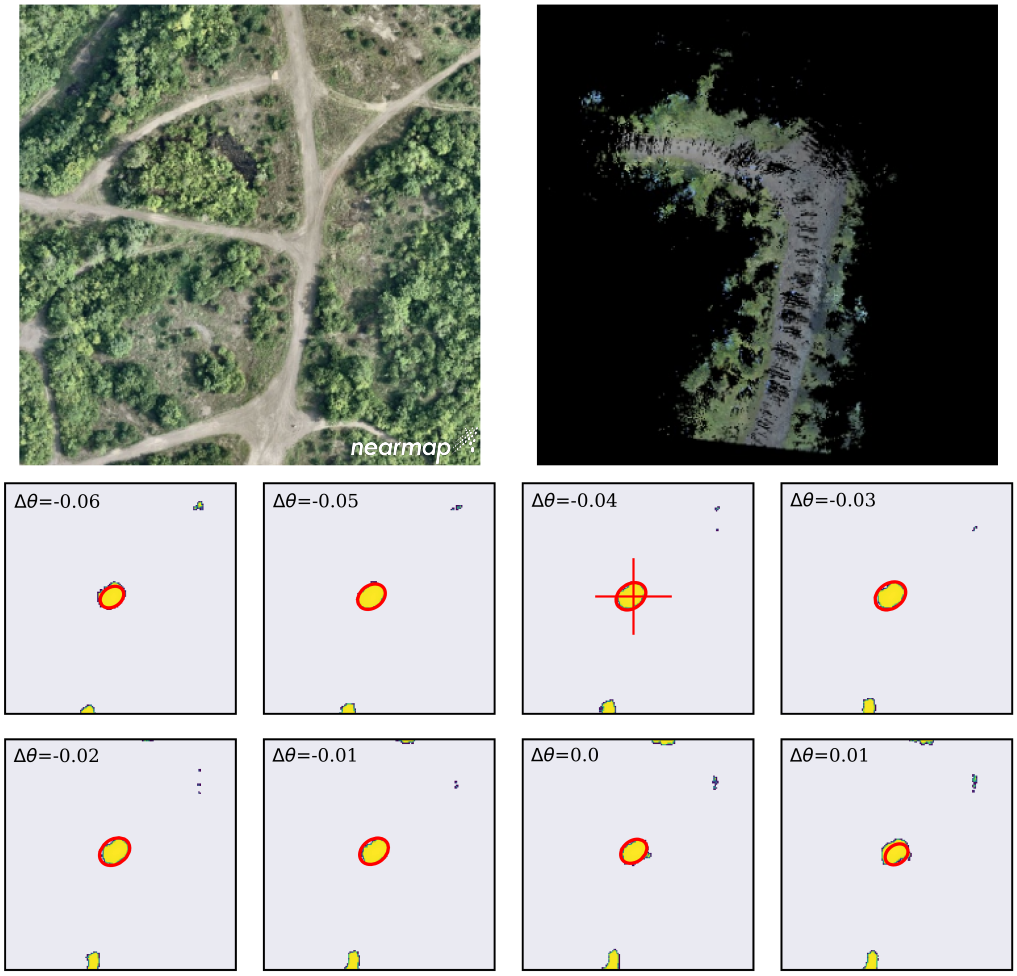

Figure 3.2: The search region in the top left is extracted from reference imagery around the current global position estimate. The local map in the top right is matched against this region with 3D scan matching. The 3D cost volume $\mathbf{C}^{th}$ after thresholding is shown on the bottom for a subset of search angles. Overlaid on the cost volume is the optimum's location and covariance denoted by the red "+" and ellipses, respectively.

strictly positive and normalized according to

$$p(\Delta\mathbf{r}_i) = \frac{\mathbf{C}^{th}_{\Delta\mathbf{r}_i}}{\sum_j \mathbf{C}^{th}_{\Delta\mathbf{r}_j}} \tag{3.3}$$

and the covariance is calculated with the mean μ as

$$\Sigma_{\Delta\mathbf{r}} = \sum_i p(\Delta\mathbf{r}_i)(\Delta\mathbf{r}_i - \mu)(\Delta\mathbf{r}_i - \mu)^\top \tag{3.4}$$

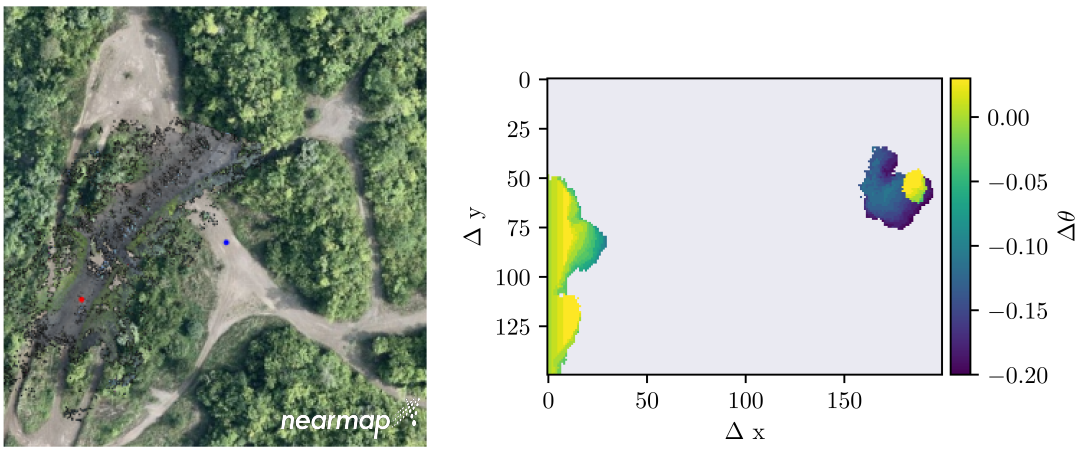

Figure 3.3: An example of an outlier registration. On the right is a top down view of the non-zero entries of the cost volume after thresholding. We can see three distinct modes in this volume indicating a poor or ambiguous measurement. On the left is the alignment according to the cost volume optimum. The correct registration would align the red and blue dots at the proper angle.

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. 3.3.

3.4 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 our estimation as a maximum a posteriori (MAP) problem where we

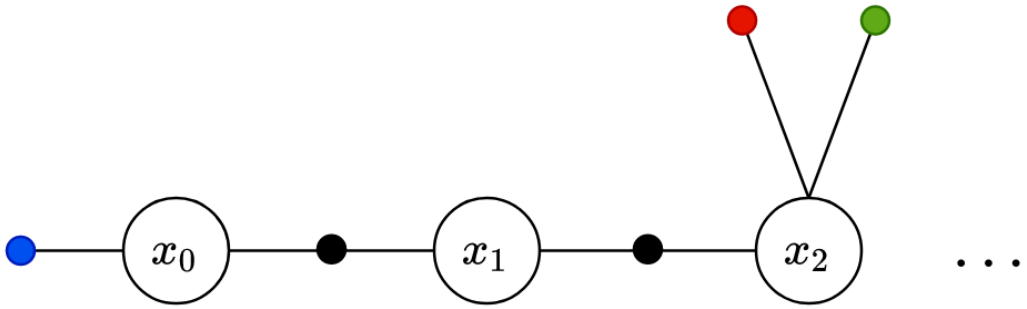

Figure 3.4: An illustration of the global pose graph. The white nodes represent the 6 DOF pose of the vehicle in the global frame, the black factors represent the MSVIO relative constraints, the red and green factors represent the registration and elevation constraints, respectively, and the blue factor is a prior on the initial state.

estimate the poses of all frames up to a time t

$$\mathcal{X}_t = \{\mathbf{x}_0, \ldots, \mathbf{x}_t\} \tag{3.5}$$

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. 3.4. 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

$$\begin{aligned}\mathcal{X}_t^\star = \underset{\mathcal{X}_t}{\operatorname{argmin}} & \underbrace{||\mathbf{x}_0||^2_{\Sigma_0}}_{\text{Prior}} + \sum_{i=1}^{t} \Big(\underbrace{||P(\mathbf{x}_{i-1}, \mathbf{x}_i)||^2_{\Sigma_P}}_{\text{MSVIO Factor}} \Big) \\ & + \sum_{i=1}^{t/N} \Big(\underbrace{||H(\mathbf{x}_{Ni}, h_{Ni})||^2_{\Sigma_H}}_{\text{Elevation Factor}} + \underbrace{||R(\mathbf{x}_{Ni}, \mathbf{r}_{Ni})||^2_{\Sigma_R}}_{\text{Registration Factor}} \Big)\end{aligned} \tag{3.6}$$

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.

Chapter 4

Experiments

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. 4.1. 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. I used a DEM from the National Elevation Dataset [34]. I generated and registered the local map against the reference imagery every 50 frames.

For comparison we implement two alternative methods based on [35] and [17], referred to as "ORB" and "CVM" respectively, owing to the basis of their implementation. For more details, the reader should refer to Sec. 2. The implementations of both methods use the same particle filter and a motion model derived for the data collection platform. We make two modifications to the 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 [17], we use the publicly available pre-trained

[1]Nearmap: `nearmap.com`

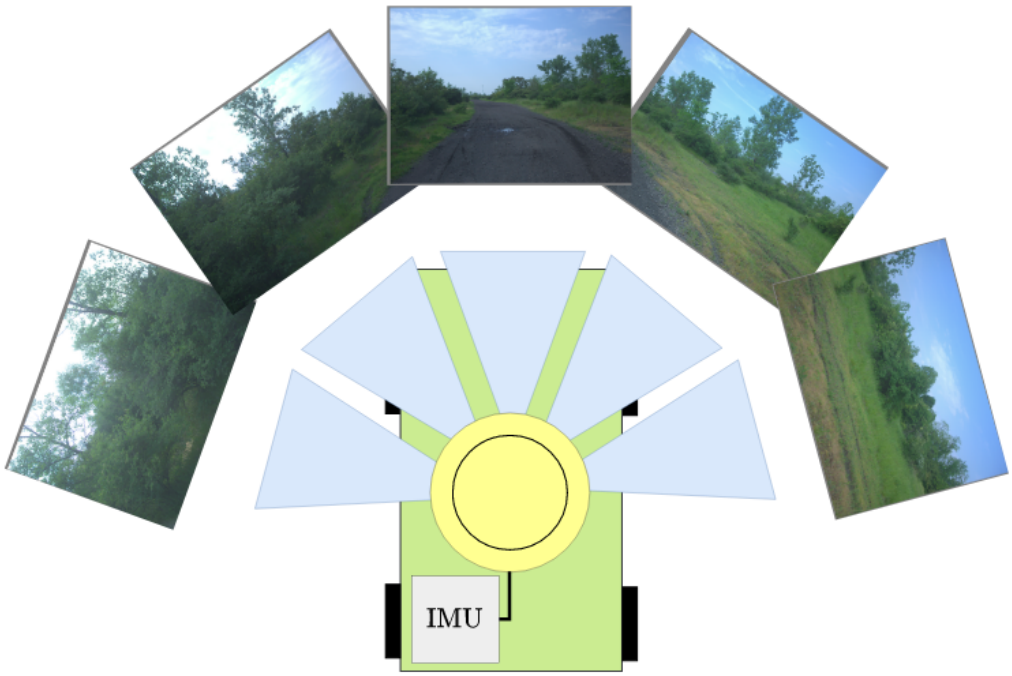

Figure 4.1: A diagram of the vehicle used for data collection, with an example of the stereo images (one from each stereo pair). The vehicle was equipped with five stereo pairs as well as an IMU module, all time synchronized with an FPGA.

weights for the CVM-Net-II model [18]. 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. 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

Table 4.1: 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

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. 4.2, Table. 4.1 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× slower than ours while CVM's was 100× 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.

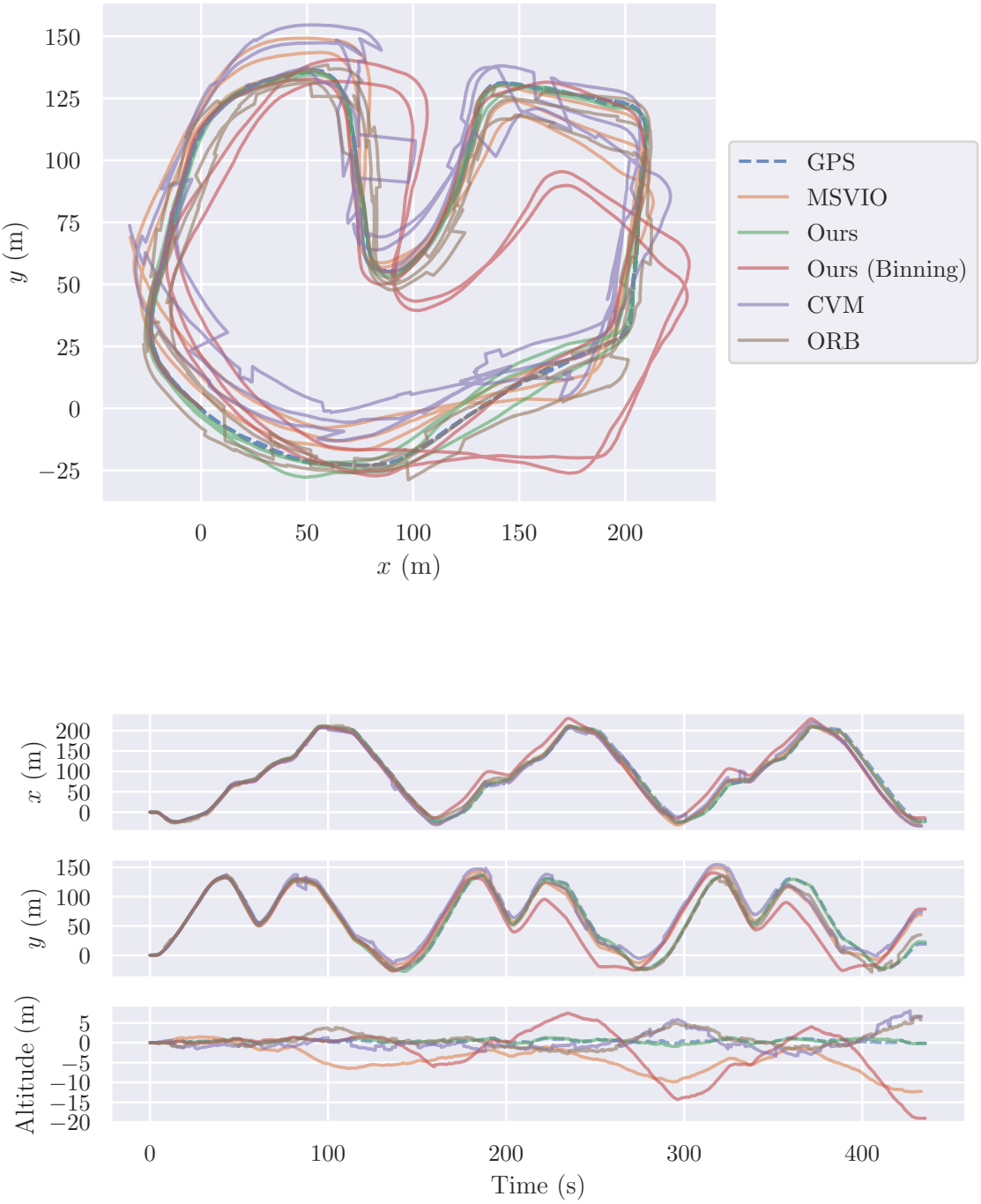

Figure 4.2: Trajectories for all methods are shown in the top image while the position on the 3 axes with respect to time is shown in the bottom 3 figures.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, I 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.

5.2 Limitations

At the moment, the pipeline's biggest limitation is the fact that it cannot be easily deployed due to difficulties with visual registration and parameter tuning. In many cases, the pipeline could not be deployed without previous parameter testing to derive the best registration performance, by tuning parameters such as occupancy threshold and cell size in the 3D grid, registration frequency, etc.

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. Because we use color information directly from the images, we could transform the color of the generated local maps as desired. Since no such transformation is applied in our formulation, our approach works best by registering the local map to georeferenced imagery of similar visual properties. Thus, if the system uses georeferenced imagery captured in a different season, it may lead to reduced performance during global registration. This is due to visual differences in colors and illuminations of what the UGV is seeing and what the georeferenced imagery contains.

5.3 Future Work

In future work, I 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. One approach to mitigate the differences between the imagery data could be through the use of a deep learning model to improve registration against visual differences. Another approach could be employing digital signal processing on the sensed imagery data or local map to make it invariant to visual qualities in captured ground images, all while keeping intact structural patterns present in the images. Additional work could be apply the pipeline to multi-robot mapping and registration for collaborative exploration and more efficient distributed computing. Lastly, future work should look at the possibility of porting the system to robots functioning beyond natural rural environments such as indoors, inside cities, forests, etc. It should also be explored how to take into account the fact that the robot could travel in between these environments.

Bibliography

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef 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. 2.2.3

[2] Mollie Bianchi and Timothy D. Barfoot. UAV localization using autoencoded satellite images. *IEEE Robotics and Automation Letters*, 6(2):1761–1768, 2021. doi: 10.1109/LRA.2021.3060397. 2.2.3

[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. doi: 10.1109/ICIP.2016.7533185. 2.2.2

[4] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James 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. 2.2.1

[5] Han-Pang Chiu, Aveek Das, Phillip Miller, Supun Samarasekera, and Rakesh Kumar. Precise vision-aided aerial navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 688–695, 2014. doi: 10.1109/IROS.2014.6942633. 2.2.3

[6] Junho Choi and Hyun 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. doi: 10.1109/IROS45743.2020.9341682. 2.2.3

[7] Giovanni Cioffi and Davide 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. doi: 10.1109/IROS45743.2020.9341697. 2.2.3, 2.2.3

[8] Onkar Dabeer, Wei Ding, Radhika Gowaiker, Slawomir K. Grzechnik, Mythreya J.

Lakshman, Sean Lee, Gerhard Reitmayr, Arunandan Sharma, Kiran Somasundaram, Ravi Teja Sukhavasi, and Xinzhou 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. doi: 10.1109/IROS.2017.8202218. 2.2.1

[9] Gerald J. Van Dalen, Daniel P. Magree, and Eric N. Johnson. Absolute localization using image alignment and particle filtering. In *AIAA Guidance, Navigation, and Control Conference*, 2016. doi: 10.2514/6.2016-0647. URL https://arc.aiaa.org/doi/abs/10.2514/6.2016-0647. 2.2.3

[10] Frank Dellaert. Factor graphs and GTSAM: A hands-on introduction. Technical report, Georgia Institute of Technology, 2012. 3.4

[11] Frank Dellaert and Michael Kaess. *Factor Graphs for Robot Perception*. Now Publishers Inc., August 2017. 3.4

[12] Eric Dexheimer, Patrick Peluse, Jianhui Chen, James Pritts, and Michael Kaess. Information-theoretic online multi-camera extrinsic calibration. *IEEE Robotics and Automation Letters*, 2022. doi: 10.1109/LRA.2022.3145061. 4

[13] Hunter Goforth and Simon Lucey. GPS-denied UAV localization using pre-existing satellite imagery. In *International Conference on Robotics and Automation (ICRA)*, pages 2974–2980, 2019. doi: 10.1109/ICRA.2019.8793558. 2.2.3

[14] Garrett Hemann, Sanjiv Singh, and Michael 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. 2.2.2

[15] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008. doi: 10.1109/TPAMI.2007.1166. 3.1

[16] Armin Hornung, Kai M. Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Autonomous Robots*, 2013. doi: 10.1007/s10514-012-9321-0. URL http://octomap.github.com. Software available at http://octomap.github.com. 3.2, 3.2

[17] Sixing Hu and Gim Hee Lee. Image-based geo-localization using satellite imagery. *International Journal of Computer Vision*, 128:1205–1219, 2019. 2.2.3, 4

[18] Sixing Hu, Mengdan Feng, Rang M. H. Nguyen, and Gim Hee 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. doi: 10.1109/CVPR.2018.00758. 4

[19] J. Jaekel, J.G. 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. doi: 10.1109/IROS45743.2020.9341604. 3.1

[20] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J.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. 2.2.3

[21] Dong-Ki Kim and Matthew R. Walter. Satellite image-based localization via learned embeddings. *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2073–2080, 2017. 2.2.3

[22] Rong Liu, Jinling Wang, and Bingqi Zhang. High definition map for automated driving: Overview and analysis. *Journal of Navigation*, 73:324–341, 2020. doi: 10.1017/S0373463319000638. 2.2.1

[23] Andreas Masselli, Richard Hanten, and Andreas Zell. Localization of Unmanned Aerial Vehicles Using Terrain Classification from Aerial Images. In Emanuele Menegatti, Nathan Michael, Karsten Berns, and Hiroaki Yamaguchi, editors, *Intelligent Autonomous Systems 13*, pages 831–842, Cham, 2016. Springer International Publishing. ISBN 978-3-319-08338-4. 2.2.3

[24] Gellert Mattyus, Shenlong Wang, Sanja Fidler, and Raquel 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. 2.2.1

[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. doi: 10.1109/ICIP.2014.7025326. 2.2.2

[26] Thien Hoang Nguyen, Thien-Minh Nguyen, and Lihua Xie. Range-focused fusion of camera-imu-uwb for accurate and drift-reduced localization. *IEEE Robotics and Automation Letters*, 6(2):1678–1685, 2021. doi: 10.1109/LRA.2021.3057838. 2.2.3, 2.2.3

[27] David Pannen, Martin Liebner, Wolfgang Hempel, and Wolfram 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. doi: 10.1109/ICRA40945.2020.9197419. 2.2.1

[28] Bhavit Patel, Timothy D. Barfoot, and Angela 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. doi: 10.1109/ICRA40945.2020.9196606. 2.2.3

[29] David Paz, Hengyuan Zhang, Qinru Li, Hao Xiang, and Henrik I Christensen. Probabilistic semantic mapping for urban autonomous driving applications. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2059–2064, October 2020. 2.2.1

[30] Tong Qin, Shaozu Cao, Jie Pan, and Shaojie Shen. A general optimization-based framework for global pose estimation with multiple sensors, 2019. Preprint arXiv:1901.03642. 2.2.3, 2.2.3, 3.4

[31] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *International Conference on Computer Vision*, pages 2564–2571, 2011. doi: 10.1109/ICCV.2011.6126544. 4

[32] Mo Shan, Fei Wang, Feng Lin, Zhi Gao, Ya Z. Tang, and Ben 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. doi: 10.1109/ROBIO.2015.7418753. 2.2.3

[33] Akshay Shetty and Grace Xingxin Gao. UAV pose estimation using cross-view geolocalization with satellite imagery. In *International Conference on Robotics and Automation (ICRA)*, pages 1827–1833, 2019. doi: 10.1109/ICRA.2019.8794228. 2.2.3

[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. 4

[35] Anirudh Viswanathan, Bernardo R. Pires, and Daniel 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. doi: 10.1109/IROS.2014.6942560. (document), 2.2.3, 2.1, 4

[36] Nam Vo, Nathan Jacobs, and James Hays. Revisiting IM2GPS in the deep learning era. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2640–2649, Oct 2017. 2.2.3

[37] Guowei Wan, Xiaolong Yang, Renlan Cai, Hao Li, Yao Zhou, Hao Wang, and Shiyu Song. Robust and precise vehicle localization based on multi-sensor fusion in diverse city scenes. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4670–4677, 2018. doi: 10.1109/ICRA.2018.8461224. (document), 2.1

[38] Xipeng Wang, Steve Vozar, and Edwin Olson. FLAG: Feature-based localization between air and ground. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3178–3184, 2017. doi: 10.1109/ICRA.2017.7989360.

(document), 2.1

[39] Aur`élien Yol, Bertrand Delabarre, Amaury Dame, Jean Èmile Dartois, and Eric Marchand. Vision-based absolute localization for unmanned aerial vehicles. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3429–3434, 2014. doi: 10.1109/IROS.2014.6943040. 2.2.3