Visual Localization with Google Earth Images for Robust Global Pose Estimation of UAVs

Bhavit Patel, Timothy D. Barfoot, and Angela P. Schoellig

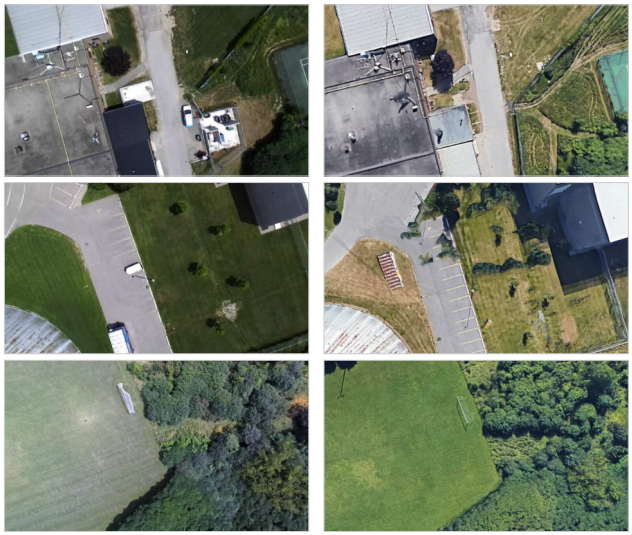
***Abstract*— We estimate the global pose of a multirotor UAV by visually localizing images captured during a flight with Google Earth images pre-rendered from known poses. We metrically localize real images with georeferenced rendered images using a dense mutual information technique to allow accurate global pose estimation in outdoor GPS-denied environments. We show the ability to consistently localize throughout a sunny summer day despite major lighting changes while demonstrating that a typical feature-based localizer struggles under the same conditions. Successful image registrations are used as measurements in a filtering framework to apply corrections to the pose estimated by a gimballed visual odometry pipeline. We achieve less than 1 m and 1° RMSE on a 303 m flight and less than 3 m and 3° RMSE on six 1132 m flights as low as 36 m above ground level conducted at different times of the day from sunrise to sunset.**

I. INTRODUCTION

Vision-based techniques involving Visual Odometry (VO) are the most popular for Unmanned Aerial Vehicle (UAV) navigation in GPS-denied environments. However, pure odometry techniques are unreliable for accurate pose estimates since they drift over time in the absence of corrections. Visual Simultaneous Localization and Mapping (SLAM) corrects these drifts through loop closure and has been sucessfully demonstrated in GPS-denied environments [1]–[3] but requires revisiting locations. On the other hand, Visual Teach and Repeat (VT&R) [4] can enable safe navigation in the absence of GPS without requiring globally accurate poses but is limited to navigation along previously traversed routes. Such a technique is suitable to perform emergency return of UAVs in the event of GPS loss [5].

The aforementioned techniques require the vehicle itself to map an area either through a human-operated manual teach phase in the case of VT&R or a carefully developed safe exploration algorithm for autonomous SLAM. However, a 3D reconstruction of many parts of the world is already available in Google Earth (GE). The ability to use this 3D reconstruction as a map would enable global pose estimation without GPS, having to worry about safe exploration, or restricting navigation to a previously traversed route. One of the main challenges to using this map is the large appearance difference between the 3D reconstruction and the true world: lighting and seasonal changes, as well as recent structural changes to the environment all present difficulties for visual localization.

In this work, we present a technique to determine the full six Degree of Freedom (DoF) global pose of a UAV in an area where the UAV itself has not mapped by using only a gimballed stereo camera, Inertial Measurement Unit (IMU), and georeferenced images from GE. These geoferenced images are rendered before the flight and stored onboard the UAV to enable navigation within the region covered by the images. The only limitations to the map size are the extents of the reconstruction coverage area and the available onboard storage.



Real from UAV Camera Rendered from Google Earth

Fig. 1: Comparison of real-world UAV images and rendered Google Earth images taken from the approximately same viewpoint at three locations along one of the flights. Large appearance changes, especially with vegetation, impermanent objects such as cars, poor 3D reconstructions (e.g., trees in middle pair), and structural changes to buildings (top pair) can all cause difficulties for visual localization.

Visual localization of the real images with the rendered images using traditional sparse features (e.g., Speeded-Up Robust Features (SURF)) is challenging due to the large appearance difference mentioned previously (see Fig. 1). Therefore, we perform image registration using a dense technique that relies on Mutual Information (MI). MI provides robustness to appearance changes allowing us to accurately register the images. We optimize the MI over warping parameters to align the real and rendered images. The result from this image registration is then fused with a pose estimated by a gimballed VO pipeline. The performance of this technique is evaluated on multiple datasets collected at the University of Toronto Institute for Aerospace Studies (UTIAS).

The contribution of this work is a method to accurately estimate the global pose of a UAV in GPS-denied environments using pre-rendered images from a 3D reconstruction

All authors are with the Institute for Aerospace Studies, Univesity of Toronto, Canada: bhavit.patel@robotics.utias.utoronto.ca, {tim.barfoot, angela.schoellig}@utoronto.ca. Accompanying video: tiny.cc/GElocalization

of the Earth. Our method allows accurate estimation at lower altitude flights compared to similar previous work described below. We also demonstrate robust estimation over an entire day in the presence of signficant lighting changes incurred from sunrise to sunset on $6.8\,\text{km}$ of real-world data.

The rest of this paper is outlined as follows: Section II presents related work on vision-based navigation aided by georeferenced imagery and MI-based techniques. Section III details our estimation pipeline. The datasets collected for this work are described in Section IV. The results are discussed in Section V. Finally, Section VI concludes the work presented in this paper.

II. RELATED WORK

Some of the earliest work using georeferenced satellite images used edges for registration. However, a simple edge detector resulted in only two successful matches along a $1\,\text{km}$ trajectory [6]. Building outlines extracted using local features were more successful in estimating the 6DoF pose of an aerial vehicle [7]. Unfortunately, this technique cannot be employed for lower altitude flights where the outlines of multiple buildings are not visible in a single image.

Some recent work using local image features use street view images to estimate the pose of a ground robot [8] and a UAV [9]. In both cases, techniques similar to bag-of-words are first used for place recognition followed by image registration using SIFT keypoints in the matched images. However, even after finding the best matching georeferenced image, the feature matching can contain 80% outliers [9] due to the large image appearance and viewpoint differences which makes it difficult to accurately localize.

Unsurprisingly, Convolutional Neural Networks (CNNs) have seen increased usage in recent years as image descriptors due to their ability to learn generic features that can be applied to a variety of tasks such as image classification and object detection. Features from the middle layers of the CNNs have been shown to be robust against appearance changes while features from the final layers are robust to viewpoint changes and provide more semantic information about the structure of the scene [10]. Often pretrained CNNs are further trained for the task of place recognition allowing topological localization [11]–[13] followed by filtering with VO in a particle filter [12] or Kalman filter [13]. These whole-image descriptors only allow finding an image match and do not provide metric information about the relative pose between the query and map image. Often the pose of the matching map image is taken as the best estimate which ultimately limits the accuracy of the localizations to the spatial resolution of the georeferenced images.

We are interested in accurate metric localization using georeferenced images. To accomplish this, we use a dense image registration technique to align images captured by a camera mounted on the UAV with pre-rendered georeferenced images. Instead of minimizing the photometric error, we use a metric computed using MI to add robustness to appearance changes.

We adopt the use of the Normalized Information Distance (NID) [14], [15] which is computed from MI (6). The NID is a value between 0 and 1 that is not as dependent on the amount of information content in the images (i.e., the amount of image overlap) as MI. It has been shown to be able to robustly register images [14], and localize a ground vehicle equipped with a monocular camera using a textured 3D map generated from a LIDAR and camera [15]. One of the reasons for the high accuracy in [15] is their ability to generate synthetic images online from the textured 3D map allowing direct optimization over the $SE(3)$ pose parameters. Since GE has no 3D view API, we pre-render images at a limited number of poses and perform a warping online for interpolation.

Similar to our work is [16], which determines the global position and heading of a UAV by finding the optimal scale-rotation-translation (sRt) warping (4) that maximizes the MI of a query image taken by a nadir-pointed camera warped into a mosaic of satellite images. An sRt warping is a 4DoF image warping that performs a scaling (zoom), 1D rotation, and 2D translation. It assumes the scene is planar and parallel to the image plane. For a nadir-pointed camera this assumption becomes more valid at higher altitudes since the building heights become small relative to the distance to the camera. In contrast to [16], we conduct lower altitude flights (e.g., $36\,\text{m}$ Above Ground Level (AGL) compared to $150\,\text{m}$) where the scene is often non-planar. Despite this, we are able to use this warping due to our method of rendering images at mulitple nearby poses in the 3D reconstruction.

III. METHODOLOGY

We estimate the global 6DoF $SE(3)$ pose of a multirotor UAV using only a gimballed stereo camera, an IMU (for vehicle attitude only), and a set of geoferenced GE images. Let

$$\mathbf{T}_{W,k} = \begin{bmatrix} \mathbf{C}_{W,k} & \mathbf{r}_W^{k,W} \\ \mathbf{0}^\top & 1 \end{bmatrix} \tag{1}$$

be the transformation from the vehicle at keyframe k to a world East-North-Up (ENU) frame. The position of the vehicle in the ENU frame is given by $\mathbf{r}_W^{k,W} = [x_W^{k,W} \; y_W^{k,W} \; z_W^{k,W}]^\top$ and the roll, pitch, and yaw ($\phi_{W,k}$, $\theta_{W,k}$, $\psi_{W,k}$, respectively) can be extracted from the 3×3 rotation matrix $\mathbf{C}_{W,k}$. Let $\mathcal{I}^q = (\mathbf{I}_1^q, \mathbf{I}_2^q, \ldots, \mathbf{I}_K^q)$ be the sequence of real UAV query images from each keyframe. We attempt to localize each keyframe image using a set of geoferenced map images, $\mathcal{I}^m = \{\mathbf{I}_1^m, \mathbf{I}_2^m, \ldots, \mathbf{I}_N^m\}$, where the global pose of map image n is denoted $\mathbf{T}_{W,n_s}$ with s indicating the sensor (camera) frame.

A. Gimballed Visual Odometry

The first step in the estimation pipeline is to perform VO on the UAV images. VO is performed using the VT&R 2.0 software system adapted for use on UAVs with gimballed cameras [5].

The inputs are rectified stereo greyscale images and a non-static vehicle-to-sensor transform, $\mathbf{T}_{f_s,f}$, computed at $10\,\text{Hz}$ for each frame. It is computed by compounding

transformations using the three gimbal angles and known translations between joints followed by a rotation into the standard camera frame. The roll and pitch axes of the gimbal are globally stabilized in a gravity-aligned inertial frame while the yaw follows the vehicle heading.

For each stereo image pair, features are extracted and SURF descriptors matched between the left and right frames to perform stereo landmark triangulation. Features that are unable to be triangulated from stereo matching are triangulated through motion between consecutive frames. The descriptors in the latest image are matched to the last keyframe to generate 2D-3D point correspondances. These are used in an Maximum Likelihood Estimation SAmple Consensus (MLESAC) estimator to determine the full $SE(3)$ incremental vehicle pose with uncertainty from the current frame to the last keyframe $\mathbf{T}_{f,k}, \mathbf{\Sigma}_{f,k}$. If the translation or rotation exceeds a threshold, or the number of inliers drops below a minimum amount, a new keyframe is added. With every new keyframe, a windowed refinement (bundle adjustment) is performed on the last 5 keyframes using the Simultaneous Trajectory Estimation And Mapping (STEAM) engine [17].

B. Image Registration

For every keyframe image, $\mathbf{I}_k^q$, the goal is to determine the relative $SE(3)$ pose between the query camera at k and a virtual GE camera that generated image n, $\mathbf{T}_{k_s,n_s}$. The global pose measurement of the vehicle is then obtained from

$$\mathbf{T}_{W,k} = \mathbf{T}_{W,n_s}\mathbf{T}_{k_s,n_s}^{-1}\mathbf{T}_{k_s,k}. \tag{2}$$

In this work, all real and rendered images are taken with the camera pointed in the nadir direction. The relative roll, ϕ_{k_s,n_s}, and pitch θ_{k_s,n_s}, are obtained from our gimbal, which keeps these angles at approximately 0 degrees. Therefore, we only need to estimate four pose parameters:

$$\boldsymbol{\eta} = [x_{k_s}^{n_s,k_s} \;\; y_{k_s}^{n_s,k_s} \;\; z_{k_s}^{n_s,k_s} \;\; \psi_{k_s,n_s}]^\top. \tag{3}$$

Since the image registration is estimating 4DoF, we use an sRt warping instead of a full homography:

$$\mathbf{x}' = w(\mathbf{x}, \boldsymbol{\mu}) = s\mathbf{R}(\psi) + \mathbf{t}, \tag{4}$$

where $\mathbf{x} = [x \;\; y]^\top$ is the query image plane coordinate warped to $\mathbf{x}' = [x' \;\; y']^\top$ for the map image, s is a scale, $\mathbf{R}(\psi)$ is a 1D rotation, $\mathbf{t} = [t_x \;\; t_y]^\top$ is a 2D translation, and $\boldsymbol{\mu} = [s \;\; \psi \;\; t_x \;\; t_y]^\top$. We can also directly warp pixel coordinates $\mathbf{u} = [u \;\; v]^\top$ from the query image into map image pixel coordinates $\mathbf{u}' = [u' \;\; v']^\top$:

$$\bar{\mathbf{u}}' = w(\bar{\mathbf{u}}, \boldsymbol{\mu}) = \mathbf{K}' \begin{bmatrix} s\mathbf{R}(\psi) & \mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix} \mathbf{K}\bar{\mathbf{u}}, \tag{5}$$

where $\bar{\mathbf{u}} = [u \;\; v \;\; 1]^\top$ and K is the camera intrinsics matrix.

The NID between a query image and warped map image is

$$NID(\mathbf{I}_k^q, \mathbf{I}_n^m, \boldsymbol{\mu}) = \frac{H(\mathbf{I}_k^q, \mathbf{I}_n^m, \boldsymbol{\mu}) - MI(\mathbf{I}_k^q; \mathbf{I}_n^m, \boldsymbol{\mu})}{H(\mathbf{I}_k^q, \mathbf{I}_n^m, \boldsymbol{\mu})}, \tag{6}$$

where the MI is

$$MI(\mathbf{I}_k^q; \mathbf{I}_n^m, \boldsymbol{\mu}) = H(\mathbf{I}_k^q) + H(\mathbf{I}_n^m, \boldsymbol{\mu}) - H(\mathbf{I}_k^q, \mathbf{I}_n^m, \boldsymbol{\mu}). \tag{7}$$

The joint entropy is given by

$$H(\mathbf{I}_k^q, \mathbf{I}_n^m, \boldsymbol{\mu}) = -\sum_{a=1}^{N}\sum_{b=1}^{N} p_{qm}(a, b, \boldsymbol{\mu}) \ln(p_{qm}(a, b, \boldsymbol{\mu})), \tag{8}$$

where $p_{qm}(a, b, \boldsymbol{\mu})$ is the joint probability distribution of image intensities in $\mathbf{I}_k^q$ and $\mathbf{I}_n^m$ for N bins with bin indices a and b. Similarly, the individual entropies are

$$H(\mathbf{I}_k^q) = -\sum_{a=1}^{N} p_q(a) \ln(p_q(a)) \tag{9}$$

$$H(\mathbf{I}_n^m, \boldsymbol{\mu}) = -\sum_{b=1}^{N} p_m(b, \boldsymbol{\mu}) \ln(p_m(b, \boldsymbol{\mu})), \tag{10}$$

where $p_q(a)$ and $p_m(b, \boldsymbol{\mu})$ are the marginal probability distributions (e.g., $p_m(b, \boldsymbol{\mu})$ gives the probability that pixel $\mathbf{u}'$ in image $\mathbf{I}_n^m$ has intensity that falls into bin b).

To register the images we determine the optimal warping parameters, $\boldsymbol{\mu}^* = [s^* \;\; \psi^* \;\; t_x^* \;\; t_y^*]^\top$, to minimize the NID between a query image and selected map image:

$$\boldsymbol{\mu}_k^* = \underset{\boldsymbol{\mu}}{\operatorname{argmin}}\, NID(\mathbf{I}_k^q, \mathbf{I}_n^m, \boldsymbol{\mu}). \tag{11}$$

Since this optimization problem is non-convex, we solve it using Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm; a quasi-Newton method. Previous work has modified the discrete marginal and joint probabilty distribution functions to be analytically differentiable by using B-spline weights [14]. We instead use a simpler approach of a central difference numerical gradient. Furthermore, we apply a two-step optimization procedure. The first step applies a Gaussian blur on both images to smooth out the cost function and gradients and optimizes with the blurred images. The second step uses the optimal warping from the blurred optimization to initialize a refined optimization that operates on the raw images.

The query-to-map pose parameters (3) are recovered from the optimal warping:

$$x_{k_s}^{n_s,k_s} = -t_x^* s^* z_{n_s}^{g,n_s} \tag{12a}$$

$$y_{k_s}^{n_s,k_s} = -t_y^* s^* z_{n_s}^{g,n_s} \tag{12b}$$

$$z_{k_s}^{n_s,k_s} = z_{n_s}^{g,n_s}(s^* - 1) \tag{12c}$$

$$\psi_{k_s,n_s} = -\psi^*, \tag{12d}$$

where $z_{n_s}^{g,n_s}$ is the distance from the nadir-pointed virtual camera to the ground. Therefore, for each successfully registered image we have an estimate for $\mathbf{T}_{k_s,n_s}$, which is used to obtain the global pose via (2).

It is important to select an appropriate map image $\mathbf{I}_n^m$ to register the query image $\mathbf{I}_k^q$. We compute the NID between the query image and unwarped map images in a radius around a predicted pose given by VO (14b). This strategy aims to provide the best aligned images before any warping. A simpler strategy is to select the spatially nearest map

image to the predicted pose but this relies on having accurate predictions and is less robust to drift. We start with a larger search radius (e.g., 10 m) until VO is scaled and then reduce (e.g., to 4 m) for subsequent registrations. If registration is unsuccessful for multiple keyframes in a row, then we once again inflate the search radius. An image registration is deemed unsuccessful if the position distance or relative yaw from the registered pose to the predicted pose (14b) is too large or the optimizer fails to converge.

C. Pose Filtering

We follow the methods in [18] to compound uncertain transforms and fuse uncertain pose estimates. The estimation is performed in a local coordinate frame where $\mathbf{T}_{W,0}$ is constructed using the RTK position and vehicle attitude at the first VO keyframe.

VO provides a relative transform between keyframes, $\mathbf{T}_{k,k-1}$, which serves as an input to our filtering framework. We found the VO uncertainties to be overconfident so we define our own. Since VO is unscaled, we use a large uncertainty, $\mathbf{Q}_k = \text{diag}(0.04, 0.04, 0.04, 0.05, 0.05, 0.01)$, until there are enough recent localizations to estimate a scale factor after which we reduce the positional uncertainty by a factor of 10. Each incremental transform is scaled with the result of a sliding-window scale estimator. The VO scale estimator determines a scale factor to minimize the uncertainty-weighted error between the incremental posterior and VO position estimates inside a window.

Our image registration provides a measurement of the vehicle pose for each keyframe, $\mathbf{T}_{k,0}$. We currently use a fixed measurement covariance where $\mathbf{R}_k = \text{diag}(0.11, 0.11, 1.0, 0.01, 0.01, 0.01)$. Therefore, the correction step fuses two uncertain poses with the error between them defined as

$$\mathbf{e}_k = \ln\left(\mathbf{T}_{k,0}\check{\mathbf{T}}_{k,0}^{-1}\right)^{\vee}. \tag{13}$$

As a result, our filtering equations are

$$\check{\mathbf{P}}_k = \mathbf{Q}_k + \boldsymbol{\mathcal{T}}_{k,k-1}\hat{\mathbf{P}}_{k-1}\boldsymbol{\mathcal{T}}_{k,k-1}^{\top} \tag{14a}$$

$$\check{\mathbf{T}}_{k,0} = \mathbf{T}_{k,k-1}\hat{\mathbf{T}}_{k-1,0} \tag{14b}$$

$$\mathbf{K}_k = \check{\mathbf{P}}_k\left(\check{\mathbf{P}}_k + \mathbf{R}_k\right)^{-1} \tag{14c}$$

$$\hat{\mathbf{P}}_k = (\mathbf{1} - \mathbf{K}_k)\check{\mathbf{P}}_k \tag{14d}$$

$$\hat{\mathbf{T}}_{k,0} = \exp\left(\left(\mathbf{K}_k \ln(\mathbf{T}_{k,0}\check{\mathbf{T}}_{k,0}^{-1})^{\vee}\right)^{\wedge}\right)\check{\mathbf{T}}_{k,0}, \tag{14e}$$

where $\boldsymbol{\mathcal{T}}_{k,k-1}$ is the adjoint of $\mathbf{T}_{k,k-1}$, the prior uncertainty $\check{\mathbf{P}}_k$ is a second-order approximation, $\mathbf{K}_k$ is the Kalman gain, and $\ln(\cdot)^{\vee}$ and $\exp(\cdot^{\wedge})$ are $SE(3)$ operators. We refer the reader to [18] for more detail. For unsuccessful registrations, the predicted position and uncertainties are propagated (i.e., $\hat{\mathbf{T}}_{k,0} = \check{\mathbf{T}}_{k,0}$ and $\hat{\mathbf{P}}_k = \check{\mathbf{P}}_k$). The posterior global vehicle pose at each keyframe is obtained by

$$\hat{\mathbf{T}}_{W,k} = \mathbf{T}_{W,0}\hat{\mathbf{T}}_{k,0}^{-1}. \tag{15}$$

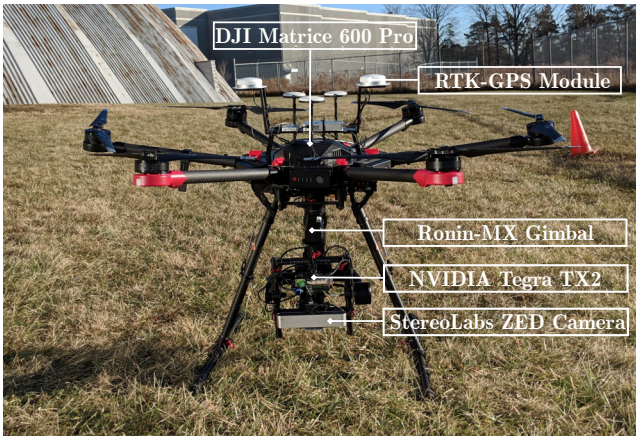


Fig. 2: A 3-axis gimballed stereo camera on a multirotor UAV with an onboard computer is used for our data collection.

IV. EXPERIMENTAL SETUP

A. UAV Dataset Collection

The experiments are conducted with data collected at UTIAS using the hardware setup shown in Fig. 2. We use the DJI Matrice 600 Pro multirotor UAV with a 3-axis DJI Ronin-MX gimbal. A StereoLabs ZED camera is connected to the onboard NVIDIA Tegra TX2 computer to provide 1280×720 RGB stereo images at 10FPS. These images are downscaled to 560×315 and converted to greyscale for VO and image registration. The gimbal connects to the flight controller to provide angular positions from joint encoders at 10 Hz. The RTK-GPS system provides the vehicle position at 5 Hz, and an IMU provides the vehicle attitude at 50 Hz.

The first dataset is a simple 303 m rectangular path flown with height variations beween $45 - 48$ m AGL. It was collected in the fall during an overcast day and is the primary dataset used for development and tuning of our method. We also collected six datasets during a sunny summer day on a more complicated 1132 m path flown with height variations between $36 - 42$ m AGL to show the ability of our method to localize a) at lower altitudes, and b) using a single map image database despite significant lighting changes in the real-world images. We collect a dataset near distinctive times of the day: sunrise (`06:17 AM`), morning (`08:50 AM`), noon (`11:54 AM`), afternoon (`02:50 PM`), evening (`05:50 PM`), and sunset (`08:24 PM`). Fig. 3 shows examples of the extreme lighting changes that occur throughout the day at two locations. These flights are over both man-made structure and significant stretches of vegetation to evaluate the performance in different environments.

B. Map Images

The set of geoferenced map images, $\mathcal{I}^m$, is generated from the 3D view in Google Earth at desired camera poses in an offline step. We define a virtual camera at each pose with the same focal length as the UAV-mounted camera so that query and map images taken at the same pose can have a nearly perfect alignment when the 3D reconstruction is precise. We also use GE elevation data to obtain the height of the camera AGL at each pose.

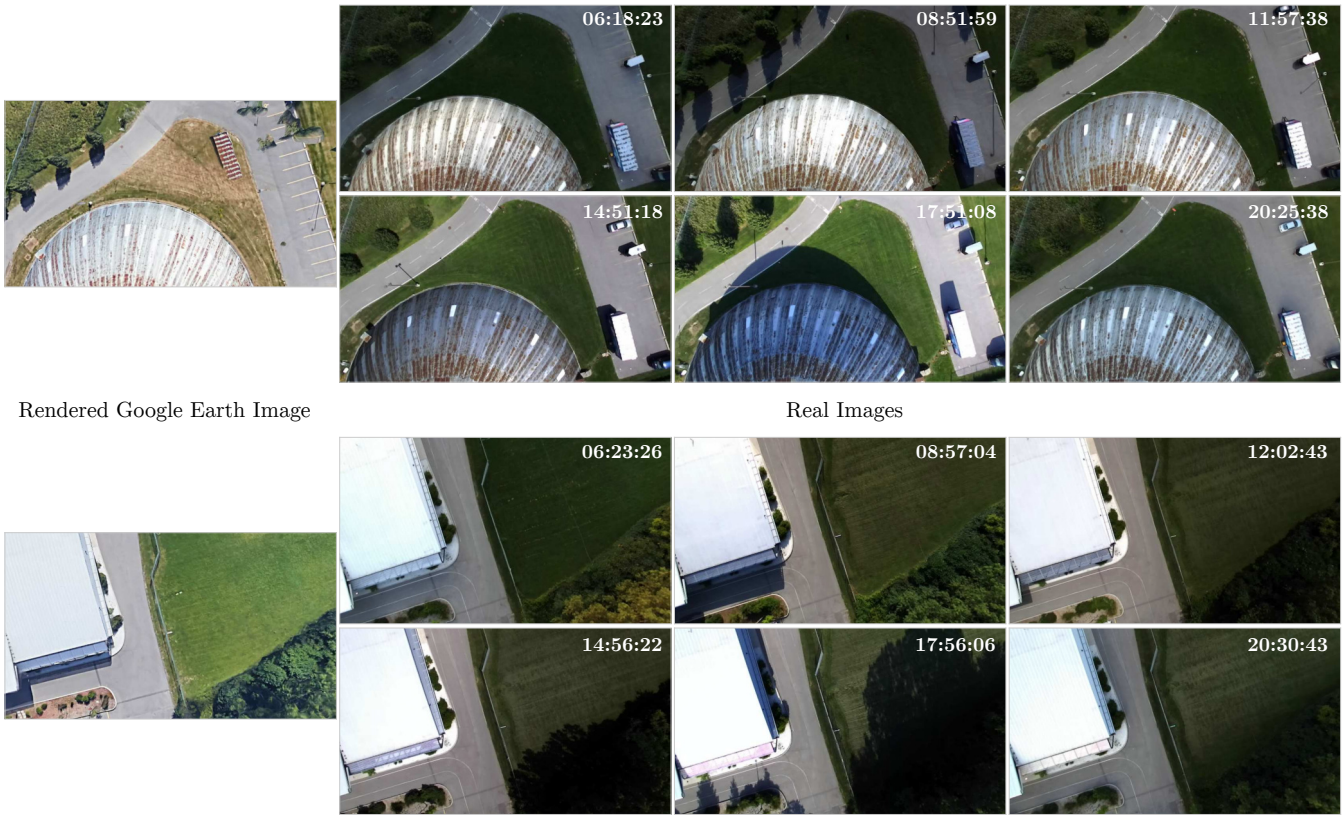


Fig. 3: Examples of lighting changes that occur from sunrise to sunset east of the dome (top) and north-west of the soccer field (bottom). The Google Earth reconstruction appears to contain late morning to early noon shadows.

After planning the UAV path, we render images at discrete poses along the nominal path. For this work, all images are rendered with the camera facing east and pointed in the nadir direction. A multirotor UAV is able to travel in all directions with the same heading allowing this restriction to be feasible for many applications such as drone delivery. It is also possible to use a second gimbal to orient an application-specific sensor. Images are generated every $3\,\text{m}$ along the nominal path to match our gimballed VO pipeline, which creates a new keyframe at approximately the same spacing. We extend the images $12\,\text{m}$ to the left and right, and above and below the path with a $6\,\text{m}$ spacing. The end result is a rectangular tube of images centered along the path with a $24 \times 24\,\text{m}$ cross-section and $6 \times 6 \times 3\,\text{m}$ spacing between images. The tube of images ensures that if the vehicle deviates off the nominal path, there is a map image taken from a nearby pose that captures the non-planar changes in the scene (e.g., side of a building becoming visible). This allows us to accurately localize with sRt warping at lower altitudes where the planar scene assumption is less valid. The spacing was chosen heuristically: the largest possible distance to any map image is $4.5\,\text{m}$ when inside the tube, which is approximately the width of our convergence basin at the altitude flown in these experiments.

The map images could be extended beyond $12\,\text{m}$ or cover an entire flight area. The only limitation is the storage available on the UAV. Although we save high-resolution RGB images, the image registration algorithm only uses 560×315 4-bit greyscale images (the NID is computed using 16-bin histograms of the greyscale intensities). Our $313\,\text{m}$ and $1.1\,\text{km}$ paths contain 2393 and 8992 map images, respectively, which would require approximately only $212\,\text{Mb}$ and $794\,\text{Mb}$ if saved in the minimum required format. With today's large capacity and inexpensive storage, map images covering several square kilometres could easily be stored onboard.

C. Ground Truth

It is important to note that the RTK-GPS and GE global coordinate frames, $\underrightarrow{\mathcal{F}}_{W'}$ and $\underrightarrow{\mathcal{F}}_{W}$, respectively, do not perfectly align. Therefore, we uniformly sample 10% of the posterior pose estimates along the path and use these to align the coordinate frames with a transform, $\mathbf{T}_{W',W}$, that is determined by miniziming the uncertainty-weighted relative pose errors between the RTK-GPS poses and posterior pose estimates. We report all image registration and filtered errors on the remaining 90% of the path.

V. RESULTS AND DISCUSSION

A. MI-based Image Registration

We first show an example of aligning two images with sRt warping using the NID cost function. Fig. 4 shows the cost function values swept over the four warping parameters for the first image in the overcast dataset. The warping

TABLE I: Summary of MI-based Image Registration Results

Lighting Condition	Registration Success (%)	Successful Registrations RMSE (m)				All Registrations RMSE (m)			
		long.	lat.	altitude	heading	long.	lat.	altitude	heading
Overcast	100	0.69	0.46	0.50	0.89	0.69	0.46	0.50	0.89
Sunrise	94.7	1.10	0.71	1.17	2.28	1.87	1.47	1.73	2.80
Morning	95.1	1.02	0.58	0.78	2.57	2.24	1.39	1.20	2.97
Noon	97.8	0.78	0.61	1.01	1.82	1.26	1.02	1.40	2.70
Afternoon	96.0	1.69	0.92	1.17	1.71	2.14	1.57	1.54	2.63
Evening	81.3	3.03	1.32	1.35	2.49	4.09	3.63	2.98	5.25
Sunset	87.5	1.95	1.12	1.55	2.64	3.03	1.95	2.54	3.06

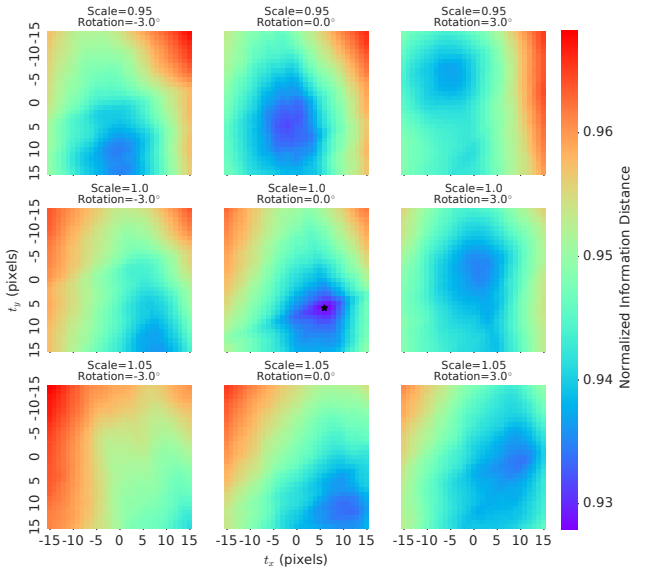
that generates the minimum NID is often the best image alignment as shown in this example. However, it is not guaranteed that there will be a single minimum or even that the global minimum corresponds to the best alignment. Furthermore, the absolute NID value highly depends on the scene. The near-perfect alignment in the example occurs at nearly 0.93 even though the NID is a value between 0 and 1 with a lower value indicating more similarity. For this reason, we use a geometric criterion for classifying image registration failures instead of thresholding the NID.

Next, we present our image registration results on all datasets using our two-step optimization approach. Table I shows the success rate and Root-Mean-Square Error (RMSE) computed using all registrations and only successful ones. The optimizer always converged to a solution thus all failures were due to poor image alignment.

For the overcast flight we successfully register every keyframe and achieve sub-metre position and sub-degree orientation errors. This is in part due to the higher altitude flight (although this is still relatively low compared to previous work), which provides more objects and boundaries to aid in the alignment. For the sunrise to sunset flights, the registration performs the best at noon as expected; the GE 3D reconstruction in our flight area resembles early noon. The image registrations alone were able to achieve nearly less than $3\,$m and 3° position and heading RMSE.

There are two types of scenes that are particularly difficult for our image registration: scenes with lots of self-similar texture (e.g., vegetation in Figs. 8b, 8c), and scenes with large shadows (e.g., Figs. 8a, 8d). Self-similar texture results in many local minima in the registration cost function. Shadows can trick the MI into associating the shadow with its caster resulting in a strong local minimum that may even be a global minimum. These shadows were most prevalent in the evening flight resulting in its lower success rate. While our blurred optimization provides robustness to shallow local minima, we depend on good initial guesses to handle the aforementioned problematic areas. Figs. 8f, 8g, 8h show examples of when the MI optimizer can settle in the correct local minimum with an initial guess given by VO near the true alignment. Another method to handle these scenes is to optimize over a window of keyframes. Although this may produce a suboptimal alignment for each individual keyframe, it prevents large jumps in the measured poses introduced by these additional minima.

The number of cost function evaluations required per image registration is presented in Fig. 5. Currently we place



(a) NID cost function swept from -15 to 15 pixel translations in t_x and t_y at three different scales and rotations.

(b) Alpha blended image of the UAV query image (prominent) and the Google Earth map image warped with the optimal sRt parameters.

Fig. 4: An example sRt alignment using the NID cost function showing the smoothness over the warping parameters with a clear optimum that corresponds to a nearly perfect alignment.

a very loose constraint on the number of iterations allowed for optimization. If necessary, we can reduce this number as, in many cases, the additional iterations do not provide a significant improvement to the accuracy. The current offline implementation uses the `SciPy` library for optimization thus is not able to run in real-time. However, we strongly believe our upcoming C++ implementation will enable the image registration to run at rate of at least $1\,$Hz.

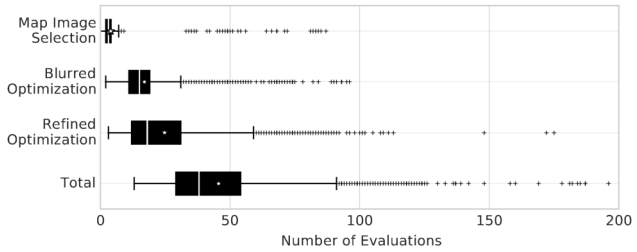


Fig. 5: Number of cost function evaluations per image registration at each step of the proposed method.

B. Comparison with Feature-based Registration

We briefly present the results from a feature-based image registration scheme for comparison. We use the aforementioned VT&R framework with SURF. The GE images along the nominal path are used for the teach run. Repeats are then attempted with each of the sunrise to sunset flights but the result is a poor registration performance. The features are only capable of producing less than 7% successes per repeat where the registration is declared a failure if the number of MLESAC inliers is below 30.

Since it is quite obvious that feature matching across the rendered and real-world images will struggle, we also briefly evaluate the performance of a typical teach-and-repeat without the use of GE images. The sunrise flight is used as the teach with subsequent flights used as repeats. This results in 34.9%, 30.3%, 15.0%, 8.4%, and 72.4% success rate for morning to sunset. It is clear that the dramatic changes in lighting makes feature matching unreliable. The sunset flight is able to localize the most frequently due to the similar brightness and minimal shadows that appear during sunrise and sunset.

C. Filtered Pose Estimation

TABLE II: Summary of Filtered Results

Lighting Condition	**RMSE (m)**					
	long.	lat.	altitude	roll	pitch	heading
Overcast	0.61	0.42	0.32	0.27	0.29	0.84
Sunrise	1.10	0.76	0.32	0.31	0.69	2.19
Morning	1.15	0.80	0.31	0.35	0.46	2.67
Noon	0.91	0.82	0.30	0.55	0.78	1.76
Afternoon	1.51	0.86	0.47	0.52	0.61	1.54
Evening	2.73	1.64	0.45	0.52	0.82	2.48
Sunset	1.78	0.76	0.51	0.52	0.84	2.55

Finally, we highlight the accuracy we can achieve by fusing VO and our MI-based real-to-rendered image registration. The 2D pure VO, image registration measurements, and filtered position estimates for the overcast flight are shown in Fig. 6 alongside the ground truth. It is clear that the combination of scaled VO to smooth out registrations and the registrations to correct for drifts in VO results in an accurate filtered global pose estimate. Fig. 7 shows the filtered positions and height AGL with the ground truth for two flights in our sunrise to sunset experiment: our best (noon) and worst (evening) performances. As we saw previously, a few particular areas were problematic for image registration in the presence of lighting changes. However, VO was able to carry the estimation through these small stretches

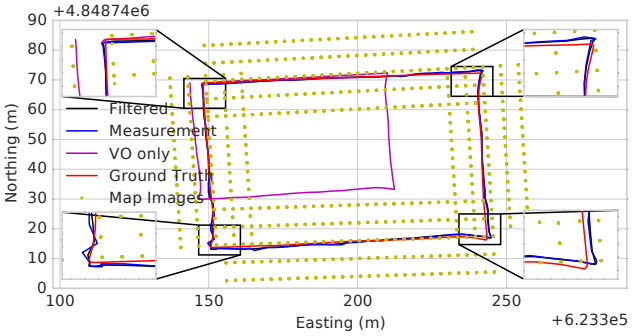


Fig. 6: Position estimates using our method for the 303 m overcast flight. The VO drifts significantly but our accurate image registrations allow us to estimate the scale and apply corrections.

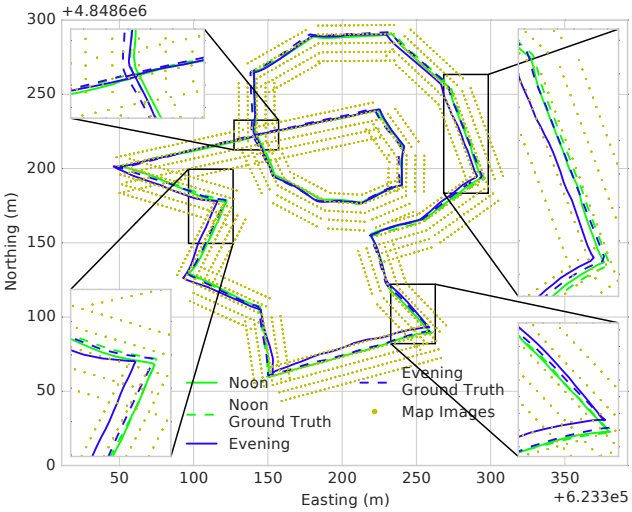
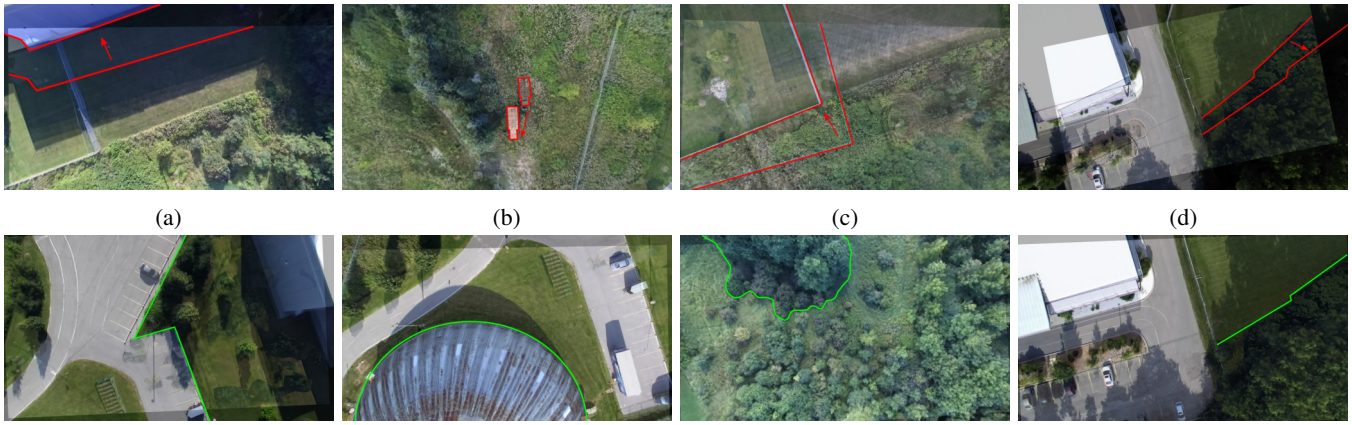


Fig. 7: Filtered estimates showing only our best (noon) and worst (evening) localization performances. The path starts at the intersection shown in the top left and spirals outward clockwise. The remaining three zoomed-in segments are where image registration is difficult for the evening flight resulting in slightly worse localization performance. Overall, however, the localization is smooth and performs well on the 1132 m path flown as low as 36 m AGL.

(5 – 10 keyframes) of failures that predominantly occurred during the evening and sunset flights. Overall, our method is able to estimate a global pose throughout the day with a position accuracy that rivals (non-differential) GPS.

In future work, we aim to show the estimation running online on the onboard computer. The localization will be tested at even lower altitudes and its performance verified in more locations. We will also focus on proper uncertainty quantification as this would allow adaptive search regions for



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 8: The top row shows examples of bad alignments due to (a) alignment of the building with its shadow, (b) and (c) almost no structure to aid in alignment, and (d) alignment of the trees with their shadows. The bottom row shows good alignments despite: (e) poor 3D reconstructions, (f) and (h) large shadows, and (g) very little structure. A better initial pose guess and slightly more structure in (h) compared to (d) allows the MI optimizer to correctly align the images.

the map image and adaptive geometric constraints to classify registration failures. In parallel, we will explore using deep learning for image registration to improve the robustness.

VI. CONCLUSIONS

We presented a method for global pose estimation of a UAV by visually localizing real-world images with pre-rendered images from a 3D reconstruction of the Earth. We used a MI-based dense image registration scheme to align the real and rendered images for metric localization. The registrations were then used to apply corrections to gimballed VO in a filtering framework. On multiple flights totaling 7.1 km of data with altitudes as low as 36 m AGL, we estimated the full pose with an accuracy on the order of a few metres and degrees. We also showed the ability to consistently localize over the course of a sunny summer day using a single database of pre-rendered images despite dramatic changes in lighting. Our method enables global pose estimation with a position accuracy on par with GPS.

ACKNOWLEDGMENT

This work was funded by NSERC Canada Graduate Scholarship-Master's (CGS-M), Defence Research and Development Canada (DRDC), Drone Delivery Canada (DDC) and the Centre for Aerial Robotics Research and Education (CARRE), University of Toronto.

REFERENCES

[1] M. Blösch, S. Weiss, D. Scaramuzza, and R. Siegwart, "Vision based MAV navigation in unknown and unstructured environments," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 21–28, 2010.

[2] S. Weiss, M. W. Achtelik, S. Lynen, M. C. Achtelik, L. Kneip, M. Chli, and R. Siegwart, "Monocular Vision for Long-term Micro Aerial Vehicle State Estimation: A Compendium," *Journal of Field Robotics*, vol. 30, no. 5, pp. 803–831, 2013.

[3] S. Shen, N. Michael, and V. Kumar, "Tightly-coupled monocular visual-inertial fusion for autonomous flight of rotorcraft MAVs," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, vol. 2015-June, pp. 5303–5310, 2015.

[4] P. Furgale and T. D. Barfoot, "Visual Teach and Repeat for Long Range Rover Autonomy," *Journal of Field Robotics*, vol. 27, no. 5, pp. 534–560, 2010.

[5] M. Warren, M. Greeff, B. Patel, J. Collier, A. P. Schoellig, and T. D. Barfoot, "There's no place like home: Visual teach and repeat for emergency return of multirotor UAVs during GPS failure," *IEEE Robotics and Automation Letters*, vol. 4, no. 1, pp. 161–168, 2019.

[6] G. Conte and P. Doherty, "An integrated UAV navigation system based on aerial image matching," in *IEEE Aerospace Conference Proceedings*, 2008.

[7] K.-h. Son, Y. Hwang, and I.-s. Kweon, "UAV global pose estimation by matching forward-looking aerial images with satellite images," in *Proc. of the International Conference on Intelligent Robots and Systems (IROS)*, pp. 3880–3887, 2009.

[8] P. Agarwal and L. Spinello, "Metric Localization using Google Street View," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3111–3118, 2015.

[9] A. L. Majdik, D. Verda, Y. Albers-Schoenberg, and D. Scaramuzza, "Air-ground Matching: Appearance-based GPS-denied Urban Localization of Micro Aerial Vehicles," *Journal of Field Robotics*, 2015.

[10] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of ConvNet features for place recognition," in *Proc. of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 4297–4304, 2015.

[11] T.-y. Lin, J. Hays, and C. Tech, "Learning Deep Representations for Ground-to-Aerial Geolocalization," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[12] D. K. Kim and M. R. Walter, "Satellite image-based localization via learned embeddings," in *Proc. of the IEEE International Conference on Robotics and Automation*, pp. 2073–2080, 2017.

[13] A. Shetty and G. X. Gao, "UAV Pose Estimation using Cross-view Geolocalization with Satellite Imagery," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[14] G. Pascoe, W. Maddern, and P. Newman, "Robust Direct Visual Localisation using Normalised Information Distance," in *British Machine Vision Conference (BMVC)*, (Swansea, Wales), pp. 70.1–70.13, 2015.

[15] G. Pascoe, W. Maddern, A. D. Stewart, and P. Newman, "FARLAP: Fast robust localisation using appearance priors," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6366–6373, 2015.

[16] A. Yol, B. Delabarre, A. Dame, and J.-e. Dartois, "Vision-based Absolute Localization for Unmanned Aerial Vehicles," in *Proc. of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 3429–3434, 2014.

[17] S. Anderson and T. D. Barfoot, "Full STEAM ahead: Exactly sparse Gaussian process regression for batch continuous-time trajectory estimation on SE(3)," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 157–164, 2015.

[18] T. D. Barfoot, *State Estimation for Robotics*. Cambridge University Press, 2017.