Towards Visual Teach & Repeat for GPS-Denied Flight of a Fixed-Wing UAV

M. Warren, M. Paton, K. MacTavish, A. P. Schoellig and T. D. Barfoot *

Abstract Most consumer and industrial Unmanned Aerial Vehicles (UAVs) rely on combining Global Navigation Satellite Systems (GNSS) with barometric and inertial sensors for outdoor operation. As a consequence, these vehicles are prone to a variety of potential navigation failures such as jamming and environmental interference. This usually limits their legal activities to locations of low population density within line-of-sight of a human pilot to reduce risk of injury and damage. Autonomous route-following methods such as Visual Teach and Repeat (VT&R) have enabled long-range navigational autonomy for ground robots without the need for reliance on external infrastructure or an accurate global position estimate. In this paper, we demonstrate the localisation component of VT&R outdoors on a fixed-wing UAV as a method of backup navigation in case of primary sensor failure. We modify the localisation engine of VT&R to work with a single downward facing camera on a UAV to enable safe navigation under the guidance of vision alone. We evaluate the method using visual data from the UAV flying a 1200 m trajectory (at altitude of 80 m) several times during a multi-day period, covering a total distance of 10.8 km using the algorithm. We examine the localisation performance for both small (single flight) and large (inter-day) temporal differences from teach to repeat. Through these experiments, we demonstrate the ability to successfully localise the aircraft on a self-taught route using vision alone without the need for additional sensing or infrastructure.

1 Introduction

With increasing use in civilian airspace, UAVs need to be able to navigate reliably and safely using a variety of redundant sensing modalities. Typically, low-cost, commercial UAVs (Figure 1) used for mapping and surveillance tasks rely on a combination of GNSS such as the Global Positioning System (GPS) in combination with barometric, airspeed and inertial sensing to navigate outdoors. However, these sensors are prone to both malicious and environmental interference (e.g., jamming, poor sky view, obstruction and mechanical stress). This means that airspace regula-

University of Toronto Institute for Aerospace Studies (UTIAS), 4925 Dufferin St. Toronto, ON M3H 5T6, {michaelwarren, mpaton, kirk.mactavish, angela.schoellig}@robotics.utias.utoronto.ca, tim.barfoot@utoronto.ca

* This work was supported by the NSERC Canadian Field Robotics Network (NCFRN) and a Mathematics of Information Technology and Complex Systems (MITACS) Accelerate Fellowship in partnership with PrecisionHawk Ltd.

tors often tightly restrict their operation to line-of-sight and low-population-density locations to minimise risk.

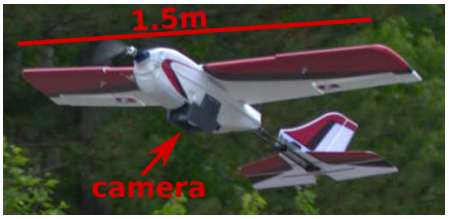


Fig. 1: The PrecisionHawk Lancaster fixed-wing UAV used in experiments, seen here during take-off. Note the payload bay in the centre of the fuselage housing the downward-looking camera used for experiments. (Image: François Pomerleau)

Autonomous route-following methods such as VT&R [12] have enabled long-range navigational autonomy for ground robots without relying on external infrastructure or an accurate global position estimate. By first building a visual map while under control of a human operator (the 'teach' phase), VT&R then allows the vehicle to autonomously re-follow the taught path (the 'repeat' phase) by matching sensor observations back to the original map in a local co-ordinate frame and providing path-tracking errors to a suitable vehicle controller [22]. We seek to adapt VT&R for use on a fixed-wing UAV as a demonstration of a low-cost navigation method in case of GPS, communications, or other navigational failure.

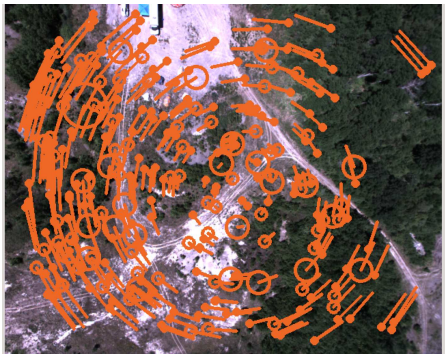


Fig. 2: Post-MLESAC matches (orange lines) and features (orange circles, size denotes octave) during localisation for a repeat flight. Large orientation differences between teach and repeat phases account for the large pixel offsets seen in matching, while a high number of inliers is representative of the short temporal difference between teach and repeat (∼20 minutes) in this case.

We see the applicability of VT&R on aircraft in two different cases: 1) a method of emergency return during an exploratory or traditional mapping flight, by following the 'visual breadcrumbs' home, and 2) acting as a complement or complete replacement of primary navigational systems when performing flights over repeat trajectories (e.g., inter-warehouse delivery or linear infrastructure inspection) in cases where GPS may be unreliable (e.g., due to poor sky view or jamming). This type of safety net could open the door to operation beyond line-of-sight in more urbanised environments, and in less than ideal physical conditions.

To date, VT&R has been primarily demonstrated on ground vehicles following restricted routes. In this paper, we demonstrate a core aspect of VT&R adapted to a fixed-wing UAV: accurately localising during a repeat flight over a pre-taught route using a downward-facing, onboard camera in a large-scale, outdoor experiment (Figure 2). This demonstration of VT&R on a UAV has some critical differences to a ground-based robot: 1) reliance on stereo for accurate scale is not feasible due to the ratio of practical baseline to altitude; 2) perspective of the scene can be radically different due to changes in altitude, orientation and position; meaning map observations can often be fleeting, 3) trajectories are no longer restricted

to specific routes as there are few traversability concerns like that of ground robots when flying at sufficient altitude.

This paper presents the first demonstration of the VT&R localisation engine, or any visual route-following method, in this scenario. We present performance statistics related to localisation robustness and algorithm speed and discuss the implications and challenges of adapting VT&R to this scenario. To sufficiently limit the scope of this paper, we leave a number of tasks to future work; including closing the control loop on navigating the aircraft along the autonomously taught route, the planning of an efficient return route, and identifying when traditional navigation has failed in order to switch over to the emergency return mode.

The rest of this paper is outlined as follows: section 2 describes related work, section 3 outlines the monocular VT&R framework and application-specific modifications, section 4 describes the vehicle, datasets, and experiments used to test the VT&R localisation engine, section 5 demonstrates the results of experiments, while section 6 discusses the outcomes and challenges of this work. The paper is concluded in section 7.

2 Related Work

Today, most small-scale UAVs utilise GPS and inertial measurements in a filtered framework for 6 Degree-of-Freedom (DOF) state estimation [20]. However, many civil aviation authorities have made clear, through statements and regulations [15, 8], that for UAVs to perform routine operations over urban and other sensitive environments, reliance solely on GPS and radio-based communication for accurate navigation is not sufficient. New technology is attempting to bridge or mitigate this gap with improved air-to-air communications, localisation from existing infrastructure, and the ability to land safely in the event of an emergency.

Non-GPS-based navigation on aerial vehicles has seen increased interest in recent years due to these regulatory and operational issues, with many demonstrations in GPS-denied environments [29, 11, 9] using LiDAR [5] and stereo [16], visual-inertial systems [1, 26], and with vision alone [18, 10]. In most cases of outdoor, vision-only navigation, the literature has mostly been restricted to visual odometry or relatively small maps with few online examples [6, 17], mostly due to the mass and compute limitations on board the aircraft, or sometimes offloading processing to a more powerful ground-based computer with a high-rate data link.

Visual route-following on pre-built maps has been studied for some time [14]. As a modern technique, VT&R has been extensively tested in ground-based applications with stereo cameras [12], with LiDAR [19], and with multiple experiences for long-term autonomous navigation [24], and has made use of colour-constant imagery to improve resistance to lighting change [23]. It has also been tested with monocular cameras by taking advantage of the ground-plane assumption [7] and has seen preliminary testing in the air on board a Micro Aerial Vehicle (MAV) [25], demonstrating its wide applicability. Our work differs from [7] in that it does not make strict assumptions about the ground plane nor require continuous knowledge of the camera altitude, as in [7] and [25].

3 Methodology

In this paper, we intend to demonstrate robust localisation on imagery gathered from a fixed-wing UAV suited to the task of autonomous route following, without requiring input from additional sensors. We use the same software system as [24], adapted to suit a monocular front end for the single camera on board the aircraft. Similar to [24], the algorithm consists of separate teach and repeat phases. During the teach phase, the aircraft flies under control of an on-board autopilot during a primary data gathering task, analagous to the human operator used in ground-vehicle demonstrations, inserting the visual observations from this privileged experience into a relative map of pose and scene structure. During the repeat phase, without reliance on GPS or other sensors, the vehicle should autonomously re-follow the route by visually localising to the map of the privileged path. The vehicle repeats a path by sending high-frequency localisation updates to a path-tracking controller [21]. While such a system has been demonstrated online on ground vehicles (by using a human operator and stereo vision to follow the privileged path) [24], in this paper we demonstrate the localisation engine of VT&R using datasets and leave closing the control loop to future work. The remainder of this section provides details on the mapping process (Section 3.1), Visual Odometry (VO) (Section 3.2) and the localisation process (Section 3.3-3.6).

3.1 Map Building: The map used in our system, which we refer to as a Spatio-Temporal Pose Graph (STPG), is depicted in Figure 3. This data structure is an undirected graph, $G = \{V, E_s, E_t\}$, where V is a set of vertices, E_t is a set of *temporal* edges, and E_s is a set of *spatial* edges. Vertices, each with an associated reference frame, $\mathcal{F}$, store raw sensor observations and triangulated 3D landmarks with associated covariances and descriptors. Landmarks and associated descriptors are stored in the first vertex at which the feature corresponding to the landmark is observed. An edge in the graph links vertices metrically with a relative 6 DOF

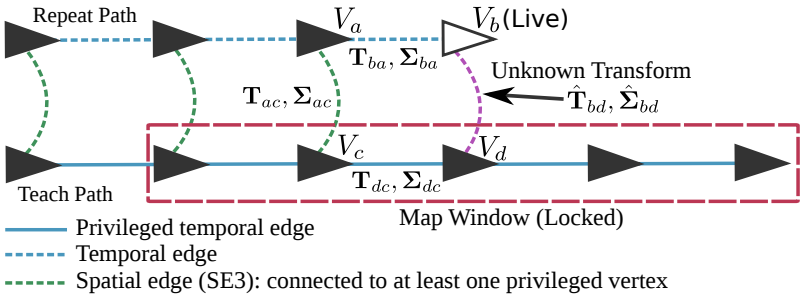


Fig. 3: Overview of the localisation problem and the spatio-temporal pose graph (STPG) data structure used as our map. We wish to estimate the unknown transform and uncertainty, $\{\hat{\mathbf{T}}_{bd}, \hat{\mathbf{\Sigma}}_{bd}\}$ (dashed, purple line), between the live vertex, V_b, and the target vertex, V_d, in the privileged path (solid blue line). This is achieved by matching all landmarks in V_b to landmarks observed in the map window (dashed, red rectangle), transformed into the coordinate frame of V_d. This setup allows for outlier rejection and a simple optimisation of $\{\hat{\mathbf{T}}_{bd}, \hat{\mathbf{\Sigma}}_{bd}\}$ against a map of locked landmarks with uncertainty.

$SE(3)$ transformation with uncertainty. Temporal edges (blue lines) link vertices that are temporally adjacent, while spatial edges (green lines) link vertices that are

temporally distant yet spatially close, i.e., from the repeat to the teach pass. Temporal edges are furthermore denoted as *privileged* if they were collected while the aircraft was self-teaching a route under autopilot control or *repeated* if the aircraft was following a privileged route; this distinction is illustrated in Figure 3 as solid and dashed lines, respectively. We define an *experience* as a contiguous collection of vertices linked by temporal edges. Mapping consists of adding either a privileged or autonomous experience to a new or existing STPG while computing data products and temporal edge transformations through a monocular VO pipeline, which is illustrated in Figure 4. For each incoming frame (the *live* frame), sparse visual features are extracted and their descriptors computed. Features are represented by a single measurement, $\{\mathbf{y}, \mathbf{Y}\}$, where $\mathbf{y}$ is the 2×1 keypoint position of the feature and $\mathbf{Y}$ is the 2×2 covariance on the measurement. We use oriented Speeded-Up Robust Features (SURF) [4] to detect and describe keypoints and calculate $\mathbf{Y}$ based on the octave and Hessian of the response. When two measurements are matched through their descriptors, the 3D landmark is triangulated via the inverse camera model and the relative transform between the cameras to obtain a 3D landmark including uncertainty, $\{\mathbf{p}, \boldsymbol{\Phi}\}$, where $\mathbf{p}$ is the 4×1 positional mean in homogeneous coordinates and $\boldsymbol{\Phi}$ is the uncertainty represented by a 3×3 covariance.

3.2 Visual Odometry: To initialise the VO, features and descriptors are extracted from the first image, and then matched against those from subsequent frames. Both an Essential matrix and 2D Homography matrix are computed using MLESAC [28] to extract a relative transformation for each new frame, subject to a Geometric Robust Information Criterion (GRIC) test [27] to select the best estimate. Once the inlier count for each frame-to-frame transformation drops below a threshold (as an analogue for translational motion), landmarks are triangulated (subject to a re-projection and plane-distance test to eliminate gross outliers) and the pair of frames placed as the first two vertices in the graph, with the computed transformation inserted as the edge. To initialise the scale appropriately, a ground plane is fitted to the triangulated landmarks from which the height from the scene is extracted, then the true height from the ground is retrieved from a GPS position at a similar time-point to find the scaling parameter. This is then applied to the transformation and landmarks. In practice, any approximate scaling data can be used, such as from a calibrated barometer, laser altimeter, or other suitable sensor. Perfect scaling is not crucial to the function of VT&R, drift of global estimates is easily handled.

For subsequent frames, extracted features from the live view are matched via their appearance to locked landmarks in the latest graph vertex (a.k.a., keyframe) and motion computed (Figure 4a) by solving the Perspective-Three-Point (PnP) problem [13], again using MLESAC. New landmarks are triangulated from new matches that are not associated with an existing landmark, subject to the same plane-distance and re-projection tests to remove outliers. A trajectory (velocity and position) estimate is produced at frame rate from the optimisation and can be queried to predict future motion. This prediction is used to project landmarks into the new frame (reducing image search space for matching) and compensates for latency between the localisation system and the path-tracking controller. If the translational or rotational motion is large, or the number of matched features between the live view and the last graph

(a) High-rate: frame-to-vertex VO. Landmark estimates (black stars) at the latest vertex are locked, and the motion estimate is solved using new matches from the live frame (purple triangle) to the last vertex (black triangle) while using a smoothing trajectory prior (black dot).

(b) Low-rate: sliding-window vertex bundle adjustment. Transforms and landmarks connected to unlocked vertices (white triangles) are optimised and those connected to locked vertices (black triangles) are locked.

Fig. 4: VO pipeline showing the parallel high-rate, approximate (a) and low-rate, accurate (b) estimators similar to [17].

vertex drops too low, the live frame is inserted as a new vertex in the graph; otherwise, it is discarded. Upon insertion of a new vertex, a temporal edge linking to the previous vertex is added. If the aircraft is in GPS-based teach mode, this edge is flagged as privileged. Following vertex insertion, bundle-adjustment is performed on a sliding window of the latest vertices in the graph (Figure 4b) using our Simultaneous Trajectory Estimation And Mapping (STEAM) [2] engine; smoothing factors are added to the relative transforms to ensure stability in the estimated trajectory during areas of poor feature tracks. After optimisation, the updated poses, landmarks, and their uncertainties are re-inserted into the graph.

3.3 Localisation: When repeating a path, the overall objective of the algorithm is to estimate the posterior transform and uncertainty, $\{\hat{\mathbf{T}}_{bd}, \hat{\boldsymbol{\Sigma}}_{bd}\}$, between the most recent vertex in the live run, V_b, and the estimated closest vertex in the privileged path, V_d. This is achieved by minimising the measurement error of landmarks in the map window (red, dashed rectangle in Figure 3) observed by V_b. Throughout the algorithm, we make use of the prior term, $\{\check{\mathbf{T}}_{bd}, \check{\boldsymbol{\Sigma}}_{bd}\}$, obtained by compounding the uncertain transforms [3],

$$\{\mathbf{T}_{ba}, \boldsymbol{\Sigma}_{ba}\},\ \{\mathbf{T}_{ac}, \boldsymbol{\Sigma}_{ac}\},\ \{\mathbf{T}_{cd}, \boldsymbol{\Sigma}_{cd}\}, \tag{1}$$

which are computed through previous VO and previous localisation estimates. The localisation pipeline consists of the following main steps: a) Landmark Transformation, b) Localisation Matching, and c) State Estimation.

3.4 Landmark Transformation: The first step of localisation is to transform all landmark means and uncertainties in the active map window from their respective coordinate frames in each vertex to $\underrightarrow{\mathcal{F}}_d$, the coordinate frame of V_d and the one in which localisation is to be computed. We use the same process as Paton *et al.* [24] to transform landmarks expressed in a nearby vertex map frame, $\underrightarrow{\mathcal{F}}_m$, with mean and covariance, $\{\mathbf{p}_m, \boldsymbol{\Phi}_m\}$, to $\underrightarrow{\mathcal{F}}_d$, giving $\{\mathbf{p}_d, \boldsymbol{\Phi}_d\}$, ensuring uncertainty is appropriately transformed along with the landmark co-ordinates. This process is carried out on all landmarks in the map window to produce a set of landmarks with 3D position and uncertainty, all expressed in the privileged frame, $\underrightarrow{\mathcal{F}}_d$. The locations and uncertainties of all landmarks are transformed, even if they are not matched, as these help refine the matching process, making it faster and more robust.

3.5 Localisation Matching: The goal of localisation matching is to associate every *feature* observed by V_b to a landmark in the map window, even if the feature is not associated with a landmark in V_b. The process begins with labeling all features in the live vertex as unmatched. Vertices in the map window are sequentially examined starting from V_d in an outward search pattern. We chose to center the search around the privileged target vertex as a heuristic for prioritising landmarks that have the lowest uncertainty in the target privileged frame. For every new vertex visited, the transformed map landmarks associated with this vertex are projected into the camera frame of vertex V_b using the prior term, $\{\check{\mathbf{T}}_{bd}, \check{\boldsymbol{\Sigma}}_{bd}\}$. Each feature associated with this vertex is then checked for matching feasibility to the unmatched live features by comparing keypoint position and descriptor appearance. This process continues until one of three criteria are met: i) a sufficient number of matches are found, ii) the amount of time has surpassed an allowed limit, or iii) the map window of vertices is exhausted. As the process of comparing visual features is costly, this process is the most computationally expensive step of localisation, but it is performed in parallel to the main VO pipeline for each new vertex on the Graphics Processing Unit (GPU), meaning online operation is possible. Upon completion of localisation matching, the problem is set up so that there are candidate features in V_b associated with landmarks in V_d. This information is sent through a MLESAC PnP estimator to initialise the relative transform between V_b and V_d (as this may be significantly different from the prior) and remove outliers.

3.6 State Estimation: We now seek the optimal posterior,

$$\{\hat{\mathbf{T}}_{bd}, \hat{\boldsymbol{\Sigma}}_{bd}\}, \tag{2}$$

given the prior term, $\{\check{\mathbf{T}}_{bd}, \check{\boldsymbol{\Sigma}}_{bd}\}$, as well as associated data between V_b and map landmarks in the coordinate frame of V_d. This can be achieved by minimising the negative-log-likelihood cost function:

$$J(\mathbf{T}_{bd}) = \frac{1}{2}\sum_{j=1}^{M} \mathbf{e}_j{}^T \mathbf{R}_j^{-1} \mathbf{e}_j + \frac{1}{2}\mathbf{e}^T \mathbf{R}^{-1} \mathbf{e}, \tag{3}$$

with the first term in J summing the squared reprojection error of map landmarks and the second term encoding the transform prior. Given a map landmark, j, with mean and uncertainty, $\{\mathbf{p}_{d,j}, \boldsymbol{\Phi}_{d,j}\}$, expressed in the co-ordinate frame of V_d and a monocular measurement of j, $\mathbf{y}_j$, with uncertainty, $\mathbf{Y}_j$, expressed in the camera frame of V_b, the reprojection error is defined by

$$\mathbf{e}_j = \mathbf{y}_j - \mathbf{g}(\mathbf{T}_{bd}\mathbf{p}_{d,j}), \tag{4}$$

$$\mathbf{R}_j = \mathbf{Y}_j + \mathbf{G}_j \mathbf{T}_{bd} \mathbf{D} \boldsymbol{\Phi}_{d,j} \mathbf{D}^T \mathbf{T}_{bd}^T \mathbf{G}_j^T, \tag{5}$$

where $\mathbf{g}(\cdot)$ is the monocular measurement model and $\mathbf{G}_j$ is its Jacobian (evaluated at $\mathbf{p}_{b,j} = \mathbf{T}_{bd}\mathbf{p}_{d,j}$), with

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}. \tag{6}$$

This weights each error by uncertainty in the measurement and the map. The second term of Eq. (3) constrains the optimisation problem by the prior with

$$\mathbf{e} = \ln(\check{\mathbf{T}}_{bd}\mathbf{T}_{bd}^{-1})^{\vee}, \quad \mathbf{R} = \check{\boldsymbol{\Sigma}}_{bd}, \tag{7}$$

where $\vee$ is the inverse operator of $\wedge$ [3]. To obtain an optimal posterior estimate, $\hat{\mathbf{T}}_{bd}$ is iteratively refined in a nonlinear least-squares optimisation using our STEAM engine [2]. In the absence of any matches between the live image and map, the prior estimate (based on VO) is returned.

4 Experimental Setup

To evaluate the performance of the modified VT&R localisation in this new application, a series of experiments were conducted offline using a monocular dataset collected using a PrecisionHawk Lancaster Rev IV. (Figure 1). This aircraft is the target system for the developed algorithms, with a take-off weight of $\sim$2.5kg and wingspan of 1.5m. The typical flight time is 30-45 minutes. This UAV was fitted with a custom payload consisting of a single Point Grey Chameleon machine-vision camera, configured to face straight down from the payload bay. This camera uses a ½″ global shutter CMOS sensor, with an approximately $90° \times 80°$ Field of View (FOV). Bayer-encoded imagery is captured at $\sim$22Hz and 1280×1024 pixel resolution (converted to grayscale and down-sampled to 640×512 for this experiment). Imagery is recorded using an on-board, 1.6Ghz Intel Atom PicoITX computer along with GPS data at 5Hz from an on-board Ublox LEA-6N receiver. The data was gathered at a disused open-pit gravel mine in Sudbury, in central Canada, during early summer. This site consists of dirt roads, both undisturbed and naturally reforested boreal forest and exposed regolith from prior mining operations. The dataset consists of multiple flights using the custom payload, covering a square box pattern (Figure 5) with segments approximately 400 m in length. This pattern is flown in sequence, multiple times per flight. Each flight is approximately 15-25 minutes in duration, with 1-7 repeats per flight, and these are split into individual 'experiences' that cover a full loop of the flown square pattern. A selection of these are used in the experiments for this paper, shown in Table 1, chosen to cover a variety of test scenarios (other flights were for different test configurations).

ID	Flight	Start Time	Conditions
		Condition	
e0	2	14/6/2016 12:53	sunny, calm
e4	2	14/6/2016 13:05	sunny, calm
e5	4	14/6/2016 14:57	sunny, calm
e11	4	14/6/2016 15:14	sunny, calm
e12	8	14/6/2016 17:54	sunny, calm
e18	8	14/6/2016 18:07	sunny, calm
e19	10	15/6/2016 12:07	sunny, windy
e24	10	15/6/2016 12:19	sunny, windy
e25	15	16/6/2016 12:02	sunny, windy

Table 1: Overview of the selected experiences in the Sudbury dataset.

Fig. 5: The configuration of the flight path. Start and end of route at bottom left corner. Note forested areas in top-left and right of image.

We test the performance of VT&R localisation on the airborne data by experimenting with these selected sets of flights, focusing on increasing time differences between the initial teach pass and repeat to evaluate the performance of the system under appearance change and compare to our well established ground-based system. For each experiment, performance is evaluated by examining both relative uncertainty of the UAV and inlier matches during localisation in the repeat phase. The makeup and included experiences in each experiment are listed in Table 2. These experiments can be grouped into three general categories: 1) same-flight repeats (g0-g3), 2) temporally close repeats (g4-g6), and 3) temporally distant repeats (g7-g8).

Experiments g0-g3 include teach and repeat from the same flight (the first pass of the pattern to the last). In all these experiments, the time difference between teach and repeat is less than 17 minutes. Experiments g4-g6 include teach and repeat from flights that are temporally close, but different days. Experiment g4 includes a repeat approximately 24 hours after the teach, but with only a 9-minute time-of-day difference. Experiments g5 and g6 are one and two days after the teach, but 43 and 54 minutes temporally distant from the teach. Finally, experiments g7-g8 are conducted on the same day, but approximately 2 and 5 hours after the teach. We use both grayscale and colour-constant imagery [23] in all experiments to ensure the best performance in VO and localisation.

By examining the performance of VT&R localisation in this way, we can establish the temporal limitations on safe and accurate repeats for emergency returns, and make comparisons to the performance of VT&R localisation in the more traditional ground-vehicle environment. We use the localisation uncertainty as the primary metric for judging localisation success, and define it as the one-standard-deviation uncertainty of our 3D translation estimate relative to the privileged path. This tells us how uncertain we are of the distance of the vehicle to the privileged path. This is plotted as a Cumulative Distribution Function (CDF), where better performance is indicated by lower uncertainties over a greater percentage of the path. It is important to note that while uncertainty is calculated at every stage of the algorithm from keypoint detection to landmark transformation, we have not yet performed a rigorous evaluation of our uncertainty estimates with respect to ground truth to ensure consistency. Therefore we treat this metric as a way to compare relative performance between experiments and do not necessarily trust the exact scale of our uncertainty estimates.

We also include the inlier count for each localisation on the repeat path, and use a count of 15 inliers as the minimum number to constitute a successful localisation. Fewer than 15 inliers generally indicates either a poor or degenerate estimate. We

plot this as inlier count vs. time since repeat start, grouped into three figures corresponding to the experiment type described above, to highlight the reliability of localisation over the course of each flight. Each experience is from 120 to 170 seconds long, and we normalise the inlier count vs. time results to 170 seconds to improve the consistency of comparison when discussing sections that cover the same area.

ID	Live exerience	Privileged experience	ΔT teach to repeat hh:mm (24hr hh:mm)
g0	e4	e0	0:09
g1	e24	e19	0:12
g2	e18	e12	0:13
g3	e11	e5	0:17
g4	e25	e19	23:51 (-0:09)
g5	e19	e0	23:17 (-0:43)
g6	e25	e0	47:04 (-0:54)
g7	e5	e0	2:07
g8	e12	e0	5:04

Table 2: Overview of the configurations used for localisation experiments.

ID	Localisation keyframes	Successful localisations	%
g0	880	880	100.0
g1	938	932	99.4
g2	824	824	100.0
g3	832	832	100.0
g4	999	978	97.9
g5	797	701	88.0
g6	775	530	68.4
g7	664	443	66.7
g8	526	230	43.7

Table 3: Localisation performance for each experiment. Success is an inlier count > 15.

5 Results

Results are presented in Figures 6-7c. In Figure 6, it can be seen that short time differences ($<$ 20 minutes) mean the probablility of successful localisation is high (Table 3). This corresponds well with our intended application of emergency return, where the repeat would typically be conducted in the same flight as the teach. Since the aircraft's typical flight time is 30-45 minutes, these results indicate that return within a single flight is feasible and reliable. With increasing temporal difference, however, the average uncertainty grows rapidly. Performance rapidly drops after 45 minutes difference (with a total time difference of 24 hours). At two hours (g7) and five hours (g8) difference from teach to repeat, uncertainty is high and localisation performance is significantly degraded (Table 3).

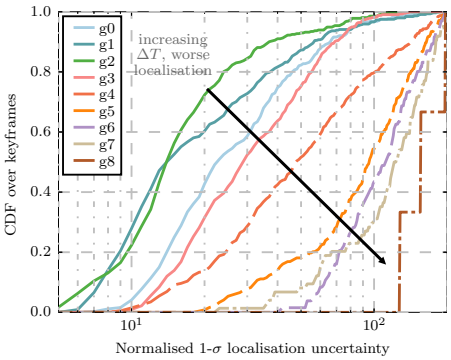
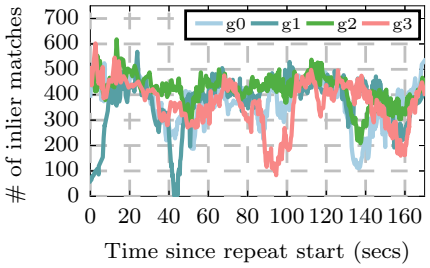


Fig. 6: The CDF of translational uncertainty for each experiment, in increasing temporal time difference. g0-g3 (solid lines) are within the same flight, g4-g6 (dashed lines) are different days but within 9-54 minutes of the teach and g7-g8 (dash-dot lines) are large temporal differences (2 and 5 hours respectively).

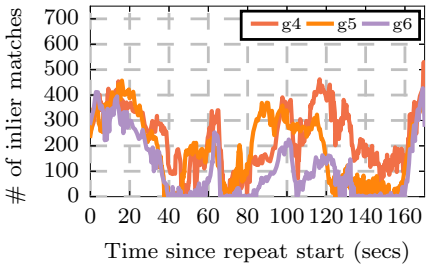
These results are corroborated by plots of localisation inliers over time (Figure 7). For experiments g0-g3 (Figure 7a), localisation performance is strong, with the number of inliers at each localisation step approximated by the number ($\sim$500) of migrated points. Figure 7b and Figure 7c highlight the reduced performance of later repeats. The algorithm uses 1GB RAM and runs at $\sim$25Hz on an NVIDIA Tegra TX2, which exceeds the framerate used in the dataset of $\sim$22Hz.

6 Discussion, Challenges & Future Work

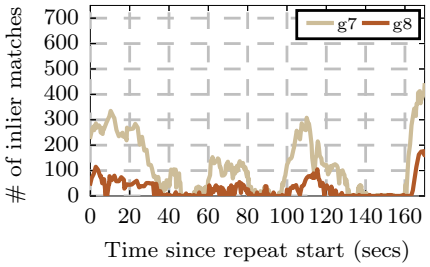
These initial results show the concept of a vision-based UAV emergency navigation system is feasible using VT&R as a basis. This first prototype has generated valuable lessons and highlighted some significant challenges for future research, as highlighted below.



(a) Temporally close experiments.



(b) Temporally close experiments, diff. days.



(c) Temporally distant flight experiences.

Fig. 7: Inlier counts for each keyframe during the repeat phase for the three grouped sets of experiments: (a) g0-g3, (b) g4-g6, (c) g7-g8.

First, the localization performance (Figure 6) when repeating over a path deteriorates an order of magnitude faster than experienced on ground vehicles [23]. Results from Paton *et al.* showed that with colour-constant imagery, strong performance in localisation during repeat was possible seven hours after the initial teach. In the airborne case, two hours saw significant loss of accuracy and reliability. This is due to a number of factors: 1) A reduced perspective constraint due to unrestricted paths. When localising, minimising the perspective change from the live to the map view enhances reliability in matching descriptor-based features. For ground vehicles, the restricted paths and locally 2D surface makes this an easier task. In the air, an un-closed control loop, unrestricted paths and 6-DoF motion mean perspective can significantly change from teach to repeat. 2) Significant shadow dependent appearance change. In the airborne perspective, particularly over forests (which have high depth variation), shadows move rapidly and the 'opposition effect' (bright halos seen around an object's shadow when illuminated directly from behind) means that features are generally tracked for shorter distances.

This latter potential cause is highlighted by certain segments of the flight path. During the first and last segments of flight (seconds 0-30 and 170-180), the imagery consists of grass and exposed regolith, which tends to show better invariance to appearance change than areas that cover forest (seconds 30-60, and 140-170). The appearance change of these areas for different experiments are highlighted in Figure 8. Contributing to the significant appearance change seen in forests are the rapidly

moving shadows generated in small regions between tall trees, and less robust descriptors generated from typically smaller octaves due to significant fine detail.

Apart from these application specific challenges, the general use of feature descriptors presents the same limitations as that for ground vehicles: rapid appearance change due to cloud shadowing or featureless environments will reduce localisation performance. We have addressed these challenges through the development of colour-constant imagery [23] and Multi-Experience Localisation (MEL) [24] (see below). However, the airborne case is less strictly reliant on continuous localisation as the return trajectory does not need to be strictly the same. We expect long periods of dead-reckoning where the aforementioned factors cause localisation failure, and during the first stage of an emergency return (the turn-around). Our tests show a VO translational error of approximately 1% over the trajectories tested, in line with current state-of-the-art. We have tested with other feature types such as Oriented FAST and Rotated BRIEF (ORB), but have seen similar localisation performance.

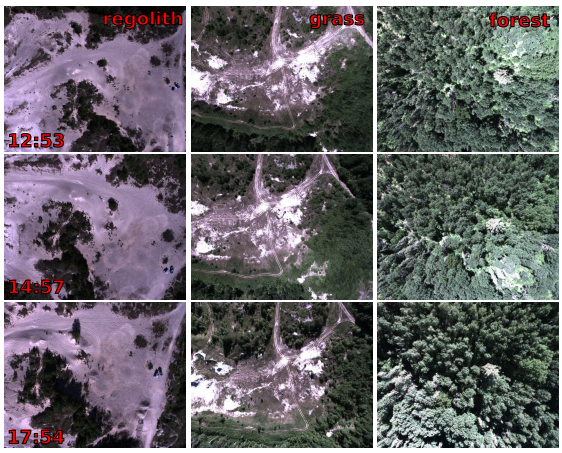


Fig. 8: Sample images from differing regions of the dataset during experiment g0 (top row), experiment g8 (centre row) and experiment g9 (bottom row). Exposed regolith (left column, approx. 70s since start of repeat) and grass (centre colum, 20s since start of repeat) are significantly more robust to appearance change than forest (right column, 40s since start of repeat). Note change in shadow and movement of halo due to the opposition effect in the right-hand column (identify the tall tree in centre right of image to assist in recognising scene).

The second major lesson is reflected in the shortage of results that leverage multiple experiences, as demonstrated for ground vehicles in [24], which is a current major focus for our lab. While not strictly required for emergency return, a multi-experience framework would remain useful in applications that require repeat trajectories (such as deliveries) in GPS-denied or GPS-intermittent areas. Given rapid appearance change in the airborne case, the need for accurately timed bridging experiences is critical, and effectively requires continuous flights with 10-15 minutes delay in order to generate enough inlier matches to successfully estimate pose relative to the original path. This is logistically difficult to implement feasibly, so such investigations are left to future work.

Since the primary application is emergency return, the algorithm will be further developed in conjuntion with a path-tracking controller suited to guidance of a fixed-wing aircraft. This will build on previous work for ground vehicles [22]. To improve localisation performance during large temporal differences, we are exploring techniques to learn place-specific binary descriptors that are more invariant to appearance change and localisation on data captured from differing sensors.

7 Conclusions

This paper presented the application of a monocular VT&R localisation engine on an outdoor, fixed-wing UAV. A key contribution is the demonstration of localisation using only a single camera in a configuration as-yet untested outdoors. Through an analysis of localisation performance and estimated uncertainty, we have shown that our algorithm is able to provide metric localisation to a privileged experience during a single flight within the capabilities of the PrecisionHawk Lancaster, such that it can be used for emergency return. Performance was also evaluated with increasing temporal difference, showing the current limitations of the algorithm given significant and rapid appearance change.

Acknowledgements Thanks to PrecisionHawk, MITACS, and the NCFRN for project funding, Ethier Sand and Gravel for property access to gather data, and Haowei Zhang for logistical support and data processing.

References

[1] Markus W. Achtelik, Michael C. Achtelik, Stephan M Weiss, and Roland Siegwart. Onboard IMU and Monocular Vision Based Control for MAVs in Unknown In- and Outdoor Environments. In *International Conference on Robotics and Automation (ICRA)*, pages 3056–3063. IEEE, may 2011.

[2] S Anderson and Timothy D. Barfoot. Full STEAM ahead: Exactly sparse gaussian process regression for batch continuous-time trajectory estimation on SE(3). In *Intelligent Robots and Systems (IROS)*, pages 157–164, sep 2015.

[3] Timothy D. Barfoot and Paul T. Furgale. Associating Uncertainty With Three-Dimensional Poses for Use in Estimation Problems. *IEEE Transactions on Robotics*, 30(3):679–693, 2014.

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded Up Robust Features. In *European Conference on Computer Vision*, 2006.

[5] Adam Bry, Abraham Bachrach, and Nicholas Roy. State Estimation for Aggressive Flight in GPS-Denied Environments Using Onboard Sensing. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2012.

[6] F. Caballero, L. Merino, J. Ferruz, and A. Ollero. Vision-Based Odometry and SLAM for Medium and High Altitude Flying UAVs. *Journal of Intelligent and Robotic Systems*, 54 (1-3):137–161, jul 2008.

[7] Lee E Clement, Jonathan Kelly, and Timothy D. Barfoot. Monocular Visual Teach and Repeat Aided by Local Ground Planarity. In David Wettergreen and Timothy D. Barfoot, editors, *Field and Service Robotics*, chapter VI, pages 547–561. Springer International Publishing, Toronto, 2015.

[8] Gerald L Dillingham. Unmanned Aircraft Systems: Continued Coordination, Operational Data, and Performance Standards Needed to Guide Research and Development. Technical report, U.S. Government Accountability Office, 2013.

[9] Zheng Fang, Shichao Yang, Sezal Jain, Geetesh Dubey, and Silvio Maeta. Robust Autonomous Flight in Constrained and Visually Degraded Environments. In David Wettergreen and Timothy D. Barfoot, editors, *Field and Service Robotics*, chapter IV, pages 411–425. Springer International Publishing, Toronto, 2015.

[10] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO : Fast Semi-Direct Monocular Visual Odometry. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2014.

[11] Christian Forster, Matthias Faessler, Flavio Fontana, Manuel Werlberger, and Davide Scaramuzza. Continuous On-Board Monocular-Vision based Elevation Mapping Applied to Autonomous Landing of Micro Aerial Vehicles. In *International Conference on Robotics and Automation (ICRA)*, Seattle, 2015. IEEE.

[12] Paul Furgale and Timothy D. Barfoot. Visual Teach and Repeat for Long-Range Rover Autonomy. *Journal of Field Robotics*, 27(5):534–560, 2010.

[13] Xiao Shan Gao, Xiao Rong Hou, Jianliang Tang, and Hang Fei Cheng. Complete Solution Classification for the Perspective-Three-Point Problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):930–943, 2003.

[14] T Goedemé, M Nuttin, T Tuytelaars, and L Van Gool. Omnidirectional Vision Based Topological Navigation. *International Journal of Computer Vision*, 74(3):219–236, 2007.

[15] TRANSPORTATION INFRASTRUCTURE. Vulnerability Assessment of the Transportation Infrastructure Relying on the Global Positioning System. Technical report, Center, John A. Volpe National Transportation Systems, 2001.

[16] Jonathan Kelly and GS Sukhatme. An Experimental Study of Aerial Stereo Visual Odometry. In *Symposium on Intelligent Autonomous Vehicles*, pages 1–6, 2007.

[17] Georg Klein and David Murray. Parallel Tracking and Mapping for Small AR Workspaces. In *6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 1–10. IEEE, nov 2007.

[18] L Majdik, Yves Albers-schoenberg, and Davide Scaramuzza. MAV Urban Localization from Google Street View Data. In *Intelligent Robots and Systems (IROS)*, pages 3979–3986. IEEE, 2013.

[19] Colin McManus, Paul Furgale, and Timothy D. Barfoot. Towards Appearance-Based Methods for LiDAR Sensors. In *International Conference on Robotics and Automation (ICRA)*, pages 1930–1935. IEEE, 2011.

[20] Philipp Oettershagen, Thomas Stastny, Thomas Mantel, Amir Melzer, Pascal Gohl, Gabriel Agamennoni, Kostas Alexis, and Roland Siegwart. Long-Endurance Sensing and Mapping using a Hand-launchable Solar-powered UAV. In David Wettergreen and Timothy D. Barfoot, editors, *Field and Service Robotics*, chapter IV, pages 441–454. Springer International Publishing, Toronto, 2015.

[21] Chris J. Ostafew, Angela P. Schoellig, and Timothy D. Barfoot. Visual Teach and Repeat, Repeat, Repeat: Iterative Learning Control to Improve Mobile Robot Path Tracking in Challenging Outdoor Environments. In *Intelligent Robots and Systems (IROS)*, pages 176–181. IEEE, 2013.

[22] Chris J Ostafew, Angela P. Schoellig, and Timothy D. Barfoot. Learning-Based Nonlinear Model Predictive Control to Improve Vision-Based Mobile Robot Path-Tracking in Challenging Outdoor Environments. In *International Conference on Robotics and Automation (ICRA)*, pages 4029–4036. IEEE, 2014.

[23] Michael Paton, Kirk Mactavish, Chris J Ostafew, and Timothy D. Barfoot. It's Not Easy Seeing Green: Lighting-Resistant Stereo Visual Teach & Repeat Using Color-Constant Images. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2015.

[24] Michael Paton, Kirk Mactavish, Michael Warren, and Timothy D. Barfoot. Bridging the Appearance Gap : Multi-Experience Localization for Long-Term Visual Teach and Repeat. In *Intelligent Robots and Systems (IROS)*, 2016.

[25] Andreas Pfrunder, Angela P. Schoellig, and Timothy D. Barfoot. A Proof-of-Concept Demonstration of Visual Teach and Repeat on a Quadrocopter Using an Altitude Sensor and a Monocular Camera. In *Conference on Computer and Robot Vision (CRV)*, pages 238–245, 2014.

[26] Shaojie Shen, Nathan Michael, and Vijay Kumar. Tightly-Coupled Monocular Visual-Inertial Fusion for Autonomous Flight of Rotorcraft MAVs. In *International Conference on Robotics and Automation (ICRA)*, pages 5303–5310, Seattle, 2015. IEEE.

[27] Phil Torr. An Assessment of Information Criteria for Motion Model Selection. In *Computer Vision and Pattern Recognition*, pages 47–52, San Juan, 1997. IEEE.

[28] Phil Torr. MLESAC: A New Robust Estimator with Application to Estimating Image Geometry. *Computer Vision and Image Understanding*, 78(1):138–156, apr 2000.

[29] Stephan M Weiss, Davide Scaramuzza, and Roland Siegwart. Monocular SLAM Based Navigation for Autonomous Micro Helicopters in GPS Denied Environments. *Journal of Field Robotics*, 28(6):854–874, 2011.