Robust Monocular Visual Teach and Repeat Aided by Local Ground Planarity and Color-Constant Imagery

Lee Clement
Institute for Aerospace Studies
University of Toronto
Toronto, ON M3H 5T6
lee.clement@mail.utoronto.ca

Jonathan Kelly
Institute for Aerospace Studies
University of Toronto
Toronto, ON M3H 5T6
jkelly@utias.utoronto.ca

Timothy D. Barfoot
Institute for Aerospace Studies
University of Toronto
Toronto, ON M3H 5T6
tim.barfoot@utoronto.ca

Abstract

Visual Teach and Repeat (VT&R) allows an autonomous vehicle to accurately repeat a previously traversed route using only vision sensors. Most VT&R systems rely on natively 3D sensors such as stereo cameras for mapping and localization, but many existing mobile robots are equipped with only 2D monocular vision, typically for teleoperation. In this paper, we extend VT&R to the most basic sensor configuration – a single monocular camera. We show that kilometer-scale route repetition can be achieved with centimeter-level accuracy by approximating the local ground surface near the vehicle as a plane with some uncertainty. This allows our system to recover absolute scale from the known position and orientation of the camera relative to the vehicle, which simplifies threshold-based outlier rejection and the estimation and control of lateral path-tracking error – essential components of high-accuracy route repetition. We enhance the robustness of our monocular VT&R system to common failure cases through the use of color-constant imagery, which provides it with a degree of resistance to lighting changes and moving shadows where keypoint matching on standard grey images tends to struggle. Through extensive testing on a combined 30 km of autonomous navigation data collected on multiple vehicles in a variety of highly non-planar terrestrial and planetary-analogue environments, we demonstrate that our system is capable of achieving route-repetition accuracy on par with its stereo counterpart, with only a modest trade-off in robustness.

1 Introduction

For many mobile robotic applications including mining, surveillance, and search and rescue, it is essential for the robot to be able to retrace its path or to traverse the same path accurately and reliably over long periods of time. For repetitive navigation tasks like these, Visual Teach and Repeat (VT&R) has proven to be an effective technology for enabling mobile robots to repeat a previously-traversed route with high accuracy. Indeed, VT&R has already found applications in autonomous tramming for mining operations

(Marshall et al., 2008) and exploration missions in planetary-analogue environments (Furgale and Barfoot, 2010).

A typical VT&R system operates in two phases: a *teach pass*, and a *repeat pass*. In the teach pass, a human operator drives the vehicle along a desired route while a vision sensor records imagery from which a map of the route can be reconstructed. In subsequent repeat passes, the system localizes against the reconstructed map and repeats the route autonomously, substituting dead-reckoned motion estimates from visual odometry (VO) when map-based localization is unavailable (Furgale and Barfoot, 2010).

Of central importance to VT&R is the selection of a map representation. On the one hand, the map may be a purely topological network of reference images or keyframes, where the navigation goal is to match the live image to a target image using a visual-homing procedure. While purely topological maps have been used with monocular (Matsumoto et al., 1996; Ohno et al., 1996; Jones et al., 1997; Goedemé et al., 2007), stereo (Matsumoto et al., 2000), and omnidirectional vision (Tang and Yuta, 2001; Argyros et al., 2005; Remazeilles et al., 2006) for teach-and-repeat navigation, visual-homing systems like these are restricted to purely heading-based control, which only loosely bounds lateral path-tracking error. On the other hand, the map may be purely metric (Baumgartner and Skaar, 1994; Kidono et al., 2002; Royer et al., 2007), which enables lateral path-tracking error to be explicitly controlled, leading to more reliably accurate route repetition. However, the large, globally consistent metric maps required to traverse long routes are prohibitively expensive to create online. Topometric maps (Simhon and Dudek, 1998; Marshall et al., 2008; Zhang and Kleeman, 2009; Furgale and Barfoot, 2010) combine the virtues of topological and metric representations by allowing the system to navigate on a network of topologically connected metric submaps rather than requiring global metric consistency. This has the effect of decoupling map size from path length while still retaining metric information where it is needed.

The choice of sensor is also important to VT&R. Furgale and Barfoot (2010) showed that a stereo camera is an effective sensor choice in a topometric VT&R system. Indeed, theirs was the first system capable of autonomously repeating multi-kilometer routes in unstructured outdoor terrain using only a stereo camera. Their system has since been extended to other natively 3D sensor configurations including appearance-based lidar (McManus et al., 2013), multiple stereo cameras (Paton et al., 2015b), and RGB-D cameras. Recently, Clement et al. (2015) investigated the use of a much simpler sensor configuration – a single 2D monocular camera – in a VT&R system, with 3D information inferred from approximations of local scene geometry and the known position and orientation of the camera relative to the vehicle.

A monocular VT&R system is particularly valuable in that it enables countless monocular robots (i.e., robots equipped with a single monocular camera) already deployed in the real world to perform previously impossible navigation tasks. Commercially available monocular robots can be found in such application domains as bomb disposal, aerial surveillance, and telepresence. Examples of monocular robots can also be found in search and rescue operations, mining, construction, and personal assistive robotics, where they are equipped with monocular vision mainly for teleoperation. By enabling such robots to autonomously navigate previously-traversed routes via a simple software upgrade, we can enhance their functionality while avoiding the potentially costly process of retrofitting them with additional sensors.

Several techniques exist for accomplishing online 3D simultaneous localization and mapping (SLAM) with monocular vision. Traditionally, these techniques have been based on sparse keypoint detection and tracking, and include filter-based approaches (Eade and Drummond, 2006; Davison et al., 2007) as well as online batch techniques that make use of local bundle adjustment to limit the computational cost of the problem (Klein and Murray, 2007; Zhao et al., 2010; Holmes and Murray, 2013). Recently, dense photometric approaches to monocular SLAM have been developed that operate directly on per-pixel intensity measurements to produce maps with a higher spatial density than is typically achievable using sparse keypoint-based methods (Newcombe et al., 2011; Engel et al., 2014; Pizzoli et al., 2014). While the use of dense visual SLAM in a VT&R system presents an interesting avenue of research, in this work we follow the tried-and-true keypoint-tracking paradigm.

Regardless of density and representation, monocular SLAM algorithms can produce accurate 3D maps only up to an unknown scale factor. This scale ambiguity complicates threshold-based outlier rejection schemes such as Random Sample Consensus (RANSAC) (Fischler and Bolles, 1981), as well as the estimation and control of lateral path-tracking error during the repeat pass, both of which are essential for high-accuracy path-following. Furthermore, monocular SLAM systems are not well-suited to land vehicles with a forward-facing camera, such as those used in our experiments, since it can be difficult to obtain measurements with sufficient disparity to triangulate keypoints accurately. While pointing the camera sideways could solve this problem, it would do so at the cost of making teleoperation and obstacle detection more difficult for the types of robots we envision.

This paper builds on the work of our recent conference paper (Clement et al., 2015), in which we proposed to represent the map as a manifold of locally planar ground surfaces with a representation of surface uncertainty, rather than using a full monocular SLAM system to create a globally consistent map. This provides a simple means of generating local metric information from the known height and orientation of the camera on the vehicle, without requiring any specific camera motion. Similar techniques have succeeded in computing VO with a monocular camera using both sparse feature tracking (Choi et al., 2011; Zhang et al., 2012; Farraj and Asmar, 2013) and dense image alignment (Lovegrove et al., 2011; Zienkiewicz and Davison, 2014), but our monocular VT&R system was the first to use such techniques for mapping and localization inside a control loop, achieving route-repetition accuracy on par with an equivalent stereo system (Furgale and Barfoot, 2010) and an autonomy rate of 99.4% on 4.3 km of autonomous navigation.

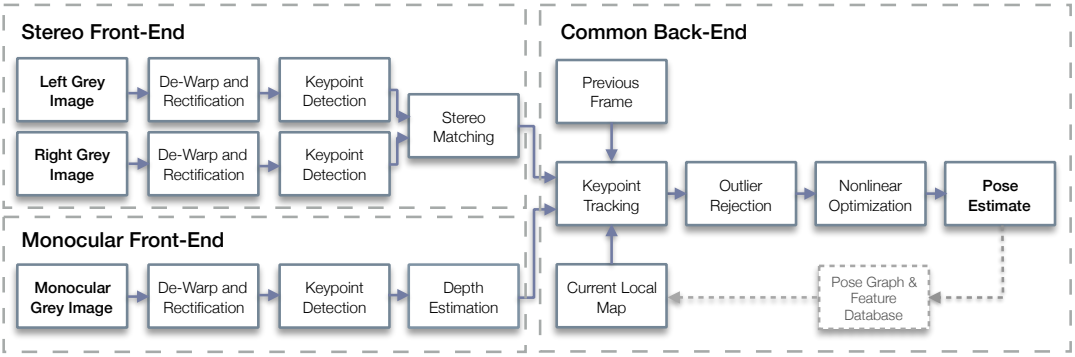
In this paper, we substantially extend our previous work by incorporating color-constant imagery (Ratnasingam and Collins, 2010; Paton et al., 2015a) in our monocular localization pipeline to improve its robustness to shadows and changing lighting conditions. In contrast to appearance-based lidar (McManus et al., 2013), which achieves lighting-resistant localization using range and intensity images from a 3D lidar sensor, our system does so using physics-based transformations of monocular RGB images based on assumptions about scene lighting and the imaging sensor. The difference is essentially one of *active* versus *passive* sensing. Furthermore, our system does not rely on native 3D information such as one might obtain from lidar or a stereo camera. Rather, we use simple assumptions about local scene geometry and the known position and orientation of the camera relative to the vehicle to obtain approximate local 3D information.

While the use of color-constant imagery has already been shown to improve precision/recall performance on visual place recognition tasks (Corke et al., 2013), as well as stereo localization quality in the presence of shadows and changing lighting conditions (McManus et al., 2014; Paton et al., 2015a), it remains to be seen whether and to what extent our monocular localization pipeline would benefit from their use. We address this gap through offline testing of a lighting-resistant version of our monocular pipeline on a further 26 km of autonomous navigation data collected at the Canadian Space Agency's Mars Emulation Terrain facility. Based on these data, we conduct a thorough comparison of our system's localization quality to that of the equivalent lighting-resistant stereo pipeline described by Paton et al. (2015a).

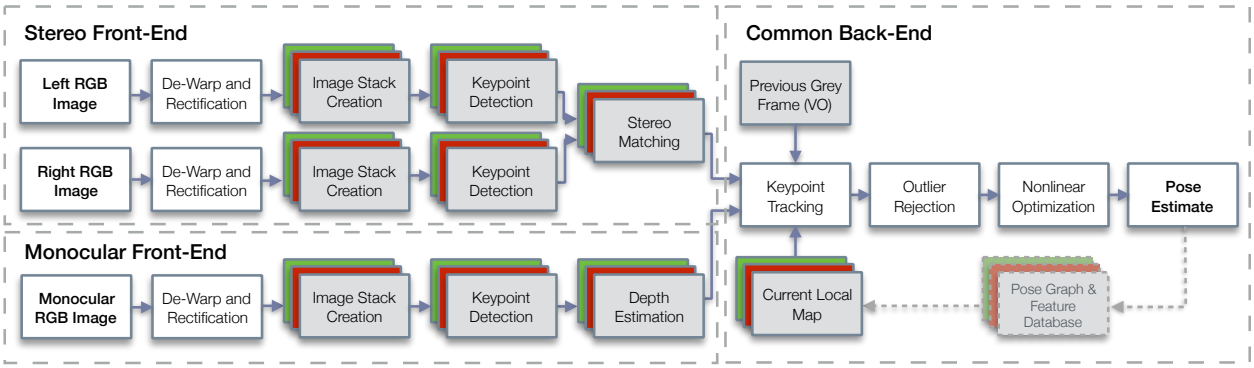
2 System Overview: Lighting-Resistant Visual Teach and Repeat

We first provide an overview of both the legacy VT&R system used in the experiments of Sections 4 and 5, and the lighting-resistant VT&R system used in the experiments of Section 5. A detailed description and analysis of the controller used in our experiments and the experiments of Paton et al. (2015a) is provided by Ostafew et al. (2015).

Both systems' localization pipelines are depicted graphically in Figure 1, in both their stereo and monocular forms. The stereo and monocular localization pipelines differ mainly in the front-end image processing used to generate 3D keypoints, while the lighting-resistant pipeline differs from the legacy pipeline in its use of color-constant images as described in Section 2.1. Each new RGB image entering the pipeline first undergoes a pre-processing step that de-warps and rectifies the image using a calibrated camera model. The system



(a) Legacy localization pipeline using grey images only (Furgale and Barfoot, 2010; Clement et al., 2015).



(b) Lighting-resistant localization pipeline using grey images combined with color-constant images (Paton et al., 2015a; this paper). The stacked grey-red-green blocks correspond to processing blocks that operate on both the grey and color-constant images.

Figure 1: The major processing blocks of the stereo and monocular localization pipelines in both their grey-only and lighting-resistant forms. The monocular pipeline shares most of the same processing blocks as its stereo counterpart, differing mainly in the front-end image processing used to generate 3D keypoints, while the lighting-resistant pipeline differs from the legacy pipeline in its use of color-constant images as described in Section 2.1. The "Current Local Map" block is only used for keypoint tracking during the repeat pass.

then generates a grey image from the green channel, and, if color-constant image creation is enabled, the two color constant images described in Section 2.1. A SURF keypoint detector (Bay et al., 2008) then generates a set of keypoints in each processed image whose 3D coordinates are estimated using either stereo triangulation or monocular depth estimation as discussed in Section 3. The system uses the grey keypoints in both the teach pass (Section 2.4) and the repeat pass (Section 2.5) to compute frame-to-frame VO using a combination of the three-point RANSAC algorithm (Fischler and Bolles, 1981) and the bundle adjustment procedure outlined in Section 2.3, with keypoint correspondences established as described in Section 2.2. Map-based localization in the repeat pass makes use of a similar procedure, but instead operates on both the grey and color-constant images over a larger window of keyframes.

2.1 Color-Constant Image Transformations

We use the same color-constant image transformations as described by Paton et al. (2015a), which are derived from the results of Ratnasingam and Collins (2010). By assuming that the environment contains a single light source that is a perfect black-body radiator, and that the response of the imaging sensor is infinitely narrow at the sensor's peak wavelength, these results show that it is possible to compute images that are resistant to shadows and changes in scene lighting. The use of such images has been shown to improve

Table 1: Parameters for generating color-constant images from a PointGrey Bumblebee XB3 stereo camera (Paton et al., 2015a)

Parameter	Description	Value
α_v	Vegetation image (F_v) α weight	0.29
β_v	Vegetation image (F_v) β weight	0.71
α_r	Rock/sand image (F_r) α weight	-1.3
β_r	Rock/sand image (F_r) β weight	2.3

(a) RGB image (b) Grey image (c) Vegetation image (F_v) (d) Rock/sand image (F_r)

Figure 2: Sample RGB image from the dataset described in Section 5, and the three corresponding processed images (Paton et al., 2015a). Note that Figures 2c and 2d have been rescaled to enhance the contrast.

precision/recall performance on place recognition tasks (Corke et al., 2013), as well as stereo localization quality in the presence of shadows and changing lighting conditions (McManus et al., 2014; Paton et al., 2015a).

Color-constant images can be generated on a per-pixel basis by taking a weighted sum of the logarithms of the sensor responses for three color channels (e.g., red, green, and blue):

$$F = \log(R_2) - \alpha \log(R_1) - \beta \log(R_3), \tag{1}$$

where R_i is the sensor response at peak wavelength λ_i, the weights (α, β) are subject to the constraints

$$\frac{1}{\lambda_2} = \frac{\alpha}{\lambda_1} + \frac{\beta}{\lambda_3} \qquad \text{and} \qquad \beta = (1 - \alpha), \tag{2}$$

and the indices $i = 1, 2, 3$ are chosen such that $\lambda_1 < \lambda_2 < \lambda_3$. We refer the interested reader to Paton et al. (2015a) for the full derivation of this result.

We generate a set of three images: a grey image from the green channel, and two color-constant images F_v and F_r with weights trained on time-lapse stereo imagery of static outdoor scenes consisting of vegetation and rocks/sand, respectively. Table 1 summarizes the parameters used to generate the two color-constant images with a PointGrey XB3 stereo camera. We refer the interested reader to Paton et al. (2015a) for a thorough discussion of the training procedure used to determine the weights. Figure 2 shows a sample RGB image from the dataset described in Section 5, along with the three corresponding processed images.

2.2 Keypoint Detection and Matching

Given the grey (green channel) image and any color-constant images that were computed, a GPU implementation of the SURF detector (Bay et al., 2008) detects a set of 2D keypoints in the processed images, and estimates their image-space covariances according to the image pyramid level at which they were detected. In order to ensure an even distribution of keypoints across the image, we divide the image evenly into an 8×6 grid and detect keypoints in each grid cell independently. The 3D coordinates and covariances of each keypoint are subsequently estimated via stereo matching and triangulation in the stereo pipeline, or by the depth estimation scheme described in Section 3 in the monocular pipeline.

For stereo matching, VO, and map-based localization, we compute correspondences between keypoints using SURF descriptor matching and a goodness test, which is a helpful criterion for rejecting ambiguous keypoint matches that are likely to be outliers. This procedure begins by ranking pairs of keypoints based on a scalar distance score $d_{i,j}$ computed from the normalized 64-element SURF descriptor vectors $\mathbf{d}_i$ and $\mathbf{d}_j$ of reference keypoint i and match candidate j, respectively:

$$d_{i,j} := 1 - \mathbf{d}_i^T \mathbf{d}_j. \tag{3}$$

We decide whether to accept the top-ranked (i.e., smallest distance) keypoint pair as a positive match by checking whether its distance score is at least a factor δ_g (i.e., the goodness ratio) smaller than the second-best keypoint match:

$$d_{i,1} < \delta_g d_{i,2}. \tag{4}$$

In all three cases (stereo matching, VO, and map-based localization), we set $\delta_g = 0.9$ since this value generally produces good results in practice.

2.3 Bundle Adjustment

We use a common bundle adjustment procedure to compute both frame-to-frame visual odometry in the teach and repeat passes, as well as to compute local metric maps from keyframes in the repeat pass. Given a set of visual observations $\mathbf{y}_{k,j}$ corresponding to an observation of keypoint $j = 1, 2, \ldots, J$ from pose $k = 1, 2, \ldots, K$ and the set of transformation matrices $\mathbf{T}_{k,0} \in SE(3)$ corresponding to vehicle pose k expressed in the base frame $\underrightarrow{\mathcal{F}}_0$ of the bundle adjustment, we can define the per-observation reprojection error,

$$\mathbf{e}_{k,j} := \mathbf{y}_{k,j} - \mathbf{g}\left(\mathbf{T}_{k,0}, \mathbf{p}_0^{j,0}\right), \tag{5}$$

where $\mathbf{p}_0^{j,0}$ is a vector from the origin of $\underrightarrow{\mathcal{F}}_0$ to keypoint j, expressed in $\underrightarrow{\mathcal{F}}_0$ (i.e., the 3D position of keypoint j expressed in the base frame), and $\mathbf{g}(\cdot)$ is the sensor model.

The objective function we seek to minimize is then

$$O := \frac{1}{2} \sum_{k=1}^{K} \sum_{j=1}^{J} \mathbf{e}_{k,j}^T \mathbf{Q}_{k,j}^{-1} \mathbf{e}_{k,j} \tag{6}$$

where $\mathbf{Q}_{k,j}$ is the covariance of $\mathbf{e}_{k,j}$. We solve this minimization problem using an iterative Gauss-Newton algorithm, taking special care with the handling of transformation matrices since $SE(3)$ is not a vector space. The details of this algorithm are beyond the scope of this paper, and we refer the interested reader to Furgale (2011) for the full solution.

2.4 Teach Pass

During the teach pass, the system continually computes frame-to-frame visual odometry (VO) to estimate the vehicle's motion. Given a pair of frames and associated keypoints, the system searches for keypoint matches based on SURF descriptors as described in Section 2.2. The set of matched keypoints then forms the input to a 3-point RANSAC algorithm (Fischler and Bolles, 1981), which rejects outlying matches and estimates the interframe vehicle motion. In our monocular VT&R system, this procedure typically rejects keypoints far from the local ground surface (e.g., walls and trees) since their motion is not adequately captured by the uncertainty model of Section 3.3. The inlying keypoint tracks and vehicle motion are then optimized using a two-frame version of the bundle adjustment procedure described in Section 2.3.

In addition to computing VO, the system also constructs a pose graph whose vertices store keyframes consisting of inlying 3D keypoints along with their covariances and SURF descriptors, and whose edges store lists of matched keypoints and 6DOF pose change estimates as computed by VO. While only the grey image

is used to compute frame-to-frame VO, any keypoints associated with color-constant images in the lighting-resistant pipeline are also stored as part of each keyframe. The system creates a new keyframe whenever a user-specified threshold on vehicle translation or rotation since the last keyframe is exceeded. Once the teach pass is complete, the pose graph serves as the map and reference path during the repeat pass.

2.5 Repeat Pass

The repeat pass begins with a manual initialization at some vertex in the pose graph, and the specification of a destination vertex. The system then reconstructs the vehicle's path from the appropriate chain of relative transformations.

During the repeat pass, the system continually computes frame-to-frame VO and simultaneously attempts to localize against the map built during the teach pass. Using the last successful localization result and the frame-to-frame VO solution as an initial guess, the system selects the nearest (in the Euclidean sense) keyframe to the current pose estimate as the active keyframe. The system then uses the bundle adjustment technique of Section 2.3 to generate a metrically-consistent local map from a user-specified number of topologically adjacent keyframes, with the active keyframe as the base frame. By projecting each 3D keypoint in the local metric map into the active keyframe, the system produces an augmented keyframe against which freshly detected keypoints may be matched. Note that the local metric map may also contain keypoints associated with color-constant images.

Keypoints in the live image are matched against keypoints in the augmented keyframe, and if the number of matches exceeds a user-specified threshold, localization is considered successful. If localization is not successful, the system will rely purely on VO until it successfully localizes against the map, or until it exceeds a user-specified distance threshold since the last successful localization against the map, in which case the system will halt the traverse, entering a search mode until it relocalizes or the operator intervenes. This behaviour prevents the system from drifting so far off the path that it will not be able to relocalize without substantial manual intervention.

Based on the localization result (whether from VO or from the map), the system estimates the vehicle's 3DOF pose relative to the projection of the reference path on the local ground plane. The error between the estimated and desired vehicle pose forms the input to a learning-based nonlinear model-predictive path-tracking controller (Ostafew et al., 2015) that uses a simple *a priori* vehicle model, a learned disturbance model, and an experience-based speed scheduler to achieve high-performance path tracking in challenging outdoor terrain. Since the focus of this paper is the localization component of VT&R, we do not discuss the details of the controller here and instead refer the interested reader to Ostafew et al. (2015) for more information. We note, however, that both our experiments and the experiments of Paton et al. (2015a) make use of the most basic form of this controller (i.e., disturbance learning and speed-scheduling disabled) so that the performance of the localization pipeline can be assessed independently of any effect it may have on the controller's experience.

3 Monocular Depth Estimation

We now describe the depth estimation scheme that we use to generate metric information from a monocular camera observing the ground. By assuming that all keypoints of interest are on the ground and approximating the local ground surface in the vicinity of the vehicle as a plane (with uncertainty), we can obtain approximate local metric information with scale determined from the known position and orientation of the camera relative to the ground. While this approximation is not *globally* valid, VT&R relies only on *local* metric information, so this type of simplification is sufficient provided the ground surface is textured enough to generate good keypoint matches.

3.1 Observation Model

Given pixel coordinates $\mathbf{y}_{k,j}$ of keypoint j observed by the camera from pose k, we want to recover the 3D coordinates $\mathbf{p}_{c_k}^{j,c_k} := \begin{bmatrix} x_{c_k} & y_{c_k} & z_{c_k} \end{bmatrix}^T$ of the keypoint, expressed in the camera frame $\underrightarrow{\mathcal{F}}_{c_k}$. In what follows, we have dropped the k subscript for notational convenience since we are only considering one camera pose.

First, let us consider how image coordinates $\mathbf{y}_j$ are formed from 3D coordinates $\mathbf{p}_c^{j,c}$. Assuming that the image has been de-warped and rectified in a pre-processing step, we can use an ideal pinhole camera model with focal lengths f_u and f_v, principal point (c_u, c_v), and camera matrix

$$\mathbf{K} = \begin{bmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{bmatrix}$$

to define the observation model $\mathbf{g}(\cdot)$:

$$\mathbf{y}_j = \mathbf{g}\left(\mathbf{p}_c^{j,c}\right) = \mathbf{h}^{-1}\left(\mathbf{K}\frac{1}{z_c}\mathbf{p}_c^{j,c}\right), \tag{7}$$

where $\mathbf{h}(\cdot)$ is the function that converts Cartesian coordinates of any dimension to their equivalent homogeneous coordinates. Note that the projective division by z_c effectively transforms $\mathbf{p}_c^{j,c}$ from a 3D Cartesian point to a 2D homogeneous point.

Unless the keypoint depth z_c is known, the function $\mathbf{g}(\cdot)$ is not uniquely invertible due to the loss of information through the projective division by z_c. In general, recovering z_c requires either multiple views of keypoint j, or assumptions about the shape of the scene. In our system, we choose to make simple assumptions about the scene in order to avoid issues with scale ambiguity and keypoint re-observation that accompany multi-view monocular reconstruction.

3.2 Locally Planar Ground Surfaces

Provided that the monocular camera is observing the ground, we can estimate the 3D coordinates of keypoints near the ground by approximating the ground surface around the vehicle as a plane. While this assumption is certainly not valid over the entirety of most routes, VT&R relies on metric information only locally, so a local approximation to the true terrain shape is sufficient.

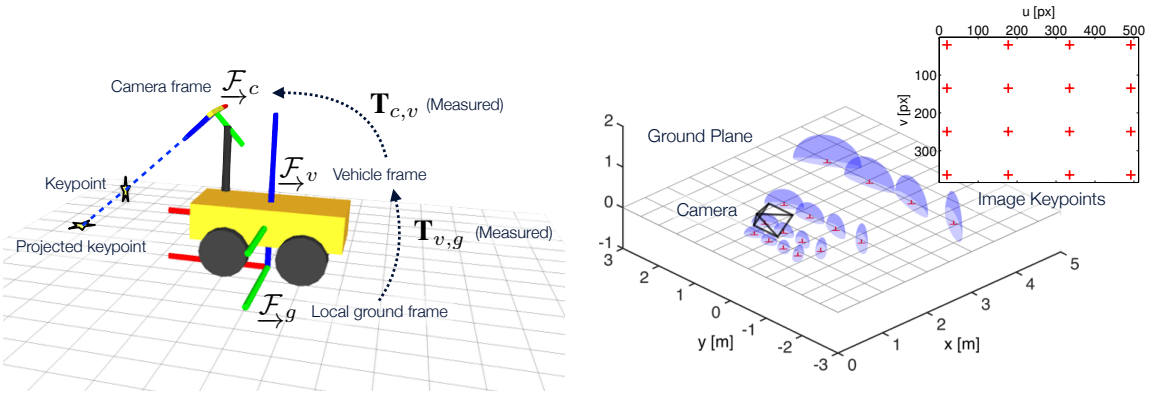
In addition to the camera-centric coordinate frame $\underrightarrow{\mathcal{F}}_c$, we define two additional coordinate frames, which are graphically depicted in Figure 3a:

$\underrightarrow{\mathcal{F}}_g$ – a local ground coordinate frame attached to the vehicle, which, for a ground vehicle, is defined such that its xy-plane contains the vehicle's footprint; and

$\underrightarrow{\mathcal{F}}_v$ – the vehicle body coordinate frame.

We then make the following assumptions about the geometry of the scene and the vehicle:

1. all keypoints lie in the xy-plane of $\underrightarrow{\mathcal{F}}_g$ (i.e., $z_g = 0$);
2. the transformation $\mathbf{T}_{c,v} \in SE(3)$ from $\underrightarrow{\mathcal{F}}_v$ to $\underrightarrow{\mathcal{F}}_c$ is known; and
3. the transformation $\mathbf{T}_{v,g} \in SE(3)$ from $\underrightarrow{\mathcal{F}}_g$ to $\underrightarrow{\mathcal{F}}_v$ is known.



(a) Coordinate frames in our monocular depth estimation scheme. The local ground frame $\mathcal{F}_g$ is defined relative to the vehicle frame $\mathcal{F}_v$ and travels with the vehicle.

(b) Evenly-spaced synthetic image keypoints (top right) and estimated 3D coordinates with 1σ uncertainty ellipses for the experimental configuration described in Section 4.

Figure 3: Geometry and uncertainty model of our monocular depth estimation scheme.

With these assumptions in mind, we can re-express our observation model as

$$\mathbf{y}_j = \mathbf{g}\left(\mathbf{p}_g^{j,g}, \mathbf{T}_{c,v}, \mathbf{T}_{v,g}\right) = \mathbf{h}^{-1}\left(\mathbf{K}\frac{1}{z_c}\underbrace{\mathbf{h}^{-1}\left(\mathbf{T}_{c,v}\mathbf{T}_{v,g}\mathbf{h}(\mathbf{p}_g^{j,g})\right)}_{\mathbf{p}_c^{j,c}}\right) \tag{8}$$

where $\mathbf{p}_g^{j,g} := \begin{bmatrix} x_g & y_g & 0 \end{bmatrix}^T$ is the 3D coordinate of keypoint j in the local ground frame. We can rearrange (8) so that the inverse observation model $\mathbf{g}^{-1}(\cdot)$ is given by

$$\mathbf{p}_c^{j,c} = \mathbf{g}^{-1}\left(\mathbf{y}_j, \mathbf{T}_{c,v}, \mathbf{T}_{v,g}\right) = z_c\mathbf{K}^{-1}\mathbf{h}(\mathbf{y}_j), \tag{9}$$

where

$$z_c\mathbf{K}^{-1}\mathbf{h}(\mathbf{y}_j) = \mathbf{T}_{c,v}\mathbf{T}_{v,g}\mathbf{h}(\mathbf{p}_g^{j,g}) \tag{10}$$

is a system of three equations in three unknowns (x_g, y_g, and z_c). Defining the (unitless) normalized image plane coordinates $\begin{bmatrix} n_x & n_y & 1 \end{bmatrix}^T := \mathbf{K}^{-1}\mathbf{h}(\mathbf{y}_j)$ and solving the third component of (10) for z_c yields

$$z_c = \frac{k_1}{k_2 + k_3 n_x + k_4 n_y}\ , \tag{11}$$

where, using $T_{m,n}$ as shorthand for the m^{th} row and n^{th} column of the transformation matrix $\mathbf{T}_{c,g} = \mathbf{T}_{c,v}\mathbf{T}_{v,g}$,

$$\begin{aligned} k_1 &= T_{1,1}\left(T_{2,2}T_{3,4} - T_{2,4}T_{3,2}\right) \\ &\quad + T_{1,2}\left(T_{2,4}T_{3,1} - T_{2,1}T_{3,4}\right) \\ &\quad + T_{1,4}\left(T_{2,1}T_{3,2} - T_{2,2}T_{3,1}\right) \end{aligned} \qquad \begin{aligned} k_2 &= T_{1,1}T_{2,2} - T_{1,2}T_{2,1} \\ k_3 &= T_{2,1}T_{3,2} - T_{2,2}T_{3,1} \\ k_4 &= T_{1,2}T_{3,1} - T_{1,1}T_{3,2}\ . \end{aligned}$$

Estimating keypoint depths in this way is useful for several reasons. First, it provides approximate local metric information without requiring multiple views with large disparity, which can be difficult to obtain on a forward-moving vehicle with a front-facing camera. Second, it resolves the scale ambiguity that makes monocular SLAM systems difficult to use for control and outlier rejection. Third, solving for keypoint depths via a chain of relative transformations provides a principled way to model the uncertainty in keypoint positions, namely as a function of the uncertainty on each transformation matrix in the chain as we show in Section 3.3. Finally, the definition of the local ground frame $\mathcal{F}_g$ allows this method to be straightforwardly extended to aerial vehicles with a downward-facing camera by estimating the transformation $\mathbf{T}_{v,g}$ using, for example, an onboard altimeter and orientation sensor.

3.3 Uncertainty Modelling

Although the local planarity assumption is a useful starting point for monocular VT&R, its rather obvious disadvantage is that it is frequently violated, especially in outdoor environments. We therefore require an appropriate model of the uncertainty in each keypoint observation $\mathbf{p}_c^{j,c}$ that accounts for deviations in ground shape. We model keypoint uncertainty by considering two important factors: uncertainty in image coordinates $\mathbf{y}_j$, and uncertainty in ground shape. In early experiments, we found that image coordinate uncertainty alone did not permit reliable keypoint tracking due to the large Mahalanobis distance between 3D keypoint estimates across multiple frames.

We model keypoint coordinates in image space as Gaussian distributions centred on $\mathbf{y}_j$ with covariance $\mathbf{R}_{\mathbf{y}_j} := \text{diag}\{\sigma_{u_j}^2, \sigma_{v_j}^2\}$. We use SURF keypoints (Bay et al., 2008) in our system and determine σ_{u_j} and σ_{v_j} from the image pyramid level at which each keypoint is detected. To incorporate uncertainty in ground shape far from the vehicle, we represent the ground-to-vehicle transformation as a Gaussian distribution on $SE(3)$ with mean $\mathbf{T}_{v,g}$ and covariance $\mathbf{R}_{\mathbf{T}_{v,g}} := \text{diag}\{\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2, \sigma_5^2, \sigma_6^2\}$, where $\sigma_1 \dots \sigma_6$ are tunable parameters corresponding to the six generators of $SE(3)$. Together these factors form an 8-dimensional Gaussian distribution with covariance $\mathbf{R}_j := \text{diag}\{\mathbf{R}_{\mathbf{y}_j}, \mathbf{R}_{\mathbf{T}_{v,g}}\}$, which we propagate via the combined Jacobian

$$\mathbf{G}_j := \left[\frac{\partial \mathbf{g}^{-1}\left(\mathbf{y}_j, \mathbf{T}_{c,v}, \mathbf{T}_{v,g}\right)}{\partial \mathbf{y}_j}\Big|_{\mathbf{y}_j, \mathbf{T}_{c,v}, \mathbf{T}_{v,g}} \quad \frac{\partial \mathbf{g}^{-1}\left(\mathbf{y}_j, \mathbf{T}_{c,v}, \mathbf{T}_{v,g}\right)}{\partial \mathbf{T}_{v,g}}\Big|_{\mathbf{y}_j, \mathbf{T}_{c,v}, \mathbf{T}_{v,g}} \right]$$

to approximate $\mathbf{p}_c^{j,c}$ as a Gaussian distribution in 3D space with covariance $\mathbf{Q}_j = \mathbf{G}_j \mathbf{R}_j \mathbf{G}_j^T$.

The individual Jacobians are given by

$$\frac{\partial \mathbf{g}^{-1}\left(\mathbf{y}_j, \mathbf{T}_{c,v}, \mathbf{T}_{v,g}\right)}{\partial \mathbf{y}_j} = \frac{z_c}{k_1} \begin{bmatrix} \left(k_1 + k_3 x_c\right)/f_u & k_4 x_c / f_v \\ k_3 y_c / f_u & \left(k_1 + k_4 y_c\right)/f_v \\ k_3 z_c / f_u & k_4 z_c / f_v \end{bmatrix} \tag{12}$$

and

$$\frac{\partial \mathbf{g}^{-1}\left(\mathbf{y}_j, \mathbf{T}_{c,v}, \mathbf{T}_{v,g}\right)}{\partial \mathbf{T}_{v,g}} = \frac{\partial \mathbf{g}^{-1}\left(\mathbf{y}_j, \mathbf{T}_{c,v}, \mathbf{T}_{v,g}\right)}{\partial \mathbf{T}_{c,g}} \frac{\partial \mathbf{T}_{c,g}}{\partial \mathbf{T}_{v,g}} = \begin{bmatrix} \mathbf{1} & (-\mathbf{p}_c^{j,c})^\wedge \end{bmatrix} \text{Ad}(\mathbf{T}_{c,v}) \tag{13}$$

where, adopting the notation of Barfoot and Furgale (2014), $\mathbf{1}$ denotes the (3×3) identity matrix, $\text{Ad}(\cdot)$ denotes the adjoint in $SE(3)$, and

$$\mathbf{p}^\wedge = \begin{bmatrix} 0 & -p_3 & p_2 \\ p_3 & 0 & -p_1 \\ -p_2 & p_1 & 0 \end{bmatrix}.$$

Figure 3b shows 1σ uncertainty ellipses for a number of evenly spaced synthetic keypoints resulting from a camera configuration similar to that used in the experiments described in Section 4. Note that we could also incorporate uncertainty in the vehicle-to-camera transformation $\mathbf{T}_{c,v}$ in this calculation, but we found that this was of limited use since ground shape uncertainty was the dominant contributor to keypoint uncertainty in our experiments.

4 UTIAS Experiments

We conducted two sets of experiments at the University of Toronto Institute for Aerospace Studies (UTIAS). The first took place outdoors on relatively flat terrain, and the second on the highly non-planar terrain of the UTIAS MarsDome indoor rover testing environment. Using grey images only, we compare the performance

Table 2: Summary of results for UTIAS experiments

				Local start time (UTC-4)			Autonomy rate [%]	
Trial	Route	Length [m]	$v_{\max}$ [m/s]	Teach	Mono	Stereo	Mono	Stereo
1	Outdoor	1370	0.6	09:56	10:35	12:08	99.71[†]	100.00
2	Outdoor	1360	0.6	11:45	12:22	13:43	99.88	100.00
3	Outdoor	1361	0.6	13:26	14:00	15:20	99.74	100.00
4	Indoor	126	0.3	13:32	13:40	14:02	96.28	100.00
5	Indoor	140	0.3	12:18	12:32	12:59	91.60	100.00

	Absolute lateral error – μ (σ) [cm]		VO matches – μ (σ)		Map matches – μ (σ)	
Trial	Mono	Stereo	Mono	Stereo	Mono	Stereo
1	1.5 (3.0)	0.68 (0.91)	260 (82)	210 (42)	150 (110)	120 (48)
2	0.19 (7.1)[‡]	0.54 (5.6)[‡]	190 (70)	190 (40)	180 (90)	86 (43)
3	0.82 (1.2)	1.4 (2.3)	210 (61)	220 (37)	240 (92)	121 (47)
4	0.65 (0.83)	0.61 (0.76)	300 (76)	250 (41)	110 (99)	89 (44)
5	0.55 (0.57)	0.57 (0.65)	310 (94)	360 (57)	160 (115)	140 (50)

	Mono	**Stereo**
Total distance driven	4298 m[†]	4357 m
Total distance autonomously traversed	99.41%	100.00%

[†] During the monocular repeat pass of Trial 1, a parked vehicle in a no-parking zone on the path forced manual driving for 59 m before successful relocalization. We exclude this segment in our analysis and report the monocular autonomy rate for Trial 1 based on a reduced path length of 1311 m.

[‡] RTK correction of our GPS data failed during the teach pass of Trial 2, so ground truth lateral path tracking error could not be computed reliably. We report the lateral path tracking error estimated by the localization pipeline as an approximation to the true value, although it tends to over-estimate these errors, leading to a larger spread in errors than we see in the ground truth data.

of our monocular VT&R system to that of the legacy stereo system (Furgale and Barfoot, 2010) over 4.3 km of autonomous navigation. We compare the two systems based on path-tracking accuracy as well as the *autonomy rate* of each traverse, which we define to be the proportion of the path, by distance, that the system drove autonomously, whether on pure VO or by localizing against the map. Table 2 reports path lengths, maximum repeat speeds ($v_{\max}$), start times, and autonomy rates for each experiment along with the mean and standard deviation of the lateral path tracking errors (from ground truth), VO keypoint matches, and map keypoint matches for each repeat pass.

4.1 Hardware

As depicted in Figure 4, our robotic platform for these experiments consisted of a four-wheeled skid-steered Clearpath Husky A200 rover equipped with a PointGrey Bumblebee XB3 stereo camera outputting 512×384 pixel grey images at 15 frames per second over a FireWire connection. A MacBook Pro 10,1 laptop running Linux and ROS (Quigley et al., 2009) interfaces with the vehicle and the stereo camera, and handles the VT&R algorithms as well as any additional payload drivers (e.g., GPS).

The camera is positioned 1.0 m above the base of the vehicle and is angled downward at 47° to the horizontal. These values were measured by hand since we do not require a precise estimate of the vehicle-to-camera transform $\mathbf{T}_{c,v}$; uncertainty in the ground-to-vehicle transform $\mathbf{T}_{v,g}$ is the dominant source of uncertainty in our system.

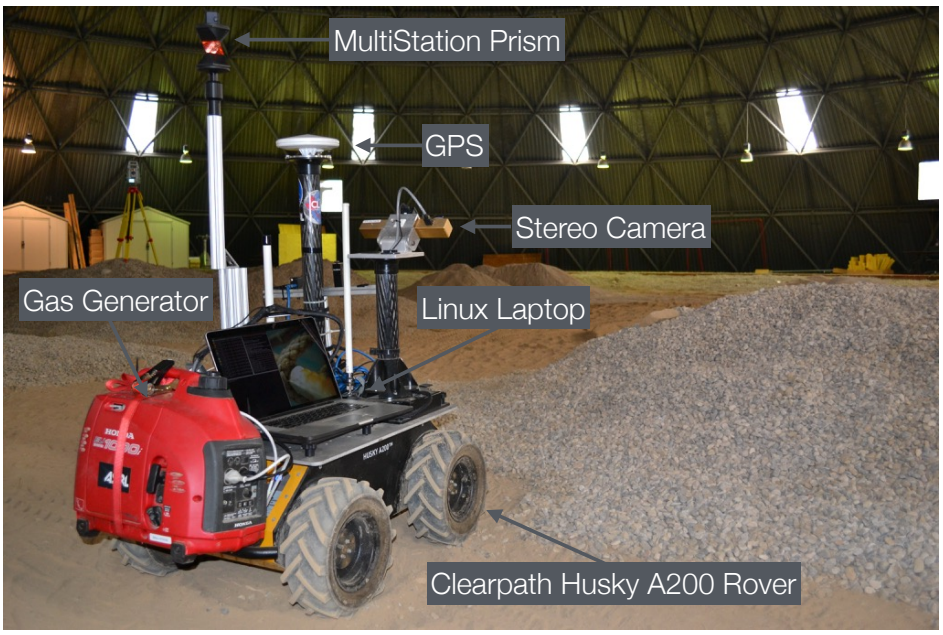


Figure 4: Our robotic platform for the UTIAS experiments was a Clearpath Husky A200 rover equipped with a PointGrey Bumblebee XB3 stereo camera, DGPS receiver, Leica Nova MS50 MultiStation prism, and a 1 kW gas generator. A MacBook Pro 10,1 laptop running Linux and ROS (Quigley et al., 2009) interfaces with the vehicle and the stereo camera, and handles the VT&R algorithms and any additional payload drivers.

4.2 Procedure

Because the intent of our experiments was to compare the accuracy and robustness of grey-only monocular and stereo VT&R in similar conditions, our experimental procedure consisted of recording stereo images during a manually-driven teach pass, and using the recorded images to teach identical routes using both the monocular and stereo pipelines. We conducted each experiment between roughly 10:00 and 14:00 when the sun was highest in the sky to minimize the effects of lighting changes and shadows. For each experiment, we repeated the route using the monocular pipeline first, using imagery from the left camera of the stereo pair only since the stereo camera coordinate frame has its origin in the left camera.

4.3 Parameter Selection and Sensitivity

Table 3 lists the important parameters used for both pipelines in these experiments. We manually tuned these parameters on separate small-scale training runs until the system achieved the desired level of performance. While a quantitative sensitivity analysis with respect to each of these parameters is infeasible, a qualitative discussion may still be valuable to researchers developing similar systems.

One of the key tradeoffs in tuning these parameters is the balance between the quantity of information available for localization and the time required to compute an accurate localization result. In particular, the number of keypoints detected and tracked, N_k, and the number of RANSAC iterations, N_r, are important considerations. With too few keypoints being inserted into the map, repeat pass localization performance declines sharply since fewer keypoints can be matched against the map. On the other hand, too many keypoints in the map results in a bundle adjustment optimization that is too slow to converge to be used for online operation. Moreover, the number of RANSAC iterations must be large enough to reliably reject

Table 3: Parameters for UTIAS experiments

Parameter	Description	Value
N_k	SURF keypoints detected and tracked	600
N_r	RANSAC iterations	400
N_m	Minimum match count for localization	10
τ	Maximum distance without localizing against the map	10 m
K	Bundle adjustment window size (keyframes)	11
δ_r	Keyframe creation threshold (translation)	25 cm
δ_θ	Keyframe creation threshold (rotation)	2.5°
$\sigma_1, \sigma_2, \sigma_3$	Ground-to-vehicle translation standard deviation†	10 cm
$\sigma_4, \sigma_5, \sigma_6$	Ground-to-vehicle rotation standard deviation†	10°
$v_{\max}$	Maximum repeat speed during data collection	See Table 2.

† These parameters were used in the monocular pipeline only.

the majority of outlier keypoint matches, and the computational effort expended per iteration scales linearly with the number of keypoint matches to evaluate.

The choice of bundle adjustment window size, K, compounds this issue, since a reasonable number of keypoint reobservations is required to obtain a good estimate of 3D keypoint positions. This is especially important to the monocular pipeline since the initial estimates of keypoint positions are more uncertain and less accurate than in the stereo pipeline. The ideal window size depends partly on keyframe creation thresholds, δ_r and δ_θ, since many features are reobserved over only a few meters before going out of view, and over an even shorter distance when the vehicle is turning. We tuned these parameters somewhat conservatively in order to ensure a good number of observations per keypoint.

The minimum match count for localization, N_m and the maximum distance without localizing against the map, τ, are important for ensuring reliable localization against the map and for preventing the vehicle from driving too far off the path due to accumulated drift in the VO estimate if map-based localization fails. While, theoretically, only three inlier keypoint matches are required to uniquely determine the vehicle's pose, in practice we require a larger number of matches to account for the fact that RANSAC is not guaranteed to identify a true set of inliers. Since grossly incorrect localization results ultimately lead to nonsensical control inputs and unpredictable vehicle behaviour, we tuned N_m rather conservatively. At the same time, we wished to prevent the vehicle from drifting so far off the path in cases of localization failure that it would be unable to relocalize without substantial manual intervention. We therefore required the vehicle to stop and enter a search mode after 10 meters, since we observed the likelihood of localization recovery to be very low beyond that point.

Finally, the parameters $\sigma_{1\ldots6}$ in the monocular pipeline, corresponding to the uncertainty in each of the six degrees of freedom in the ground-to-vehicle transformation, were tuned in response to the expected roughness of the terrain to be traversed. In contrast to a naively exploring vehicle that would need to learn these parameters online, we know the roughness of the terrain *a priori* since the route is taught by a human. For very flat surfaces, these parameters can be set quite low, perhaps to a tenth of their reported values, but for rougher terrain with more deviations from the local planarity assumption, they must be increased to account for the scene geometry. On the other hand, setting them much higher than the reported values can increase the difficulty of keypoint matching, especially in cases where the keypoints have similar descriptors and can only be distinguished in terms of their spatial location.

4.4 Outdoor Experiments

We evaluated the performance of our monocular VT&R system over three 1.4 km routes through the parking lots and driveways of UTIAS. These paths consisted mainly of flat pavement, but they also included several

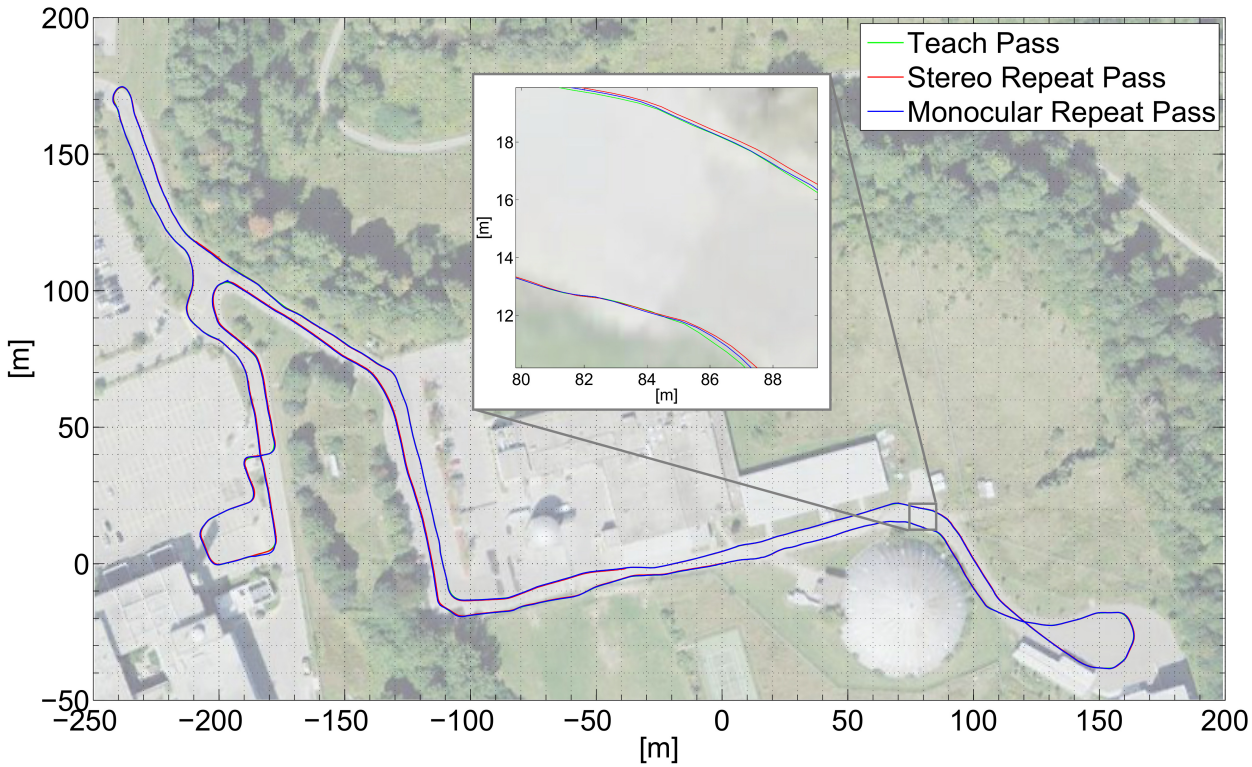
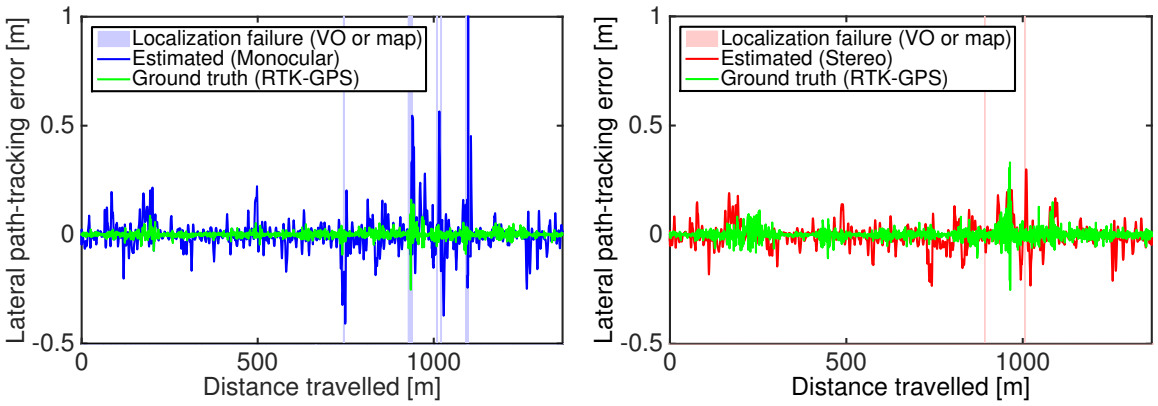


Figure 5: Comparison of RTK-corrected GPS tracks of the teach pass, stereo repeat pass, and monocular repeat pass of a 1.4 km outdoor route at the University of Toronto Institute for Aerospace Studies (Trial 3 in Table 2). The zoomed-in section highlights the centimeter-level accuracy of both pipelines. (Map data: Google, DigitalGlobe.)



(a) Monocular repeat pass

(b) Stereo repeat pass

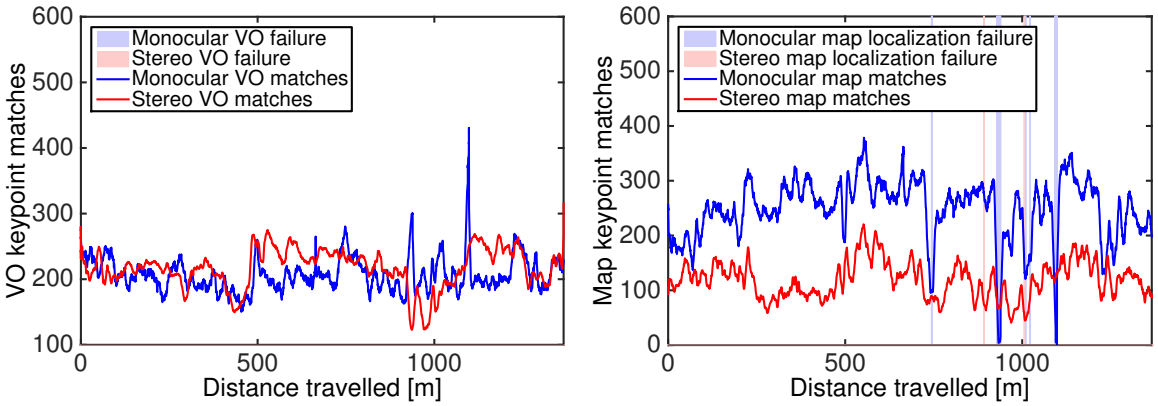
Figure 6: Estimated and measured lateral path-tracking error during the monocular and stereo repeat passes of the 1.4 km outdoor route shown in Figure 5 (Trial 3 in Table 2). GPS tracking shows that both monocular and stereo VT&R achieve centimeter-level accuracy, although estimated lateral path-tracking error (blue and red lines) tends to diverge from the true value (green lines) in cases of localization failure due to accumulated error in the VO estimate.

non-planar features such as speed bumps, side slopes, deep puddles, and rough shoulders, as well as a variety of other terrain types including gravel, sand, and grass.

We established ground truth for each teach pass and each repeat pass by equipping the rover with an Ashtech DG14 Differential GPS unit used in tandem with a second stationary DG14 unit to obtain centimeter-accuracy RTK-corrected GPS data. Figure 5 shows GPS tracks and satellite imagery of the teach and repeat passes for one of these routes.

Figure 6 shows estimated lateral path-tracking errors during the monocular and stereo repeat passes, as well as the lateral path-tracking error measured from RTK-GPS ground truth. Both pipelines achieved centimeter-level accuracy in their respective repeat passes and produced similar estimates of lateral path-tracking error when successfully localized against the map. In cases of map localization failure (i.e., when the system relied on pure frame-to-frame VO), the monocular pipeline's estimated lateral path-tracking error diverged from ground truth more quickly than that of the stereo pipeline. This means that the interframe pose estimates from monocular VO were more error-prone, which in turn led to the dead-reckoned motion estimate accumulating drift error more rapidly. This behaviour is a result of the comparatively large uncertainties associated with keypoints in the monocular pipeline. Since keypoint positions are poorly constrained by only two measurements, the two-frame bundle adjustment is less likely to converge to an accurate solution in the monocular pipeline than in the stereo pipeline. Note, however, that the ground truth lateral path tracking error (the green lines in Figure 6) has 3σ bounds of 3.6 cm for the monocular pipeline and 6.9 cm for the stereo pipeline, indicating that both systems stayed well within 10 cm of the taught path for the vast majority of the traverse (see Table 2).

Figure 7 compares the number of successful keypoint matches for frame-to-frame VO and map-based localization for monocular and stereo VT&R. Both pipelines track similar numbers of keypoints from frame to frame, but as shown in Table 2, the monocular pipeline generally tracks twice as many map keypoints as its stereo counterpart and the spread of map match counts is much wider. This behaviour is most likely due to a combination of two factors: keypoint rejection during the stereo matching and triangulation step (from which the monocular pipeline is necessarily exempt); and incorrect data association during local map construction in the monocular pipeline, which is a result of the comparatively large positional uncertainties of distant keypoints.



(a) VO keypoint matches

(b) Map keypoint matches

Figure 7: Keypoint matches during the monocular and stereo repeat passes of the 1.4 km outdoor route shown in Figure 5 (Trial 3 in Table 2), with localization failures highlighted. A localization failure is defined as fewer than 10 keypoint matches. There were no VO failures during either repeat pass. For clarity, we have applied a 20-point sliding-window mean filter to the raw data.

4.5 MarsDome Experiments

We also conducted similar experiments in the UTIAS MarsDome indoor rover testing facility, a 1,100 m^2 fully enclosed testing facility whose interior has been modified using various materials to create hills, valleys, and other obstacles (Figure 8). Although these routes were only one-tenth of the length of the outdoor routes, they covered substantially more difficult terrain.

Since the MarsDome is an enclosed facility, GPS tracking is not available and we instead made use of a Leica Nova MS50 MultiStation to track the position of the rover with millimeter-scale accuracy. Figure 9 shows MultiStation tracks of the teach pass and the two repeat passes of one MarsDome route. Figure 9 also highlights several highly non-planar features along the route such as side slopes, large bumps, valleys, ramps, and hills.

In spite of the difficulty of the terrain and clear violations of the planarity assumption, even locally, Figure 10 shows that both monocular and stereo VT&R achieved centimeter-level lateral path-tracking error. Again, note that although the monocular pipeline's estimated lateral path-tracking error diverged significantly from ground-truth during localization failures, the MultiStation tracks show that the vehicle remained within a few centimeters of the taught path throughout the traverse. Indeed, Table 2 shows that the 3σ bounds on the ground truth lateral path tracking error (the green lines in Figure 10) are 1.7 cm for the monocular pipeline and 2.0 cm for the stereo pipeline.

Figure 11 shows VO and map keypoint matches for both repeat passes. The monocular pipeline suffered map localization failures more frequently than the stereo pipeline, with the worst failures occurring in the valley and hill regions (see Figure 9) where the lighting was especially poor. Poor lighting led to increased motion blur (see Figure 12b) and unreliable keypoint matching, which was more problematic for the monocular pipeline due to greater uncertainty in keypoint positions compared to the stereo pipeline. Both failures necessitated manual intervention over a few meters, but the system successfully relocalized once the lighting improved. Similarly to the outdoor experiments, Table 2 shows that the monocular pipeline tracked more keypoints than the stereo pipeline on average, but also had much more variable tracking performance.

Figure 8: The UTIAS MarsDome is a 1,100 m^2 fully enclosed testing facility whose interior has been modified using various materials to create hills, valleys, and other obstacles.

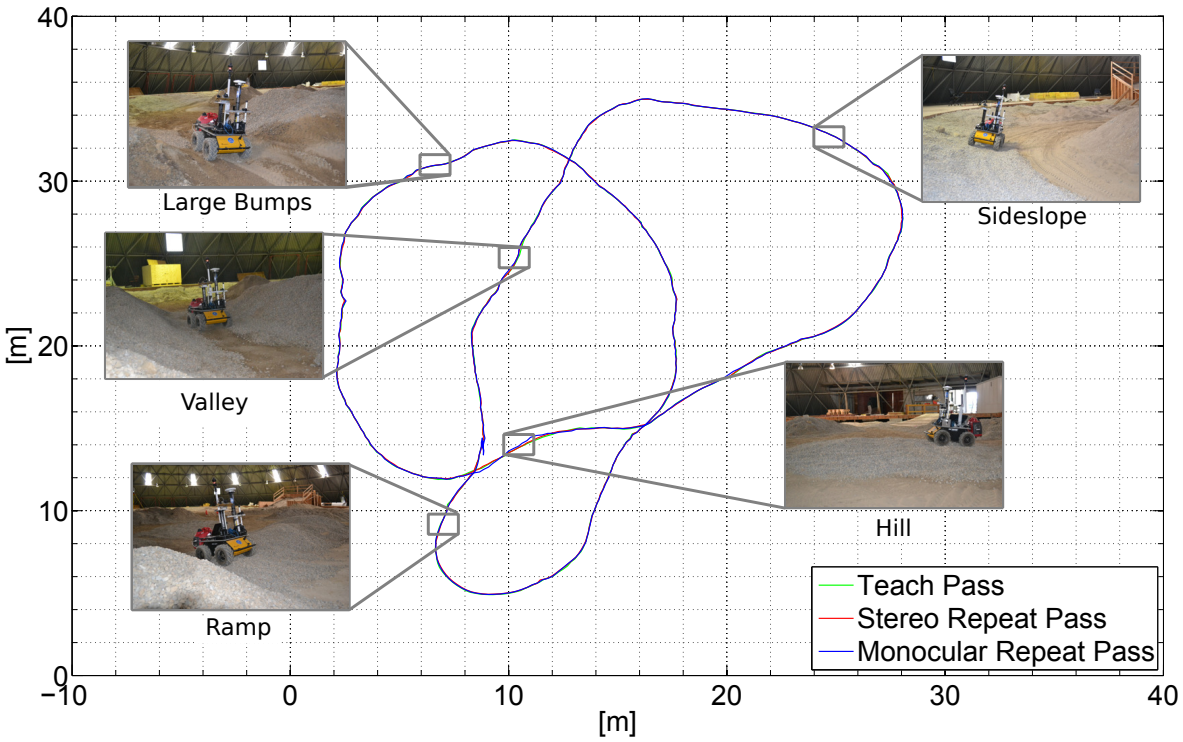
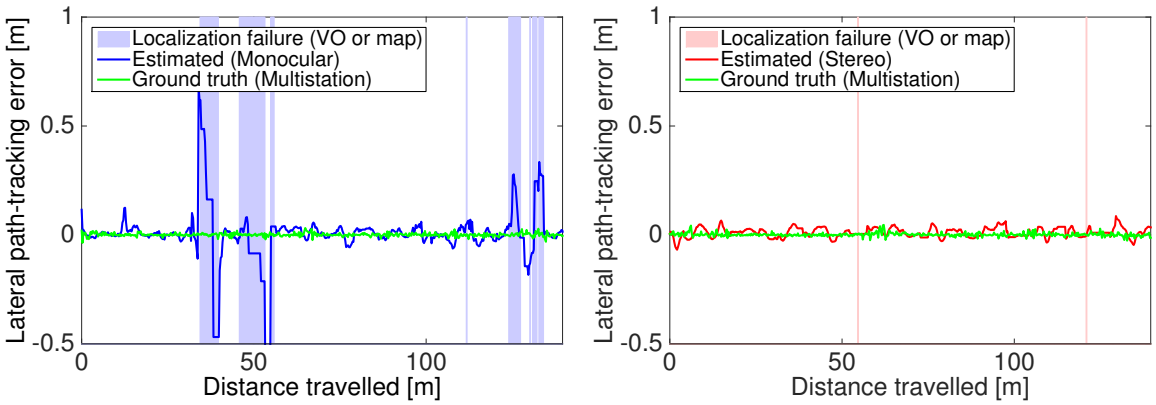


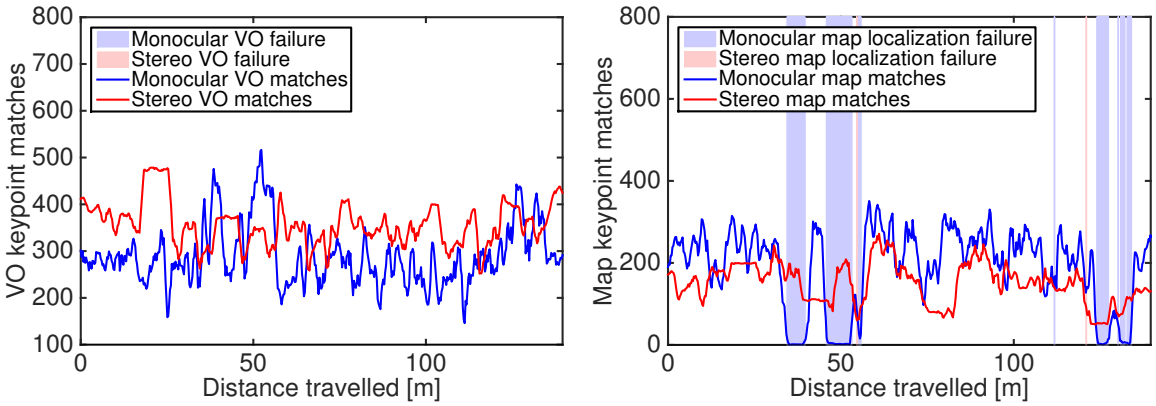
Figure 9: Comparison of MultiStation tracks of the teach pass, stereo repeat pass, and monocular repeat pass of a 140 m MarsDome route (Trial 5 in Table 2), with some interesting segments highlighted.



(a) Monocular repeat pass

(b) Stereo repeat pass

Figure 10: Estimated and measured lateral path-tracking error during the monocular and stereo repeat passes of the 140 m indoor route shown in Figure 9 (Trial 5 in Table 2). MultiStation tracking shows that both monocular and stereo VT&R achieve centimeter-level accuracy in highly non-planar terrain, although estimated lateral path-tracking error (blue and red lines) tends to diverge from the true value (green lines) in cases of localization failure due to accumulated error in the VO estimate. Note the difference in scale between the two plots.



(a) VO keypoint matches

(b) Map keypoint matches

Figure 11: Keypoint matches during the monocular and stereo repeat passes of the 140 m indoor route shown in Figure 9 (Trial 5 in Table 2), with localization failures highlighted. A localization failure is defined as less than 10 keypoint matches. There were no VO failures during either repeat pass. For clarity, we have applied a 5-point sliding-window mean filter to the raw data.

(a) Self-similar terrain (b) Motion blur

Figure 12: The most common causes of localization failure in these experiments were highly self-similar terrain and motion blur. Neither stereo nor monocular VT&R is immune to these conditions, but their effects were exacerbated by high spatial uncertainty in the monocular case.

4.6 Limitations

Localization failure in the monocular pipeline was chiefly due to difficulty in finding good keypoint matches between the map and the live image, especially in regions of highly self-similar terrain (e.g., Figure 12a) and in poorly-lit regions where motion blur is noticeable (e.g., Figure 12b). While the stereo pipeline is not immune to these effects, it is more resistant to them than the monocular pipeline since keypoint uncertainties are much larger in the monocular pipeline. With fewer correctly associated measurements, the bundle adjustment procedure will not maximally constrain keypoint positions, reducing the reliability of the map and increasing the risk of localization failures. Indeed, Figures 7 and 11 show that the monocular pipeline suffered more serious map localization failures than the stereo pipeline, although these were not often severe enough to force manual intervention.

5 CSA Dataset Experiments

While the experiments presented in Section 4 demonstrated that, when operating nominally, our monocular VT&R system can achieve route-repetition accuracy on par with its stereo counterpart using only standard grey images, we also saw that the monocular pipeline was more sensitive to lighting and self-similar textures and had a higher incidence of localization failure. In this section we examine the use of color-constant images (Paton et al., 2015a) to improve the monocular pipeline's robustness to lighting changes. We compare this lighting-resistant monocular pipeline to the equivalent stereo system presented by Paton et al. (2015a) through offline testing on an additional 26 km of autonomous navigation data collected at the Canadian Space Agency (CSA) Mars Emulation Terrain facility. As shown in Figure 13, this dataset consists of a 1.04 km route covering a variety of terrain types including rocks, sand, grass, gravel, and wooded areas. The route was taught in sunny conditions, and successfully repeated 25 times from sunrise to sunset over the course of four days in both sunny and cloudy conditions (Table 4).

5.1 Hardware

The robotic platform in this dataset is a Clearpath Grizzly Robotic Utility Vehicle (RUV). Like the Clearpath Husky rover described in Section 4, the Grizzly RUV is equipped with a PointGrey Bumblebee XB3 stereo camera, GPS receiver, and gas generator, as well as a suite of additional sensors that were not used in these

Table 4: Summary of CSA dataset traverses †

Traverse	Day	Local start time (UTC-4)	Lighting
Teach	1	10:50	Sunny
Repeat 1	1	11:40	Sunny
Repeat 2	1	12:53	Sunny
Repeat 3	1	13:35	Sunny
Repeat 4	1	14:00	Sunny
Repeat 5	1	16:06	Sunny
Repeat 6	1	17:27	Sunny
Repeat 7	1	18:14	Sunny
Repeat 9	1	19:29	Sunny
Repeat 10	1	20:06	Sunset
Repeat 11	2	06:20	Cloudy
Repeat 12	2	07:05	Cloudy
Repeat 13	2	08:00	Cloudy
Repeat 14	2	09:00	Cloudy
Repeat 15	2	10:00	Cloudy
Repeat 16	2	11:00	Cloudy
Repeat 17	2	12:00	Cloudy
Repeat 18	2	13:00	Cloudy
Repeat 19	2	14:00	Cloudy
Repeat 20	2	15:10	Cloudy
Repeat 21	2	16:00	Cloudy
Repeat 22	2	17:00	Cloudy
Repeat 23	2	18:00	Cloudy
Repeat 24	2	19:00	Cloudy
Repeat 25	2	20:00	Cloudy
Repeat 27	4	08:50	Sunny

† Repeat passes 8 and 26 have been omitted due to data loss and repeat pass failure during data collection, respectively.

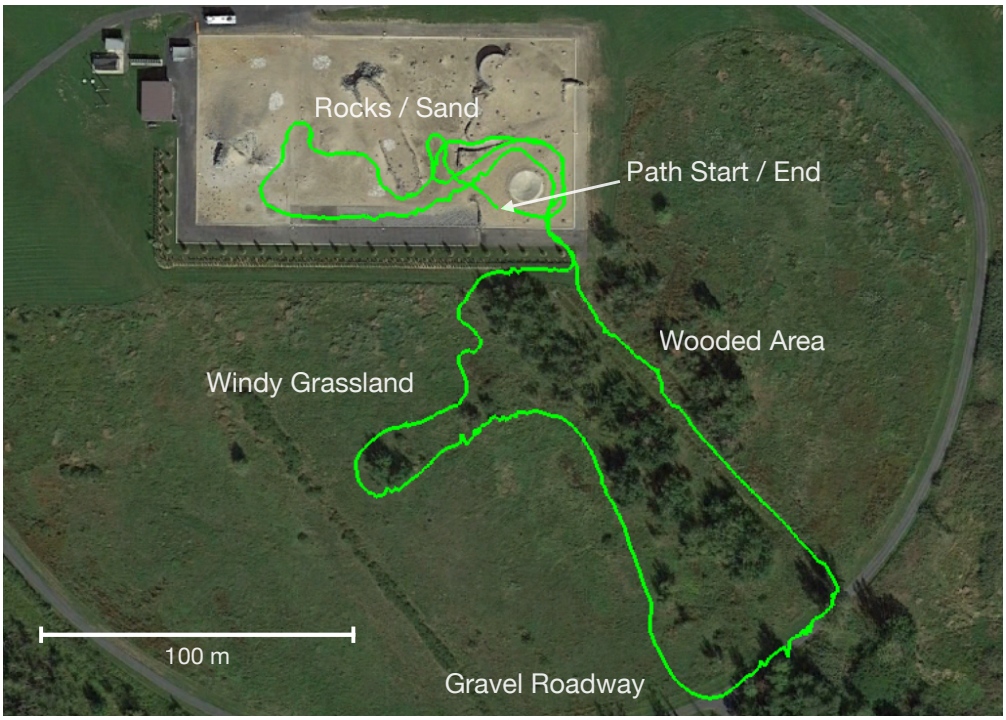


Figure 13: GPS tracks (not RTK-corrected) of the 1.04 km teach route at the Canadian Space Agency Mars Emulation Terrain facility and surrounding area. The route covers a variety of terrain types including rocks, sand, grass, gravel, and wooded areas. (Map data: Google, DigitalGlobe.)

experiments. An embedded Linux computer running ROS (Quigley et al., 2009) handles motor control and safety features, while a Lenovo W540 laptop also running Linux and ROS handles the VT&R algorithms and interfaces with the onboard computer and the stereo camera.

The camera on the Grizzly RUV is mounted 1.5 meters above the ground and is angled downwards at 29° to the horizontal, significantly higher and at a much shallower angle than on the Husky rover. As with the UTIAS experiments, these values were measured by hand.

5.2 Procedure

Since we are performing an offline analysis of teach and repeat pass data originally collected using a vehicle controlled by the lighting-resistant stereo pipeline, it is impossible to compare path tracking accuracy for the monocular and stereo pipelines as we did in Section 4. Instead, we choose to compare localization quality across pipelines in terms of the distribution of localization failures over each traverse. We discuss this comparison in more detail in Section 5.4.

We generated the data for these experiments by first training the monocular pipeline on the teach pass data using color-constant images as described in Section 2.1, then testing the pipeline on each of the repeat pass datasets. We recorded the number of keypoint matches for the grey image and each of the two color-constant images, as well as the distances driven on VO in each repeat pass. We then repeated this procedure using the grey-only monocular pipeline, as well as both the lighting-resistant and grey-only versions of the stereo pipeline.

Figure 14: The robotic platform for the CSA experiments was a Clearpath Grizzly Robotic Utility Vehicle (RUV) equipped with, among other things, a PointGrey Bumblebee XB3 stereo camera, DGPS receiver, and gas generator. An embedded Linux computer running ROS (Quigley et al., 2009) handles motor control and safety features, while a Lenovo W540 laptop handles the VT&R algorithms and interfaces with the onboard computer and the stereo camera.

Table 5: Parameters for CSA dataset experiments[‡]

Parameter	Description	Value
N_k	SURF keypoints detected and tracked	600
N_r	RANSAC iterations	400
N_m	Minimum match count for localization	10
τ	Maximum distance without localizing against the map	10 m
K	Bundle adjustment window size	11
δ_r	Keyframe creation threshold (translation)	20 cm
δ_θ	Keyframe creation threshold (rotation)	2.5°
$\sigma_1, \sigma_2, \sigma_3$	Ground-to-vehicle translation standard deviation[†]	5 cm
$\sigma_4, \sigma_5, \sigma_6$	Ground-to-vehicle rotation standard deviation[†]	10°
$v_{\max}$	Maximum repeat speed during data collection	1.0 m/s

[†] These parameters were used in the monocular pipeline only.
[‡] These are not precisely the parameters used by Paton et al. (2015a) during dataset collection. We selected these parameters for our own experiments to be consistent with the experiments of Section 4, with small empirically determined adjustments to improve localization performance on the dataset.z

5.3 Parameter Selection and Sensitivity

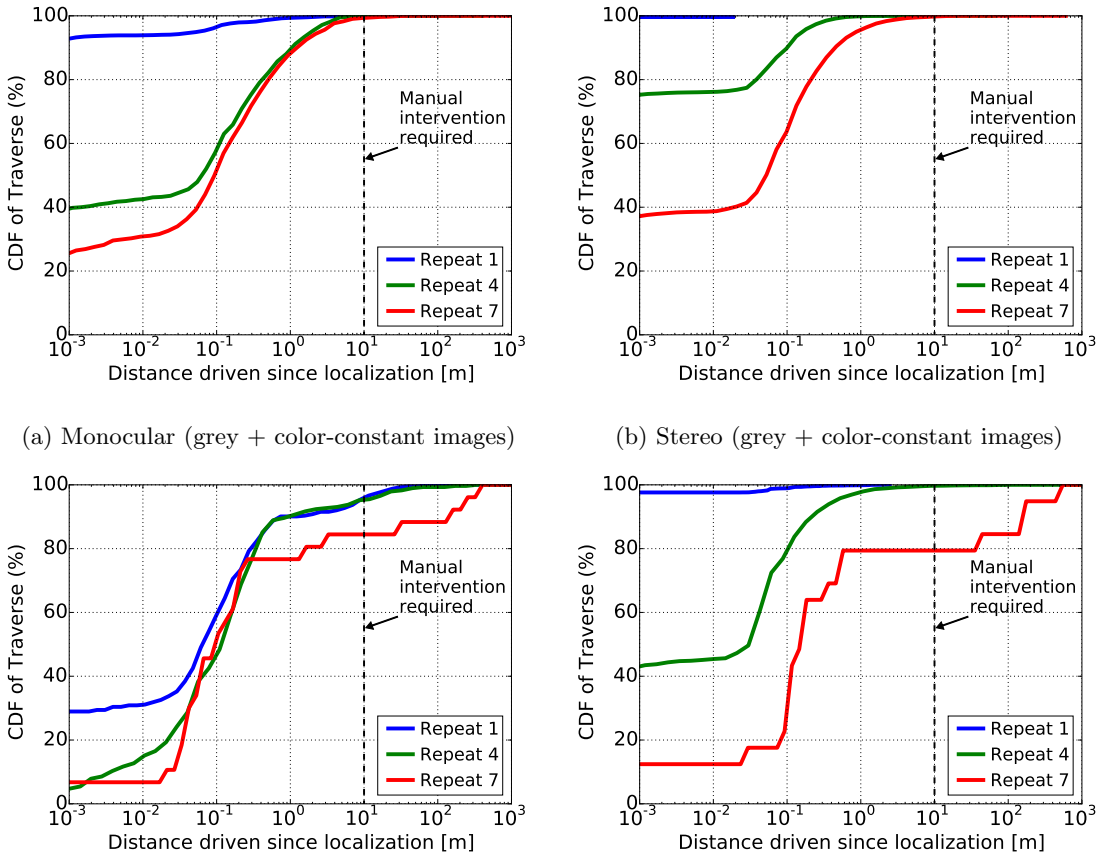
Table 5 lists the important parameters used for both pipelines in these experiments. These parameter values were chosen to be consistent with the experiments of Section 4, with small empirically determined adjustments to improve localization performance on the dataset. These adjustments were necessary mainly due to differences in camera placement on the vehicle, which had an impact on the reliability of the monocular pipeline's repeat pass localization performance. The comments of Section 4.3 pertaining to parameter tuning and the system's sensitivity to these parameters are relevant here as well.

5.4 Comparison of Localization Quality

We evaluate the quality of repeat-pass localization by examining the cumulative distribution function (CDF) of the distances over which map-based localization failed and the system drove on pure VO. This is a useful tool for assessing repeat-pass quality because it takes into account the statistical distribution of localization failures, and provides a means of predicting whether and to what extent a particular localization pipeline would have failed on a particular dataset with a particular set of parameters.

A VO CDF plot reads "for Y% of the traverse, the system drove less than X meters on VO". Intuitively, a system for which the distribution of distances driven on VO is skewed towards small (or zero) values performs better than a system for which the distribution is skewed towards larger values. This is equivalent to saying that the first system successfully localized against the map for a larger proportion of the route than the second system. An ideal repeat pass is one in which the system localized against the map 100% of the time (i.e., the distance travelled on VO is zero), so the closer the CDF is to the top-left of the plot, the better the performance of the repeat pass.

One important use of the VO CDF is to assess whether, for a given maximum distance since localizing against the map (parameter τ in Table 5), a particular repeat pass *would have* required manual intervention *if* the vehicle had been driving autonomously under the same conditions. Equivalently, we can use the VO CDF to predict whether the system *would have* achieved an autonomy rate of less than 100% on the repeat pass in question. This allows us to easily compare different localization pipelines on the same dataset without directly comparing repeat pass autonomy rates as we did in the experiments of Section 4. It also provides a means of identifying the smallest manual intervention threshold that would allow for perfect or near-perfect



(a) Monocular (grey + color-constant images)

(b) Stereo (grey + color-constant images)

(c) Monocular (grey image only)

(d) Stereo (grey image only)

Figure 15: CDF of the distance over which the system had to localize using VO for repeat passes occurring 1 hour (Repeat 1), 3 hours (Repeat 4), and 7.5 hours (Repeat 7) after the teach pass, both with and without color-constant imagery as described in Section 2.1. The plots read, "for Y% of the traverse, the system drove less than X meters on VO". CDFs closer to the top-left corner indicate better localization performance during the repeat pass. In our experiments, we consider localization to have failed and manual intervention to be required if the system relies on pure VO for more than 10 meters.

autonomy during the repeat pass, although this goal must be weighed against the likelihood of excessive VO drift as discussed in Section 4.3.

Figure 15 shows the VO CDF plots for three reconstructed repeat passes from the CSA dataset, conducted 1 hour, 3 hours, and 7.5 hours after the teach pass, using both the stereo and monocular versions of the legacy and lighting-resistant pipelines. As reported by Paton et al. (2015a), the stereo pipeline enjoys significant gains in robustness by using color-constant images in addition to the standard grey image (compare Figures 15b and 15d). This benefit is especially apparent for repeat passes 4 and 7, conducted 3 hours and 7.5 hours after the teach pass, respectively. For both of these repeat passes, the system traversed substantially more of the route while successfully localizing against the map than it would have using grey images only. For repeat pass 7, the use of color-constant images was actually sufficient to prevent the manual intervention that would have been required using grey images only.

The results for the monocular pipeline are perhaps even more striking. While Figure 15c shows that the legacy monocular pipeline would have performed much worse than the stereo pipeline and would have required manual intervention even on repeat pass 1 (1 hour after the teach pass), Figure 15a shows that the addition

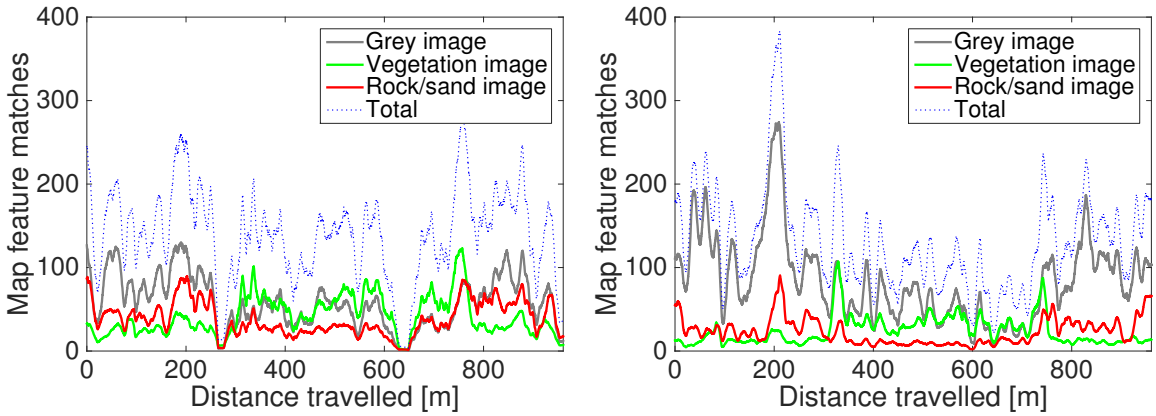
Table 6: Comparison of lighting-resistant monocular and stereo VO CDFs at selected distances

	CDF (0.01 m) [%]		CDF (0.1 m) [%]		CDF (1 m) [%]		CDF (10 m) [%]†	
Repeat	Mono	Stereo	Mono	Stereo	Mono	Stereo	Mono	Stereo
1	93.91	99.62	96.66	99.87	99.37	100.00	99.91	100.00
2	84.49	95.94	90.75	98.91	98.05	100.00	100.00	100.00
3	53.62	87.89	66.80	95.93	92.57	99.97	99.85	100.00
4	42.55	76.15	58.06	89.98	89.43	99.86	99.93	100.00
5	43.12	63.32	59.32	80.86	90.38	99.30	99.65	100.00
6	31.58	38.10	53.05	62.25	89.08	96.34	99.69	99.68
7	30.79	38.70	51.19	63.94	88.36	95.54	99.31	99.78
9	49.89	64.61	64.38	82.36	93.43	99.54	99.88	99.98
10	45.34	55.79	63.20	75.26	91.80	98.71	100.00	99.98
11	52.18	80.40	66.27	92.63	92.83	99.93	99.88	100.00
12	51.04	79.74	66.06	92.14	93.03	99.85	99.79	100.00
13	52.50	86.04	65.80	95.23	91.05	99.92	99.91	100.00
14	45.97	79.49	61.00	91.99	91.67	99.88	99.93	100.00
15	59.17	84.00	70.47	93.85	93.84	99.92	100.00	100.00
16	49.08	85.86	61.93	94.43	92.05	99.94	100.00	100.00
17	56.96	82.79	70.19	93.15	93.81	99.94	99.88	100.00
18	57.98	84.56	70.86	94.11	95.46	99.85	100.00	100.00
19	54.88	78.04	68.13	91.69	94.29	99.89	100.00	100.00
20	50.46	75.36	65.09	89.23	93.71	99.86	100.00	100.00
21	53.37	75.90	67.96	90.20	93.88	99.82	100.00	99.99
22	43.99	55.28	58.92	76.00	92.12	98.83	100.00	100.00
23	55.18	77.56	69.46	90.52	94.54	99.81	99.94	100.00
24	51.35	66.74	66.43	83.63	93.92	99.45	100.00	100.00
25	51.53	61.60	67.67	80.84	94.18	99.07	100.00	100.00
27	37.05	55.53	55.37	76.02	91.41	97.53	99.94	99.89
Mean	51.92	73.16	66.20	87.00	92.97	99.31	99.90	99.97
Stdev.	13.49	15.81	9.89	10.04	2.49	1.17	0.16	0.08

† These are the expected autonomy rates given the parameters we used in our experiments (see Table 5).

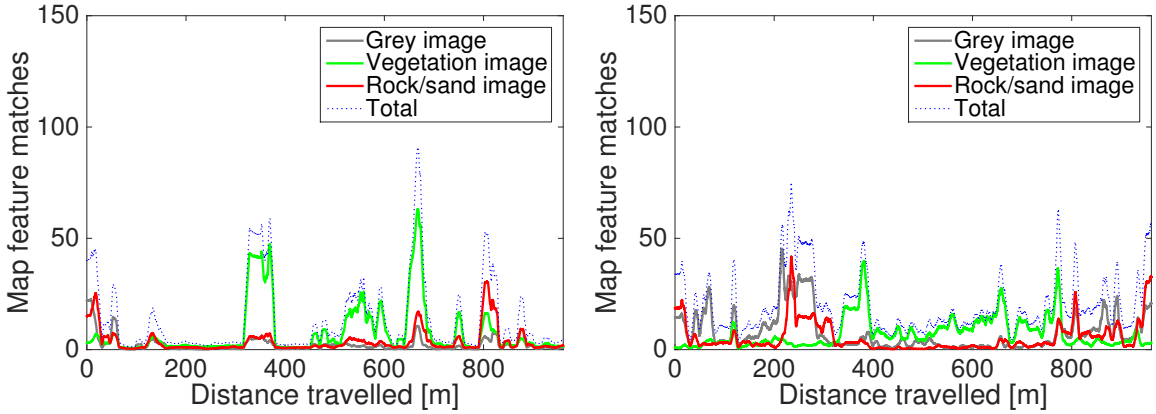
of color-constant imagery improves localization quality to the extent that repeat pass 1 could have been traversed nearly as reliably as with stereo. This improvement is less pronounced for repeat passes 4 and 7, but is still sufficient to avoid manual intervention on repeat pass 4, and to nearly avoid manual intervention on repeat pass 7.

Table 6 summarizes the VO CDFs for all 25 repeat passes using both the monocular and stereo lighting-resistant pipelines. We compare the localization performance of the two pipelines by reporting the value of the VO CDFs (in percent) at several representative distance scales. This comparison gives an idea of the expected autonomy rate of each system if it were prevented from travelling more than the reported distance without localizing against the map. From the table we can see that the expected autonomy rates of the lighting-resistant stereo pipeline are generally higher than those of the lighting-resistant monocular pipeline, although with the threshold we used in all of our experiments (10 m), we would expect to see similar levels of autonomy from both systems.



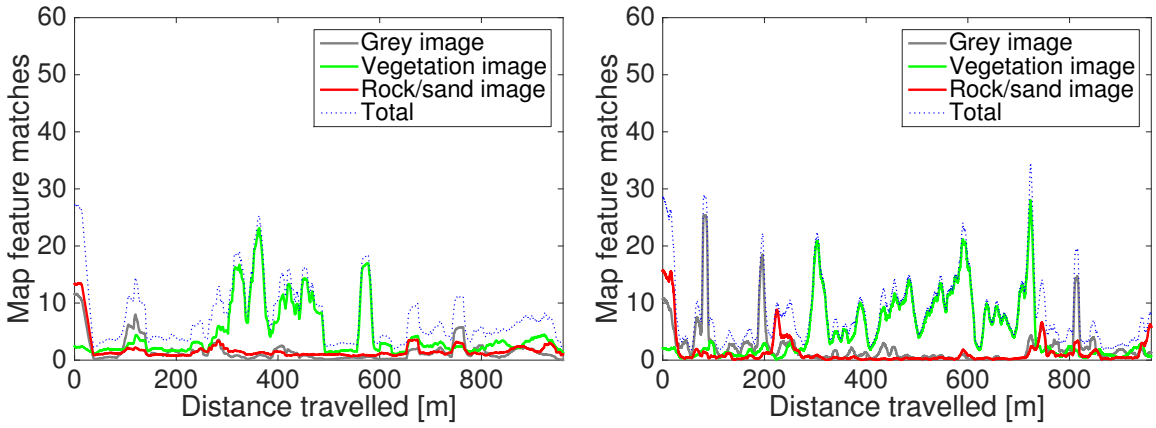
(a) Monocular (b) Stereo

Figure 16: Keypoint match counts for Repeat 1 (1 hour after teach pass) for the grey image and the two color-constant images. For clarity, we have applied a 200-point sliding-window mean filter to the raw data.



(a) Monocular (b) Stereo

Figure 17: Keypoint match counts for Repeat 4 (3 hours after teach pass) for the grey image and the two color-constant images. For clarity, we have applied a 200-point sliding-window mean filter to the raw data.



(a) Monocular (b) Stereo

Figure 18: Keypoint match counts for Repeat 7 (7.5 hours after teach pass) for the grey image and the two color-constant images. For clarity, we have applied a 200-point sliding-window mean filter to the raw data.

Table 7: Map keypoint match count statistics for each lighting-resistant pipeline on the CSA dataset

	Grey – μ (σ)		Vegetation – μ (σ)		Rock/sand – μ (σ)		Total – μ (σ)	
Repeat	Mono	Stereo	Mono	Stereo	Mono	Stereo	Mono	Stereo
1	59 (47)	85 (60)	42 (35)	24 (21)	39 (31)	23 (19)	140 (95)	130 (77)
2	24 (26)	33 (35)	26 (25)	14 (14)	21 (21)	13 (15)	71 (58)	60 (47)
3	9.6 (17)	19 (24)	10 (16)	10 (11)	10 (16)	8.3 (11)	30 (41)	38 (34)
4	2.2 (4.9)	7.8 (11)	8.0 (15)	7.3 (8.8)	3.7 (7.1)	5.4 (8.3)	14 (22)	20 (18)
5	1.4 (3.9)	3.7 (7.3)	4.1 (7.6)	6.2 (7.6)	2.9 (6.0)	3.2 (6.5)	8.4 (14)	13 (14)
6	0.9 (2.7)	1.6 (3.8)	6.2 (10)	4.7 (6.2)	2.2 (4.3)	1.7 (4.4)	9.3 (12)	8.0 (8.2)
7	1.6 (3.5)	2.2 (5.4)	4.5 (8.3)	5.0 (6.8)	1.7 (2.7)	1.3 (3.1)	7.8 (10)	8.5 (9.3)
9	3.8 (7.6)	9.1 (14)	8.0 (13)	7.4 (9.3)	3.0 (5.3)	3.2 (6.4)	15 (20)	20 (19)
10	3.5 (8.3)	7.4 (13)	6.1 (9.8)	6.0 (7.6)	2.6 (4.6)	2.8 (5.8)	12 (17)	16 (17)
11	6.2 (11)	16 (20)	7.5 (12)	8.4 (10)	6.4 (11)	6.4 (9.7)	20 (27)	30 (27)
12	6.7 (12)	14 (19)	7.3 (12)	8.1 (9.6)	5.5 (8.9)	5.7 (8.5)	20 (26)	28 (25)
13	10 (18)	22 (29)	9.1 (14)	9.4 (10)	7.9 (13)	6.9 (10)	27 (37)	39 (37)
14	7.4 (13)	17 (21)	6.7 (11)	8.0 (9.2)	6.7 (11)	7.0 (10)	21 (29)	32 (29)
15	9.4 (18)	22 (30)	11 (16)	8.9 (9.5)	7.5 (12)	7.3 (10)	27 (37)	38 (38)
16	8.1 (16)	22 (29)	8.5 (14)	9.3 (9.6)	6.4 (11)	7.7 (11)	23 (35)	39 (38)
17	8.5 (13)	21 (23)	8.7 (14)	7.8 (8.6)	8.7 (13)	9.5 (13)	26 (32)	38 (34)
18	8.4 (15)	19 (24)	10 (15)	8.8 (8.9)	7.6 (12)	8.1 (11)	26 (34)	36 (34)
19	6.4 (12)	14 (20)	7.6 (12)	7.9 (8.8)	5.9 (10)	6.0 (9.0)	20 (29)	28 (28)
20	5.8 (11)	14 (18)	6.5 (11)	7.2 (8.4)	5.6 (9.2)	6.3 (9.2)	18 (26)	27 (26)
21	5.3 (10)	12 (16)	8.0 (13)	7.1 (8.1)	5.6 (9.5)	5.5 (8.5)	19 (27)	25 (23)
22	2.7 (5.6)	6.7 (10)	5.7 (10)	5.5 (6.7)	3.4 (5.7)	4.2 (7.4)	12 (18)	16 (17)
23	5.5 (9.0)	13 (18)	8.4 (12)	7.0 (8.0)	4.6 (7.1)	5.2 (8.5)	18 (23)	25 (25)
24	3.4 (6.8)	8.7 (12)	7.1 (13)	6.0 (7.5)	3.3 (5.7)	3.7 (6.6)	14 (21)	18 (18)
25	4.9 (9.6)	12 (16)	6.0 (9.4)	5.1 (6.0)	3.5 (6.6)	3.8 (7.1)	14 (20)	20 (22)
27	2.2 (6.1)	9.0 (14)	3.6 (5.1)	3.6 (3.8)	2.3 (4.3)	3.8 (6.3)	8.1 (13)	16 (20)
Mean	8.3 (12)	16 (16)	9.5 (7.9)	8.1 (3.9)	7.1 (7.7)	6.4 (4.3)	25 (27)	31 (24)

5.5 Comparison of Keypoint Match Stability

In order to understand the effect that color-constant images have on the robustness of the monocular pipeline, it is illustrative to compare the number of keypoint matches across repeat passes (see Figures 16 to 18 and Table 7). Although each individual image (grey, vegetation, and rock/sand) yields a modest number of map keypoint matches, the combination of matches from all three images is sufficient to permit substantially more reliable localization than using the grey image alone, even without accounting for lighting changes. Consulting Table 7, we see that on average, the stereo pipeline tends to match twice as many map keypoints in the grey image as the monocular pipeline, but both pipelines match approximately the same number of color-constant map keypoints on average, yielding total match counts that are more similar on average than those obtained from the grey images alone. We also see that the spread of average keypoint matches is smaller for the two color-constant images than for the grey images, indicating that they are generally more reliable sources of stable keypoint matches than the grey image. This is the main reason for the improved robustness of the lighting-resistant monocular pipeline seen in Figure 15.

5.6 Limitations

Although the use of color-constant imagery provides monocular VT&R with significant gains in robustness, its performance remains worse than its stereo counterpart, especially when the lighting has changed substantially after map creation. Consulting Figures 17 and 18, we notice that the total number of keypoint matches is

consistently lower in the monocular case than in the stereo case. We believe this is due mainly to the orientation of the camera in this dataset. In Section 4, we showed that the monocular pipeline generally matches similar, if not somewhat higher, numbers of keypoints to the map when the camera is mounted at 47° to the horizontal (see Figures 7 and 11). However, since the camera on the Grizzly RUV was mounted at a fairly shallow angle, 29° to the horizontal, much of the camera's field of view is taken up by objects not on the ground, especially trees and rock faces. Since keypoints detected on non-ground objects severely violate the planarity assumption in our monocular pipeline, even with uncertainty, they are typically rejected as outliers, thereby reducing the total number of inlying keypoint matches. Although it is impossible to verify on this dataset, we believe that the monocular pipeline would have enjoyed even greater gains in robustness if the camera had been angled more steeply towards the ground.

6 Conclusions and Future Work

We have presented a Visual Teach and Repeat (VT&R) system that is capable of autonomously repeating kilometer-scale routes with centimeter-scale accuracy in rough terrain, using only monocular vision. By approximating the scene geometry as a manifold of uncertain local ground planes, we relax the requirement for true 3D sensing that had prevented the deployment of Furgale and Barfoot's (2010) VT&R system on a wide range of vehicles equipped with monocular cameras. Field tests on 4.3 km of autonomous navigation in which our monocular VT&R system was used to control a robotic vehicle have demonstrated that our system is capable of achieving route-repetition accuracy on par with its stereo counterpart, achieving an overall autonomy rate of 99.4% in these experiments.

In our previous conference paper (Clement et al., 2015), we noted that our monocular VT&R system is less robust to lighting and self-similar textures than its stereo counterpart. In this work, we address this trade-off through the use of color-constant imagery that is robust to shadows and illumination changes in scenes with vegetation, rocks, and sand (Ratnasingam and Collins, 2010; Paton et al., 2015a). In contrast to lighting-resistant active sensing techniques such as appearance-based lidar (McManus et al., 2013), our system achieves lighting-resistance using only physics-based transformations of images obtained from a passive RGB sensor. We demonstrate through offline testing on an additional 26 km of autonomous navigation data that the addition of color-constant imagery to the monocular pipeline results in vastly improved localization quality compared to the use of standard grey images only. However, the localization performance of the lighting-resistant monocular pipeline still falls short of that of the lighting-resistant stereo pipeline.

This work is novel in several ways. First, it is the only VT&R system capable of repeating kilometer-scale routes using only 2D monocular vision for motion estimation. Second, it is the first monocular localization system that both builds and localizes against maps generated from color-constant images inside a control loop. Finally, our system has been extensively tested on multiple vehicles in a variety of terrestrial and planetary-analogue environments, which is an important contribution in itself since it has allowed us to identify the strengths and weaknesses of our approach in realistic settings.

One important lesson learned from our experiments is that our monocular VT&R system is sensitive to a number of tuning parameters, especially those pertaining to keypoint uncertainty and outlier rejection. Rather than searching manually through this high-dimensional parameter space for a configuration that may or may not be optimal, future extensions to this system could incorporate iterative learning algorithms such as those proposed by Ostafew et al. (2013) to learn these parameters from experience.

Another important lesson is that our system is sensitive to camera placement, apparently performing better when the camera is angled more sharply towards the ground. Further experimentation to determine an optimal camera orientation could prove fruitful, and need not be limited to varying the camera's angle to the horizontal. For example, orienting the camera perpendicular to the direction of travel has been shown to improve the accuracy of stereo visual odometry (Peretroukhin et al., 2014). In future we would like to experiment with with alternative camera orientations to determine whether our monocular localization

pipeline exhibits similar improvements in accuracy.

As mentioned in Section 3.2, our depth estimation scheme could easily be extended to incorporate information from other sensors to estimate the ground-to-vehicle transformation. For ground vehicles, an inertial measurement unit (IMU) could be used to smooth out the transition between local ground planes, particularly when driving up or down small hills where the hill surface does not occupy enough of the camera field of view to constitute the dominant ground plane. This could potentially allow us to reduce the uncertainty in the ground plane orientation and achieve more reliable feature tracking. For aerial vehicles, an altimeter and IMU could be used to ensure that the local ground plane is always orthogonal to the gravity direction, irrespective of the vehicle orientation. This presents an interesting avenue for future work on improving our technique and extending it to a wider range of autonomous vehicles, although this relies on the availability of an additional sensor (an IMU).

In summary, we have shown that centimeter-accurate autonomous visual route-repetition is possible using only monocular vision and simple assumptions about scene geometry. We have validated this approach through online and offline testing on a combined 30 km of autonomous navigation data collected in multiple environments using two different robotic vehicles. While our monocular VT&R system is less robust to common failure cases than its stereo counterpart, we have shown that its robustness can be significantly improved by introducing computationally cheap color-constant imagery to the localization pipeline. Regardless, we believe that the benefit of deploying VT&R on existing monocular vehicles without requiring additional sensors far outweighs the modest reduction in robustness compared to an equivalent stereo system.

Acknowledgments

The authors would like to thank Matthew Giamou (now at the MIT Aerospace Controls Lab) and Valentin Peretroukhin of the Space and Terrestrial Autonomous Robotic Systems (STARS) lab for their assistance with field testing, the Autonomous Space Robotics Lab (ASRL) for their guidance in interacting with the VT&R code base and for the use of the CSA dataset, Leica Geosystems for providing the MultiStation, and Clearpath Robotics for providing the Husky and Grizzly rovers. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) through the NSERC Canadian Field Robotics Network (NCFRN).

References

Argyros, A. A., Bekris, K. E., Orphanoudakis, S. C., and Kavraki, L. E. (2005). Robot homing by exploiting panoramic vision. *Autonomous Robots*, 19(1):7–25.

Barfoot, T. and Furgale, P. (2014). Associating uncertainty with three-dimensional poses for use in estimation problems. *IEEE Trans. Robot. (T-RO)*, 30(3):679–693.

Baumgartner, E. T. and Skaar, S. B. (1994). An autonomous vision-based mobile robot. *IEEE Trans. Automat. Control (TAC)*.

Bay, H., Ess, A., Tuytelaars, T., and Gool, L. V. (2008). Speeded-up robust features (SURF). *Comput. Vision and Image Understanding (CVIU)*, 110:346–359.

Choi, S., Joung, J., Yu, W., and Cho, J. (2011). Monocular visual odometry under planar motion constraint. *Proc. Int. Conf. Control, Autom. and Syst. (ICCAS)*, pages 1480–1485.

Clement, L., Kelly, J. S., and Barfoot, T. D. (2015). Monocular visual teach and repeat aided by local ground planarity. In *Proc. Field and Service Robot. (FSR)*. To appear.

Corke, P., Paul, R., Churchill, W., and Newman, P. M. (2013). Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robot. and Syst. (IROS)*, pages 2085–2092.

Davison, A. J., Reid, I. D., Molton, N. D., and Stasse, O. (2007). MonoSLAM: Real-time single camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 29(6):1052–1067.

Eade, E. and Drummond, T. (2006). Scalable monocular SLAM. In *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*.

Engel, J., Schöps, T., and Cremers, D. (2014). LSD-SLAM: Large-scale direct monocular SLAM. In *Proc. European Conf. Comput. Vision (ECCV)*, pages 834–849, Cham. Springer International Publishing.

Farraj, F. and Asmar, D. (2013). Non-iterative planar visual odometry using a monocular camera. In *Proc. Int. Conf. Advanced Robot. (ICAR)*, pages 1–6.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6).

Furgale, P. and Barfoot, T. D. (2010). Visual teach and repeat for long-range rover autonomy. *J. Field Robot. (JFR)*, 27(5):534–560.

Furgale, P. T. (2011). *Extensions to the Visual Odometry Pipeline for the Exploration of Planetary Surfaces*. PhD thesis, University of Toronto, Toronto, ON.

Goedemé, T., Nuttin, M., Tuytelaars, T., and Gool, L. V. (2007). Omnidirectional vision based topological navigation. *Int. J. Comput. Vision (IJCV)*, 74(3):219–236.

Holmes, S. A. and Murray, D. W. (2013). Monocular SLAM with conditionally independent split mapping. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 35(6):1451–1463.

Jones, S. D., Andresen, C., and Crowley, J. L. (1997). Appearance based process for visual navigation. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robot. and Syst. (IROS)*, pages 551–557.

Kidono, K., Miura, J., and Shirai, Y. (2002). Autonomous visual navigation of a mobile robot using a human-guided experience. *Robot. and Autonomous Syst. (RAS)*, 40(2-3):121–130.

Klein, G. and Murray, D. (2007). Parallel tracking and mapping for small AR workspaces. In *Proc. IEEE/ACM Int. Symp. Mixed and Augmented Reality (ISMAR)*.

Lovegrove, S. J., Davison, A. J., and Ibanez-Guzman, J. (2011). Accurate visual odometry from a rear parking camera. In *Proc. IEEE Intelligent Vehicles Symp. (IV)*, pages 788–793.

Marshall, J., Barfoot, T. D., and Larsson, J. (2008). Autonomous underground tramming for center-articulated vehicles. *J. Field Robot. (JFR)*, 25:400–421.

Matsumoto, Y., Inaba, M., and Inoue, H. (1996). Visual navigation using view-sequenced route representation. In *Proc. IEEE Int. Conf. Robot. and Autom. (ICRA)*, pages 83–88.

Matsumoto, Y., Sakai, K., Inaba, M., and Inoue, H. (2000). View-based approach to robot navigation. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robot. and Syst. (IROS)*, pages 1702–1708.

McManus, C., Churchill, W., Maddern, W. P., Stewart, A. D., and Newman, P. M. (2014). Shady dealings: Robust, long-term visual localisation using illumination invariance. In *Proc. IEEE Int. Conf. Robot. and Autom. (ICRA)*, pages 901–906.

McManus, C., Furgale, P., Stenning, B., and Barfoot, T. D. (2013). Lighting-invariant visual teach and repeat using appearance-based lidar. *J. Field Robot. (JFR)*, 30(2):254–287.

Newcombe, R. A., Lovegrove, S. J., and Davison, A. J. (2011). DTAM: Dense tracking and mapping in real-time. In *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, pages 2320–2327.

Ohno, T., Ohya, A., and Yuta, S. (1996). Autonomous navigation for mobile robots referring pre-recorded image sequence. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robot. and Syst. (IROS)*, volume 2, pages 672–679.

Ostafew, C., Schoellig, A., and Barfoot, T. (2013). Iterative learning control to improve mobile robot path tracking in challenging outdoor environments. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robot. and Syst. (IROS)*, pages 176–181.

Ostafew, C. J., Schoellig, A. P., Barfoot, T. D., and Collier, J. (2015). Learning-based nonlinear model predictive control to improve vision-based mobile robot path tracking. *J. Field Robot. (JFR)*, 33(1):133–152.

Paton, M., MacTavish, K., Ostafew, C. J., and Barfoot, T. D. (2015a). It's not easy seeing green: Lighting-resistant stereo visual teach & repeat using color-constant images. In *Proc. IEEE Int. Conf. Robot. and Autom. (ICRA)*, pages 1519–1526.

Paton, M., Pomerleau, F., and Barfoot, T. D. (2015b). Eyes in the back of your head: Robust visual teach & repeat using multiple stereo cameras. In *Proc. Conf. Comput. and Robot Vision (CRV)*, pages 46–53.

Peretroukhin, V., Kelly, J., and Barfoot, T. (2014). Optimizing camera perspective for stereo visual odometry. In *Proc. Conf. Comput. and Robot Vision (CRV)*, pages 1–7.

Pizzoli, M., Forster, C., and Scaramuzza, D. (2014). REMODE: Probabilistic, monocular dense reconstruction in real time. In *Proc. IEEE Int. Conf. Robot. and Autom. (ICRA)*, pages 2609–2616.

Quigley, M., Conley, K., Gerkey, B. P., Faust, J., Foote, T., Leibs, J., Wheeler, R., and Ng, A. Y. (2009). ROS: An open-source robot operating system. In *Proc. Int. Conf. Robot. and Autom. (ICRA) Workshop on Open Source Software*.

Ratnasingam, S. and Collins, S. (2010). Study of the photodetector characteristics of a camera for color constancy in natural scenes. *J. Optical Soc. America A*, 27(2):286–294.

Remazeilles, A., Chaumette, F., and Gros, P. (2006). 3D navigation based on a visual memory. In *Proc. IEEE Int. Conf. Robot. and Autom. (ICRA)*, pages 2719–2725.

Royer, E., Lhuillier, M., Dhome, M., and Lavest, J.-M. (2007). Monocular vision for mobile robot localization and autonomous navigation. *Int. J. Comput. Vision (IJCV)*, 74(3):237–260.

Simhon, S. and Dudek, G. (1998). A global topological map formed by local metric maps. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robot. and Syst. (IROS)*, pages 1708–1714.

Tang, L. and Yuta, S. (2001). Vision based navigation for mobile robots in indoor environment by teaching and playing-back scheme. In *Proc. IEEE Int. Conf. Robot. and Autom. (ICRA)*, pages 3072–3077.

Zhang, A. M. and Kleeman, L. (2009). Robust appearance based visual route following for navigation in large-scale outdoor environments. *Int. J. Robot. Research (IJRR)*, 28(3):331–356.

Zhang, J., Singh, S., and Kantor, G. (2012). Robust monocular visual odometry for a ground vehicle in undulating terrain. In *Proc. Field and Service Robot. (FSR)*, pages 311–326.

Zhao, L., Huang, S., Yan, L., Jianguo, J., Hu, G., and Dissanayake, G. (2010). Large-scale monocular SLAM by local bundle adjustment and map joining. In *Proc. IEEE Int. Conf. Control, Autom., Robot. and Vision (ICARCV)*, pages 431–436.

Zienkiewicz, J. and Davison, A. J. (2014). Extrinsics autocalibration for dense planar visual odometry. *J. Field Robot. (JFR)*, 32(5):803–825.