Monocular Visual Teach and Repeat Aided by Local Ground Planarity

Lee Clement, Jonathan Kelly and Timothy D. Barfoot

Abstract Visual Teach and Repeat (VT&R) allows an autonomous vehicle to repeat a previously traversed route without a global positioning system. Existing implementations of VT&R typically rely on 3D sensors such as stereo cameras for mapping and localization, but many mobile robots are equipped with only 2D monocular vision for tasks such as teleoperated bomb disposal. While simultaneous localization and mapping (SLAM) algorithms exist that can recover 3D structure and motion from monocular images, the scale ambiguity inherent in these methods complicates the estimation and control of lateral path-tracking error, which is essential for achieving high-accuracy path following. In this paper, we propose a monocular vision pipeline that enables kilometre-scale route repetition with centimetre-level accuracy by approximating the ground surface near the vehicle as planar (with some uncertainty) and recovering absolute scale from the known position and orientation of the camera relative to the vehicle. This system provides added value to many existing robots by allowing for high-accuracy autonomous route repetition with a simple software upgrade and no additional sensors. We validate our system over 4.3 km of autonomous navigation and demonstrate accuracy on par with the conventional stereo pipeline, even in highly non-planar terrain.

1 Introduction

Visual Teach and Repeat (VT&R) is an effective tool for autonomously navigating previously traversed paths using only on-board visual sensors. In an initial *teach pass*, a human operator manually drives an autonomous vehicle along a desired route while

L. Clement (✉) · J. Kelly · T.D. Barfoot
Institute for Aerospace Studies, University of Toronto, Toronto, Canada
e-mail: lee.clement@mail.utoronto.ca

J. Kelly
e-mail: jkelly@utias.utoronto.ca

T.D. Barfoot
e-mail: tim.barfoot@utoronto.ca

D.S. Wettergreen and T.D. Barfoot (eds.), *Field and Service Robotics*,
Springer Tracts in Advanced Robotics 113, DOI 10.1007/978-3-319-27702-8_36

the VT&R system uses imagery from a camera to build a map of the route. In the subsequent *repeat pass*, the system localizes against the stored map to autonomously repeat the route, sometimes combining map-based localization with visual odometry (VO) to estimate relative motion in cases where map-based localization is temporarily unavailable (Fig. 1). VT&R is well-suited to repetitive navigation tasks where GPS is unavailable or insufficiently accurate, and has found applications in autonomous tramming for mining operations [14] and sample return missions [8].

The map representation in a VT&R system may be purely topological, purely metric, or a mixture of the two (sometimes called topometric). Purely topological VT&R [9, 15, 20] uses a network of reference images (keyframes) where the navigation goal is to match the current image to the nearest keyframe using a visual homing procedure. These systems are restricted to heading-based control, which only loosely bounds lateral path-tracking error. Purely metric maps are uncommon in VT&R systems due to the high computational cost of creating globally consistent maps for long routes, but successful applications do exist [11, 21]. Topometric systems [8, 14, 22, 23] reap the benefits of both mapping strategies by decoupling map size from path length while still retaining metric information.

Furgale and Barfoot [8] developed the first VT&R system capable of autonomously repeating multi-kilometre routes in unstructured outdoor terrain using only a stereo camera. Their system creates a topometric map of metric keyframes connected by 6DOF VO estimates, which are combined via local bundle adjustment into locally consistent metric submaps for localization in the repeat pass.

Furgale and Barfoot's system has been extended to other 3D sensors such as lidar [16] and RGB-D cameras, but a monocular implementation has not been forthcoming. While monocular cameras are appealing in terms of size, cost, and simplicity, perhaps the most compelling motivation for using monocular vision for VT&R is the plethora of existing mobile robots that would benefit from it. Indeed, vehicles equipped with monocular vision, typically for teleoperation, run the gamut of robotics applications,

Fig. 1 Our field robot during a 140 m autonomous traverse in the UTIAS MarsDome indoor rover testing environment, with the path overlaid for illustration. In order to compare the performance of stereo and monocular VT&R with the same hardware, we equipped our rover with a stereo camera and used only the left image stream for our monocular traverses

and in many cases—search and rescue, mining, construction, and personal assistive robotics, to name a few—would benefit from accurate autonomous route-repetition, especially if it were achievable with existing sensors.

Several techniques exist for accomplishing online 3D simultaneous localization and mapping (SLAM) with monocular vision, ranging from filter-based approaches [4, 5] to online batch techniques that make use of local bundle adjustment [10, 12, 25]. Such algorithms are capable of producing accurate 3D maps, but only up to an unknown scale factor. This scale ambiguity complicates threshold-based outlier rejection, as well as the estimation and control of lateral path-tracking error during the repeat pass, which are essential for achieving high-accuracy route-following.

In this paper, we extend Furgale and Barfoot's VT&R system to monocular vision by using the approximately known position and orientation of a camera mounted on a rover to estimate the 3D positions of keypoints near the ground with absolute scale. Similar techniques have succeeded in computing VO with a monocular camera using both sparse feature tracking [3, 6, 24] and dense image alignment [13], but have not considered the problem of map construction. We show that by treating the ground surface near the vehicle as approximately planar and applying an appropriate uncertainty model, we can generate local metric maps that are accurate enough to achieve centimetre-level accuracy during the repeat pass, even on highly non-planar terrain. Although the flat-ground assumption is not globally valid, it is sufficient for our purposes since VT&R uses metric information only locally.

The main contribution of this paper is an extensive comparison of the performance of monocular and stereo VT&R in a variety of conditions, including an evaluation of their robustness to common failure cases. To this end, we present experimental results comparing the two systems over 4.3 km of autonomous navigation. While our results show that both systems achieve similar path-tracking accuracy when functioning normally, the monocular system suffers a reduction in robustness compared to its stereo counterpart in certain conditions. We argue that, for many applications, the benefit of deploying VT&R without a potentially costly sensor upgrade far outweighs the associated reduction in robustness.

2 Monocular Depth Estimation

We estimate the 3D coordinates of features observed by a camera pointed downward, but not directly at the ground surface, by approximating the local ground surface near the vehicle as planar and recovering absolute scale from the known position and orientation of the camera relative to the vehicle. We account for variations in terrain shape by applying an appropriate uncertainty model. In what follows, $\mathbf{z}_j^i$ denotes the 3D coordinates of feature i expressed in coordinate frame $\mathscr{F}_j$.

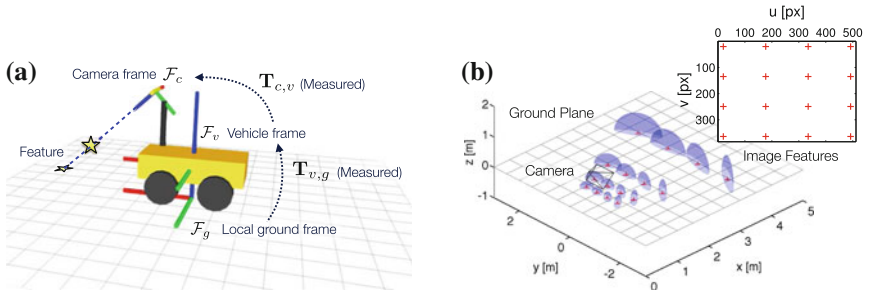

Fig. 2 Geometry and uncertainty model of our monocular depth estimation scheme. **a** Coordinate frames in our monocular depth estimation scheme. The local ground frame $\mathcal{F}_g$ is defined relative to the vehicle frame $\mathcal{F}_v$ and travels with the vehicle. **b** Evenly-spaced synthetic image features (*top right*) and estimated D coordinates with 1σ uncertainty ellipses for the experimental configuration described in Sect. 4

2.1 Locally Planar Ground Surfaces

For a monocular camera observing the ground, we can estimate the 3D coordinates of features near the ground by making the following assumptions (see Fig. 2a):

1. all features of interest lie in the xy-plane of a local ground frame $\mathcal{F}_g$ defined such that its z-axis is normal to the ground and always intersects the origin of the vehicle coordinate frame $\mathcal{F}_v$ (for a ground vehicle, this is the vehicle's footprint);
2. the transformation $\mathbf{T}_{c,v} \in \mathrm{SE}(3)$ from $\mathcal{F}_v$ to the camera-centric coordinate frame $\mathcal{F}_c$ is known; and
3. the transformation $\mathbf{T}_{v,g} \in \mathrm{SE}(3)$ from $\mathcal{F}_g$ to $\mathcal{F}_v$ is known.

Assuming that incoming images have been de-warped and rectified in a pre-processing step, we can model the camera as an ideal pinhole camera with calibrated camera matrix $\mathbf{K}$ such that the image coordinates $\mathbf{y}^i$ of $\mathbf{z}_c^i$ are given by

$$\mathbf{y}^i := \begin{bmatrix} u^i & v^i & 1 \end{bmatrix}^T = \mathbf{K}\mathbf{p}^i \,, \tag{1}$$

where

$$\mathbf{p}^i := \begin{bmatrix} p_x^i & p_y^i & 1 \end{bmatrix}^T = \frac{1}{z_c^i} \begin{bmatrix} x_c^i & y_c^i & z_c^i \end{bmatrix}^T \tag{2}$$

represents the (unitless) normalized coordinates of $\mathbf{z}_c^i$ on the image plane. Note that although u^i, v^i represent pixel coordinates, they are not necessarily integer-valued.

By assumption 1, $z_g^i = 0, \forall i$, so we can write

$$\mathbf{z}_c^i := \begin{bmatrix} x_c^i & y_c^i & z_c^i & 1 \end{bmatrix}^T = \mathbf{T}_{c,g} \begin{bmatrix} x_g^i & y_g^i & 0 & 1 \end{bmatrix}^T \,, \tag{3}$$

where $\mathbf{T}_{c,g} = \mathbf{T}_{c,v}\mathbf{T}_{v,g}$. We can therefore obtain the feature depth z_c^i as a function of $\mathbf{p}^i$ by substituting $x_c^i = z_c^i p_x^i$ and $y_c^i = z_c^i p_y^i$ according to Eq. (2), and solving the third component of Eq. (3) for z_c^i, yielding

$$z_c^i = \frac{k_1}{k_2 + k_3 p_x^i + k_4 p_y^i} , \tag{4}$$

where, using T_{mn} as shorthand for the mth row and nth column of $\mathbf{T}_{c,g}$,

$$\begin{aligned} k_1 = T_{11}\,(T_{22}T_{34} - T_{24}T_{32}) \\ + T_{12}\,(T_{24}T_{31} - T_{21}T_{34}) \\ + T_{14}\,(T_{21}T_{32} - T_{22}T_{31}) \end{aligned} \qquad \begin{aligned} k_2 &= T_{11}T_{22} - T_{12}T_{21} \\ k_3 &= T_{21}T_{32} - T_{22}T_{31} \\ k_4 &= T_{12}T_{31} - T_{11}T_{32} . \end{aligned}$$

Finally, using Eqs. (1) and (2) with z_c^i as in Eq. (4), we can express the Cartesian coordinates of $\mathbf{z}_c^i$ in terms of $\mathbf{y}^i$ as

$$\mathbf{z}_c^i = z_c^i \mathbf{K}^{-1}\mathbf{y}^i . \tag{5}$$

2.2 Uncertainty Considerations

A crucial component of enabling monocular VT&R using this depth estimation scheme is an appropriate model of the uncertainty in each observation $\mathbf{z}_c^i$. We consider two important factors: uncertainty in image coordinates $\mathbf{y}^i$, and uncertainty in ground shape far from the vehicle. In early experiments, we found that image coordinate uncertainty alone did not permit reliable feature tracking since there was little overlap in 3D feature coordinate estimates across multiple frames.

We model feature coordinates in image space as Gaussian distributions centred on $\mathbf{y}^i$ with covariance $\mathbf{R}_{\mathbf{y}^i} := \mathrm{diag}\{(\sigma_u^i)^2, (\sigma_v^i)^2\}$. We use SURF features [2] in our system and determine σ_u^i, σ_v^i from the image pyramid level at which each feature 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 $\mathbf{R}_{\mathbf{T}_{v,g}} := \mathrm{diag}\{\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2, \sigma_5^2, \sigma_6^2\}$, where $\sigma_1 \ldots \sigma_6$ are tunable parameters corresponding to the six generators of SE(3). Together these factors form an 8-dimensional Gaussian distribution with covariance $\mathbf{R}_i := \mathrm{diag}\{\mathbf{R}_{\mathbf{y}^i}, \mathbf{R}_{\mathbf{T}_{v,g}}\}$, which we propagate via the combined Jacobian

$$\mathbf{G}_i := \begin{bmatrix} \dfrac{\partial \mathbf{z}_c^i}{\partial \mathbf{y}^i} & \dfrac{\partial \mathbf{z}_c^i}{\partial \mathbf{T}_{v,g}} \end{bmatrix}$$

to approximate $\mathbf{z}_c^i$ as a Gaussian in 3D space with covariance $\mathbf{Q}_i = \mathbf{G}_i \mathbf{R}_i \mathbf{G}_i^T$.

Using the Cartesian coordinates of $\mathbf{z}_c^i$ and $\mathbf{y}^i$ to compute the Jacobian, we have

$$\frac{\partial \mathbf{z}_c^i}{\partial \mathbf{y}^i} = \frac{z_c^i}{k_1} \begin{bmatrix} \left(k_1 + k_3 x_c^i\right)/f_u & k_4 x_c^i / f_v \\ k_3 y_c^i / f_u & \left(k_1 + k_4 y_c^i\right)/f_v \\ k_3 z_c^i / f_u & k_4 z_c^i / f_v \end{bmatrix} \tag{6}$$

and

$$\frac{\partial \mathbf{z}_c^i}{\partial \mathbf{T}_{v,g}} = \frac{\partial \mathbf{z}_c^i}{\partial \mathbf{T}_{c,g}} \frac{\partial \mathbf{T}_{c,g}}{\partial \mathbf{T}_{v,g}} = \left[\mathbf{1} \ (-\mathbf{z}_c^i)^\times\right] \mathrm{Ad}(\mathbf{T}_{c,v}) \ . \tag{7}$$

In the above, we adopt the notation of [1]: $\mathbf{1}$ denotes the (3×3) identity matrix, $\mathrm{Ad}(\cdot)$ the adjoint in SE(3), and $(\cdot)^\times$ the skew-symmetric cross-product matrix.

Figure 2b shows 1σ uncertainty ellipses for a number of evenly spaced synthetic image features resulting from a camera configuration similar to that used in the experiments described in Sect. 4.

3 System Overview

This section provides an overview of the VT&R system as it pertains to the methods of the previous section. In particular, we discuss the generic localization pipeline used for both online mapping in the teach pass and local map construction in the repeat pass. Figure 3 shows the stereo and monocular versions of the pipeline, which differ mainly in the front-end image processing used to generate 3D keypoints.

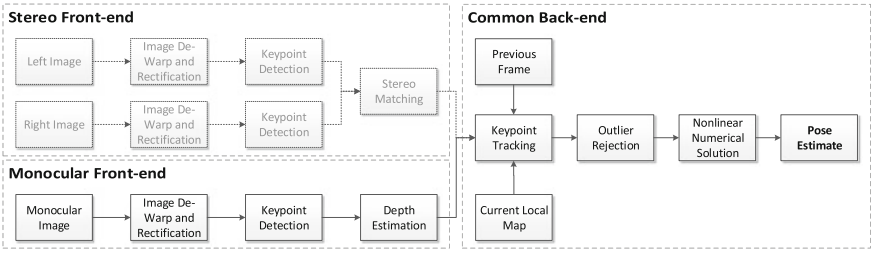

Fig. 3 The major processing blocks of the stereo and monocular localization pipelines. The monocular pipeline shares most of the same processing blocks as its stereo counterpart, differing mainly in the front-end image processing used to generate 3D keypoints. The "Current Local Map" block is only used for keypoint tracking during the repeat pass

3.1 Keypoint Generation

Raw images entering the pipeline first pass through a pre-processing step that uses a calibrated camera model to make them appear as though they were produced by an ideal pinhole camera. A GPU implementation of the SURF detector [2] then identifies keypoints in the de-warped and rectified images. The pipeline estimates the 3D coordinates of each keypoint in the camera frame using a matching procedure in the stereo case or the technique of Sect. 2 in the monocular case. The subsequent behavior of the pipeline differs slightly between the teach pass and the repeat pass.

3.2 Teach Pass

In the teach pass, the system constructs a pose graph whose vertices store lists of 3D keypoints with associated uncertainty and SURF descriptors, and whose edges store lists of matched keypoints and 6DOF pose change estimates. The system first tracks 3D keypoints in the current image against those in the most recent keyframe to generate a list of keypoint matches. These matches form the input to a 3-point RANSAC algorithm [7] that generates hypotheses for the 6DOF interframe pose change and rejects outlying feature tracks. In the context of monocular VT&R, this procedure typically rejects features far from the local ground surface (e.g., walls) since their motion is not adequately captured by the uncertainty model described in Sect. 2.2. The resulting pose change estimate serves as the initial guess in an iterative Gauss-Newton that refines the estimate based on inlying tracks.

3.3 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.

At every timestep, the system identifies the nearest keyframe in the path and performs a local bundle adjustment over a user-specified number of topologically adjacent keyframes, generating a local metric map in the reference frame of the nearest keyframe. The system then forms an augmented keyframe from the adjusted map keypoints against which freshly detected features may be matched. As in the teach pass, the system performs frame-to-frame VO to obtain an initial 6DOF pose estimate at each time step, which it uses as an initial guess to localize against the current local map and refine its pose estimate.

If the system fails to localize against the map, it may rely purely on VO until either a successful localization occurs or the vehicle exceeds some preset distance

threshold since the last successful localization. In the latter case, the system will halt the traverse and enter a search mode until it relocalizes or the operator intervenes.

4 Experiments

We conducted two sets of experiments at the University of Toronto Institute for Aerospace Studies (UTIAS), the first outdoors on relatively flat terrain, and the second on the highly non-planar terrain of the UTIAS MarsDome indoor rover testing environment. We compare the performance of our monocular VT&R system to that of the established stereo system [8] over 4.3 km of autonomous navigation. Table 1 reports path lengths, repeat speeds, start times, and autonomy rates for each experiment. We repeated each route using the monocular pipeline first, and 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.

4.1 Hardware

We used a four-wheeled skid-steered Clearpath Husky A200 rover equipped with a PointGrey Bumblebee XB3 stereo camera, which outputs 512 × 384 pixel greyscale images at 15 frames per second. The camera is mounted 1.0 m above the ground and is angled downwards at 47° to the horizontal (Fig. 4). These values were measured by hand since our system functions well even without an especially accurate estimate of $\mathbf{T}_{c,v}$. Small errors in $\mathbf{T}_{c,v}$ are simply absorbed by the uncertainty in $\mathbf{T}_{v,g}$.

Table 1 Summary of experimental results

Trial	Route	Path length (m)	Repeat speed (m/s)	Local start time (UTC-4)			Autonomy rate	
				Teach	Mono	Stereo	Mono (%)	Stereo (%)
1	Outdoor	1370	0.6	09:56:46	10:35:10	12:08:30	99.71[a]	100.00
2	Outdoor	1360	0.6	11:45:40	12:22:26	13:43:49	99.88	100.00
3	Outdoor	1361	0.6	13:26:41	14:00:12	15:20:12	99.74	100.00
4	Indoor	126	0.3	13:32:23	13:40:53	14:02:46	96.28	100.00
5	Indoor	140	0.3	12:18:57	12:32:20	12:59:11	91.60	100.00
				Mono	**Stereo**			
Total distance driven				4298 m[a]	4357 m			
Total distance autonomously traversed				99.41 %	100.00 %			

[a]During the monocular repeat pass of Trial 1, a parked vehicle 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

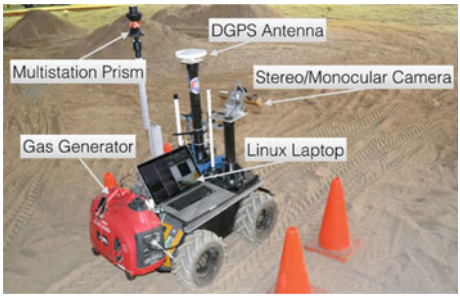

Fig. 4 Clearpath Husky A200 rover equipped with PointGrey Bumblebee XB3 stereo camera, DGPS receiver, Leica Nova MS50 MultiStation prism, 1 kW gas generator, and Linux laptop running ROS [19]

During the teach pass, we recorded stereo images and used them to teach identical paths using both the monocular and stereo pipelines. For the monocular pipeline, we used imagery from the left camera of the stereo pair only. The system detects 600 SURF keypoints in each incoming image and creates new keyframes every 25 cm in translation or 2.5° in rotation. For the monocular pipeline, we assigned standard deviations of 10 cm to the translational components of $\mathbf{T}_{v,g}$ and 10° to its rotational components as these values generally worked well in practice.

4.2 Outdoor Experiments

To evaluate the performance of the monocular VT&R system over long distances, we taught three 1.4 km paths through the parking lots and driveways of UTIAS. While these paths consisted mostly of flat pavement, they included many non-planar features such as speed bumps, side slopes, deep puddles, and rough shoulders, as well as other terrain types including gravel, sand, and grass.

We equipped the rover with an Ashtech DG14 Differential GPS unit used in tandem with a second stationary DG14 unit to obtain centimetre-accuracy RTK-corrected GPS data during the outdoor experiments. We used these data purely for ground-truthing purposes; they had no bearing on the behaviour of either pipeline. Figure 5 shows GPS tracks of the teach and repeat passes of one outdoor route.

Figure 6 shows estimated and measured lateral path-tracking errors during the monocular and stereo repeat passes. Both pipelines achieved centimetre-level accuracy in their respective repeat passes and produced similar estimates of lateral path-tracking error. In cases of map localization failure (i.e., when the system relied on pure VO), the monocular pipeline's estimated lateral path-tracking error diverged from ground truth more quickly than that of the stereo pipeline since keypoint position uncertainties are poorly constrained by only two measurements. Note, however, that the vehicle remained within about 20 cm of the taught path at all times.

Figure 7 compares the number of successful feature matches for frame-to-frame VO and map-based localization for both pipelines. Both pipelines track similar numbers of features from frame to frame, but the monocular pipeline generally tracks

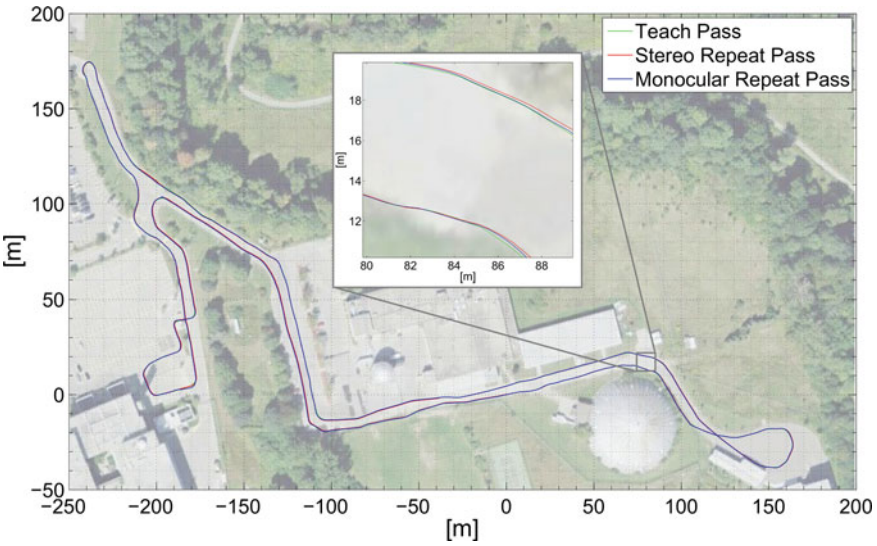

Fig. 5 Comparison of RTK-corrected GPS tracks of the teach pass, stereo repeat pass, and monocular repeat pass of a 1.4 km outdoor route (Trial 3 in Table 1). The zoomed-in section highlights the centimetre-level accuracy of both pipelines (Map data: Google, DigitalGlobe.)

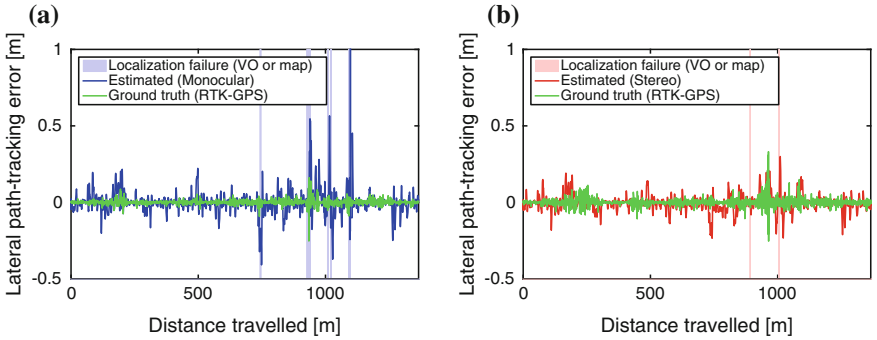

Fig. 6 Estimated and measured lateral path-tracking error during the monocular and stereo repeat passes of the 1.4 km outdoor route shown in Fig. 5 (Trial 3 in Table 1). GPS tracking shows that both monocular and stereo VT&R achieve centimetre-level accuracy, although estimated lateral path-tracking error tends to diverge from the true value in cases of localization failure. **a** Monocular repeat pass. **b** Stereo repeat pass

twice as many map features as its stereo counterpart. This result is most likely due to bad data association during local map construction in the monocular pipeline, which stems from the comparatively large positional uncertainties of distant keypoints.

Bad data association is especially problematic in regions of highly self-similar terrain (e.g., Fig. 11a) since large positional uncertainties exacerbate ambiguity in feature matches. With fewer correctly associated measurements, the bundle adjustment procedure will not maximally constrain the positions of map keypoints, which

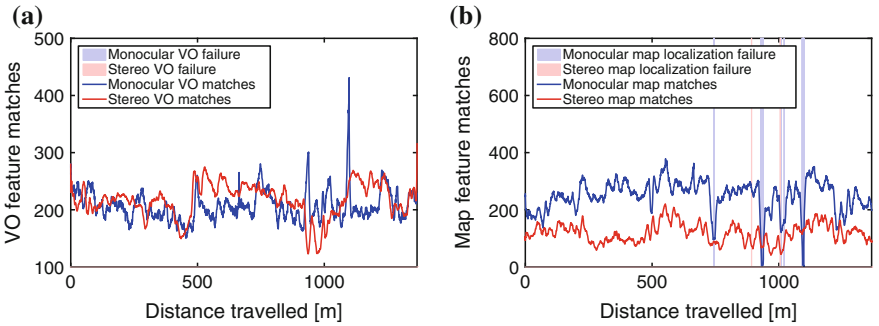

Fig. 7 Keypoint matches during the monocular and stereo repeat passes of the 1.4 km outdoor route shown in Fig. 5 (Trial 3 in Table 1), with localization failures highlighted. A localization failure is defined as less than 10 feature matches. There were no VO failures during either repeat pass. For clarity, we have applied a 20-point sliding-window mean filter to the raw data. **a** VO feature matches. **b** Map feature matches

we would expect to increase the risk of localization failures. Indeed, Fig. 7b shows that the monocular pipeline suffered more serious map localization failures than the stereo pipeline, although these forced manual intervention only once.

4.3 Indoor Experiments

The second set of experiments took place in the more challenging terrain of the UTIAS MarsDome. These routes included a number of highly non-planar features such as hills, large bumps, valleys, and slopes of a similar scale to the vehicle.

Since the MarsDome is an enclosed facility, GPS tracking was not available, and we instead made use of a Leica Nova MS50 MultiStation to track the position of the rover with millimetre-level accuracy. Similarly to the outdoor experiments, we used these data for ground-truthing purposes only. Figure 8 shows MultiStation data of the teach and repeat passes of a 140 m route through the MarsDome, along with images of some of the more challenging terrain features on the route.

Figure 9 shows estimated and measured lateral path-tracking errors for the monocular and stereo repeat passes. As in the outdoor case, both pipelines achieved centimetre-level accuracy, even in difficult terrain. 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 centimetres of the path throughout the traverse.

Figure 10 shows VO and map feature matches for both repeat passes. The monocular pipeline suffered map localization failures more often than the stereo pipeline, the worst failure occurring in the valley and hill regions (see Fig. 8) where the lighting

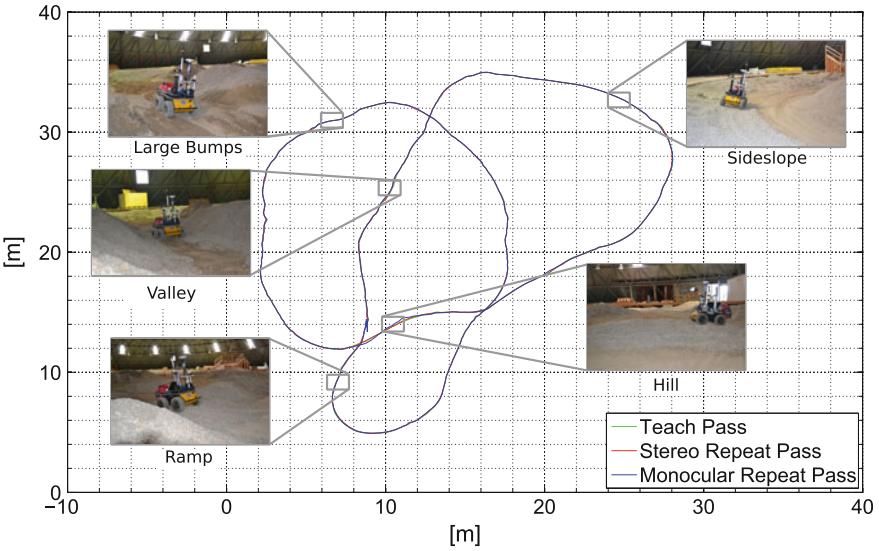

Fig. 8 Comparison of MultiStation tracks of the teach pass, stereo repeat pass, and monocular repeat pass of a 140 m indoor route (Trial 5 in Table 1), with some interesting segments highlighted

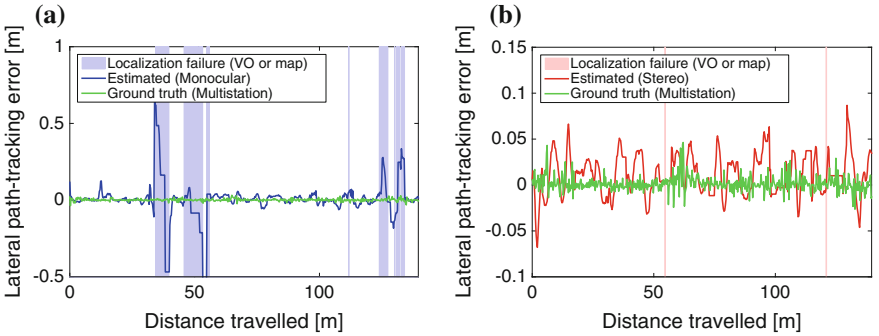

Fig. 9 Estimated and measured lateral path-tracking error during the monocular and stereo repeat passes of the 140 m indoor route shown in Fig. 8 (Trial 5 in Table 1). MultiStation tracking shows that both monocular and stereo VT&R achieve centimetre-level accuracy in highly non-planar terrain, although estimated lateral path-tracking error tends to diverge from the true value in cases of localization failure. Note the difference in scale between the two plots. **a** Monocular repeat pass. **b** Stereo repeat pass

was especially poor. This led to increased motion blur (see Fig. 11b) and poor feature matching due to greater uncertainty in keypoint positions. Both failures necessitated manual intervention over a few metres, however, the system successfully relocalized once the lighting improved.

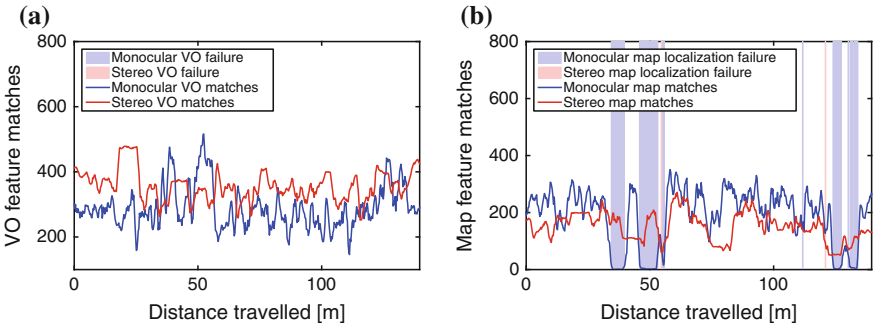

Fig. 10 Keypoint matches during the monocular and stereo repeat passes of the 140 m indoor route shown in Fig. 8 (Trial 5 in Table 1), with localization failures highlighted. A localization failure is defined as less than 10 feature matches. There were no VO failures during either repeat pass. For clarity, we have applied a 5-point sliding-window mean filter to the raw data. **a** VO feature matches. **b** Map feature matches

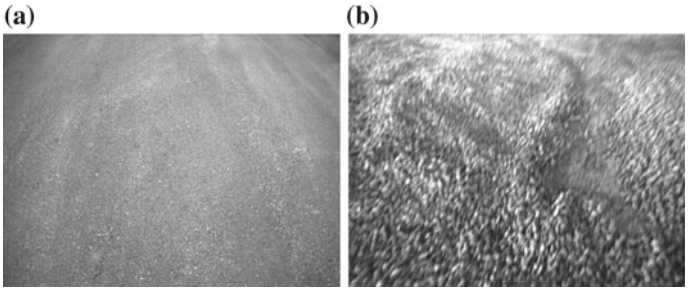

Fig. 11 The most common causes of localization failure were highly self-similar terrain and motion blur. Neither stereo nor monocular VT&R is immune to these conditions, but their effects were exacerbated by high spatial uncertainty in the monocular case. **a** Self-similar terrain. **b** Motion blur

5 Lessons Learned and Future Work

Experiments with our systems led to several useful lessons and possible extensions:

1. With sufficient spatial uncertainty, the flat-ground assumption seems to be usable even in rough driving conditions, provided the scene is well-lit and reasonably textured. Steep hills were problematic for monocular VT&R since the camera would observe features mainly on the horizon or on walls during the ascent.
2. The performance our systems depends on a search (often manual) through a high-dimensional space of tuning parameters, and it is difficult to be certain that an optimal configuration has been found. Iterative learning algorithms such as [17] may present a solution by learning optimal parameters from experience.
3. Data association quality is not a monotonic function of observation uncertainty. Too little uncertainty and good feature matches get rejected; too much and all

matches are equally good (or bad). Both cases result in tracking failure. This reinforces the need for an accurate model of a system's noise properties.
4. Experimenting with camera orientation could improve the accuracy of monocular VT&R, particularly on hills. For example, orienting the camera perpendicular to the direction of travel has been shown to improve the accuracy of stereo visual odometry [18].
5. By using stereo vision in the teach pass and monocular vision in the repeat pass, we could forgo the flat-ground assumption for mapping, which should result in fewer localization failures in the repeat pass.

6 Conclusions

This paper has described a Visual Teach and Repeat (VT&R) system capable of autonomously repeating kilometre-scale routes in rough terrain using only monocular vision. By constraining features of interest to lie on 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 VT&R system [8] on a wide range of vehicles equipped with monocular cameras. Extensive field tests have demonstrated that this system is capable of achieving centimetre-level accuracy on par with its stereo counterpart, but that there is an associated trade-off in robustness. Nevertheless, we believe that the benefit of deploying VT&R on existing vehicles without requiring the installation of additional sensors far outweighs the associated reduction in robustness.

Acknowledgments The authors would like to thank Matthew Giamou 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, Leica Geosystems for providing the MultiStation, and Clearpath Robotics for providing the Husky rover. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) through the NSERC Canadian Field Robotics Network (NCFRN).

References

1. Barfoot, T., Furgale, P.: Associating uncertainty with three-dimensional poses for use in estimation problems. IEEE Trans. Robot. (T-RO) **30**(3), 679–693 (2014)
2. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. (CVIU) **110**, 346–359 (2008)
3. Choi, S., Joung, J., Yu, W., Cho, J.: Monocular visual odometry under planar motion constraint. In: Proceedings of the International Conference on Control, Automation and Systems (ICCAS), pp. 1480–1485 (2011)
4. Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O.: MonoSLAM: real-time single camera SLAM. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **29**(6), 1052–1067 (2007)
5. Eade, E., Drummond, T.: Scalable monocular SLAM. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2006)

6. Farraj, F., Asmar, D.: Non-iterative planar visual odometry using a monocular camera. In: Proceedings of the International Conference on Advanced Robotics (ICAR), pp. 1–6 (2013)
7. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6) (1981)
8. Furgale, P., Barfoot, T.D.: Visual teach and repeat for long-range rover autonomy. J. Field Robot. (JFR) **27**(5), 534–560 (2010)
9. Goedemé, T., Nuttin, M., Tuytelaars, T., Gool, L.V.: Omnidirectional vision based topological navigation. Int. J. Comput. Vision (IJCV) **74**(3), 219–236 (2007)
10. Holmes, S.A., Murray, D.W.: Monocular SLAM with conditionally independent split mapping. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **35**(6), 1451–1463 (2013)
11. Kidono, K., Miura, J., Shirai, Y.: Autonomous visual navigation of a mobile robot using a human-guided experience. Robot. Auton. Syst. (RAS) **40**(2–3), 121–130 (2002)
12. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: Proceedings of IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR) (2007)
13. Lovegrove, S., Davison, A.J., Ibanez-Guzman, J.: Accurate visual odometry from a rear parking camera. In: Proceedings of the Intelligent Vehicles Symposium (IV) (2011)
14. Marshall, J., Barfoot, T.D., Larsson, J.: Autonomous underground tramming for center-articulated vehicles. J. Field Robot. (JFR) **25**, 400–421 (2008)
15. Matsumoto, Y., Inaba, M., Inoue, H.: Visual navigation using view-sequenced route representation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 83–88 (1996)
16. McManus, C., Furgale, P., Stenning, B., Barfoot, T.D.: Lighting-invariant visual teach and repeat using appearance-based Lidar. J. Field Robot. (JFR) **30**(2), 254–287 (2013)
17. Ostafew, C., Schoellig, A., Barfoot, T.: Iterative learning control to improve mobile robot path tracking in challenging outdoor environments. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 176–181 (2013)
18. Peretroukhin, V., Kelly, J., Barfoot, T.: Optimizing camera perspective for stereo visual odometry. In: Proceedings of the Conference on Computer and Robot Vision (CRV), pp. 1–7 (2014)
19. Quigley, M., Conley, K., Gerkey, B.P., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y.: ROS: an open-source robot operating system. In: Proceedings of the ICRA Workshop Open Source Software (2009)
20. Remazeilles, A., Chaumette, F., Gros, P.: 3D navigation based on a visual memory. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 2719–2725 (2006)
21. Royer, E., Lhuillier, M., Dhome, M., Lavest, J.M.: Monocular vision for mobile robot localization and autonomous navigation. Int. J. Comput. Vision (IJCV) **74**(3), 237–260 (2007)
22. Simhon, S., Dudek, G.: A global topological map formed by local metric maps. In: Proceedings of the IEEE/RSJ Intrnational Conference on Intelligent Robots and Systems (IROS), pp. 1708–1714 (1998)
23. Zhang, A.M., Kleeman, L.: Robust appearance based visual route following for navigation in large-scale outdoor environments. Int. J. Robot. Research (IJRR) **28**(3), 331–356 (2009)
24. Zhang, J., Singh, S., Kantor, G.: Robust monocular visual odometry for a ground vehicle in undulating Terrain. In: Proceedings of the Field and Service Robotics (FSR), pp. 311–326 (2012)
25. Zhao, L., Huang, S., Yan, L., Jianguo, J., Hu, G., Dissanayake, G.: Large-scale monocular SLAM by local bundle adjustment and map joining. In: Proceedings of the IEEE International Conference on Control, Automation Robotics and Vision (ICARCV), pp. 431–436 (2010)