2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

Level-Headed: Evaluating Gimbal-Stabilised Visual Teach and Repeat for Improved Localisation Performance

Michael Warren, Angela P. Schoellig, and Timothy D. Barfoot[1]

Abstract— **Operating in rough, unstructured terrain is an essential requirement for any truly field-deployable ground robot. Search-and-rescue, border patrol and agricultural work all require operation in environments with little established infrastructure for easy navigation. This presents challenges for sensor-based navigation such as vision, where erratic motion and feature-poor environments test feature tracking and hinder the performance of repeat matching of point features. For vision-based route-following methods such as Visual Teach and Repeat (VT&R), maintaining similar visual perspective of salient point features is critical for reliable odometry and accurate localisation over long periods. In this paper, we investigate a potential solution to these challenges by integrating a gimbaled camera with VT&R on a Grizzly Robotic Utility Vehicle (RUV) for testing at high speeds and in visually challenging environments. We examine the benefits and drawbacks of using an actively gimbaled camera to attenuate image motion and control viewpoint. We compare the use of a gimbaled camera to our traditional fixed stereo configuration and demonstrate cases of improved performance in Visual Odometry (VO), localisation and path following in several sets of outdoor experiments.**

I. INTRODUCTION

In order for field-robotic systems to be effective tools in applications such as search-and-rescue, border patrol and agricultural work, the sensor systems they use must be reliable in their intended environments. Rough, unstructured terrain is a frequent encounter in these deployments, and can challenge the ability of state-estimation algorithms that rely on sensing with limited update rates and limited perspective, such as Light Detection And Ranging (LiDAR) and vision. Additionally, vision-based systems generally rely on complex, textured terrain invariant to environmental change in order to accurately track features reliably over long time periods. Generally, research demonstrations of visual navigation on ground robots are heavily biased towards smooth trajectories and carefully planned viewpoints to avoid poorly textured surfaces in order to achieve robust results.

VT&R is a path-following algorithm capable of autonomously driving a robot by following a previously traversed route [1]. By extracting features taken from a monocular or stereo camera in the live view and matching them back to those from a perspectively similar 'teach' view, relative path tracking error can be computed and sent to a path tracking controller to drive the robot along the path. Computational complexity is constant with respect to map size,

[1]All authors are with the University of Toronto Institute for Aerospace Studies (UTIAS), University of Toronto, 4925 Dufferin St, Ontario, Canada {michaelwarren, angela.schoellig}@robotics.utias.utoronto.ca, tim.barfoot@utoronto.ca

Fig. 1: The experimental setup for gimbal-stabilised VT&R: (1) DJI Matrice 600 multirotor body, (2) DJI Ronin-MX 3-axis gimbal, (3) StereoLabs Zed stereo camera, (4) Lenovo W541 Intel Core-i7 laptop, (5) Clearpath Grizzly RUV.

which enables long-distance operation, but the algorithm is highly dependent on matching Speeded-Up Robust Features (SURF) or other similar point features, and hence is sensitive to both perspective and appearance change. One of the key advantages of VT&R, however, is that it exploits certain strengths of computer vision by keeping the viewpoint as similar as possible between repeat traverses of a path, i.e. as a crude version of active perception.

In extended field deployments, however, we have encountered situations that challenge our reliance on fixed cameras to achieve the goal of true long-term autonomy. These include areas of poor features directly in front of the vehicle (such as smooth ashphalt, sandy or snow-covered areas), where looking in an alternative direction would improve matching performance. More regularly in our outdoor deployments, driving on steep inclines causes the sky to fill a large portion of the field of view, meaning feature tracking is impossible. Finally, driving at high speeds results in general VO failures on rough terrain due to loss of feature tracks and poor localisation performance, which are highly

978-1-5386-3081-5/18/$31.00 ©2018 IEEE

coupled effects. For most activities, the robot is driven at well below walking pace and is still subject to a significant amount of careful perspective planning for reliable long-term operation. Addressing these deficiencies will result in a more reliable VT&R system and assist in bridging the gap between research testing and real-world deployments.

This paper presents a gimbaled VT&R system, where the on-board camera is stabilised against pitch, roll and yaw motions, and actuated when necessary during the teach pass to improve future localisation performance. We describe our gimbaled system, including both hardware and software changes, and we investigate the performance of this gimbaled setup in comparison to our traditional fixed-camera VT&R when performing high-speed manouvres and in perceptually difficult situations.

The rest of this paper is outlined as follows: Section II discusses related work in vision-based navigation, focusing on the active perception problem, Section III describes the VT&R methodology and the improvements implemented to use a gimbaled system with the framework, Section IV presents the experimental configuration and online testing framework for the gimbaled VT&R algorithm and shows the results of this testing. Section V discusses the impact of these results and the paper is concluded in Section VI.

II. PREVIOUS WORK

A large and well-known number of visual feature detectors and descriptors exist [2], [3], [4], [5] for robustly matching keypoints in imagery. However, they are all subject to performance degradation due to lighting and perspective change. Many features have rotationally invariant options, but their reliance on computing a consistent orientation means matching often suffers when using these versions. By restricting some dimensions of motion when matching, such as driving on smooth ground or using active perception, performance is often significantly improved. VT&R already performs a crude form of active perception when repeating, maintaining as similar a viewpoint for matching as possible. This ensures perspective differences are minimised when repeating a route.

For vision-only odometry systems, breaks in feature tracks and erratic odometry due to rotational motion cause significant error build up in pose estimates over long trajectories. Visual-Inertial Navigation Systems (VINS) [6], [7], [8] can address the problems caused by rough or extreme motions very well by using complementary measurements to address the deficiencies of each sensor. By updating state through inertial measurements during breaks in visual tracking, accuracy over very long and extremely irregular trajectories are possible. Similarly, semi-dense methods such as SVO [9] allow accurate pose tracking under fast motions through very high frame-rates and the ability to track very weak features on uniform surfaces.

However, both VINS and semi-dense odometry methods do not solve the perspective problem and the inherent brittleness of matching point features under perspective change. Even non-feature-based methods are susceptible to perspective changes (e.g., driving in a different lane) [10]. Topological methods that are more robust to appearance change exist [11], but at the expense of metric localisation, which is critical to the online performance of VT&R. Using hardware can also be an effective way to improve robustness. We have investigated this before by using multiple stereo cameras [12] to look in opposing directions, countering difficulties cause by sun-glare and uniform surfaces. Alternatively, omnidirectional or catadioptric cameras improve invariance to perspective, but with an associated loss of metricity and consequent accuracy.

Surprisingly, there is little (to our knowledge) published work that includes active gimballing in the performance of visual navigation on ground robots. Of course, visual servoing is a large and related field that addresses the issue of control to achieve a desired viewpoint of robot manipulators [13], [14]. Several approaches apply these concepts to mobile robots [15]. Our standard fixed-camera system can be considered a form of visual servoing, where, given the current pose and feedback from a visual sensor, a control system actuates the plant to pass through a set of desired viewpoints.

III. METHODOLOGY

In this paper, we use our well-established VT&R 2.0 software system as presented in [16] and extend this to take advantage of the gimbaled camera. As in [16], the autonomous driving algorithm consists of separate teach and repeat phases. During the teach phase, the vehicle is manually driven by a human operator along a desired route, while the VT&R algorithm performs passive visual odometry; inserting the visual observations from this privileged experience into a relative map of pose and scene structure. During the repeat phase, without reliance on Global Positioning System (GPS) or other sensors, the vehicle should autonomously re-follow the route by visually localising to the map of the privileged path. The vehicle repeats a path by sending high-frequency localisation updates to a path-tracking controller [17]. In the following sections, we describe our VT&R system extensions to use a gimbaled camera setup.

A. *Gimbaled Visual Odometry*

During both teach and repeat phases, image pairs are captured by a calibrated stereo camera at a frame rate of $\sim$15-20Hz (depending on setup), while the gimbal state (read as roll-, pitch-, and yaw-axis angular positions) is captured at approximately 30Hz. The gimbal state gives the pose of the camera in the vehicle frame by compounding the captured gimbal angles through a series of transforms with known translations. We denote the vehicle-to-sensor (camera) transform at time t as $\mathbf{T}_{sv(t)}$.

For each stereo image pair captured at time t, upright SURF features are extracted, descriptors generated and landmarks triangulated. Then, each feature in this latest frame-pair is matched to the last keyframe via SURF descriptor matching. The raw matches are then matched to landmarks using Maximum Likelihood Estimation SAmple Consensus

(MLESAC) to find the relative (temporal) transform between the current frame and latest keyframe's poses in the vehicle frame, using $\mathbf{T}_{sv}$ at their respective time points. Finally, the temporal transform is optimised using our Simultaneous Trajectory Estimation And Mapping (STEAM) bundle adjustment engine. A trajectory estimate is also estimated to extrapolate pose for the next frame, which is important for robust performance of our gimbaled system.

If a certain criterion is met, such as the number of inliers drops below a threshold or a component of 6-Degree of Freedom (DoF) motion exceeds a threshold, the frame is set as a keyframe and the features, new landmarks, and $\mathbf{T}_{sv}$ are stored in a vertex in a pose graph for future retrieval. The relative transform is stored as an edge to the previous vertex. The vertex is marked as privileged if generated during the teach phase. Windowed bundle adjustment is then performed on the last 5-10 vertexes. For a more thorough explanation of this component, we direct the reader to our previous work [16].

B. Localisation

During the repeat phase, after a new vertex is created and windowed bundle adjustment completes, a separate localisation process is run to estimate the spatial transform and tracking errors between the current vehicle's pose and the closest privileged vertex in the graph.

Here we introduce the 'localisation chain', a conceptual representation of the current robot pose, the spatially closest privileged vertex, the compounded transforms and the path between them through the graph. These are visualised in Figure 2 as a snapshot of a repeat run. The latest vertex (a.k.a., keyframe) in the repeat is V_b, the live frame is V_c, and the current closest privileged vertex to V_b is V_e. The chain is the transform $\mathbf{T}_{eb}$, compounded through the shortest path from V_e to V_b. If V_b has just been added to the graph but not yet localised, the chain would follow e, d, a, b. The chain is updated after every live frame to keep track of the current transform, $\mathbf{T}_{ce}$, and the current spatially closest privileged vertex is updated to V_f by searching forward along the priviledged edges' graph and finding when the translational component of $\mathbf{T}_{cf}$ is smaller than $\mathbf{T}_{ce}$.

Localisation follows a similar basic process to the keyframing component of visual odometry, but with some distinct differences. We describe this process as if V_b has just been added to the graph as a new vertex. First, a subgraph or window of vertexes (2-5 frames in both the forward and reverse directions) is extracted on the privileged (teach) path centered around the closest privileged vertex V_e. Within this window, all the landmarks from each vertex are migrated through their respective transforms to the privileged vertex V_e. This places them in a single local Euclidean frame centered at V_e. The respective descriptors of each migrated landmark are then matched in order of vertex hops from V_e to those in V_b. Positive matches and migrated landmarks are then used to find the new spatial transform $\mathbf{T}_{eb}$ via MLESAC. This transform is then inserted in the graph, and the localisation chain from V_e to V_b is updated with the new transform following the shorter path. While this description assumes a single live pass and single privileged run, we leverage multiple experiences and use landmarks from selected intermediate runs to improve localisation, as described in [16].

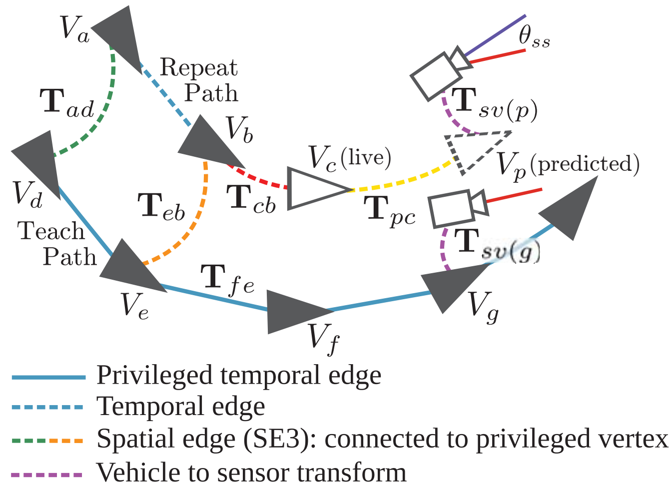


Fig. 2: Active gimbal control requires knowledge of the transform between V_g and V_p, denoted $\mathbf{T}_{pg}$, which was estimated by compounding the temporal edges through the last successfully localised transform $\mathbf{T}_{eb}$ (orange), and the estimated vehicle to sensor transforms $\mathbf{T}_{sv(g)}$ and $\mathbf{T}_{sv(p)}$, provided by the gimbal state. The uncertainties and some estimated transforms are omitted here for clarity.

C. Path-Tracking Control

We use the path-tracking controller first presented in [17] for following the teach pass during autonomous repeats[1]. For the current section of the path to which the robot is estimated to be closest, the path-tracking controller sets a desired forward speed based on the curvature of the privileged path at that pose. Typically, higher curvatures (smaller radius of turn) require slower speeds. For regular path following using our Clearpath Grizzly, we set a conservative speed profile, which can be described as a 'strolling' pace. Generally, faster speeds mean the likelihood of path-following errors (and consequently localisation failures) increase, as the fixed latency of the VO and localisation algorithms mean a delay in sending up-to-date cross-track errors to the path-tracking controller. For higher-speed operation, we set a speed profile with faster desired velocities for each curvature value. To achieve smooth and consistent performance, the desired speeds are subject to an acceleration constraint of 0.25m/s^2. This prevents large accelerations and decelerations when transitioning from tight turns to long, straight sections and vice versa, which tend to be a significant contributor to poor path-tracking accuracy.

[1]To remove the influence of the Iterative Learning Control (ILC) algorithm of the path tracker on localization results, the learning component is turned off for the experiments presented in this paper. By using the ILC algorithm, later experiments would be biased to improved path tracking by learning terrain abberations. However, the standard path-tracking controller at the core of our other VT&R work is still used.

D. Gimbal Control

There are two potential approaches to gimbal control from the perspective of VT&R; passive or active. In the passive strategy, the gimbal's internal controller will stabilise pitch and roll in the camera frame, while using a smoothed open-loop controller on yaw to maintain an approximately forward facing viewpoint. This requires no additional control inputs from the VT&R system.

In the active strategy, the gimbal can be commanded to reduce angular error between the current (live) view and nearest privileged view, if knowledge of the transform between the current and the privileged poses is known. For this to occur, we leverage the methodology described in the previous subsections and pictorialise the strategy in Figure 2. First, we utilise STEAM to extrapolate the estimated pose at the current time, given latency in the visual processing algorithm, and include an additional fixed timestep to account for transmission delays. Typically, the latency between frame capture and output transform estimate, $\mathbf{T}_{cb}$, is on the order of 60-100ms, but there is also a fixed delay of approximately 200ms between the sending of command and active motion of the gimbal due to message transmission latencies. We defined this extrapolated (predicted) pose as V_p.

Given $\mathbf{T}_{pc}$, the localisation chain is queried for the nearest privileged pose in relation to V_p, denoted as V_g. The relative pose between V_g and V_p is then compunded through the transforms:

$$\mathbf{T}_{pc},\ \mathbf{T}_{cb}, \mathbf{T}_{be},\ \mathbf{T}_{ef},\ \mathbf{T}_{fg} \tag{1}$$

This is the equivalent of the prior as stated in [16]. Of course, V_f and V_g could be equivalent, and transforms are compounded as needed on the shortest path between V_g and V_p.

The desired transform denoting the positional and rotational error between the closest privileged view and the live view in the sensor frame of V_g is defined as:

$$\mathbf{T}_{ss} = \mathbf{T}_{sv(p)}\mathbf{T}_{pg}\mathbf{T}_{sv(g)}^{-1} \tag{2}$$

Given the rotational component of this transform, the desired gimbal angle at each of the control axes is sent to the gimbal controller. In the active control case, the roll is left as open-loop control, but pitch and yaw are closed-loop.

Both the active and passive strategies are trialled in Section IV to justify why active gimbal control is required in order to outperform a static camera system.

IV. EXPERIMENTS

We used a Clearpath Grizzly RUV (Figure 1) as the base platform for two separate experiments. In the first experiment, we examine the performance of the gimbaled system specifically in high-speed driving. In the second experiment, we examine the performance of the gimbaled system in a long-term localisation scenario in tough outdoor conditions.

A. High-Speed Driving

In this experiment, the Grizzly RUV was fitted with our standard Point Grey XB3 camera system, placed facing forwards on a central mast, as this will form the baseline to which we will compare our gimbaled system. Specific to this paper, we also rigidly mounted a DJI Matrice 600 multirotor body and DJI Ronin-MX gimbal to the mast of the Grizzly (Figure 1). The Matrice 600 provides the interface to the gimbal via a serial connection and sends the gimbal state (joint-angles) at up to 100Hz to a resolution of $0.1°$. A Point Grey Bumblebee2 (BB2) stereo camera was placed on the gimbal, facing forwards. The gimbaled system was placed specifically to closely match the mounting position of the XB3 to maintain as similar a perspective as possible, and the gimbal pitch was angled approximately $20°$ below horizontal to match that of the XB3. Each camera was connected to a separate Lenovo W541 laptop (8-core Intel Core-i7) for data processing purposes.

For this experiment, we used the short baseline of the XB3 camera (central and right cameras) to ensure fair comparison with the BB2 camera, which has the same 12cm baseline. The gimbal was commanded to actively stabilise pitch and roll, and both passive and active yaw control strategies were tested. For the active strategy, the gimbal was not controlled by the operator during the teach phase but used the closed-loop controller during repeats. Both the BB2 and XB3 were configured to generate images at 16Hz for comparative purposes, and both have an approximately $65°$ horizontal Field Of View (FOV).

The comparative performance of the fixed and gimbaled systems were evaluated through a set of high-speed driving tests (Table II). For this comparison, a route covering approximately 100m was initially taught to both the fixed (XB3) and gimbaled (BB2) camera systems covering the same trajectory (shown in Figure 3). Once the teach pass was complete and the endpoints of the path merged manually, the path was autonomously repeated multiple times over several hours while switching between actively gimbaled, passively gimbaled, and fixed camera systems, each controlling the vehicle during their respective tests. We deonte these different strategies as 'fixed' for the traditional fixed XB3 rig, 'passive' for the passive gimballing strategy, and 'active', which uses our active gimbal control algorithm. The system was driven at the highest possible target speed given acceptable safety limits,

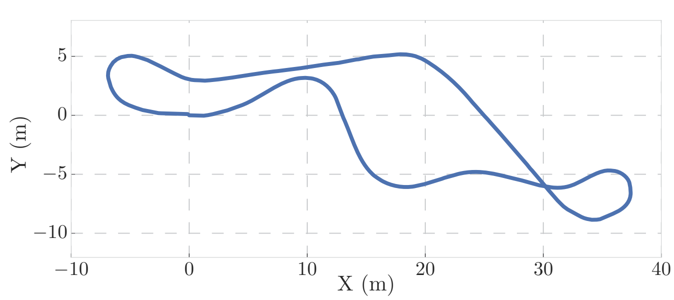


Fig. 3: The path driven for the high-speed driving experiment. The robot is driven to cover a range of path curvatures and rotational directions, including hairpins, slaloms and straight sections.

TABLE I: Target speed profiles for differing radius of path curvature. The maximum speed (at minimum curvature) can be described as a 'running' pace.

Curvature	0.01	0.2	1.5	5.0	10.0	m^{-1}
Target Speed	2.25	2.75	3.5	4.25	5.0	m/s

TABLE II: Summary of the trials used for the experimental results

Trial Number	Time started	Configuration
Teach	13:00	passive+fixed
1, 2, 3	13:34, 13:36, 13:38	fixed
4, 5, 6, 7, 8	13:(46, 49, 51, 54, 56)	active
9, 10, 11	14:00, 14:05, 14:08	fixed
12, 13, 14, 15, 16	14:(21,26,29,30,32)	passive

up to 5m/s (corresponding to path curvature as described in Section III-C). Each experiment or trial was repeated several times for consistency of results.

B. High-Speed Driving Results

To compare results and quantify the improvement of performance of the gimbaled camera system, we evaluate several different metrics: average feature track length during VO, total localisation matches, average localisation uncertainty, and camera actuation error.

Figure 4 shows the distribution of feature track lengths (the number of consecutive frames over which a landmark is matched) during VO for the three different strategies over all repeat runs. This figure shows that a gimbaled system improves the average track length over a static system by attenuating large image motions that cause tracks to break. The passive scheme can be seen to have a slight advantage over the active gimballing scheme, potentially due to the smoother operation of the gimbal's yaw control.

Turning to localisation, Figure 5 shows the mean number of MLESAC inliers for each keyframe for the three different strategies at each of the velocity profiles, and their 1-sigma standard deviation. While the mean number of inliers is consistent across all three strategies, this is merely reflective of the matching strategy in use, which collects a minimum number of feature matches (by expanding the vertex search space) before progressing to the MLESAC stage. The more important metric is the variance of the inliers for each strategy. Clearly, the passive strategy is inferior to the standard fixed-camera configuration, while active gimballing shows a small improvement over the same.

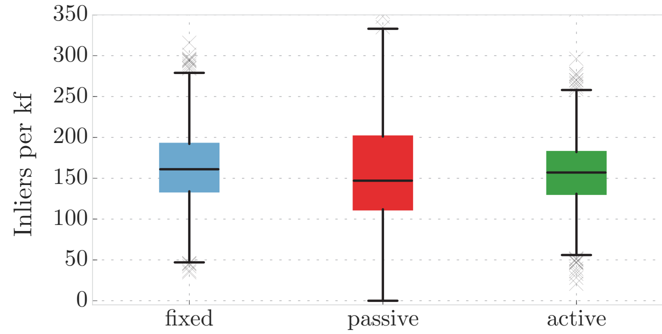


Fig. 5: Mean localisation inliers at each localised keyframe for the three strategies. The markers are offset for clarity. While the mean inliers remains roughly equal on average, the standard deviation is significantly larger under the passively gimbaled configuration, whereas the actively gimbaled configuration achieves a slightly better variance than the fixed setup at the fastest speed profile.

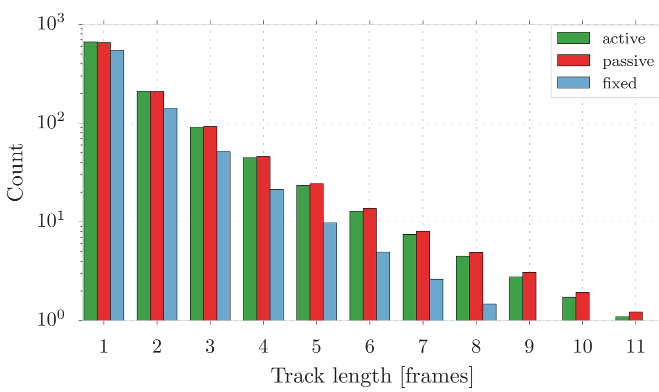


Fig. 4: Histogram of total VO feature track lengths for the three strategies during the high speed experiment.

Figure 6 shows the relative localisation yaw error taken from $\mathbf{T}_{ss}$ (Eq. 2) for the different strategies over the same path. While the fixed and passive strategies show large angular errors due to the open-loop control, the active gimballing strategy successfully attenuates large angular deviations during sections of poor path following.

Finally, the CDF of the localisation uncertainty for the same experiments is plotted in Figure 7. In this setup, the actively gimbaled system marginally improves localisation uncertainty throughout the dataset over both a fixed and passively gimbaled strategy. This can be attributed to two major components, the improved tracking performance of features in VO, meaning that landmarks are better triangulated with less average uncertainty compared to the fixed camera strategy, and more consistent perspective for localisation matching over the passive strategy. While the passive strategy will occasionally achieve reduced perspective error, performance in this metric is generally less consistent. In both the active and fixed strategies there were no localisation failures, but the passive strategy exhibited a 99.7% success rate.

C. Long-Term Experiment

In this experiment, the ability of the gimbaled system to decrease the chance of localisation failure was tested during a 'grand-tour' of deliberately challenging conditions. A dataset was gathered on the robot over a period of two days within the meadows surrounding the University of Toronto Institute for Aerospace Studies (UTIAS) campus during mid-winter. At this time of year, the sun remains low in the sky, meaning sun glare and consequent image washout is a frequent occurrence. Also, snow cover is often significant and variable from day to day, meaning that certain areas have little salient texture and features change rapidly as snow falls and melts.

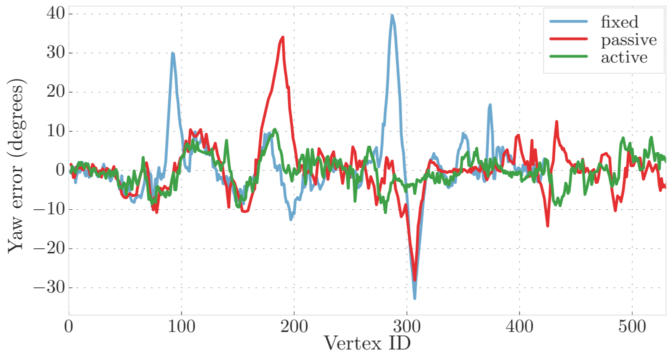


Fig. 6: Plot of relative localisation yaw error for the different strategies over three selected runs. While the fixed and passive strategies show large angular errors due to the open-loop control, the active gimballing strategy successfully attenuates large angular deviations during sections of poor path following.

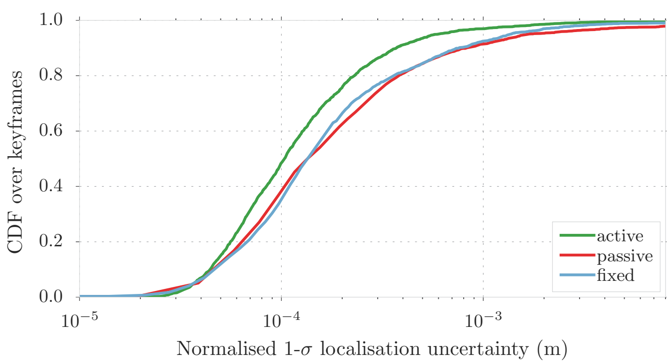


Fig. 7: The CDF of translational localisation uncertainty during the high-speed experiment. The active gimballing strategy outperforms both the fixed and passive strategies.

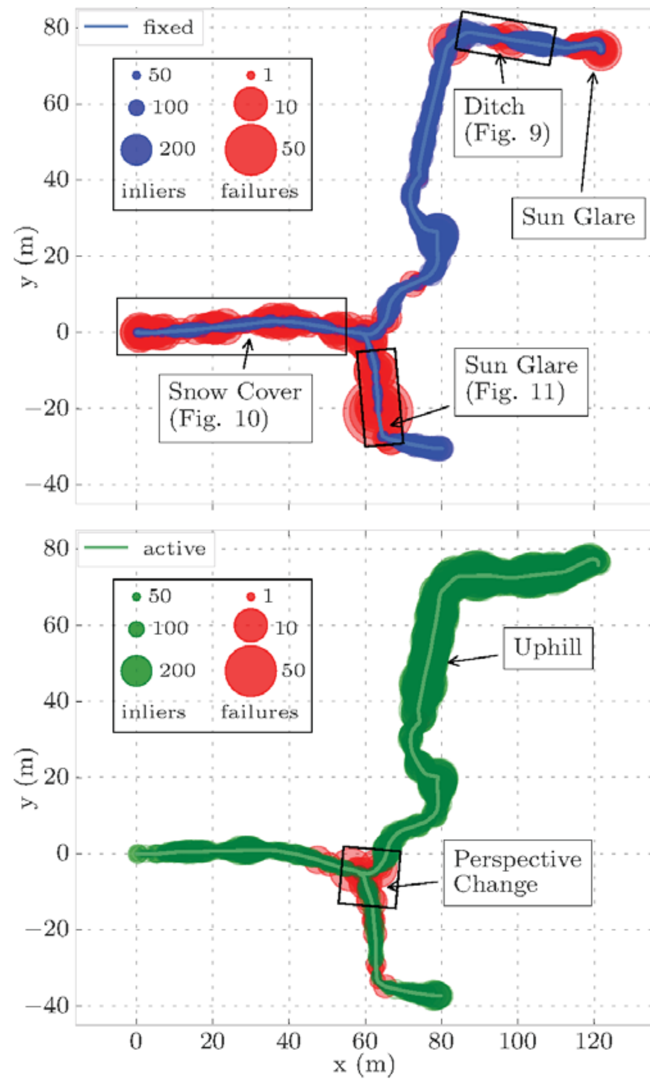


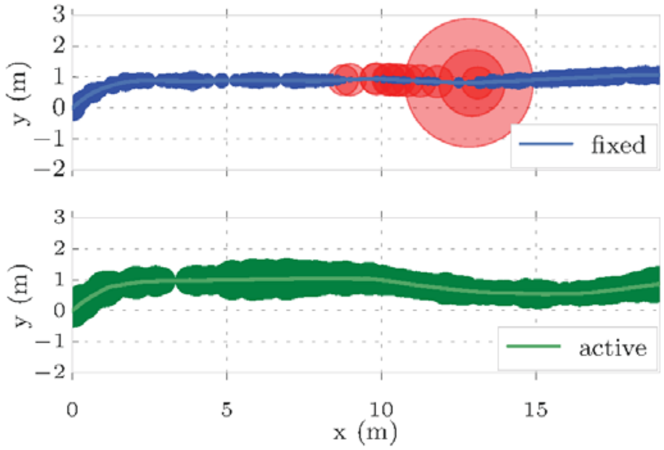
Fig. 8: Visualisation of the route traversed for experiment 2 for the fixed (top) and active (bottom) camera strategies, offset to highlight differences. Path thickness denotes average localisation inliers over all repeats. Red circles denote number of localisation failures per vertex (log_2 scale). Individual segments are highlighted for further examination.

For this experiment, a StereoLabs Zed stereo camera was mounted on the gimbaled system, and the robot was taught a series of routes during the afternoon of the first day. The robot was first driven with the actively gimbaled camera across a snow-covered field where little permanent vegetation or structure was visible directly in front of the camera, followed by driving directly towards the sun, and then across a highly textured area to a large ditch through which a seasonal creek runs. In each of these sections, the camera was actuated by the operator (i.e., manually steered) to avoid the source of the degeneracy during the teach phase. This is in contrast to the high-speed experiments, where the gimbal was not actuated by the operator. In the snow-covered section, the camera was pointed towards vegetation, avoiding the poorly textured terrain covered by snow. When driven towards the sun, the camera was aimed below the horizon to avoid sun glare. In the ditch, the camera was pointed down to reduce the percentage of the image covered in sky, but also allowed to automatically control pitch to maintain a stable orientation. These sections are visible in Figure 8.

The same set of routes was immediately re-taught while fixing the camera statically to face forward. In none of the aforementioned conditions was the fixed camera's orientation changed. The robot was then commanded to re-traverse the route multiple times to build a set of experiences from which to match features using both the fixed and active strategies. The next morning, the robot was commanded to re-follow the path using the experiences from the previous day. We record specific statistics for localisation throughout the experiment, such as the number of localisation inliers and localisation failures at each vertex of the traversed route. In total, each section of the route was traversed autonomously at least three times over the two day experiment. The passive strategy was not used in this experiment.

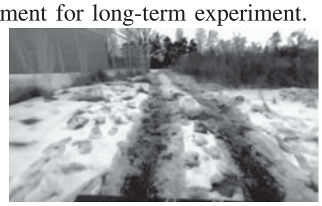
D. Long-Term Experiment Results

Results of the experiment are presented in Figures 8-11. Figure 8 shows the overall path followed by the robot for the two strategies. Each path is annotated with a circle at each vertex whose size denotes the average localisation inliers including all repeats. Larger-diameter circles show a greater average number of inliers at that vertex. Additionally, red circles are placed at each vertex where a localisation



(a) Zoomed view of ditch segment for long-term experiment.

(b) Fixed camera view.



(c) Active camera view.

Fig. 9: Examination of ditch segment for the fixed (left) and active (right) camera strategies. The gimbaled camera actively avoids looking at the sky and maintains a fixed pitch during the traversal.

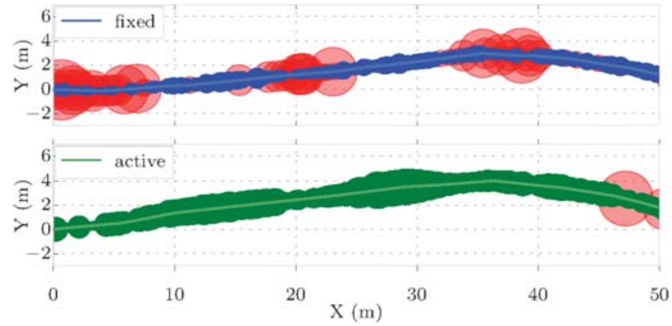
failure occurred, whose size denotes the average number of localisation failures, following a log_2 scale. During repeats, occasional failures do not necessarily require manual intervention due to reliance on VO for forward propogation, but a series of failures will ultimately cause the algorithm to exceed certainty bounds and stop the vehicle. Interesting segments of the path as described in Section IV-C are examined in Figures 9-11.

In Figure 9, the traversal of the ditch is highlighted. In all attempts, the fixed camera system failed to localise on the upward side of the path due to a significant view of sky. Each traversal required a manual intervention. Using the gimbaled system, large sky views were avoided and no failures occur. Additionally, the average number of localisation inliers at each vertex remained high and stable.

In Figure 10, the snow-covered section is highlighted. Similarly to the ditch example, the average number of localisation inliers at each vertex remained high and stable using the gimbaled system. In contrast, the fixed system again shows reduced inliers and far more frequent failures.

Finally, Figure 11 highlights the section affected by sun glare. Both strategies suffered in this section due to image saturation from directly viewing the sun. However, the actively gimbaled system, by pointing away from the horizon, showed better performance and was able to successfuly repeat the path without manual intervention. In contrast, the static camera system failed in all cases and required manual driving over a 4m section.

Overall, the active strategy resulted in fewer localisation failures and generally higher averages of feature inliers compared to the passive strategy. In the uphill section of



(a) Zoomed view of snow-covered segment for long-term experiment.

(b) Fixed camera view.

(c) Active camera view.

Fig. 10: Examination of low-texture segment for the fixed (left) and active (right) camera strategies. The active strategy shows reliably more inliers over the segment, with far fewer localisation failures compared to the fixed camera strategy.

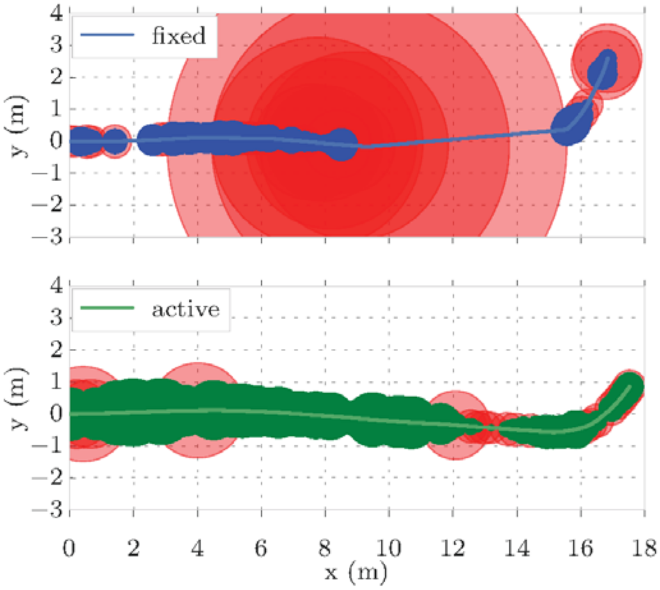
Figure 8, even small amounts of sky-view caused the passive strategy to suffer due to fewer features in the upper portion of the image typically covered by trees.

V. DISCUSSION

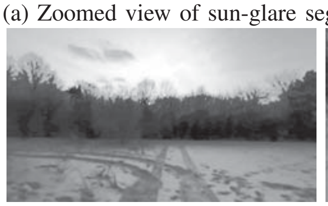
From these experiments, the use of a gimbal to address operational limits has some intriguing outcomes. Surprisingly, the ability of the gimbal to improve performance during high-speed maneuvres was marginal. While such a system is able to attenutate low-rate disturbances, the control envelope is insufficient to address large shocks that cause motion blur, which was a motivating use case of this system.

However, significant improvements are obvious when taking advantage of the gimbal's ability to account for perspective. In the high-speed experiments, the gimbal was able to attenuate yaw error caused by poor path tracking (particularly on corners), while in the long-term experiment, active targeting of feature-rich and feature-stable areas drastically improved reliability over a static system. It is less effective in areas of rich texture. Additionally, little improvement was seen during fast turns.

Importantly, during the teach phase the gimbal must be controlled smoothly by the operator while avoiding sudden perspective changes. This is highlighted in the junction of Figure 8, where discontinuities in perspective caused frequent localisation failures for the active strategy. This requires the operators to be careful in planning where to point the camera during the teach phase. Our results suggest that further research into an autonomous attention model to actively point the camera may be fruitful.



(a) Zoomed view of sun-glare segment for long-term experiment.



(b) Fixed camera view.

(c) Active camera view.

Fig. 11: Examination of sun-glare segment for the fixed (left) and active (right) camera strategies. Path thickness denotes average localisation inliers over all repeats. Red circles denote number of localisation failures per vertex (log scale). The active strategy shows reliably more inliers over the segment, with far fewer localisation failures compared to the fixed camera strategy.

VI. CONCLUSIONS

In this paper, we have shown the integration of a gimbaled camera to VT&R and evaluated the performance of localisation and path following in various conditions, including high-speed driving and difficult localisation experiments. It has been shown that an actively gimbaled setup assists marginally in both VO performance and localisation uncertainty during high-speed driving, but shows significant improvement when faced with specific difficult cases of perspective that degrade the performance of a fixed camera.

Future work will focus on reducing the computational load for deployment on low-power embedded hardware and subsequent demonstration of the gimbaled VT&R on-board the DJI M600 multirotor vehicle.

ACKNOWLEDGMENT

This work was funded by Smart Computing for Innovation Consortium (SOSCIP), Defense Research and Development Canada (DRDC), Drone Delivery Canada (DDC) and the Centre for Aerial Robotics Research and Education (CARRE), University of Toronto.

REFERENCES

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

[2] D. G. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision, The Proceedings of the*, pp. 1150–1157, 1999.

[3] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robust Features," in *European Conference on Computer Vision*, 2006.

[4] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary Robust invariant scalable keypoints," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2548–2555, 2011.

[5] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *International Conference on Computer Vision*, pp. 2564–2571, 2011.

[6] M. Li and a. I. Mourikis, "High-precision, consistent EKF-based visual-inertial odometry," *The International Journal of Robotics Research*, vol. 32, pp. 690–711, jun 2013.

[7] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "IMU Preintegration on Manifold for Efficient Visual-Inertial Maximum-a-Posteriori Estimation," in *Robotics: Science and Systems*, 2015.

[8] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual-inertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, 2014.

[9] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "SVO: Semidirect Visual Odometry for Monocular and Multicamera Systems," *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 249–265, 2017.

[10] E. Pepperell, P. I. Corke, and M. J. Milford, "Automatic Image Scaling for Place Recognition in Changing Environments," in *International Conference on Robotics and Automation*, (Seattle), 2015.

[11] N. Sünderhof, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free," *Robotics Science and Systems*, 2015.

[12] M. Paton, F. Pomerleau, and T. D. Barfoot, "Eyes in the Back of Your Head: Robust Visual Teach & Repeat Using Multiple Stereo Cameras," in *Proceedings -2015 12th Conference on Computer and Robot Vision, CRV 2015*, pp. 46–53, 2015.

[13] P. I. Corke, "Visual control of robot manipulators - a review," *Visual Servoing*, vol. 7, pp. 1–31, 1994.

[14] B. Espiau, F. Chaumette, and P. Rives, "A new approach to visual servoing in robotics," *Robotics and Automation, IEEE Transactions on*, vol. 8, no. 3, pp. 313–326, 1992.

[15] L. Mejías, S. Saripalli, P. Campoy, and G. S. Sukhatme, "Visual servoing of an autonomous helicopter in urban areas using feature tracking," *Journal of Field Robotics*, vol. 23, no. 3-4, pp. 185–199, 2006.

[16] M. Paton, K. Mactavish, M. Warren, and T. D. Barfoot, "Bridging the Appearance Gap : Multi-Experience Localization for Long-Term Visual Teach and Repeat," in *Intelligent Robots and Systems (IROS)*, 2016.

[17] C. J. Ostafew, A. P. Schoellig, and T. D. Barfoot, "Visual Teach and Repeat, Repeat, Repeat: Iterative Learning Control to Improve Mobile Robot Path Tracking in Challenging Outdoor Environments," in *Intelligent Robots and Systems (IROS)*, pp. 176–181, IEEE, 2013.