Eyes in the Back of Your Head: Robust Visual Teach & Repeat
Using Multiple Stereo Cameras

Michael Paton, François Pomerleau, and Timothy D. Barfoot
University of Toronto Institute for Aerospace Studies
{mpaton,francois.pomerleau}@robotics.utoronto.ca, tim.barfoot@utoronto.ca

Abstract—Autonomous path-following robots that use vision-based navigation are appealing for a wide variety of tedious and dangerous applications. However, a reliance on matching point-based visual features often renders vision-based navigation unreliable over extended periods of time in unstructured, outdoor environments. Specifically, scene change caused by lighting, weather, and seasonal variation lead to changes in visual features and result in a reduction of feature associations across time. This paper presents an autonomous, path-following system that uses multiple stereo cameras to increase the algorithm field of view and reliably navigate in these feature-limited scenarios. The addition of a second camera in the localization pipeline greatly increases the probability that a stable feature will be in the robot’s field of view at any point in time, extending the amount of time the robot can reliably navigate. We experimentally validate our algorithm through a challenging winter field trial, where the robot autonomously traverses a 250 m path six times with an autonomy rate of 100% despite significant changes in the appearance of the scene due to lighting and melting snow. We show that the addition of a second stereo camera to the system significantly increases the autonomy window when compared to current state-of-the-art path-following methods.

Keywords—Computer Vision, Field Robotics, Localization, Long-Term Autonomy, Autonomous Path Following.

I. INTRODUCTION

Autonomous path-following robots that operate without the use of a Global Positioning System (GPS) over large distances and times are appealing for many potential mobile robot applications. Vision-based algorithms have allowed scalable navigation using inexpensive, commercial sensors, but suffer from appearance change in outdoor environments.

An example is Stereo Visual Teach & Repeat (VT&R) [1], a path-following algorithm that is capable of autonomously repeating a previously driven path using a stereo camera. Computational complexity is constant with respect to map size enabling long-distance operation, but the algorithm is highly susceptible to scene change. This issue of scene change due to variation in lighting has been partially overcome through a lighting-resistant version of VT&R [2], which extends the algorithm’s operational window from a few hours to multiple days in varying lighting conditions. However, this algorithm does not account for general appearance change and was trained and tested for a specific environment.

Localization across appearance change is especially difficult for vision-in-the-loop navigation systems such as VT&R, where the control loop of the robot demands constant, metric localization results from the vision system to keep the robot driving. Recent methods that address appearance change are not suitable for this requirement, providing only topological localization [3] or requiring offline collection of the scene in multiple appearances [4, 5].

This paper proposes the addition of a second stereo camera to the VT&R pipeline to increase the robustness of the system against appearance change by providing a wider field of view to detect visual features. This multi-stereo VT&R system independently tracks SURF features from each stereo camera and combines data associations into a single state estimation problem. Combined data associations are used to estimate the state of the front camera through outlier rejection and nonlinear refinement.
We experimentally validate our algorithm through a field trial during the Canadian winter, where the low elevation of the sun and melting snow accelerate appearance change. Using the multi-stereo VT&R algorithm, the robot autonomously repeated a 250 m path six times, between noon and sunset, with an autonomy rate of 100% despite significant appearance change. We furthermore show through post-field analysis that this performance would not have been possible using the previously published VT&R methods. An example of the winter environment and hardware setup used in this field trial is displayed in Figure 1.

The remainder of the paper is organized as follows. Work related to VT&R and navigation with multiple cameras is outlined in Section II. The multi-camera VT&R system is detailed in Section III. Experimental work is described in Section IV. Results of the field trial are in Section V. The paper finishes with a conclusion and future work in Section VI.

II. RELATED WORK

A. Multi-Camera State Estimation

Work on multi-camera state estimation can be broken into two categories: systems that model multiple cameras as a single generalized camera, and systems that treat each camera independently. Our VT&R system falls into the latter category, where multiple stereo cameras are treated independently, and are used together to solve for a single pose estimate.

Systems that model multiple cameras as one are typically based on the Generalized Camera Model formulated by Pless [6]. The use of plucker lines to model point correspondences between cameras allow this model to solve for extrinsic calibration parameters using a generalized essential matrix. Lee et al. [7] estimate the motion of self-driving cars with four non-overlapping monocular cameras. With inter- and intra-point correspondences, they solve for the generalized essential matrix with a 2-Point RANSAC scheme and nonlinear refinement. Heng et al. [8] present a full Micro Aerial Vehicle (MAV) SLAM system including autonomous calibration of extrinsic parameters between cameras. Their system setup consists of an IMU and four monocular cameras placed in a dual-stereo configuration. To calibrate, they fly the vehicle in a pattern while performing dual-stereo bundle adjustment. The generalized camera model is used for the SLAM problem, when all four cameras are treated as one. Kneip et al. [9] formulate a general solution to multi-camera state estimation that is computationally more efficient than previous methods. They present a parametrization of the generalized camera model that is non-iterative and linear in complexity with respect to the number of points. They show tests in simulation and on a real camera system.

The alternative to a Generalized Camera Model for multi-camera state estimation systems is to formulate the system as a set of independent camera sensors. Oskiper et al. [10] perform Visual Odometry (VO) using dual stereo-cameras and an Inertial Measurement Unit (IMU). Motion is estimated through independent stereo pipelines. Using known extrinsic parameters, pose estimates from each camera are evaluated on all point correspondences. At each step, the estimate with the smallest reprojection error is used. Clipp et al. [11] build a six-degree-of-freedom (6DOF) motion estimation system using clusters of non-overlapping monocular cameras. Each camera performs independent state estimation through a 5-Point RANSAC algorithm. Any inter-camera correspondences are then used to solve for scale using a 1-Point RANSAC solution. At each step, the best estimate is used. Kazik et al. [12] estimate motion with two non-overlapping monocular cameras. Monocular VO is first performed individually on each camera up to scale. Enforcing the known transforms between cameras, they derive a linear least-squares problem to solve for the scale of the VO transformations on each camera. They use multi-frame estimation to improve accuracy. Motion estimates from each camera are then fused to obtain the final 6DOF motion estimate.

Our multi-stereo system is similar to the dual-stereo VO setup described in Oskiper et al. [10], with the exception that we use point correspondences from both stereo cameras to form a single pose estimate. This allows for the minimum number of required features to be spread across both cameras, allowing for localization in feature-limited environments. We also employ our vision system inside of a control loop for autonomous path following.

B. Increasing Robustness in Autonomous Path Following

The autonomous path-following algorithm presented in this paper is built upon the Stereo VT&R work presented by Furgale and Barfoot [1]. During the teach phase, the robot drives a path while building a pose graph consisting of Speeded Up Robust Features (SURF) [13] with 3D information linked by relative transformations obtained from VO. To autonomously repeat the path, features from the live view are associated to those in the map. While this system’s complexity is constant with respect to the size of the map, performance outdoors quickly degrades due to scene change.

The issue of lighting change can be mitigated with the use of color-constant images, which partially remove the effects of lighting change from the scene. Paton et al. [2] build a lighting-resistant VT&R pipeline using these images. By fusing data correspondences between greyscale images and two experimentally found color-constant images, they show a significant increase in reliability. They provide experimental results of a robot autonomously driving a 1 km path 26 times over the course of four days across significant lighting change in an unstructured, outdoor environment. While this method is resistant to variation in lighting, it is still susceptible to general appearance change from factors such as weather and seasons.
Another method of dealing with lighting is the use of an active sensor. McManus et al. [14] build a lighting-invariant VT&R pipeline using intensity images generated from a LiDAR. While this method is capable of operation over a complete 24 hour cycle, it suffers from motion distortion. Krusi et al. [15] perform autonomous path following through dense, point-cloud registration at the cost of potential failure cases in open spaces that lack geometric features. Vision-based path-following algorithms do not share these limitations, but are less stable in terms of appearance change.

III. MULTI-STEROE VT&R

This section provides a brief overview of the multi-stereo VT&R system, displayed in Figure 2, which involves the following components: (i) Independent Stereo Tracking, (ii) Map Representation, (iii) Localization and VO, and (iv) Autonomous Navigation.

A. Independent Stereo Tracking

The multi-stereo VT&R algorithm fuses synchronized inter-camera data correspondences observed from multiple stereo cameras into one state estimation problem. This is performed through independent detection and tracking of visual features in each stereo camera. This process is detailed in the upper section of Figure 3. The input to the system is a left-right stereo image pair observed at time \( k \) and the output is a set of data correspondences between the input stereo pair and either the previous frame in the case of VO or a map frame in the case of localization.

The first step of stereo tracking is the extraction of keypoints representing SURF visual features with associated 3D position information and uncertainty. Given a left-right image pair, extracted at time, \( k \), from a stereo camera with the reference frame, \( \mathcal{F}_{ki} \), where \( i \in \{1, 2\} \), keypoints are extracted by detecting SURF features, and matching them between left and right to obtain image coordinates, \( \{y_{j,k}, p_{j,k}^{i,ki}\} \), and a 64-dimensional feature descriptor, \( d_{j,k} \). Coordinates of the \( j \)th feature at time \( k \) are of the following form:

\[
y_{j,k} = \begin{bmatrix} u \\ v \\ d \end{bmatrix}, \quad p_{j,k}^{i,ki} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}
\]

where \( y_{j,k} \) represents the disparity coordinates such that \( u \) and \( v \) are, respectively, the horizontal and vertical pixel coordinates, and \( d \) is the disparity between the left and right image, \( p_{j,k}^{i,ki} \) is the physical location of landmark \( j \) in the coordinate frame of camera \( ki \). This value is a vector from the origin of \( \mathcal{F}_{ki} \), to the origin of \( \mathcal{F}_{j} \) (denoted by the superscript) and expressed in \( \mathcal{F}_{ki} \) (denoted by the subscript). Camera coordinates can be converted between disparity and physical location using the standard stereo camera model and its inverse.

In order to fuse data correspondences between cameras, they must be in the same coordinate frame. Because all state estimation is performed in the reference frame of the front stereo camera, keypoints in the coordinate frame of the rear camera, \( \mathcal{F}_{ki} \), are converted to the coordinate frame of the front camera, \( \mathcal{F}_{ki} \), using the transformation, \( T_{1,2} \):

\[
p_{j,k}^{i,ki} = T_{1,2}p_{j,k}^{i,k2},
\]

where \( T_{1,2} \) is the manually calculated extrinsic transformation assumed to be known a priori. Keypoints in the coordinate frame of the front camera remain unmodified.

The final step of the stereo tracking process is to match keypoints from the current view to either the previous frame in the case of VO, or the map in the case of localization. In both cases, the end result is a list of corresponding keypoints with 3D position information in the coordinate frame of the front camera.

B. Map Representation

The map scheme employed in the multi-stereo VT&R algorithm is similar to the one described by McManus et al. [14]. A map consists of a topometric pose graph of keypoints linked by relative transformations. Vertices in this pose graph are synchronized sets of keypoints obtained via independent stereo tracking from each camera. Edges in the map are relative transformations between vertices computed by multi-stereo VO. Vertices are constructed when the robot's motion exceeds a specified threshold, forcing an evenly distributed map.

During localization, the robot matches keypoints from the live view to a small submap constructed from the pose graph. This subgraph is a set number of vertices centered at the estimated closest vertex relaxed into a single global coordinate...
Stereo Tracking:

![Diagram of Stereo Tracking]

Localization / VO:

![Diagram of Localization / VO]

Figure 3. Pipelines of the multi-stereo VT&R system. Top: Independent stereo tracking. The input to the system is a pair of greyscale stereo images. Tracking begins with extraction of SURF features that are matched between stereo images to obtain depth. Features are then transformed into the coordinate frame of the front stereo camera. Keypoints are matched between either the previous frame in the case of VO, or the map in the case of localization to obtain a set of data correspondences. Bottom: Localization / VO Pipeline. The input to the system is a set of synchronized stereo image pairs from the front and rear stereo camera. Each stereo pair performs independent stereo tracking to obtain data correspondences, in the coordinate frame of the front camera. These correspondences are then fused together to solve for the relative motion of the front camera through outlier rejection and nonlinear refinement.

C. Localization and VO

The goal of localization and VO is to estimate the relative motion of the front camera between the current view at time $k$, $\mathcal{F}_k$, and a reference frame, $\mathcal{F}_{m_1}$. This motion can be represented by a Transformation matrix, $T_{k_1,m_1}$, which takes points from $\mathcal{F}_{m_1}$ into $\mathcal{F}_k$. In the case of VO, the reference frame is the previous frame while in the case of localization, the reference frame is a local submap. In both cases, we wish to minimize the reprojection error of all of the landmark observations after they are transformed by $T_{k_1,m_1}$ and reprojected into the image plane. For a given keypoint measurement of landmark $j$, $y_{j,k}$, and an observation of the landmark from the reference frame, $p_{j,m_1}^T$, the error term, $e_{j,k}$ is given by

$$e_{j,k} = y_{j,k} - g(T_{k_1,m_1}p_{j,m_1}^T).$$

(3)

Each keypoint also contains an uncertainty of the measurement of landmark $j$, $Q_j$, based on stereo geometry.

The localization-and-VO pipeline is depicted in the lower half of Figure 3 and consists of the following steps: (i) Stereo Tracking, (ii) Outlier Rejection, and (iii) Nonlinear Refinement. The inputs to the state estimation system are sets of synchronized stereo images from each camera.

Each stereo pair first undergoes independent stereo tracking to obtain data correspondences. Because correspondences from both cameras are formulated in the reference frame of the front camera they can be concatenated. These fused correspondences are then sent to an outlier rejection algorithm. Keypoints tracked by both cameras are sent through a RANSAC implementation using Horn’s 3-point method [16]. This provides a set of inliers as well as an initial estimate of the camera’s pose.

The final step of the state estimation pipeline is the nonlinear solver. The goal of the solver is to minimize the following objective function with respect to the camera transformation, $T_{k_1,m_1}$:

$$J_k = \frac{1}{2} \sum_{j=1}^{n} e_{j,k}^T Q_j^{-1} e_{j,k},$$

(4)

where $(e_{j,k}, \ldots, e_{j,k})$ is the set of errors associated with data correspondences from both stereo cameras. To minimize, this objective function is linearized and then iteratively refined through the Levenberg-Marquardt algorithm. The result is a camera transformation for the front camera that minimizes the sum of reprojection errors in both cameras.

D. Autonomous Navigation

The multi-stereo VT&R system allows a robot to autonomously repeat a previously driven path using stereo cameras. Autonomous navigation consists of a teach phase and a repeat phase. During the teach phase, the robot performs motion estimation through the stereo VO pipeline detailed in Section III-C and uses the output of VO to build the topometric pose graph detailed in Section III-B.
To autonomously repeat the path driven during the teach phase, the robot first loads the map into memory and performs localization to the desired section of the map. Once the robot is localized to the path, it can begin repeating. Localization is achieved by comparing the live stream to a local submap as described in Section III-B.

At every iteration of the pipeline, the robot performs VO and localization to obtain a relative transformation between the current position and the map. In the case of a localization success, the VO solution is used as a prior, in the case of a localization failure, the VO solution is propagated from the last localization success. This information is fed to a path-tracking controller to keep the robot on the path. Path tracking is accomplished using a Model Predictive Control algorithm [17].

E. System Failure Conditions

A localization failure is defined as obtaining less than six inlier feature correspondences between both cameras. While it is possible to obtain a state estimate using only three points, we note that in practice we found that any less than six correspondences results in a state estimate that is too noisy. In the case of a localization failure, the motion estimate from VO is propagated to obtain a state estimate. As with any dead reckoning solution, there is an associated error drift. If the robot drives too far without a localization success, the uncertainty will grow too large and the system will declare that the robot is lost and begin searching the map. In the case of these field trials, this distance was set to 20 m.

IV. METHODOLOGY

A. Field Trial

We conducted a field trial in the meadows surrounding the University of Toronto Institute for Aerospace Studies (UTIAS) campus in the winter before heavy snowfall. The environment, shown in Figures 1 and 7, consists of open fields covered by dead vegetation and sparse snow patches with buildings and trees on the horizon. This environment is particularly challenging for vision-based systems for a number of reasons: (i) the elevation of the sun in the winter is low, accelerating the effects of lighting change, (ii) sparse snow patches on the ground melt throughout the day, causing ephemeral features, and (iii) tall grass patches move significantly with the wind. The result is an environment with rapid appearance change over the course of hours.

The field trial proceed by teaching a 250 m loop through this winter environment. This path was taught when the sun was at its highest elevation point of the day and was autonomously repeated six times between noon and sunset, testing the algorithm’s ability to handle scene change.

B. Hardware

The robot used during our field trials is the Clearpath Robotics Grizzly RUV. The Grizzly is a large land rover with a suite of interoceptive and exteroceptive sensors. For the purpose of this field trial, the only sensors used were the front and rear facing Point Grey Research (PGR) Bumblebee XB3 Stereo Cameras, shown in Figure 1.

C. Calibration and Synchronization

The extrinsic calibration parameters of the stereo cameras, in particular the transformation, $T_{1,2}$, was carefully manually measured. Because this paper is an evaluation of the system’s robustness to feature matches, and not accuracy of path-tracking control with respect to ground truth, we leave the use of more sophisticated, data-driven calibration techniques for future work.

Our multi-stereo camera system assumes stereo pairs from both cameras are temporally synchronized. In practice, this is true up to the frame rate of the camera. To achieve synchronized pairs of stereo images, we use the approximate time filter that is part of the Robot Operating System (ROS) [18] software package. We furthermore synchronize the clocks on both cameras’ computers using the Chrony [19] software package. Because our cameras are 16Hz, this results in stereo image pairs with timestamps that are at the most 32ms apart. Because of the relatively slow speed of the platform, we deemed this discrepancy as noncritical and leave exactly synchronized cameras as future work.

D. Evaluation Metrics

To evaluate the performance of our new multi-stereo VT&R system we compare our method to two previously published VT&R variants labelled here: Legacy [1] and Color-Constant [2]. We judge the performance of each system based on its ability to reliably localize in difficult outdoor environments. At any point in time, we judge a localization success as the ability to match at least six post-RANSAC features between the live view and the map. While in theory the minimum number of matches needed to obtain a pose estimate is three, we found in practice that this estimate becomes too noisy below six features. We evaluate the performance of each system based on two quantitative metrics: Feature Quantity and Feature Sparsity.

Feature Quantity measures the median number of visual features seen over an entire traverse. This is calculated for all six traverses of the field trial and gives a notion of how quickly stable feature matches decline over time, using each method. Feature matches naturally decline over time in outdoor environments due to the changing appearance of the scene. In this outdoor environment, the primary causes of appearance change are lighting, melting snow, windy grass, and sun glare.

Feature Sparsity measures the distance the robot had to travel on VO between successful localizations. This is
a notion of how reliably the robot can navigate through the environment. The farther the robot travels on dead reckoning, the less certain it will become about its position on the path. Thus, this metric is the Cumulative Distribution Function (CDF) of the distance the robot would have had to drive on VO for the entirety of a single traverse. In this work, we present detailed evaluations of the fifth repeat (4.5 hours after map creation) in particular, because it highlights an autonomous traversal with significant appearance change. Specifically, the sun was low on the horizon, casting long shadows onto the scene, and some of the snow seen in the map had melted in the sun throughout the day.

V. RESULTS

A. Feature Quantity

In outdoor environments, the number of visual features matched to the map naturally declines over time, primarily due to the changing appearance of the scene. In this field trial, the causes of appearance change were lighting, melting snow, windy grass, and sun glare. These factors cause an accelerated rate of feature loss, which is displayed in Figure 4 over a period of five hours for each VT&R algorithm. This figure shows the general pattern that the multi-stereo algorithm maintains a higher number of stable features.

In unstructured and semi-structured environments, there are areas of the map that contain stable features including: trees, buildings, rocks, and bushes. As the time between the map and live view increases, these stable features are all that remains in environments such as the winter meadow studied in this field trial. The success of the multi-stereo algorithm is in part due to the wider field of view. With cameras facing forwards and backwards, there is higher probability of stable features to be in the map at any point in time. This point is illustrated in Figure 5, an examination of the number of feature matches seen in each camera during repeat 5, which occurred a little over four and a half hours after map creation. It can be seen that there are local maxima of matches for each camera through the repeat. These are stable features in the environment seen by the cameras. There are also areas where no stable features are found but each camera retains a small number of feature matches. These are not enough to localize alone but because the multi-stereo system fuses the matches into a single pose estimate, it is able to reliably localize over these difficult sections.

B. Feature Sparsity

If a localization failure occurs at a specific timestep, the system will use the propagated VO estimate as a best guess. While this is acceptable over short distances, the error associated with dead reckoning grows with respect to distance travelled, which will cause the robot to eventually lose its way. Through post-field analysis, we examine the distance the robot would have had to drive on VO using the different VT&R methods on the hardest repeat of the field trial. This repeat occurred 4.5 hours after map creation when the lighting was significantly different. Figure 6 shows the CDF of distance driven since localization for each algorithm on this repeat. It reads as: “for Y% of the traverse, the robot drove less than X m on VO”. This figure shows that in this feature-limited environment, the multi-stereo algorithm outperforms both the Legacy and Color-Constant algorithms.
Figure 6. CDF of the distance travelled on VO for the entirety of the traverse of repeat 5 for each VT&R algorithm. Repeat 5 occurred 4.5 hours after map creation with significant change in the appearance of the scene due to lighting, melting snow, and windy grass. Despite this difficult environment, the multi-stereo algorithm was able to reliably navigate due to the wider field of view. The dashed, black line in the figure represents the distance the robot can reliably travel on VO before a localization failure. In the case of our system, this value is set to 20 m.

Using the new system, the robot never drove over 1 m on dead reckoning at any point in time. In contrast, the Color-Constant and Legacy systems would have had to drive over 38 m and 41 m on dead reckoning, respectively. The black, horizontal line in the figure denotes the distance we allow the robot to drive on VO before declaring a localization failure. In the case of these field trials, this distance is set to 20 m. It is worth noting that both color-Constant and legacy systems would have declared at least one localization failure during the course of the repeat. Localization failures in the color-constant and legacy systems can be attributed to a lack of stable feature matches due to the rapidly changing appearance of the scene. Examples of this can be seen in Figure 7. These performance measures are directly related to the sparsity of stable visual features over the map. The primary reason for the multi-stereo system’s superior performance is the ability to detect an increased number of stable features through a wider field of view.

VI. CONCLUSION

We have shown that using two stereo cameras (as opposed to one) greatly increases the robot’s ability to reliably navigate in difficult, unstructured environments. Furthermore, we have shown that this method is outperforming the lighting-resistant method presented in [2]. This is possibly due to the color-constant images under performing in winter environments. Further analysis of this environment is warranted. Additional future work in this area may include an investigation into the optimal placement of the stereo cameras, and the trade-off of additional cameras vs. computational cost with respect to localization performance.

ACKNOWLEDGEMENT

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) through the NSERC Canadian Field Robotics Network (NCFRN). We would also like to thank Chris Ostafew for assisting in the field trials leading up to this paper.

REFERENCES

Figure 7. Two examples of appearance change over a short period of time in the winter meadow environment. Top: Appearance change due to sun glare and melting snow. Bottom: Appearance change due to long shadows and melting snow. Left: The appearance of the scene during the teach pass. This route was taught when the sun was at its highest elevation of the day, which results in little shadows in the environment. Right: The appearance of the scene roughly 4 hours after map creation. The low elevation of the sun accelerates lighting change and sparse snow patches melt throughout the day making this winter environment exceedingly difficult for vision-based systems.