Expanding the Limits of Vision-based Localization for Long-term Route-following Autonomy

Michael Paton, François Pomerleau, Kirk MacTavish, Chris J. Ostafew, and Timothy D. Barfoot
Institute for Aerospace Studies, University of Toronto, Toronto, ON, Canada
e-mail: mpaton@robotics.utias.utoronto.ca, f.pomerleau@gmail.com, kirk.mactavish@mail.utoronto.ca, chris.ostafew@mail.utoronto.ca, tim.barfoot@utoronto.ca

Received 2 December 2015; accepted 29 May 2016

Vision-based, autonomous, route-following algorithms enable robots to autonomously repeat manually driven routes over long distances. Through the use of inexpensive, commercial vision sensors, these algorithms have the potential to enable robotic applications across multiple industries. However, in order to extend these algorithms to long-term autonomy, they must be able to operate over long periods of time. This poses a difficult challenge for vision-based systems in unstructured and outdoor environments, where appearance is highly variable. While many techniques have been developed to perform localization across extreme appearance change, most are not suitable or untested for vision-in-the-loop systems such as autonomous route following, which requires continuous metric localization to keep the robot driving. In this paper, we present a vision-based, autonomous, route-following algorithm that combines multiple channels of information during localization to increase robustness against daily appearance change such as lighting. We explore this multichannel visual teach and repeat framework by adding the following channels of information to the basic single-camera, gray-scale, localization pipeline: images that are resistant to lighting change and images from additional stereo cameras to increase the algorithm's field of view. Using these methods, we demonstrate robustness against appearance change through extensive field deployments spanning over 26 km with an autonomy rate greater than 99.9%. We furthermore discuss the limits of this system when subjected to harsh environmental conditions by investigating keypoint match degradation through time. © 2016 Wiley Periodicals, Inc.

1. INTRODUCTION

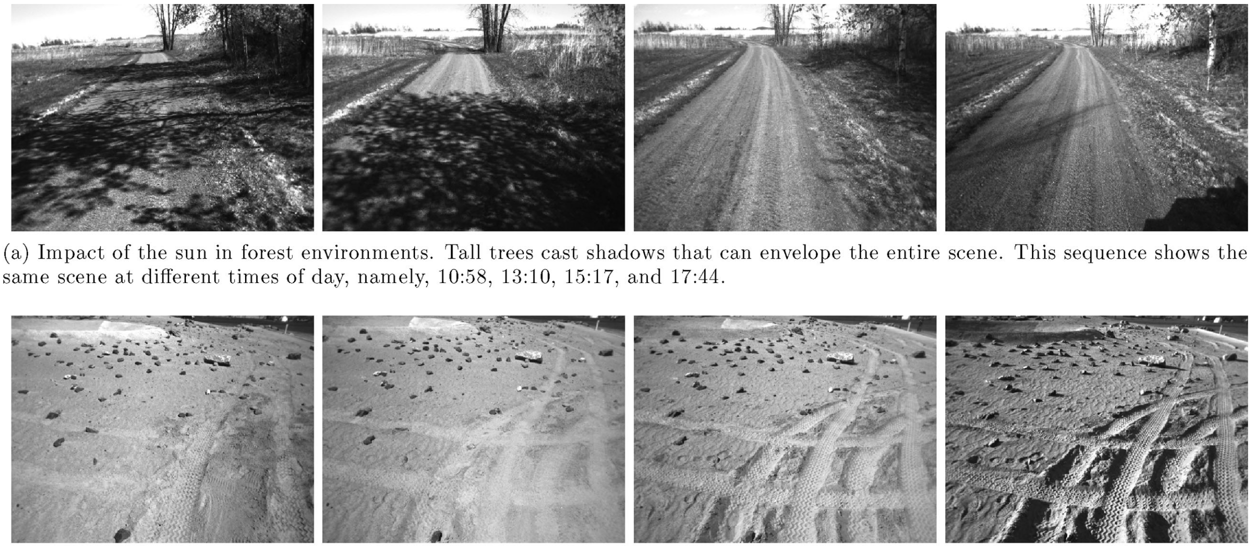
Vision-based route-following algorithms have enabled mobile robots to navigate autonomously through large-scale environments using inexpensive commercial sensors without the need for an accurate global map (Furgale & Barfoot, 2010). This technology opens the door for many applications that benefit from repeated traversals over constrained paths, such as factory floors, orchards, mines, and urban road networks. Furthermore, this method can be used in hazardous-exploration, sample-return, and search-and-rescue missions where the robot can autonomously return to a previously driven location without the need for a globally consistent map. However, in order for these applications to succeed, robots must have the ability to navigate reliably through their environments over long periods of time. This poses a serious problem for robots that operate in outdoor environments where lighting, weather, and seasonal change dramatically alter the appearance of the scene. An example of daily appearance change can be seen in Figure 1, which shows the varying appearance of one of our primary field-testing sites due to lighting change.

Environments with a variable appearance are difficult for autonomous route-following algorithms, which require vision-in-the-loop navigation. This specific task relies on a vision system to provide continuous, accurate metric localization to the control loop to keep the robot driving. Most current methods that meet this criterion are examples of single-channel, single-experience localization systems. In the context of this work, we use the term *channel* as a stream of information used to localize a robot's position and *experience* as a collection of channel data obtained during a robot traverse, or run. Methods that rely on localizing to a map collected from a single experience with a single channel of vision information are highly susceptible to appearance change. As a result, the operational domain of these methods is typically limited to only a few hours outdoors, as highlighted by Furgale and Barfoot (2010), due directly to appearance change.

This paper explores the specific challenge of intraseasonal, daily appearance change, and it provides solutions that extend the operational domain of route-following algorithms from a few hours to multiple days. This is achieved through the use of multichannel localization, where

Direct correspondence to: Michael Paton, email: mpaton@robotics.utias.utoronto.ca

View this article online at wileyonlinelibrary.com • DOI: 10.1002/rob.21669



(a) Impact of the sun in forest environments. Tall trees cast shadows that can envelope the entire scene. This sequence shows the same scene at different times of day, namely, 10:58, 13:10, 15:17, and 17:44.

(b) Impact of the sun and robot in desert environments. Small textures such as rocks and wavy sand cast shadows and vehicles driving in sand significantly alter the terrain in a short amount of time. This sequence shows the same scene at different times of day, namely, at 11:03, 13:15, 15:12, 17:49

Figure 1. Examples of the daily appearance change seen in unstructured outdoor environments due to lighting change and terrain modification.

multiple information channels are used together to solve a single localization problem. Different channels can emerge from the same sensor, from the same type of sensor, or from different visual sensors. We propose a multichannel visual teach and repeat (VT&R) algorithm in order to gain robustness against environmental changes and provide two concrete instantiations: a lighting-resistant variant that uses environmentally tuned color-constant images (Paton, McTavish, Ostafew, & Barfoot, 2015a), and a variant with an extended field of view through multiple stereo cameras (Paton, Pomerleau, & Barfoot, 2015b). We validate our algorithms through field deployments covering over 26 km of autonomous driving with an autonomy rate of over 99.9% of distance traveled. Through post-field analysis, we demonstrate significant robustness gain against the wide variety of lighting change seen during the trials, and we posit that the autonomy rate achieved would not have been possible in a single-channel framework. We furthermore explore the trends related to a decrease in localization performance in winter environments (Paton, Pomerleau, & Barfoot, 2015c).

This paper presents our prior conference publications (Paton et al., 2015a,b,c) with the following supplementary contributions: (i) a generic multichannel Visual Teach & Repeat (VT&R) framework incorporating our previous systems under the same architecture, (ii) an in-depth explanation on how to adapt color-constant parameters for a specific environment, and (iii) a comprehensive investigation of the limiting factor for a single-experience autonomous route-following algorithm. The remainder of this paper is organized as follows. Section 2 overviews related work. Details on the multichannel VT&R framework and its specific variants are presented in Section 3. Experimental setups and field trials are explained in Section 4. Results of the field trials are presented in Section 5. The paper ends with a discussion and conclusion in Sections 6 and 7, respectively.

2. RELATED WORK

This paper presents a VT&R framework that allows for the use of multiple information channels to perform metric localization. This framework is used to increase localization robustness against lighting change through color-constant images and expand the field of view of the algorithm using multiple stereo cameras. We furthermore explore the trends related to a decrease in localization performance in winter environments. As such, work related to this paper spans the following topics: (i) autonomous route-following techniques, (ii) theory of color-constancy, (iii) color-constancy in robotics, (iv) localization in dynamic environments, (v) localization and Visual Odometry (VO) using multiple cameras, and finally (vi) localization and VO in extreme environments.

2.1. Autonomous Route-following

The autonomous route-following algorithms presented in this paper are built upon the stereo VT&R work presented by Furgale and Barfoot (2010). Further work by Van Es and Barfoot (2015) has shown that fully connected networks of graphs can be built with no single reference frame or global optimization, while allowing for smooth transitions over loop closures. However, because this system navigates by comparing descriptors from gray-scale images, it is highly susceptible to lighting change. This can be overcome by using an active sensor. McManus, Furgale, Stenning, and Barfoot (2012) perform VT&R using keypoints formed from lidar-generated intensity images and range data. While this method is invariant to lighting conditions, it suffers from motion distortion issues, and it relies on an expensive sensor. Krüsi et al. (2014) performed autonomous route-following through dense point-cloud registration, a technique that is not well equipped for open spaces that lack geometric information. Vision-based path-following algorithms use both visual texture and coarse geometry to avoid these limitations, but they are highly susceptible to lighting change. One method to overcome this issue is the use of color-constant images.

2.2. Color-constancy Theory

Color-constancy can be defined as the ability to observe an object's color largely independent of the varying illumination. It is a property of our human perceptual system, and it has been a topic of research in the optics and computer vision communities. Recent research has developed a simple, fast transformation from an RGB image to a gray-scale image that is partially invariant to lighting conditions. If assumptions are made about the sensor and the light source, a gray-scale, lighting-invariant image can be obtained from a three-channel camera. Finlayson, Hordley, Cheng, and Drew (2006) calculated a two-dimensional (2D) colorspace that moves along a known direction as the lighting in the environment changes. By projecting the 2D colorspace onto a line that is orthogonal to this direction, a 1D colorspace that is invariant to lighting conditions is obtained. Ratnasingam and Collins (2010) extract a gray-scale image by taking the weighted log difference of the three channels to cancel out the effects of illumination.

2.3. Color-constancy in Robotics

Lighting change is the first issue vision-based navigation systems need to face when operating outdoors. Fortunately, the theories of color-constancy have had great success increasing the robustness of vision-based localization and place-recognition systems. Corke, Paul, Churchill, and Newman (2013) tested the image transformation described by Finlayson et al. (2006) on a dataset of images captured under varying illumination conditions. They show an increase in precision/recall performance versus gray-scale images when whole-image place recognition is performed on this dataset. MacTavish, Paton, and Barfoot (2015) further showed that color-constant images boost the performance of feature-based place-recognition systems such as FAB-MAP. Maddern, Stewart, and Newman (2014) localize monocular images against a prior map of colored 3D point clouds to obtain a six-degree-of-freedom (6DOF) pose estimate. By using color-constant images based on Ratnasingam and Collins (2010) during the day and gray-scale images at night, they show successful localization over a 24-h period. McManus, Churchill, Maddern, Stewart, and Newman (2014a) run two separate localizers in parallel, one that uses gray-scale images and one that uses color-constant images based on Ratnasingam and Collins (2010). Localization with color-constant images only occurs when the gray-scale localizer first fails. This work was shown to improve localization against a map that was collected in different lighting conditions. This *Best Fit* approach is directly compared in Section 5.2 to our multichannel framework (Section 3.4). We show an increase in localization performance by combining data correspondences from color-constant and gray-scale images to solve a single-state estimation problem.

2.4. Localization in Dynamic Environments

While color-constant images help to overcome issues with lighting, a general appearance change over time remains an issue for vision-based navigation. Recent work can be broken down into two categories: single- and multiexperience localization. Single-experience localization consists of associating live data with a single visual map, while multi-experience localization consists of associating live data with a series of previously recorded maps.

Topological localization across an appearance change as extreme as day and night can be achieved through image sequence alignment. First introduced by Milford and Wyeth (2012) as SeqSLAM, this method first computes a confusion matrix between an array of recent images and the array of map images using whole image matching, and then it searches for the best diagonal path through the matrix. While effective, this method is highly susceptible to scale and viewpoint changes, requiring almost perfect alignment of images. In Pepperell, Corke, and Milford (2015), SeqSLAM was further improved to handle viewpoint changes through automatic image scaling. Naseer, Spinello, Burgard, and Stachniss (2014) align sequences of images through a probabilistic network flow problem, allowing for potential loop closures in the traverse. While these methods are effective at topological localization across appearance change, topological localization is not well suited for full vision-in-the-loop navigation on its own, because it lacks precise metric localization.

Most methods that deal with appearance change and provide metric localization do so by using multiple experiences. Neubert, Sunderhauf, and Protzel (2013) build a dictionary that encodes the transformation of a scene between winter and summer. In the work of McManus, Upcroft, and Newman (2014b), scene-specific keypoint descriptors are learned by training support vector machine (SVM) classifiers on a collection of the scene in multiple appearances. This method provides coarse metric localization. The experience-based navigation (EBN) technique developed by Churchill and Newman (2013) is the most promising option for vision-in-the-loop navigation, as it provides metric localization across interseasonal appearance change, although no published work has tested it in this context. This method performs point-based keypoint detection and tracking while building a multiexperience map. When localization fails, the live VO output is kept as a new parallel experience in that place that can be used for future localization. The system is built upon parallel localizers, one for each experience. The state estimation from the best performing localizer is kept at any one point in time. To improve computational costs, recent improvements detailed in Linegar, Churchill, and Newman (2015) allow for the intelligent selection of which experiences to localize against based on historical performance. EBN has also been used with 2D push-broom lidars to provide accurate localization across significant appearance change over a period of a year (Maddern, Pascoe, & Newman, 2015). To ensure safe navigation, the majority of real-world route-following applications need to localize metrically to a *single* manually taught route. Despite successful interseason localization, EBN only provides localization to an arbitrary experience at any given point in time.

2.5. Localization and VO using Multiple Cameras

Work on multicamera state estimation can be broken down into two categories: systems that model multiple cameras as a single generalized camera, and systems that treat each camera independently. Our multichannel 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 (2003). The use of plucker lines to model point correspondences between cameras allows this model to solve for extrinsic calibration parameters using a generalized essential matrix. Lee, Faundorfer, and Pollefeys (2013) estimate the motion of self-driving cars with four nonoverlapping monocular cameras. With inter- and intrapoint correspondences, they solve for the generalized essential matrix with a two-point RANdom SAmple Consensus (RANSAC) scheme and nonlinear refinement. Heng, Lee, and Pollefeys (2014) present a full micro-aerial vehicle (MAV) simultaneous localization and mapping (SLAM) system including autonomous calibration of extrinsic parameters between cameras. Their system setup consists of an inertial measurement unit (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, Furgale, and Siegwart (2013) formulate a general solution to multicamera state estimation that is computationally more efficient than previous methods. They present a parametrization of the generalized camera model that is noniterative 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 multicamera state estimation systems is to formulate the system as a set of independent camera sensors. Oskiper, Zhu, Samarasekera, and Kumar (2007) perform VO using dual stereocameras and an 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, Kim, Frahm, Pollefeys, and Hartley (2008) build a 6DOF motion estimation system using clusters of nonoverlapping monocular cameras. Each camera performs independent state estimation through a five-point RANSAC algorithm. Any intercamera correspondences are then used to solve for scale using a one-point RANSAC solution. At each step, the best estimate is used. Kazik, Kneip, Nikolic, Pollefeys, and Siegwart (2012) estimate motion with two nonoverlapping 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 multiframe estimation to improve accuracy. Motion estimates from each camera are then fused to obtain the final 6DOF motion estimate.

Our multistereo system is similar to the dual-stereo VO setup described in Oskiper et al. (2007), 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 keypoints to be spread across both cameras, allowing for localization in keypoint-limited environments. More similar to our method is the nonoverlapping multicamera parallel tracking and mapping (MCPTAM) algorithm formulated in Tribou, Harmat, Wang, Sharf, and Waslander (2015). In this method, localization is achieved by running two parallel processing threads: one for frame-to-keyframe VO, and one for full keyframe bundle adjustment. While effective, this method's reliance on a single privileged reference frame and a global bundle adjustment solution is not well suited for large-scale, outdoor navigation targeted in autonomous route-following applications.

2.6. Localization and VO in Extreme Environments

The performance of vision-based state estimation systems is dependent in part on the environment in which the robot is operating. Two environment-dependent factors significantly affect vision-based systems: the rate of appearance change and the amount of contrast in the scene. This makes vision-based navigation in winter environments especially difficult as the elevation of the sun is perpetually low on the horizon, and snow rapidly accumulates, melts, and provides little contrast to the scene. Williams and Howard (2010) improve VO in snowy environments by applying contrast-limited adaptive histogram equalization (CLAHE) to increase keypoint matches in images with snowy foregrounds. They show an increase in keypoint match count by an order of magnitude. Volcanic fields are similar to snowy landscapes in their lack of contrast. Otsu, Otsuki, and Kubota (2015) extract and track different keypoints depending on the volcanic terrain, and they show an improvement in keypoint count and computation speed. An often ignored environment for vision sensors is night. Nelson, Churchill, Posner, and Newman (2015) perform vision-based, nighttime localization through the tracking of artificial light sources such as street lights.

3. THEORY

This section presents the details of the multichannel VT&R system, which uses parallel information channels to increase the robustness of metric localization across appearance change. The section starts with a sensor-generic formulation of the multichannel system. Next, the following VT&R methods are reformulated in this multichannel framework: (i) the legacy VT&R system originally published by Furgale and Barfoot (2010), (ii) the lighting-resistant VT&R system originally published by Paton et al. (2015a), and (iii) the dual-stereo VT&R system originally published by Paton et al. (2015b).

3.1. Multichannel State Estimation

The multichannel VT&R state-estimation system, depicted in Figure 2, increases robustness against appearance change by combining landmarks from multiple channels of visual information into a single-state estimation problem. This process can be broken down into two steps: independent channel tracking and multichannel state estimation.

3.1.1. Independent Channel Tracking

In the context of this system, a *channel* defines a stream of visual information used to localize a robot's position. A key innovation behind the multichannel VT&R paradigm is that landmarks independently detected and tracked in channels can be used to solve a single-state estimation problem. A multichannel VT&R system can use channels from the same sensor (i.e., gray-scale images, color-constant images), from the same type of sensor (multiple cameras), or from different sensors (stereo and lidar). In this paper, we focus on channels that originate from stereo cameras. Our multichannel localization requires the channels to be temporally synchronized with sensor transforms known *a priori.* Channel tracking consists of the detection and matching of point-based descriptors. This process is detailed in the upper section of Figure 2. The input to the system is visual data with the ability to extract depth information. The output is a set of data correspondences between the input and either the previous frame (VO) or a map frame (localization).

The first step of channel tracking is the extraction of keypoints with descriptors, 3D position, and uncertainty associated with the image measurement. This algorithm is agnostic to the methods associated with depth extraction, keypoint detection, and keypoint description. Coordinates of the jth keypoint at time k are of the following form:

$$\mathbf{y}_{j,k_i} = \begin{bmatrix} u \\ v \end{bmatrix}, \qquad \mathbf{p}_{k_i}^{j,k_i} = \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \tag{1}$$

where $\mathbf{y}_{j,k_i}$ represents the keypoint measurement coordinates, and $\mathbf{p}_{k_i}^{j,k_i}$ is the physical location of landmark j, in homogeneous coordinates, in the coordinate frame of channel i. This value is a vector from the origin of $\underrightarrow{\mathcal{F}}_{k_i}$ to the origin of $\underrightarrow{\mathcal{F}}_{j}$ (denoted by the superscript) and expressed in $\underrightarrow{\mathcal{F}}_{k_i}$ (denoted by the subscript).

To fuse data correspondences between channels, they must be in the same coordinate frame. Because all state estimation is performed in the reference frame of the single master channel, keypoints in the coordinate frame of the ith channel, $\underrightarrow{\mathcal{F}}_{k_i}$, are converted to the coordinate frame of the master channel, $\underrightarrow{\mathcal{F}}_{k_1}$, using the transformation, $\mathbf{T}_{1,i}$:

$$\mathbf{p}_{k_1}^{j,k_i} = \mathbf{T}_{1,i}\mathbf{p}_{k_i}^{j,k_i}, \tag{2}$$

where $\mathbf{T}_{1,i}$ is the extrinsic transformation assumed to be known *a priori*. Keypoints in the coordinate frame of the master channel remain unmodified. The final step of the channel-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 master channel.

3.1.2. Multichannel Estimation Framework

The goal of both localization and VO is to estimate the relative motion of the master-channel sensor between the current view at time $k, \underrightarrow{\mathcal{F}}_{k_1}$, and a reference frame, $\underrightarrow{\mathcal{F}}_{m_1}$. This motion can be represented by a transformation matrix, $\mathbf{T}_{k_1,m_1}$, which takes points from $\underrightarrow{\mathcal{F}}_{m_1}$ into $\underrightarrow{\mathcal{F}}_{k_1}$. 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 find the estimate of $\mathbf{T}_{k_1,m_1}$ that minimizes the reprojection error of all of the landmark observations after they are transformed and reprojected into the image plane. For a given keypoint measurement of

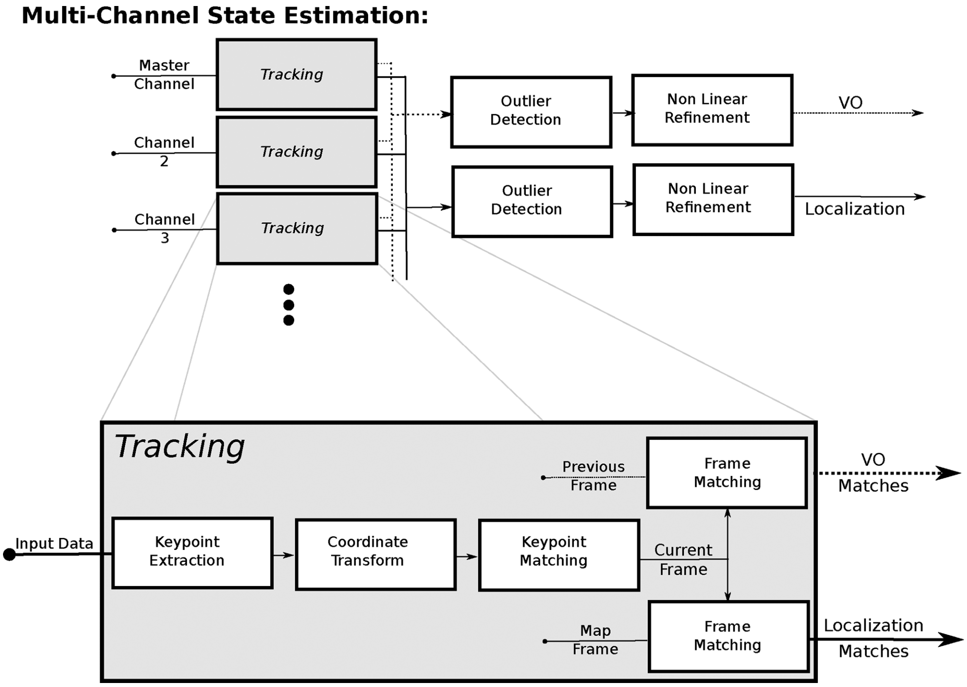


Figure 2. Pipelines of the multichannel VT&R system. Top: Multichannel State Estimation Pipeline. The input to the system is a set of synchronized data from all channels. The output to the system is a transformation relative to a reference frame. Each channel performs independent stereo tracking to obtain data correspondences in the coordinate frame of the master channel. These correspondences are then fused together to solve for the relative motion of the master channel through outlier rejection and nonlinear refinement. Bottom: Independent channel tracking. The input to the system is vision data with the ability to extract depth information. The output to the system is a set of keypoint matches with depth in the coordinate frame of the master channel. 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.

landmark j, $\mathbf{y}_{j,k}$, and an observation of the landmark from the reference frame, $\mathbf{p}_{m_1}^{j,m_1}$, the error term $\mathbf{e}_{j,k}$ is given by

$$\mathbf{e}_{j,k} = \mathbf{y}_{j,k} - \mathbf{g}(\mathbf{T}_{k_1,m_1}\mathbf{p}_{m_1}^{j,m_1}), \tag{3}$$

where $\mathbf{g}(\cdot)$ is the inverse sensor model that transforms points into the image sensor plane. Each keypoint also contains an uncertainty, $\mathbf{Q}_j$, of the measurement of landmark j.

The localization-and-VO pipeline is depicted in the lower half of Figure 2 and consists of the following steps: (i) channel tracking, (ii) outlier rejection, and (iii) nonlinear refinement. The inputs to the state-estimation system are sets of image data from each channel. Each channel first undergoes keypoint tracking to obtain data correspondences. Because correspondences from all channels are formulated in the reference frame of the master channel, they can be concatenated. These fused correspondences are then sent to an outlier rejection algorithm.

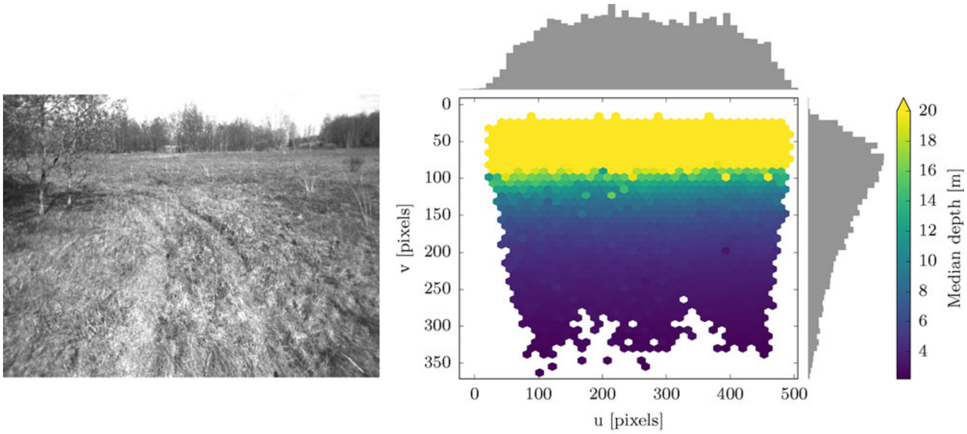
Keypoints tracked by all channels are sent through a RANSAC implementation using Horn's three-point method (Horn, 1987). This provides a set of inliers as well as an initial estimate of the master channel's pose. The goal of the solver is to minimize the following objective function with respect to the camera transformation, $\mathbf{T}_{k_1,m_1}$:

$$J_k = \frac{1}{2}\sum_{j=1}^{n} \mathbf{e}_{j,k}^T \mathbf{Q}_j^{-1} \mathbf{e}_{j,k} + J_{\text{pos}}, \tag{4}$$

where $(\mathbf{e}_{1,k}, \ldots, \mathbf{e}_{n,k})$ is the set of errors associated with data correspondences from all channels, and J_{pos} is a prior term on motion. J_{pos} minimizes the error between the posterior transform, $\mathbf{T}_{k_1,m_1}$, and a prior transform, $\check{\mathbf{T}}_{k_1,m_1}$. In the case of VO, $\check{\mathbf{T}}_{k_1,m_1}$ is a no-motion prior, and in the case of localization, $\check{\mathbf{T}}_{k_1,m_1}$ is the result of VO. To minimize this objective function, the equation is linearized and then iteratively refined through the Levenberg-Marquardt algorithm. The result is a transformation that minimizes the sum of reprojection errors in all channels.

3.2. Multichannel VT&R

This section provides an overview of the full multichannel VT&R system, which relies on the multichannel state estimation pipeline detailed in the previous section. This



(a) Image from the front stereo camera.

(b) Histograms of the keypoint positions.

Figure 3. Typical distribution of inlier keypoint matches to the map, aggregated from a 1 km repeat run. Side histograms represent the distributions of matches projected on the vertical (v-axis) and horizontal (u-axis) fields of views. Colors represent the median depth value of inliers per cell. The depth value is estimated from a single stereo-camera pair. The color is saturated for 20 m and more, and cells with fewer than 50 inliers are represented in white.

section covers the human-in-the-loop teaching phase and the autonomous vision-in-the-loop repeating phase.

3.2.1. Teaching Phase

During the teach phase, the robot is manually driven while building a map based on the multichannel VO pipeline detailed in Section 3.1. This map consists of a topometric pose graph of keyframes linked by relative transformations as described in Furgale and Barfoot (2010). Vertices in this pose graph contain synchronized sets of keypoints obtained via independent tracking from each channel. Edges in the map are the relative transformations between vertices computed by multichannel VO. While all computed keypoints from all channels are stored to the pose graph for future localization, the multichannel VO pipeline may only use a subset of the channels for state estimation in the teach phase, depending on the configuration of the system. Vertices are constructed when the robot's motion exceeds a specified threshold, forcing an evenly distributed map.

3.2.2. Repeating Phase

To autonomously repeat the taught route, the robot performs VO and localization to obtain a relative transformation between the current position and the map. Localization is achieved by comparing the stream of multichannel data to a local submap pulled from the closest vertex. This submap is computed from a fixed amount of vertices centered at the estimated closest vertex and relaxed into a single coordinate frame. Doing this allows the localization complexity to be constant with respect to the size of the total 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 estimate. This information is fed to a path-tracking controller to keep the robot on the route. Path tracking is accomplished using model predictive control (Rawlings & Mayne, 2009). At the start of a repeat, the robot performs a localization search to find its position relative to the closest vertex in the pose graph.

3.3. Legacy VT&R

The VT&R algorithm presented in Furgale and Barfoot (2010) can be reformulated in the multichannel paradigm as a single-channel method that uses rectified, gray-scale stereo images as an input to the master channel. Keypoints are detected and described using the 256-byte, upright speeded-up robust features (SURF) descriptor, and depth information is obtained through left-right image matching and stereo triangulation.

This method is effective at providing long-range autonomy with constant-time localization when there is minimal appearance change in the scene. In such an environment, a sufficient amount of inlier keypoint matches between the live view and the map can be recovered to provide centimeter-level metric localization to the path-tracking controller. An example of the distribution of inlier matches found during nominal operation can be seen in Figure 3. This shows the aggregation of inlier matches over a 1 km traverse approximately 1 h after map creation. This keypoint distribution is typical for a forward-looking camera on a moving robot. When moving forward, keypoints close to the lower image border are typically not in the field of view of both the live keyframe and the map keyframe, leading to a skewed distribution of points on the vertical axis. In addition, the platform moves through the environment, generating changes in the reobserved images. On soft ground, a heavy vehicle will generate ruts that modify the deployment area over time; this is clearly visible in Figure 3. In dynamic environments, this method is highly susceptible to lighting change (Furgale & Barfoot, 2010), with the number of inlier matches to the map approaching zero after a few hours on sunny days.

3.4. Lighting-resistant VT&R

This section presents the lighting-resistant VT&R method originally presented in Paton et al. (2015a), reformulated in the multichannel paradigm. This method uses color-constant images to increase its robustness against lighting change. We start with a primer on color-constancy theory and end with the details of the VT&R system.

3.4.1. Color-constancy

We introduce the theory of the color-constant image transformations used in this algorithm at a high level, and we refer the reader to Ratnasingam and Collins (2010) for a detailed derivation. A camera's response for a specific point, x, in the environment is described by the illuminant, the sensor response, the reflecting surface, and the geometry of the scene and camera. The light originates from an illuminant, is reflected by a surface toward the camera, and is focused onto an image sensor consisting of an array of filtered pixel sensors. This process results in the sensor response, R^x, describing the power of the light incident on the pixel sensor after being reflected and filtered. The illuminant is described by its intensity, I, and spectral power distribution, $E(\lambda, T)$, as a function of wavelength, λ, and temperature, T. At a specific point, x, the light is reflected according to the incident direction, $\underrightarrow{a}^{x}$, the surface normal, $\underrightarrow{n}^{x}$, and the surface reflectance, $S^x(\lambda)$. This light is filtered according to the sensor's channel, described by the spectral sensitivity, $F(\lambda)$. Integrating over the desired spectrum, ω, results in the image sensor response:

$$R^x = \underrightarrow{a}^{x} \cdot \underrightarrow{n}^{x} I \int_{\omega} S^x(\lambda) E(\lambda, T) F(\lambda) d\lambda. \tag{5}$$

Images that are resistant to the variation in the illumination of an outdoor scene can be calculated from a three-channel camera by making assumptions about the imaging sensor and environment (Ratnasingam & Collins, 2010). These assumptions allow cancellation of the factors of Eq. (5) that are dependent on the scene's illumination: the spectral power distribution, $E(\lambda, T)$, and the intensity of the illuminant, I. If the assumptions are that the spectral sensitivity function, $F(\lambda)$, is infinitely narrow at the sensor's peak wavelength, λ_i, and the sole illuminant of the scene is a black-body radiator, then the logarithm of Eq. (5) can be reformulated as

$$\log(R_i^x) = \log(\underrightarrow{a}^{x} \cdot \underrightarrow{n}^{x} I) + \log[S^x(\lambda_i) C_1 \lambda_i^{-5}] - \frac{C_2}{T\lambda_i}, \tag{6}$$

where C_1 and C_2 are constants. The result is a sensor response equation that separates the effect of illumination on the scene from properties of the reflected surface material. A weighted linear combination of three channel responses can be constructed to effectively cancel out the first and third terms, providing an illumination-invariant sensor response that is affected only by the properties of the surface materials. This difference of log responses is provided on a per-pixel basis by the following equation:

$$F = \log(R_2) - \alpha \log(R_1) - \beta \log(R_3), \tag{7}$$

where $\log(R_i)$ is Eq. (6) with peak wavelength, λ_i, and weights α and β subject to the following constraints:

$$\frac{1}{\lambda_2} = \frac{\alpha}{\lambda_1} + \frac{\beta}{\lambda_3}, \quad \beta = (1 - \alpha), \tag{8}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the theoretical peak sensor responses ordered from lowest to highest wavelength. If these constraints are met, the weighted difference of the log responses will cancel out the effect of the spectral power distribution of the light source, $E(\lambda)$, and the illuminant intensity, $\underrightarrow{a}^{x} \cdot \underrightarrow{n}^{x} I$. In the context of a digital RGB camera, $\{R_1, R_2, R_3\}$ are the pixel response values for the blue green and red channels, respectively.

In theory, if the peak wavelength values of the sensor, $(\lambda_1, \lambda_2, \lambda_3)$, are known, then the weights can be calculated based on the constraints in Eq. (8) with the following:

$$\alpha = \frac{(\lambda_1 \lambda_3)/\lambda_2 - \lambda_1}{\lambda_3 - \lambda_1}, \quad \beta = 1 - \alpha. \tag{9}$$

If the assumptions are met that the sole illuminant of the scene is a black-body radiator and the sensor channels of the camera are infinitely narrow, centered at their peak values, then the resulting image will be free of the effects of illumination. The first assumption is reasonably close to the truth, as long as the only illuminant is the sun. The accuracy of the second assumption varies for each camera, but will never be exactly true. An example of this can be seen in the sensor response curves for the Sony ICX446 CCD imaging sensor (see Figure 4). While each channel has distinct peaks, they are far from infinitely narrow with significant overlap between channels. Using the theoretical wavelengths seen in the response curves, the color-constant image transformation would be Eq. (7) with weights $\alpha = 0.467$ and $\beta = 0.533$. While these parameters have a theoretical basis, they do not always produce the best results. In Section 5.1, a method to find the color-constant transformation parameters that provide the best results for specific biomes is detailed.

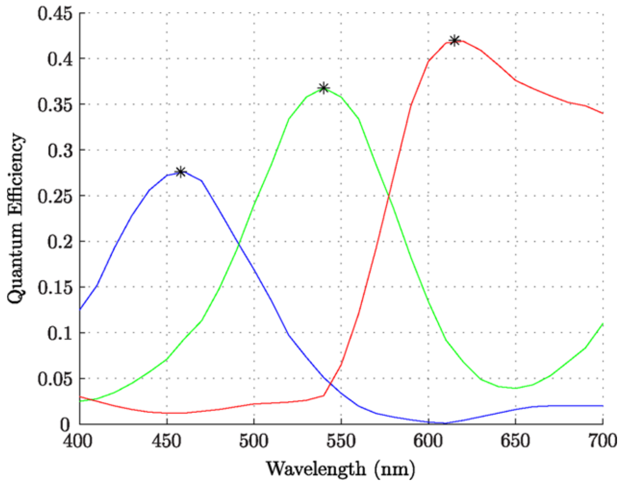


Figure 4. Sensor response of the Sony ICX445 CCD, with theoretical peak wavelengths denoted by stars.

3.4.2. System Overview

The lighting-resistant VT&R system, first published in Paton et al. (2015a), adds robustness against lighting change through the use of color-constant images. The algorithm is formulated in a multichannel paradigm by setting the input to the master channel to gray-scale stereo images and the inputs to subsequent channels to color-constant stereo images experimentally tuned for varying biomes. Inputs to all channels originate from the same RGB stereo pair. We posit that the majority of imaging sensors severely violate the assumptions put forward in Eq. (8) and may produce inferior results if the peak theoretical wavelengths are used. Furthermore, if a robot is traveling across multiple biomes, it may be necessary to provide the algorithm with multiple color-constant transformations to provide reliable localization across lighting changes.

Each channel performs tracking by extracting SURF keypoints with descriptors and 3D positions. Because all of the channels in this system originate from the same sensor, the transformations that take keypoints into the coordinate frame of the master channel can be assumed to be identity transformations. In this system, VO is performed with the master channel only, while localization uses all available channels. This is because color-constant images are inherently noisier than their gray-scale counterparts, and lighting change is not a factor for VO. Adding additional channels to the state-estimation problem of VO will only add computation cost with no benefit.

3.5. Dual-stereo VT&R

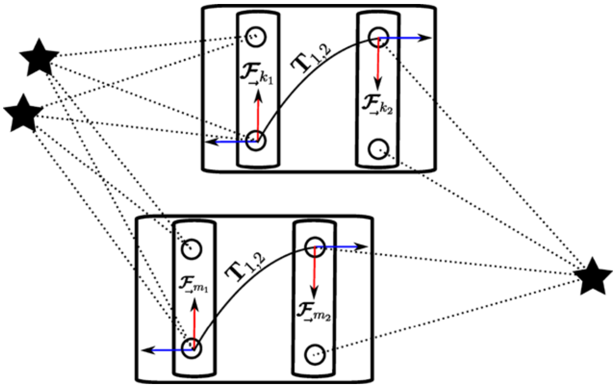

$\underrightarrow{\mathcal{F}}_{k_1}$
$\mathbf{T}_{1,2}$
$\underrightarrow{\mathcal{F}}_{k_2}$
$\underrightarrow{\mathcal{F}}_{m_1}$
$\mathbf{T}_{1,2}$
$\underrightarrow{\mathcal{F}}_{m_2}$

Figure 5. Diagram of the multistereo setup. Given a time step k, the system is defined by a robot with two stereo cameras with respective coordinate frames, $\{\underrightarrow{\mathcal{F}}_{k_1}, \underrightarrow{\mathcal{F}}_{k_2}\}$. The transformation $\mathbf{T}_{1,2}$, which takes points from $\underrightarrow{\mathcal{F}}_{k_2}$ to $\underrightarrow{\mathcal{F}}_{k_1}$, is assumed be known *a priori*. Localization is achieved through intercamera point correspondences, depicted as black stars. To localize, all point correspondences are transformed into the frame, $\underrightarrow{\mathcal{F}}_{k_1}$ and used in a joint state estimation problem. Shown here is localization between two time stamps: k and m.

The Dual-stereo VT&R method, first published in Paton et al. (2015b), extends the field of view of the legacy method by adding a channel consisting of gray-scale images originating from a second stereo camera. In this formulation, the system assumes two temporally synchronized, nonoverlapping stereo cameras with coordinate frames, $\{\underrightarrow{\mathcal{F}}_{k_1}, \underrightarrow{\mathcal{F}}_{k_2}\}$. The transformation, $\mathbf{T}_{1,2}$, which takes points from $\underrightarrow{\mathcal{F}}_{k_2}$ to $\underrightarrow{\mathcal{F}}_{k_1}$, is assumed to be known *a priori*. This algorithm is formulated in a multichannel paradigm by setting the input to the master channel to gray-scale stereo images from camera 1 and setting the input to the additional channel to gray-scale stereo images from camera 2. During independent channel tracking, keypoints originating from the second camera are transformed into the coordinate frame of the first camera with $\mathbf{T}_{1,2}$. The camera setup for this system is illustrated in Figure 5.

While this method is no more resistant to lighting than the original Legacy VT&R method, it essentially doubles the amount of inlier matches found during localization, and it tracks stable keypoints seen in the environment for longer periods of time. This increase of the algorithm's field of view greatly increases the ability to safely localize.

4. EXPERIMENTAL SETUPS

This section details our autonomous route-following experiments as well as the static image experiments used to tune the color-constant images. In total, three separate experiments covered over 26 km of driving spanning over a year, covering multiple biomes and seasonal conditions. Overall, our analysis is relying on approximately 1.5 TB of sensor data and 25,000 images.

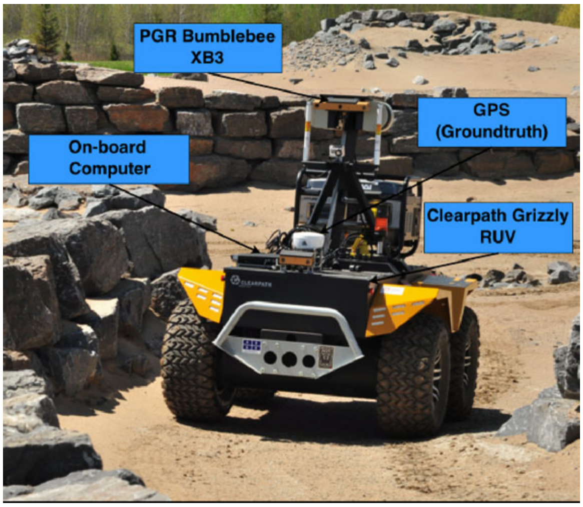


Figure 6. Clearpath Grizzly RUV and its sensor configuration. The robot is equipped with a forward- and rear-facing PGR Bumblebee XB3 camera, a Global Positioning System (GPS) receiver, and a Hyundai generator. The robot contains a ROS-enabled embedded computer that controls its motors and safety monitors.

4.1. Hardware

The hardware configuration for the field trial is displayed in Figure 6. A Clearpath Robotics Grizzly Robotic Utility Vehicle (RUV) serves as our mobile robot platform. The Grizzly is equipped with a payload that includes a suite of interoceptive and exteroceptive sensors. For the autonomous route-following field trials, the only sensors used for localization and mapping were the forward and rear Point Grey Research (PGR) Bumblebee XB3 stereo cameras labeled in Figure 6. We collected GPS data during the route for the purpose of visualization only. All of our VT&R code ran on a Lenovo W540 laptop with an Intel® Core™ i7-4800MQ CPU. The static experiment reused the same stereo camera (i.e., PGR Bumblebee XB3 stereo) used on the robot, only mounted on a tripod.

4.2. Environments

This section provides an overview of the varying environments where we collected our datasets. They are organized into static time-lapse imagery and autonomous route-following field trials.

4.2.1. Static Time-lapse Imagery

We performed a static experiment in order to tune the color-constant images, collecting stereo time-lapse imagery from sunrise to sunset in environments related to Forest and Desert biomes. Figure 7 shows key examples of those two biomes. For each dataset, a rectified stereo image pair was collected every 10 min from sunrise to sunset, resulting in a collection of approximately 60–70 stereo image pairs. The Desert dataset was collected on May 24, 2014, on a sunny day, between the hours of 07:00 and 21:00. The second dataset representing a Forest biome was recorded on November 20, 2015, on a partly sunny day, between the hours of 07:00 and 17:00. Over the course of the day, large fast-moving clouds were passing over the sun, causing the scene to constantly switch between sunny and overcast conditions.

4.2.2. Autonomous Route-following Field Trials

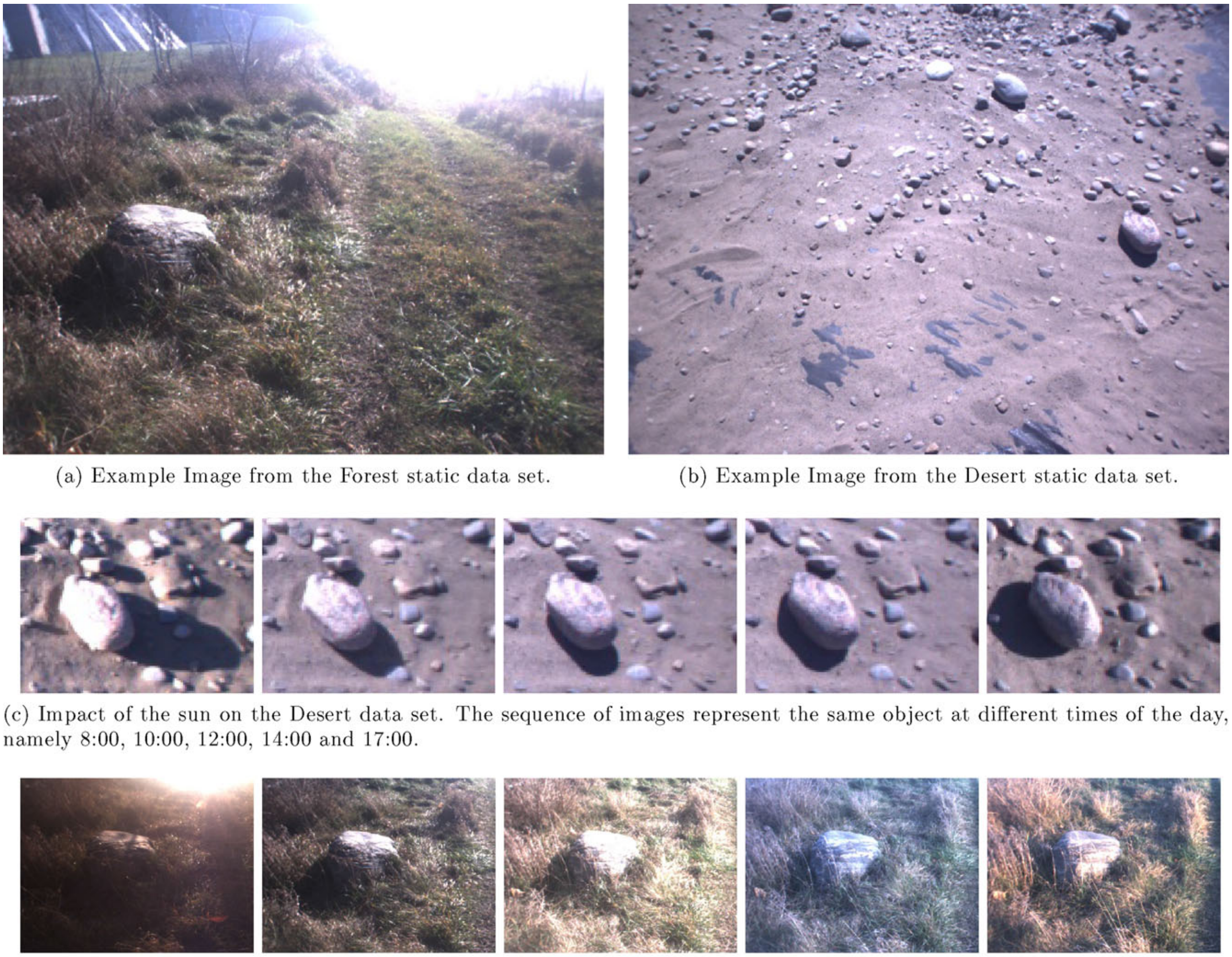
We also conducted a series of extensive field trials to properly test the different variations of the multichannel VT&R algorithms in realistic, outdoor settings. Over the course of a year, three distinct field trials were performed spanning multiple biomes and seasonal conditions: (i) summer, (ii) winter (no snow), and (iii) winter (with snow). These biomes are on display in Figure 8.

Summer. The first field trial was held at the CSA's MET in Montreal, Quebec, with the purpose of stress-testing the lighting-resistant VT&R method. The MET is a 60 m by 120 m environment consisting of sand and rocks, emulating the surface of Mars. We chose the MET as an ideal testing environment for our algorithm due to the proximity of rock/sand and grass/forest regions, providing the possibility for a single route to contain both biomes. To test our algorithm, we taught an approximately 1 km route in sunny conditions and repeated the route 26 times over the course of two days, testing localization from sunrise to sunset. Information specific to each repeat is detailed in Table I. An illustration of the sun's elevation, which effects the length of shadows on the scene, for each repeat of the field trial can be seen in Figure 9.

The 1 km route is displayed in Figure 10(a). It begins in the MET and travels approximately 300 m through sand and rocks before entering into the adjacent field. The route then snakes through the field passing by grass, trees, and a gravel roadway. The route then passes through a wooded area, traveling alongside a stream. The route reenters the MET, makes a short loop, and finishes where the route begins. This route was taught at 10:50 a.m. on the first day of the trial during bright sunny conditions with strong shadows.

The next two experiments demonstrate the impact of seasonal changes on visual navigation systems as well as investigating the dual-camera VT&R system. Both trajectories are represented in Figure 10(b). We conducted a set of trials in a meadow and a field covered by snow surrounding the UTIAS campus with the purpose of testing the limits of vision-based navigation algorithms in challenging winter environments.

Winter (no snow). This dataset was designed to test our system's robustness against lighting change and sun-stare in a challenging environment. The recording occurred in



(a) Example Image from the Forest static data set.

(b) Example Image from the Desert static data set.

(c) Impact of the sun on the Desert data set. The sequence of images represent the same object at different times of the day, namely 8:00, 10:00, 12:00, 14:00 and 17:00.

(d) Impact of the sun on the Forest data set. The sequence of images represent the same object at different times of the day, namely 8:00, 10:00, 12:00, 14:00 and 16:00.

Figure 7. Example images from the static experiments. These experiments were performed to find color-constant image transformations that maximize the descriptor-matching performance across lighting changes in different biomes. For these experiments, two scenes were selected: the first is an environment with green vegetation, which can be associated with a Forest biome, and the second is an environment with rocks and sand, which can be associated with a Desert biome.

the early winter, before large snowstorms covered the entire landscape. Displayed in Figure 8(b), this environment consists of a large field containing dead vegetation and sparse snow patches surrounded by trees and buildings in the background. Winter environments are difficult for vision systems for a number of reasons: (i) dead vegetation is uniform in color and often matted to the ground, producing little contrast; (ii) tall grass moves with the wind, resulting in keypoint matches that are inconsistent to the movement of the robot; (iii) small patches of snow shrink and change shape as they melt; and (iv) low sun elevation accelerates lighting change between traverses and is often directly in the camera's field of view, which significantly changes the exposure of the image. This field trial proceeded by teaching an approximately 100 m loop through this environment. The path was taught when the sun was at its highest elevation point. The robot autonomously repeated the path seven times between 15:20 and sunset (16:50) when the sun was setting (i.e., sunset happens much earlier during winter).

Winter (with snow). This dataset was designed to test our system's robustness against autonomous navigation through snowy environments. Snow is an especially difficult environment for vision-based systems as it is practically

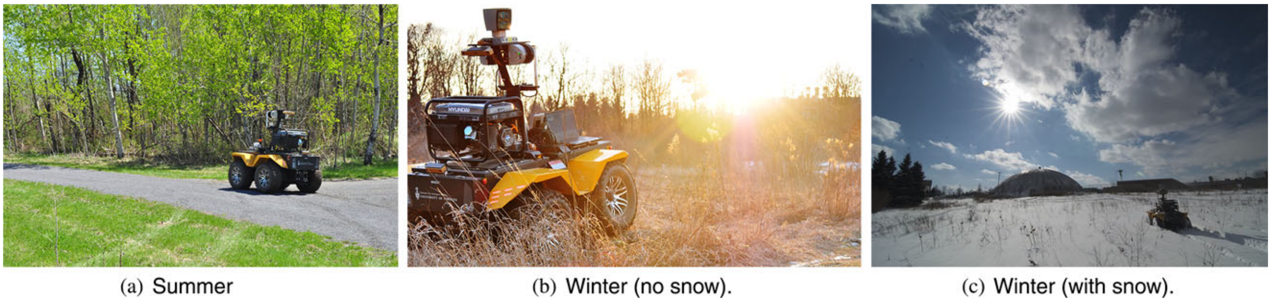

(a) Summer (b) Winter (no snow). (c) Winter (with snow).

Figure 8. Overview of the biomes covered in the autonomous route-following data sets. (a) The Canadian Space Agency (CSA) Mars Emulation Terrain (MET) in the summer, which consists of lush vegetation as well as rocks and sand. (a) A winter meadow consisting of dead vegetation, sparse snow patches, and trees at the horizon. (b) An open field with dead vegetation breaking through a 30 cm snow cover.

contrast-free, resulting in a lack of visual keypoints in most of the scene. Snow cover changes shape quickly as well. It accumulates, melts, turns to ice, and can be blown by the wind, changing the shape of the ground within minutes. Snow is also highly reflective; on sunny days this can lead a camera's autoexposure to generate images that are overexposed. An example of this environment can be seen in Figure 8(c), where the Grizzly is traversing through a snow-covered field. A 250 m path was manually driven through a large field with fresh snow cover as a teaching pass. During the teaching, the sun was at its highest point in the sky, causing significant overexposure of the camera. The path was autonomously repeated approximately 3 h later, when the elevation of the sun was significantly different. The complexity of the deployment and hardware limitations during this cold and windy day led to a smaller number of repeats compared to the other dataset. Nonetheless, it is enough to draw a comparison with other environments.

4.3. System Configuration

This section provides a brief overview of our system configuration and algorithm parameters since they can influence our results. More precisely, we detail the following steps of the localization pipeline: keypoint detection, keypoint matching, and outlier rejection. Details on the relevant parameters used can be found in Table II.

4.3.1. Keypoint Creation

Keypoints in all field trials were detected and described with upright SURF using the GPU SURF library (Furgale & Tong, 2010). In this implementation, detected keypoints are binned to ensure a uniform distribution across the image.

4.3.2. Keypoint Matching

Given a query keyframe, K_q, and a reference keyframe, k_r, our matching method seeks to find a match for every keypoint in k_r. Candidate matches are considered if the descriptor distance is below a threshold and they are within the current searching window, which filters out matches that are too far from the keypoint in pixel space. This window expands during localization failures for a more thorough search for candidate matches. For each keypoint in k_r, the candidate match with the highest matching score is selected.

4.3.3. Outlier Rejection

Our outlier rejection method uses a simple RANSAC implementation that uses the Horn three-point method (Horn, 1987) as its model. Potential inliers are evaluated based on the reprojection error after being transformed to the candidate solution.

4.4. Evaluation Metrics

To evaluate the impact of appearance change on visual navigation, we selected three quantitative metrics: keypoint quantity, keypoint sparsity, and keypoint quality.

4.4.1. Keypoint Quantity

This is a notion of the amount of total inlier matches observed at any point in time between the live keyframe and the map keyframe during an autonomous traverse. Over the course of a day, this number decreases; if it drops too low, the system will be forced to rely on VO, and eventually fail to localize relative to the taught route.

If the system is unable to relocalize to the taught route, within a prespecified distance, the system is set to stop. This metric is analyzed in Section 5.2 with respect to color-constant images, and in Section 5.3 with respect to multiple stereo cameras.

4.4.2. Keypoint Sparsity

The keypoint count alone is an insufficient metric to ensure precise route following. During an autonomous traverse,

Table I. Overview of all runs realized in three different data sets. The column Δt corresponds to the duration between the teach run and the repeat run. For all data sets, the teach run is always the first one and was manually driven.

	ID	Start time	Duration (hh:mm)	Δt (hh:mm)	Sky condition	Autonomy (%)	Distance (m)	Nb images
Summer	c0	2014/05/12 10:35	00:34	00:00	sunny	Teach Pass	954	32,347
	c1	2014/05/12 11:40	00:28	01:05	sunny	100%	960	25,721
	c2	2014/05/12 12:53	00:27	02:18	sunny	100%	952	24,863
	c3	2014/05/12 13:35	00:26	03:00	sunny	100%	947	24,553
	c4	2014/05/12 14:55	00:31	04:20	sunny	100%	960	30,036
	c5	2014/05/12 16:06	00:32	05:31	cloudy	100%	959	30,323
	c6	2014/05/12 17:27	00:26	06:52	sunny	100%	955	25,520
	c7	2014/05/12 18:14	00:27	07:39	sunny	99.2%	962	25,815
	c9	2014/05/12 19:29	00:25	08:54	sunny	100%	952	23,808
	c10	2014/05/12 20:06	00:25	09:31	sunset	100%	956	24,215
	c11	2014/05/13 06:20	00:28	19:45	cloudy	100%	955	25,126
	c12	2014/05/13 07:05	00:27	20:30	cloudy	100%	956	23,166
	c13	2014/05/13 08:00	00:28	21:25	cloudy	100%	959	23,350
	c14	2014/05/13 09:00	00:25	22:25	cloudy	100%	961	23,830
	c15	2014/05/13 10:00	00:23	23:25	cloudy	100%	955	23,185
	c16	2014/05/13 11:00	00:25	24:25	cloudy	100%	954	23,212
	c17	2014/05/13 12:00	00:29	25:25	cloudy	100%	956	26,847
	c18	2014/05/13 13:00	00:25	26:25	cloudy	100%	944	23,164
	c19	2014/05/13 14:00	00:29	27:25	cloudy	100%	948	24,050
	c20	2014/05/13 15:10	00:26	28:35	cloudy	100%	956	24,587
	c21	2014/05/13 16:00	00:27	29:25	cloudy	100%	951	23,775
	c22	2014/05/13 17:00	00:24	30:25	cloudy	100%	960	23,022
	c23	2014/05/13 18:00	00:32	31:25	cloudy	100%	960	25,582
	c24	2014/05/13 19:00	00:25	32:25	cloudy	100%	959	25,036
	c25	2014/05/13 20:00	00:25	33:25	sunset	99.9%	959	23,686
	c26	2014/05/13 20:30	00:07	33:55	dark	failed	312	9,441
Winter (no snow)	m0	2015/01/28 12:28	00:03	00:00	sunny	Teach Pass	113	7,028
	m1	2015/01/28 12:37	00:03	00:09	sunny	100%	114	6,638
	m2	2015/01/28 15:23	00:04	02:55	sunny	100%	114	7,084
	m3	2015/01/28 15:39	00:04	03:11	sunny	100%	114	7,152
	m4	2015/01/28 16:07	00:03	03:39	sunny	100%	114	5,842
	m5	2015/01/28 16:22	00:04	03:54	sunny	100%	114	6,834
	m6	2015/01/28 16:34	00:03	04:06	sunny	100%	114	5,910
	m7	2015/01/28 16:45	00:04	04:17	sunny	100%	114	6,814
Winter	w0	2015/01/30 13:44	00:06	00:00	sunny	Teach Pass	225	12,868
	w1	2015/01/30 13:55	00:05	00:11	sunny	100%	226	10,204
	w2	2015/01/30 15:57	00:06	02:13	sunny	100%	226	10,446
Total:	38	—	12h 11m	—	—	—	26k	725k

keypoint matches to the map can be distributed unevenly through a given route. The previously mentioned metric of keypoint quantity aggregates data through a full repeat trajectory, limiting the analysis on consecutive successful localizations. We can indirectly observe the sparsity of keypoint matches by observing the distance the robot relied on VO before being able to localize to the map. A short distance driven while relying on VO is a sign of a robust solution for the environment traversed. A system relying entirely on VO for a long period of time will increase its position uncertainty and will drift away from its reference trajectory, leading to a mission failure. In our system, if the dead-reckoning (VO) distance is over 20 m, the system will stop and the run is considered a failure. This metric is analyzed in Section 5.2 with respect to color-constant images, and in Section 5.3 with respect to multiple stereo cameras.

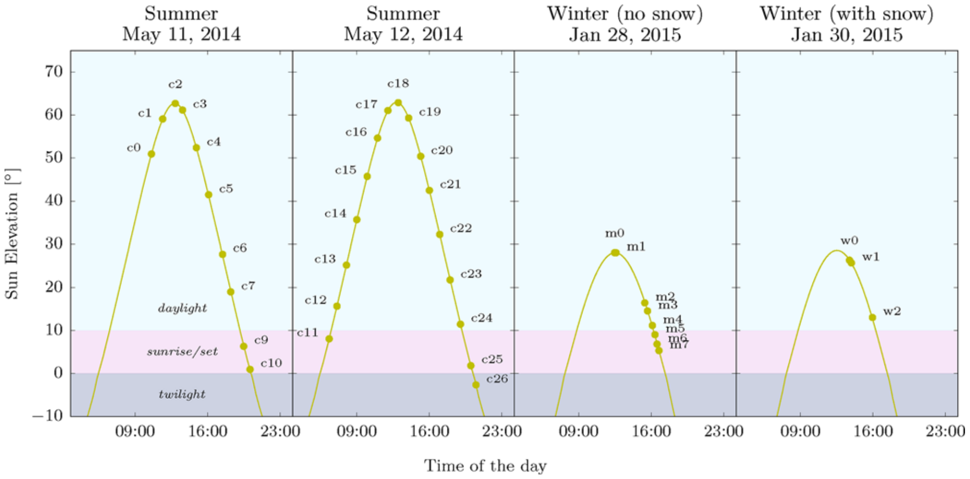
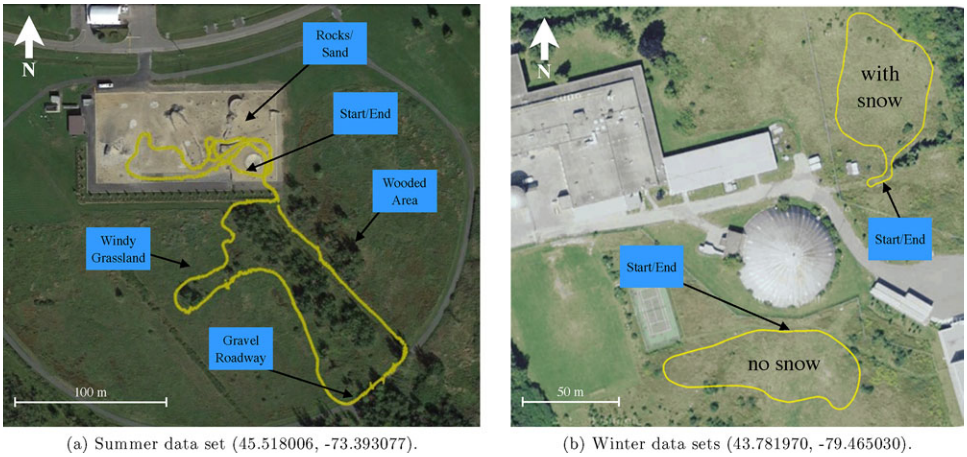


Figure 9. Overview of all recorded paths with respect to their time of day and their sun elevation. The shaded areas correspond to different elevations where the luminosity changes significantly.



(a) Summer data set (45.518006, -73.393077).

(b) Winter data sets (43.781970, -79.465030).

Figure 10. Annotated satellite imagery of the dynamic datasets recoded in different seasons. (a) Route recorded around the CSA’s Mars Emulation Terrain. Two biomes of interest are present in this dataset: Forest and Desert. (b) Routes recorded on the campus of the University of Toronto Institute for Aerospace Studies (UTIAS). The trajectory selected when there was no snow is at the bottom, and the route with snow is at the top of the image. Credit for the satellite imagery: Imagery ©2015 Google, Map data ©2015 Google.

4.4.3. Keypoint Quality

Apart from quantity and sparsity, keypoint *quality* is equally important in judging the accuracy of localization. 3D landmarks measured with a stereo camera have depth uncertainty associated between the left and right keypoint matches. As this disparity decreases, the depth becomes more sensitive to these changes, and uncertainty associated with the depth reconstruction increases. High uncertainty is correlated to keypoints observed far from the camera (i.e., in the background of the image). A reliance on background keypoints will lead to an inaccurate translation estimate, providing poor information to the path tracker and potentially endangering the vehicle. This metric is analyzed in Section 5.4 with respect to expected keypoint quality in varying environments.

5. RESULTS

The goal of this section is to demonstrate significantly improved robustness to temporal and environmental change when multichannel localization systems are used. This is achieved through metric analysis covering hourly changes, weather changes, seasonal changes, and different biomes using the different evaluation metrics described in the previous section. An overview of the evaluated techniques is provided in Table III.

Table II. Relevant parameters for the autonomous route-following field trials.

Parameter	Value
Minimum match count for localization	6
Maximum distance allowed on dead reckoning	20 m
Translation to add keyframe	0.2 m
Rotation to add keyframe	2.0°
Target number of keypoints	600
RANSAC iterations	600
RANSAC inlier threshold	4.0 std. dev.
Matching search window (prior localization success)	11 × 8 pixels
Matching search window (prior localization failure)	133 × 100 pixels

5.1. Color-constant Images During Static Experiments

Color-constant images are used in the context of localization to increase the number of descriptor matches across lighting change. Therefore, we are motivated to find the set of weights that yield the best performance with regard to keypoint detection and matching. This can be achieved through a series of static time-lapse experiments. With a collection of stereo time-lapse data across significant lighting change, we can search for the set of weights that maximizes the number of keypoint correspondences between images.

This search can be achieved by relaxing the first constraint in Eq. (8). We decided to relax the first constraint for two reasons: (i) this constraint is based on the assumption of infinitely narrow peak wavelength responses, which is far from the truth for typical sensors; and (ii) the peak wavelengths needed to calculate the weights are often unknown. Relaxing the constraint provides a free variable, α, with the weight, $\beta = 1 - \alpha$. From here, we can perform a brute force search of a discrete set of α values centered at zero.

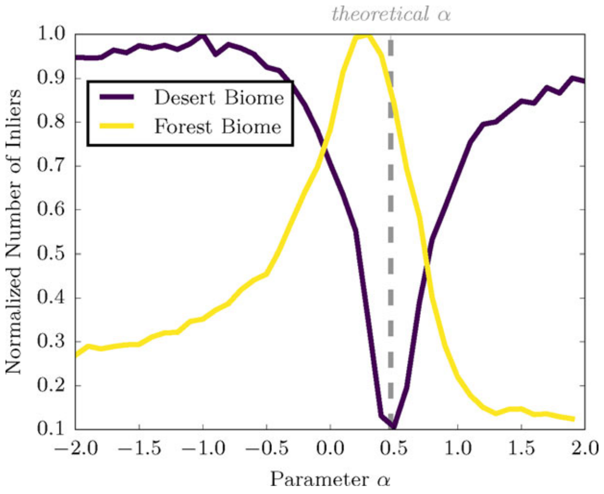


Figure 11. Performance of color-constant image transformations for the forest, desert, and snow-covered biomes. For each α value, the color-constant image transformation was tested on its ability to perform descriptor matching across lighting change. Note the position of the theoretical peak α-value.

To experimentally find the color-constant image transformation that yields the highest number of keypoint matches, we applied the following procedure for each biome. Color-constant images for the time-lapse data were calculated for each α value ranging from −2.0 to 2.0 in increments of 0.1. Each possible image pairing underwent the same localization process detailed in Section 3.1, including left-right matching, triangulation, query-map matching, and outlier detection. Inlier matches were then summed to produce a total amount of matches for each α value. Results from the experiment for both biomes are detailed in Figure 11, with the theoretical weight value highlighted. It is

Table III. Overview of the solutions evaluated.

Solution name	No. Channels	Precision
Legacy	1	Relies on the green channel of an image to extract keypoints, as in Furgale and Barfoot (2010).
Forest-CC	1	Converts the RGB channels to a single color-constant channel with the parameters tuned for Forest biomes.
Desert-CC	1	Converts the RGB channels to a single color-constant channel with the parameters tuned for Desert biomes.
Best Fit	1	Selection of either the Legacy or Forest-CC channel based on the number of keypoints extracted as in McManus et al. (2014a).
Lighting-resistant	3	Combination of Legacy, Forest-CC, and Desert-CC channels.
Dual Legacy	2	Combination of the Legacy channel for the rear and front camera.
Dual Lighting-resistant	6	Combination of the Lighting-resistant channel for the rear and front camera.

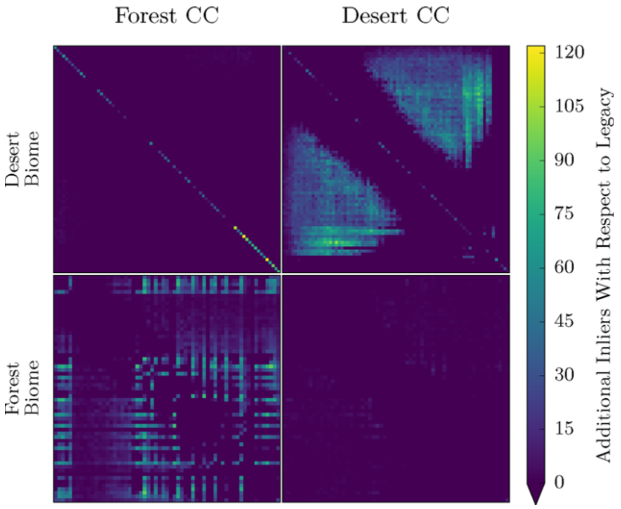


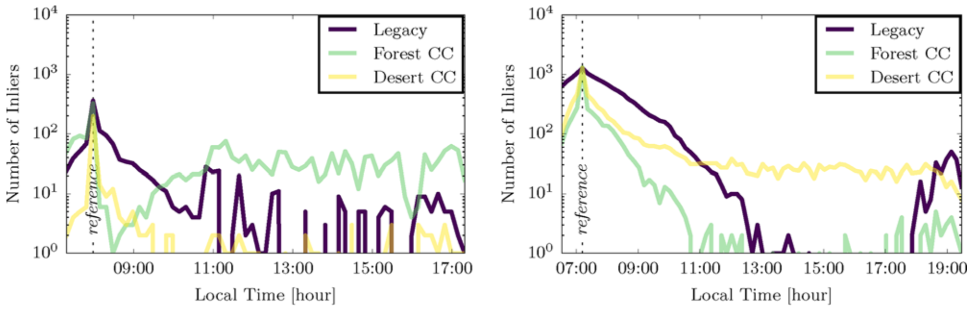
Figure 12. Performance gain over the Legacy system for the experimentally tuned color-constant images for each biome. Each quadrant represents the results of the specified color-constant image (horizontal label) in the specified biome (vertical label). Rows and columns in each quadrant represent images separated by 10 min. The matrix represents the comparison of a given image (rows) with all other images in the dataset (columns), subtracted by the amount of inlier matches found in the Legacy, gray-scale images.

interesting to note that the Forest and Desert biomes yielded nearly inverse results. The results of the search found two complimentary color-constant images, *Forest CC* and *Desert CC*, where Forest CC is Eq. (7) with weights $\alpha = 0.3$ and $\beta = 0.7$, and Desert CC is Eq. (7) with weights $\alpha = -1.3$ and $\beta = 2.3$. It is worth noting that, while close to the forest biome peak, the theoretical α value of 0.467 produces inferior results in both biomes. A possible explanation is the significant overlap of the sensor response channels seen in Figure 4.

These findings are backed by experimental results in Figure 12, which shows the gain of inlier matches over the Legacy system (i.e., gray-scale images) for both the Forest CC and Desert CC images over all possible image matches in both datasets. It can be seen from the results that the color-constant images that have been tuned for their respective environments outperform both the standard gray-scale images and the other color-constant images. A more detailed look at color-constant performance is highlighted in Figure 13. This figure shows the amount of inlier matches between a reference image and all other images in the experiment for the color-constant images and the Legacy image. Both of the graphs represent the fourth row of the matrices represented in Figure 12. It can be seen from the results that the color-constant images that have been tuned for their respective environments outperform both the standard gray-scale images and the other color-constant images. While the parameters for the Forest CC image found are close to the theoretical transformation, the Desert CC parameters are not. It is unclear why the Forest color-constant image would underperform in this context. Further investigation of this question is warranted.

5.2. Color-constant Images During Route-following

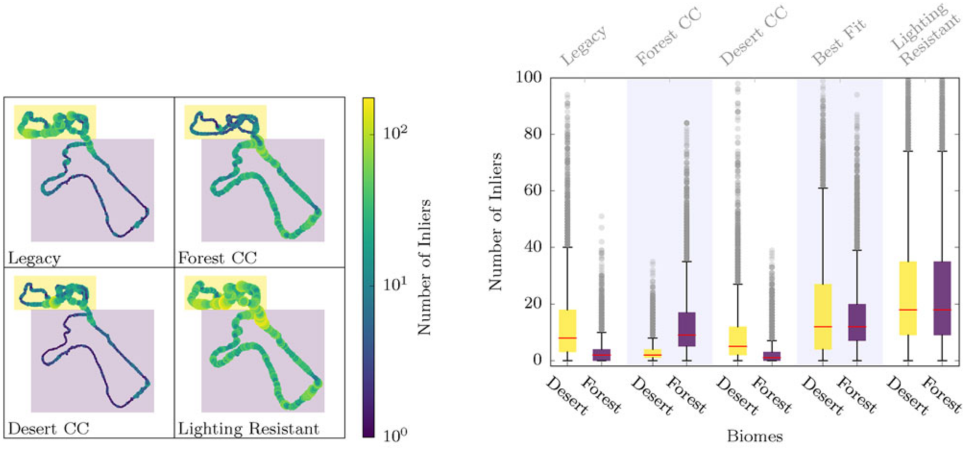
Results of the static experiments demonstrated a significant improvement in terms of an increase in descriptor matches for the different color-constant solutions (i.e., Forest CC and Desert CC). However, these experiments do not cover the performance of color-constant images when the viewpoint is not perfectly aligned, which is the typical case for autonomous route-following. To test this sensitivity, we analyze the performance of our multichannel VT&R system using the summer dataset, where the route was autonomously repeated 26 times. We use the results to validate our trained



(a) Results for the Forest data set.

(b) Results for the Desert data set.

Figure 13. Influence of the predominant biome on the generation of color-constant images. The graphs represent the evolution of matched keypoints though time, with a static image taken every 10 min. Results show that the number of inliers for different systems (Legacy, Forest Color Constant, and Desert Color Constant) is influenced by the type of biome (Forest and Desert). Note the log scale on the y-axis.



(a) Qualitative results.

(b) Quantitative results.

Figure 14. Impact of different biomes (Desert and Forest) on the number of inlier matches for a single run 03:10 h after the original run. The quantity of inlier matches found is directly related to the stability of the localization system. (a) Maps representing the number of inliers at different locations on the path. The size and color of the points give a visual representation of the number of inliers at that location. The different biomes are represented by the shaded areas, with yellow and purple boxes representing forest and desert, respectively. (b) Box plots showing the distribution of inliers for different solutions and biomes. Results are paired by solution with the colors of the box plots representing the two biomes under evaluation.

parameters, where other factors could influence the quantity of keypoints.

We first focus on the impact of the Desert and Forest biomes on different solutions in term of keypoint quantity. We use the single repeat run c4 (03:10 h after the initial run) as a representative example. The full 1 km path was clustered in two groups, represented in Figure 14(a) as shaded areas. The color and size of the points in the figure give a qualitative representation of the behavior of the different solutions in each type of biome. We can observe that the lighting-resistant solution presents a larger stability over the full trajectory when compared to individual channels. A more quantitative evaluation is presented in Figure 14(b), where Tukey box plots are used to depict the medians and the interquartile ranges. These distributions give an idea of the expected number of inliers of an image in different biomes. We can observe that the Forest CC performs significantly better[1] in the Forest biome and inversely for the Desert CC solution. The lighting-resistant solution performs similarly across the biomes, demonstrating the complementarity of the different channels when combined together. The best-fit solution performs better compared to each individual channel (i.e., Legacy, Forest CC, and Desert CC), but it generates 33% fewer keypoints than the lighting-resistant solution due to its switching behavior.

The results of Figure 14 only describe a single run. To analyze the stability of the number of keypoints through time, we computed the median value and interquartile range for all 26 runs of the summer dataset, independent of the specific biome. These results are displayed in Figure 15, with lines representing median values and shaded areas representing interquartile ranges. Results are divided between the first and second day, where weather conditions were mostly sunny and overcast, respectively. For clarity, only the results from the Legacy and lighting-resistant solutions are presented, but the lighting-resistant solution produces more inliers than all solutions presented in Figure 14 at all points in time. In Figure 15, we can observe two major points. First, there is a stark contrast between deploying a visual route-following algorithm on overcast and sunny days. On overcast days, the sun elevation has less of an impact on the shadow positions, producing a more constant number of matches throughout the day. Second, the lighting-resistant solution always produces a larger number of inliers. Its strength is most apparent on day 1 between 14:00 and 19:00, where the reference points date from a few hours since map creation, and the interquartiles of the methods do not overlap. It is worth noting that the Legacy system drops to values close to an expected value of zero inliers per image during this time period.

The number of inliers presents only a partial view of the stability of a solution. The second part of our analysis focuses on keypoint sparsity, which can only be evaluated when a camera is moving. During an autonomous traversal, the robot attempts localization at the frame rate of the camera. If a localization attempt is unsuccessful, then the robot will use the VO output for its state estimate; keypoint

[1] We use the median outside the interquartile range of the compared distributions as a simple significance test.

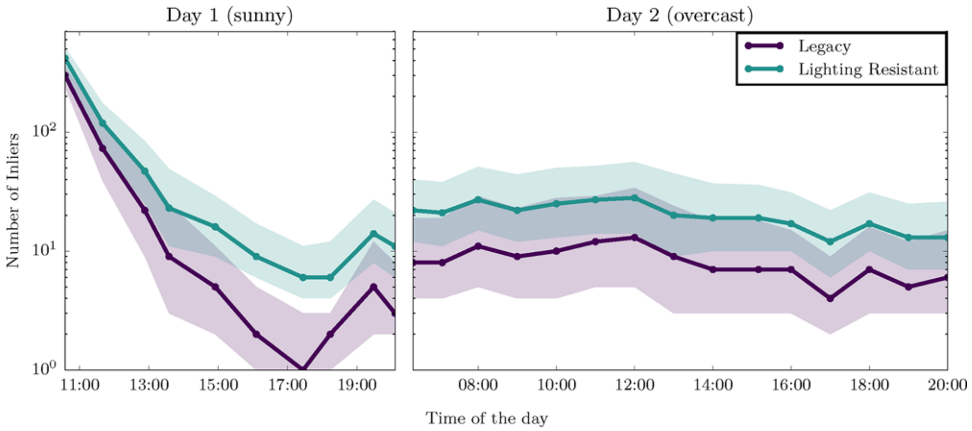
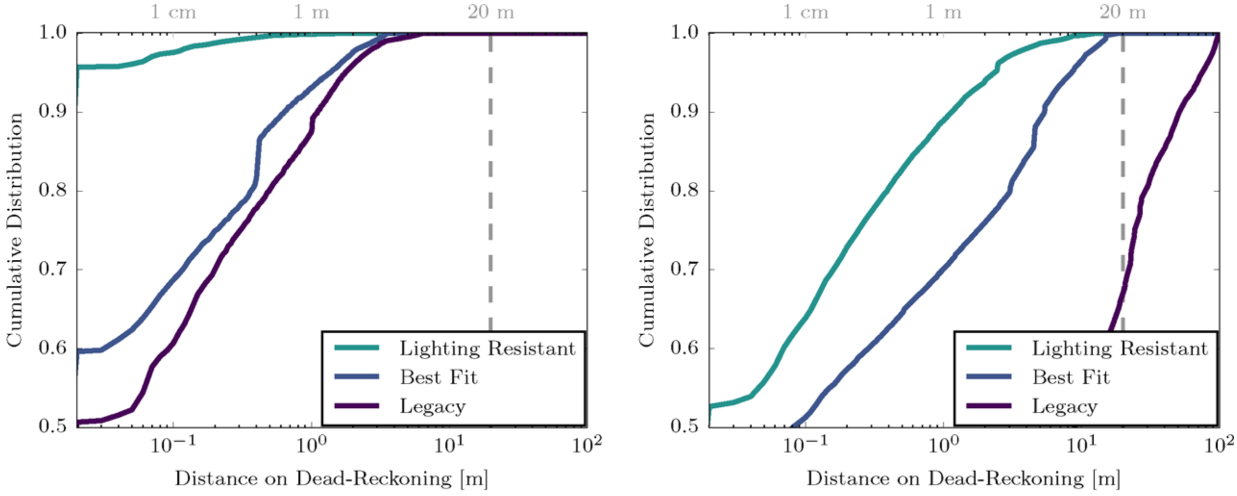


Figure 15. Evolution of the number of inlier matches through both experimental days in summer. The quantity of inlier matches found is directly related to the stability of the localization system. The advantage of the lighting-resistant solution is more significant during day one, when conditions were sunny. The lines correspond to the median number through the full 1-km-long path realized at that time of the day, and the shaded area defines the interquartile distances of 25–75 %. Note the log scale on the y-axis.



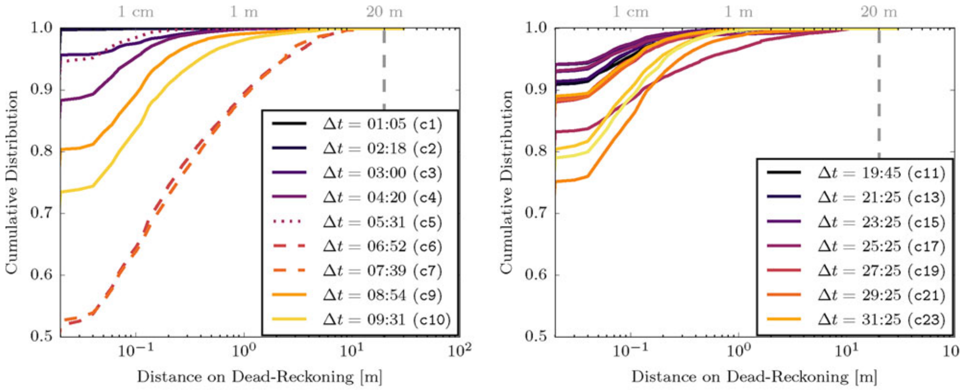
(a) Summer data set, 3 hours (c4) after map creation.

(b) Summer data set, 7 hours (c7) after map creation.

Figure 16. Comparative results between different localization systems based on the distance the vehicle would have had to travel on dead reckoning. A shorter distance indicates a more robust localization system. The dashed vertical gray line corresponds to a threshold where the autonomous drive is stopped for safety reasons, leading to a mission failure. Note the log scale on the x-axis.

sparsity measures how often this occurs. Figure 16 shows the cumulative fractional distances the vehicle traveled on dead reckoning before being able to find enough inlier matches between images from the original path and images from a repeat 3 h [Figure 16(a)] and 7 h [Figure 16(b)] after map creation. For example, the dark purple line in Figure 16(a) represents the results of the legacy system and shows that for 10% of the traverse, the robot would have driven more than 1 m on dead reckoning. This distance increases to 50 m on dead reckoning in Figure 16(b). Short distances on dead reckoning demonstrate stability through the full path and indicate a safer traverse. The graph shows that both the lighting-resistant and best-fit solutions maintain a dead-reckoning distance under our safety threshold of 20 m. The improvement of the lighting-resistant solution over the best fit is more apparent later in the day (i.e., around 20% after 7 h), where the light changes are more critical.

Keeping the same evaluation metric (i.e., keypoint sparsity), we investigate how the distance on dead reckoning evolves through time more specifically for the lighting-resistant solution. Figure 17 shows results for all 26 runs. When looking at the results of Day 1 [Figure 17(a)], we see



(a) Summer data set, day 1.

(b) Summer data set, day 2

Figure 17. Evolution of the distance traveled on dead reckoning over the course of two days for the lighting-resistant solution. A shorter distance indicates a more robust localization system. The different curves represent different runs spaced by roughly an hour. In the legend, Δt corresponds to the time separating the teach and the repeat path, and in (b) only every second label is displayed for space reasons. Note the log scale on the x-axis.

a continuous degradation of the performance through time except for three curves that stand out. During the run c5, which is represented with a dotted line, thin clouds were passing rapidly over the sun. This happened only during that run and significantly changed the temporal trend observed with the other runs, which were under a bright sun. When comparing c5 with the curves from the second day, which was overcast, we can observe a similar trend. The two curves with the worst performances (c6 and c7) occurred just before sunset, when long shadows produced images that were very different in appearance from the map images. The runs c9 and c10 were captured after sunset, when the light is very similar to an overcast day. Figure 17(b) shows the 16 runs of day 2 all clustered in the same location with slightly better results for runs close to 24 h after map creation. This shows again the positive impact of overcast days on visual route-following algorithms.

The last evaluations presented stable results through different weather conditions and biomes for the lighting-resistant solution. To push the analysis further, we investigated the impact of different seasons on the number of matches. Figure 18 shows the same solution (i.e., lighting-resistant) for the summer, winter (no snow), and winter (with snow) datasets. We can observe that the number of keypoints quickly reduces to a critical number of matches, reducing the temporal workspace of the solution. The next section investigates a proposed solution to cope with this situation.

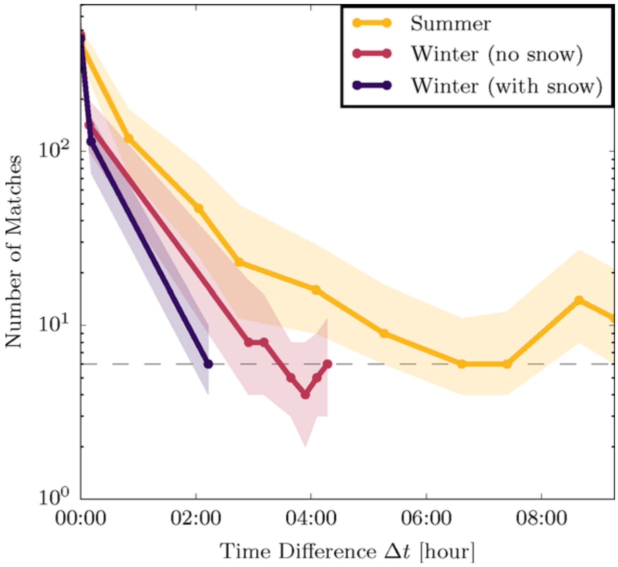


Figure 18. Evolution of the number of inlier matches for the lighting-resistant solution for multiple VT&R field trials spanning multiple seasons. All trials were conducted on sunny days when the advantages of the lighting-resistant solution are most apparent. The thick lines correspond to the median number through a full repeat path, and the shaded area defines the interquartile distance 25–75 %. The x-axis represents the time difference, Δt, between the teach and the repeat path. Note the log scale on the y-axis.

5.3. Extended Field of View

One of the problems encountered in the winter datasets is that parts of the environment that change rapidly cause large areas with few to no keypoint matches. This is partly due to the fact that during winter, the sun is lower on the horizon for a longer period of time (when away from the equator; see Figure 9). This increases the chance of sun glare completely or partially saturating an image, and it accelerates lighting change due to quickly moving, long

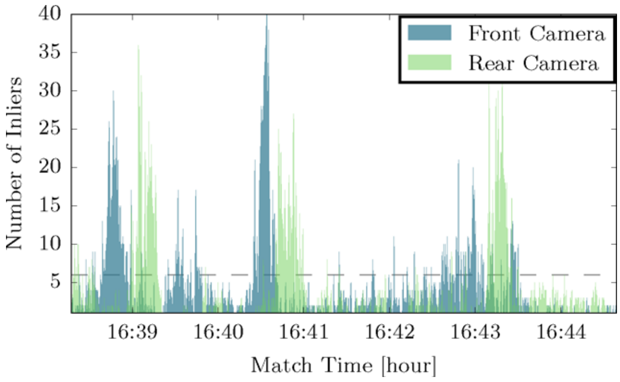


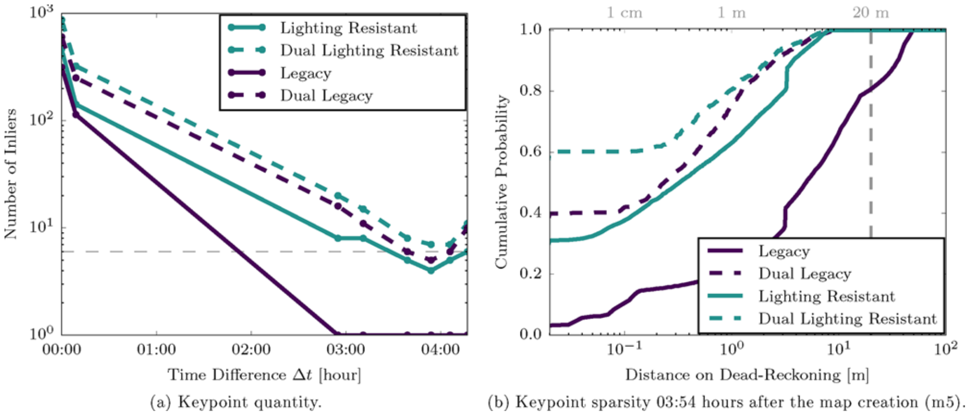
Figure 19. Evolution of the number of inliers for the front and rear cameras during a single repeat path. The dashed horizontal gray line corresponds to a safety threshold under which the vehicle is not localizing against the teach image and moves on dead reckoning. A higher number of inlier matches indicates a more robust localization system. Combining inliers from both cameras significantly increases this robustness.

shadows. Furthermore, the presence of snow leads to areas that are free of texture, which results in extremely sparse keypoint generation. With only small areas of the environment generating most of the keypoints, extending the field of view of the algorithm augments the chances of tracking a safe amount of keypoints through the whole run. In this section, we investigate the impact of adding a rear camera to cope with these harsh seasons. Figure 19 shows the impact on the number of inlier matches for an autonomous traverse. The results show a large amount of inliers being picked up by the front camera before moving to the rear camera. This handoff of keypoint matches essentially doubles the amount of time that the stable feature in the environment is observed.

Using the combination of both cameras in parallel increases the chance of maintaining enough inlier matches to safely traverse these problematic environments. This ability is analyzed by investigating the keypoint quantity in Figure 20(a) for the winter (no snow) dataset. We compare the Legacy and lighting-resistant solutions when using only the front camera (solid lines) and when using the front and rear camera (dashed lines). There is a failure case for the Legacy system using only the front camera near $\Delta t = 02{:}55$ (m2) because of the sun shining directly into the lens, which caused the system to completely lose track of its location. This repeat was not a failure case in the field because we were using the dual-camera solution. Also, we can observe that adding an extra camera to the system has a greater impact for the legacy system when compared to the lighting-resistant solution. The median number of inliers approaches the critical threshold (dashed line) for the best solution at $\Delta t = 03{:}54$ where the high contrast of the images was mostly generating silhouettes on the horizon. We used this particular run to also look at the keypoint sparsity as depicted in Figure 20(b). The dual-lighting-resistant method still performs the best, but the marginal performance improvement does not warrant the extra computation cost.

5.4. Keypoint Quality

This paper presented evidence that seasonal changes, in particular the movement of the sun through the sky, accelerates the rate at which the number of keypoint matches decays through time (recall Figure 18). A second problem amplifying the difficulties of autonomous route-following algorithms in winter is that, on top of losing keypoints, the



(a) Keypoint quantity.

(b) Keypoint sparsity 03:54 hours after the map creation (m5).

Figure 20. Impact of adding a second camera on the number of inliers and dead-reckoning distance for the winter (without snow) dataset with respect to different solutions. (a) The number of inliers through time. More inlier matches correspond to a more stable localization system. Note the log scale on the y-axis. (b) Evolution of the distance traveled on dead reckoning for multiple solutions. A shorter distance indicates a more robust localization system. The dashed vertical gray line corresponds to a safety threshold where the autonomous drive is stopped for safety issues. Note the log scale on the x-axis.

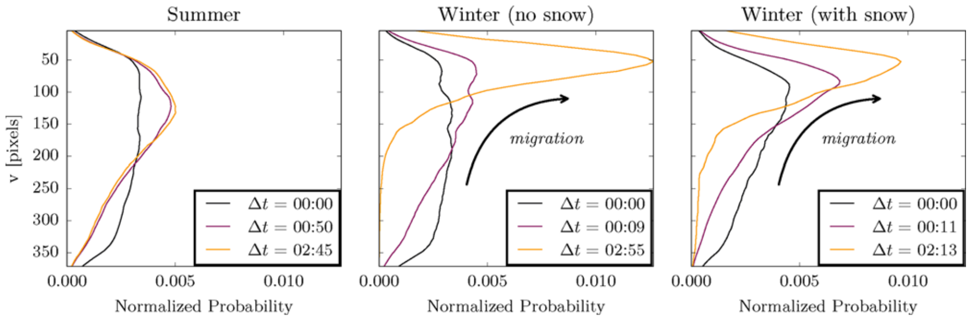


Figure 21. Vertical distribution of the matched inlier keypoints in the image coordinate frame. On the v-axis, zero corresponds to a keypoint at the top of the image and 360 at the bottom of the image. All distributions are normalized and represented over a time period of several hours for different datasets.

keypoints remaining tend to cluster on the horizon line. This phenomenon is illustrated in Figure 21, where vertical distributions of keypoint coordinates over the v-axis are presented for the three seasonal datasets. We can observe a rapid migration of the keypoints to the top of the image for both winter data sets. In all environments, it is expected that keypoints on the ground (i.e., lower in the v-axis of the image) will decay faster than higher points due to a number of reasons: (i) features seen here are observed at the cm level, while horizon features are observed at the m level; (ii) shadows have a more pronounced effect on the ground plane near the camera; and (iii) terrain modification caused by the robot. The rapid decay of close matches in the winter can be attributed to accelerated lighting change, melting snow, and the high reflectivity of snow.

This keypoint migration greatly impacts the accuracy of the localization system as keypoints with large depth uncertainty (i.e., at the horizon) reduce the accuracy of the translation estimation. The impact of the keypoints moving up to the horizon line is explained with Figure 22, where the median and the interquartile distance of keypoint depths are plotted. The expected depth for the winter dataset increases to 42.7 m, reducing the localization capability to that of a visual compass.

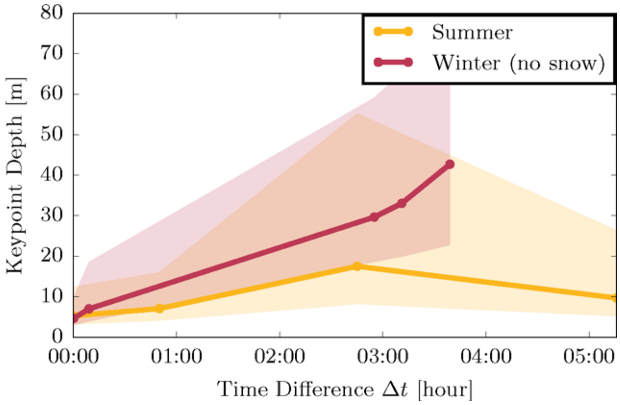


Figure 22. Comparison of the expected depth values in summer and winter. The lines represent the median, and the shaded areas represent the interquartile ranges of 25–75 %. High depth values augment uncertainty on translation estimations.

5.5. Summary

Our results have shown a significant improvement in robustness to temporal and environmental change when multichannel localization systems are used. In Section 5.1, using static time-lapse imagery of specific environments, we were able to experimentally tune color-constant image transformations to achieve superior performance with respect to SURF keypoint matching. In Section 5.2, we used these color-constant image transformations to experimentally validate our lighting-resistant, multichannel VT&R framework. We have shown a significant increase in both keypoint quantity and sparsity using our lighting-resistant method when compared to other methods [Furgale & Barfoot (2010); McManus et al. (2014a)]. Section 5.3 demonstrated a further increase in performance when multiple stereo cameras are used in a multichannel framework. By fusing data correspondences from multiple cameras into a single-state estimation problem, we essentially extend the field-of-view of the navigation system. Finally, in Section 5.4, we explored the impact of our navigation system when used in the summer and winter. Our analysis of keypoint quality shows that winter environments are still challenging for appearance-based localization due to accelerated lighting change and a lack of contrast in the scene.

6. DISCUSSION

6.1. Multichannel Localization

Our multichannel localization system performs independent detection and tracking of keypoints for multiple information channels and combines matches from all channels

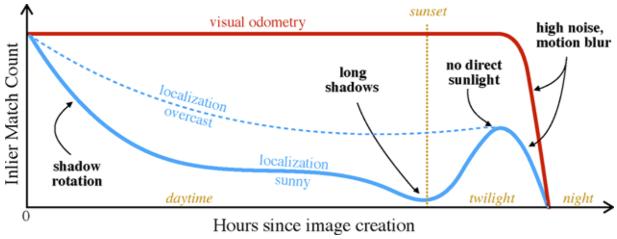


Figure 23. Illustration of the evolution of the number of inlier matches through a nominal day. Time zero corresponds to when the reference images are collected (teaching phase), and the blue line represents the typical slow degradation of the number of matches when matching current images to the teaching phase. The difference between a sunny day (solid line) and an overcast day (dashed line) is also included. The red line represents the number of keypoints used during VO, which stays constant up to the limit of the sensor. Yellow annotations refer to time events, and black annotations refer to the main causes of inlier decreases or increases.

into a single-state estimation problem. This is an important distinction from the best-fit method described in McManus et al. (2014a), where the full state estimation problem is performed in parallel for each channel, and only the best result is used as the final estimate. We argue that combining keypoint matches from multiple channels into a single-state estimation solution greatly increases the autonomy capabilities of a vision-based system when the appearance begins to change. In this case, the minimum number of required keypoints can be spread across all channels, allowing for localization in keypoint-limited environments. This is backed by our postfield analysis results (see Figures 14 and 16).

6.2. Hourly Changes

If the number of inlier matches drops too low, the system will be forced to rely on VO, and eventually it will fail at following the taught trajectory. Figure 23 shows an illustration of the trend associated with the number of inlier matches typically observed over the course of a day. This figure sums up the experience collected over all of the field trials. On overcast days there is a gradual decline in keypoint matches, because the appearance of the scene is generally constant. This is not true on sunny days, where an early drop is caused by the sun changing position and creating sharp, moving shadows on the ground. Keypoint quantity begins to rise again at the beginning of twilight, when the light from the sun is not directly observable, generating a shadowless environment similar to an overcast day. The duration and time of sunrise, sunset, and twilight are dependent on the environment. For example, if the robot is in a canyonlike environment, the sun may disappear faster than usual. As a result, modeling the correlation between factors such as sun elevation and keypoint quantity is a nontrivial task.

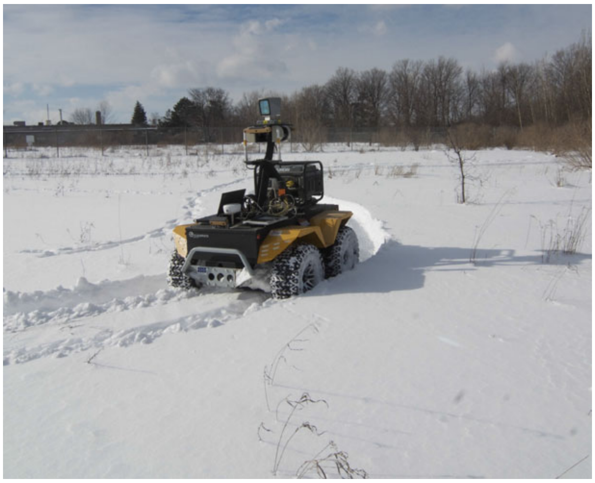


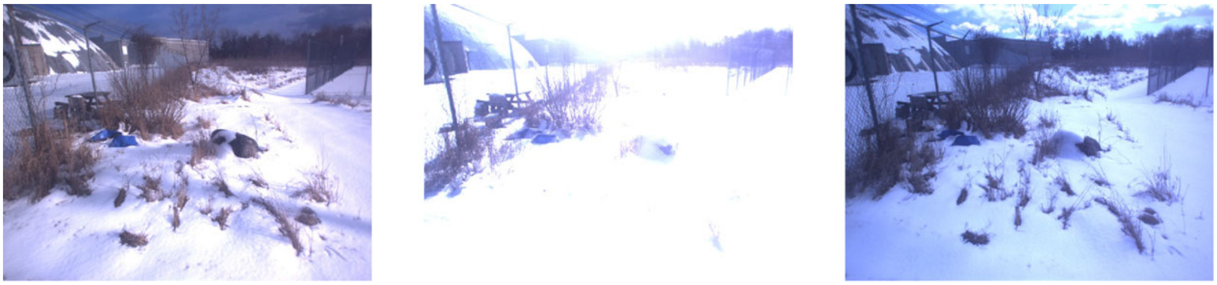
Figure 24. Photograph from an attempt in the deep snow. A lack of visual keypoints in the foreground resulted in poor localization and VO estimates. Grizzly RUV autonomously traversing in the deep snow before the failing point.

6.3. Snow

During the teaching phase of the winter (with snow) dataset, it was bright and sunny. Due to the high reflectivity of the snow, this caused unforeseen issues for our stereo cameras. The brightness of the scene brought the factory settings of the autoexposure algorithm of the PGR Bumblebee XB3 to the limit. The result was saturated images, which reduced details in the foreground. The winter (with snow) dataset was collected when there was light snow cover. We also attempted to perform autonomous path following in deep snow conditions with unsatisfactory results (see Figure 24). In light snow, small vegetation is often visible in the foreground, providing visual keypoints with high contrast. In deep snow, these keypoints are gone and what remains in the foreground is nearly featureless. The only usable matched keypoints were on the horizon, not only for localization but also for VO. This led to frequent inaccurate pose estimates, which caused issues for the path tracker. The problem of keypoint migration explained in Section 5.4 is even more apparent when a large quantity of snow is present. Furthermore, snow produces vehicle tracks that constantly change when driven over. Given the lack of other keypoints, RANSAC can easily catch those local changes, leading to a large pose estimation error.

6.4. Glare

An initial hypothesis motivating the dual-camera field deployments was the assumption that the low elevation of the sun would cause glare in the camera, making localization impossible. Due in part to the attitude of the stereo cameras, glare was not the main issue. With the cameras tilted to the ground by 20 degrees, the sun was, in the worst case, only



(a) Balanced exposure.

(b) Over exposure causing high saturation.

(c) Wrong white balance causing blue-color shift.

Figure 25. Images extracted from a static dataset recorded in a snow-covered environment. The autosettings of the PGR Bumblebee XB3 is causing artefacts in most of the images, limiting our interpretation of the calibration results in snow.

at the top of the image. Furthermore, we observed cases in which sun glare increased the contrast of horizon keypoints, providing a significant boost in keypoint count. This has an indirect impact on the keypoint quality, but it did not completely blind the camera. However, glare would be an issue if the cameras were pointed at the horizon.

6.5. Color Constancy in Winter

The color-constant image transformations are designed to remove the effects of lighting from an image. These were used to great success in the summer field trials. In these trials, the robot repeated a 1 km route 26 times with an autonomy rate of 99.9% of distance traveled in nearly every daylight condition. With this prior knowledge, the color transformations were expected to boost performance in the winter field trials as well, but this was not the case. A hypothesis is that the color-constant images were tuned to perform in green vegetation and red rocks and sand. Further investigation was performed to experimentally tune a color-constant transformation for snowy environments using the techniques described in Section 4. Results from the experiment were inconclusive, with the experimentally found Snow-CC transformation underperforming compared to traditional gray-scale images. This is primarily due to poor testing conditions, which are displayed in Figure 25. Ideally, images have a balanced exposure, as seen in Figure 25(a). Unfortunately, the majority of the images captured during the experiment were either overexposed [Figure 25(b)] or incorrectly white-balanced [Figure 25(c)]. Because we were using the automatic settings of the PGR Bumblebee XB3, which performed poorly, our confidence in the results of the experiment is low and thus not reported here.

6.6. Learning Color-constant Transformations

A concern with color-constant images is the need to tune the transformation based on the expected environment using the methods described in Section 4. The ability to learn the optimal color-constant transformation parameters for a given environment without the need for an *a priori* model is an appealing avenue of research. Without any prior data on the given environment, a good first choice is the transform obtained by choosing the camera's channel wavelength values that minimize overlap between each other. However, if the robot collects data that capture the varying appearance of the scene while it is performing autonomous traverses, it is perhaps possible to use this training data to learn the optimal color-constant transform in an introspective fashion. This, however, would likely require a full multiexperience framework and is outside of the scope of this paper.

7. CONCLUSION AND FUTURE WORK

This paper presented an autonomous route-following algorithm that takes advantage of multiple channels of information to aid localization across appearance change. A key contribution of this algorithm is that landmarks independently tracked in all channels can be used to solve a single-state estimation problem. We presented two instances of this algorithm: the first used color-constant images to increase resistance against lighting change, and the second used multiple stereo cameras to extend the algorithm's field of view. Through a series of field trials, we have shown that, through use of our multichannel localization scheme, we are able to effectively extend the autonomy rate of single-experience, vision-based route-following systems from a few hours to multiple days in realistic outdoor environments. We furthermore explored the effects of lighting change over a diurnal cycle in multiple seasons, and we quantified their influences on our localization system.

While our results show that we can extend the autonomy window of autonomous route-following algorithms through multichannel localization, this will not solve the problem of localization across longer time periods, when issues such as seasonal appearance change are a factor. A subfield of research that could potentially push autonomous route-following algorithms to true long-term autonomy solutions is multiexperience localization. Multiple experiences can be linked together to support localization, in much the same way as multichannel localization. Early work on this topic has shown promising results (Churchill & Newman, 2013; Linegar et al., 2015). We intend to explore the possibility of performing vision-in-the-loop navigation with a multichannel, multiexperience localization scheme to support navigation across seasonal changes. It is our intent to use the results of this paper to help define when new experiences are needed and what can characterize a new experience.

ACKNOWLEDGMENTS

This work was supported by the Sciences and Engineering Research Council of Canada GS501100000038 (http://dx.doi.org/10.13039/501100000038) (NSERC) through the NSERC Canadian Field Robotics Network (NCFRN).

REFERENCES

Churchill, W., & Newman, P. (2013). Experience-based navigation for long-term localisation. The International Journal of Robotics Research, 32(14), 1645–1661.

Clipp, B., Kim, J.-H., Frahm, J.-M., Pollefeys, M., & Hartley, R. (2008). Robust 6DOF motion estimation for non-overlapping, multi-camera systems. In Proceedings of the 2008 IEEE Workshop on Applications of Computer Vision.

Corke, P., Paul, R., Churchill, W., & Newman, P. (2013). Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation. In Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

Finlayson, G., Hordley, S., Cheng, L., & Drew, M. (2006). On the removal of shadows from images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(1), 59–68.

Furgale, P., & Barfoot, T. (2010). Visual teach and repeat for long-range rover autonomy. Journal of Field Robotics, 27(5), 534–560.

Furgale, P., & Tong, C. (2010). Speeded up speeded up robust features (online). Avaliable: http://asrl.utias.utoronto.ca/code/gpusurf/ (accessed: 3 March 2016).

Heng, L., Lee, G. H., & Pollefeys, M. (2014). Self-calibration and visual slam with a multi-camera system on a micro aerial vehicle. In Proceedings of Robotics: Science and Systems (RSS), Berkeley, CA, USA.

Horn, B. K. P. (1987). Closed-form solution of absolute orientation using unit quaternions. Journal of the Optical Society of America A, 4(4), 629–642.

Kazik, T., Kneip, L., Nikolic, J., Pollefeys, M., & Siegwart, R. (2012). Real-time 6D stereo visual odometry with non-overlapping fields of view. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Kneip, L., Furgale, P., & Siegwart, R. (2013). Using multi-camera systems in robotics: Efficient solutions to the npnp problem. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).

Krüsi, P., Bücheler, B., Pomerleau, F., Schwesinger, U., Siegwart, R., & Furgale, P. (2014). Lighting-invariant adaptive route following using ICP. Journal of Field Robotics, 32(4), 534–564.

Lee, G. H., Faundorfer, F., & Pollefeys, M. (2013). Motion estimation for self-driving cars with a generalized camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Linegar, C., Churchill, W., & Newman, P. (2015). Work smart, not hard: Recalling relevant experiences for vast-scale but time-constrained localisation. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).

MacTavish, K., Paton, M., & Barfoot, T. (2015). Beyond a shadow of a doubt: Place recognition with colour-constant images. In Proceedings of the International Conference on Field and Service Robotics (FSR), Toronto, ON, Canada.

Maddern, W., Pascoe, G., & Newman, P. (2015). Leveraging experience for large-scale LIDAR localisation in changing cities. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).

Maddern, W., Stewart, A., & Newman, P. (2014). LAPS-II: 6-DoF day and night visual localisation with prior 3D structure for autonomous road vehicles. In Proceedings of the IEEE Intelligent Vehicles Symposium.

McManus, C., Churchill, W., Maddern, W., Stewart, A., & Newman, P. (2014a). Shady dealings: Robust, long-term visual localisation using illumination invariance. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).

McManus, C., Furgale, P., Stenning, B., & Barfoot, T. (2012). Visual teach and repeat using appearance-based lidar. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).

McManus, C., Upcroft, B., & Newman, P. (2014b). Scene signatures: Localised and point-less features for localisation. In Proceedings of Robotics: Science and Systems (RSS), Berkely, CA, USA.

Milford, M., & Wyeth, G. (2012). Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).

Naseer, T., Spinello, L., Burgard, W., & Stachniss, C. (2014). Robust visual robot localization across seasons using network flows. In Proceedings of the AAAI Conference on Artificial Intelligence.

Nelson, P., Churchill, W., Posner, I., & Newman, P. (2015). From dusk till dawn: Localisation at night using artificial light sources. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).

Neubert, P., Sunderhauf, N., & Protzel, P. (2013). Appearance change prediction for long-term navigation across seasons. In Proceedings of the European Conference on Mobile Robots (ECMR), Barcelona, Spain.

Oskiper, T., Zhu, Z., Samarasekera, S., & Kumar, R. (2007). Visual odometry system using multiple stereo cameras and inertial measurement unit. In Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition (CVPR).

Otsu, K., Otsuki, M., & Kubota, T. (2015). Experiments on stereo visual odometry in feature-less volcanic fields. In Proceedings of the International Conference on Field and Service Robotics (FSR), Brisbane, Australia.

Paton, M., McTavish, K., Ostafew, C., & Barfoot, T. (2015a). It's not easy seeing green: Lighting-resistant visual teach & repeat using color-constant images. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).

Paton, M., Pomerleau, F., & Barfoot, T. (2015b). Eyes in the back of your head: Robust visual teach & repeat using multiple stereo cameras. In Proceedings of the 12th Conference on Computer and Robot Vision (CRV), Halifax, NS, Canada.

Paton, M., Pomerleau, F., & Barfoot, T. (2015c). In the dead of winter: Challenging vision-based path following in extreme conditions. In Proceedings of Field and Service Robotics (FSR), Toronto, ON, Canada.

Pepperell, E., Corke, P., & Milford, M. (2015). Automatic image scaling for place recognition in changing environments. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).

Pless, R. (2003). Using many cameras as one. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Ratnasingam, S., & Collins, S. (2010). Study of the photodetector characteristics of a camera for color constancy in natural scenes. J. Opt. Soc. Am. A, 27(2), 286–294.

Rawlings, J., & Mayne, D. (2009). Model predictive control: Theory and design. Nob Hill Publishers, Madison, Wisconsin, USA.

Tribou, M. J., Harmat, A., Wang, D. W., Sharf, I., & Waslander, S. L. (2015). Multi-camera parallel tracking and mapping with non-overlapping fields of view. The International Journal of Robotics Research, 34(12), 1480–1500.

Van Es, K., & Barfoot, T. (2015). Being in two places at once: Smooth visual path following on globally inconsistent pose graphs. In Proceedings of the 12th Conference on Computer and Robot Vision (CRV), Halifax, NS, Canada.

Williams, S., & Howard, A. M. (2010). Developing monocular visual pose estimation for arctic environments. Journal of Field Robotics, 27(2), 145–157.