Towards lighting-invariant visual navigation: An appearance-based approach using scanning laser-rangefinders

Colin McManus *,1, Paul Furgale 2, Timothy D. Barfoot

University of Toronto Institute for Aerospace Studies, Toronto, Ontario, M3H 5T6, Canada

HIGHLIGHTS

• We apply sparse, appearance-based computer vision techniques to laser intensity images.
• We show that descriptive features are stable in laser intensity images over a 24 h period outdoors.
• We demonstrate that laser-based VO is comparable to stereo-based VO.
• Promising visual teach and repeat results are shown (teaching during the day and matching at night).

ABSTRACT

In an effort to facilitate lighting-invariant exploration, this paper presents an appearance-based approach using 3D scanning laser-range finders for two core visual navigation techniques: visual odometry (VO) and visual teach and repeat (VT&R). The key to our method is to convert raw laser intensity data into greyscale camera-like images, in order to apply sparse, appearance-based techniques traditionally used with camera imagery. The novel concept of an image stack is introduced, which is an array of azimuth, elevation, range, and intensity images that are used to generate keypoint measurements and measurement uncertainties. Using this technique, we present the following four experiments. In the first experiment, we explore the stability of a representative keypoint detection/description algorithm on camera and laser intensity images collected over a 24 h period outside. In the second and third experiments, we validate our VO algorithm using real data collected outdoors with two different 3D scanning laser-range finders. Lastly, our fourth experiment presents promising preliminary VT&R localization results, where the teaching phase was done during the day and the repeating phase was done at night. These experiments show that it is possible to overcome lighting sensitivity encountered with cameras, yet continue to exploit the heritage of the appearance-based visual odometry pipeline.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

In the absence of a Global Positioning System (GPS), robotic systems must rely on on-board sensors to navigate and explore. Appearance-based computer vision techniques offer accurate and robust methods for localization and mapping and have been at the heart of many successful mobile robotic systems, such as the Mars Exploration Rovers (MER) [1]. On the MERs, stereo cameras were used for an appearance-based motion estimation technique called visual odometry (VO) [2], which was an essential tool for correcting motion estimates after severe wheel slip. Over the past decade, there has been a surge in the development of local appearance-based computer vision techniques for robotic vision, such as visual odometry, which offer great promise for improving the navigation potential of mobile robots.

Local appearance-based approaches search for a sparse set of descriptive regions in an image and describe these regions based on the intensity patterns of the surrounding area in image space (e.g., Scale Invariant Feature Transforms (SIFT) by Lowe (2004)). A unique fingerprint is then assigned to each of these descriptive regions, called keypoints, so they can be identified and matched in subsequent images. Sparse appearance-based techniques have been combined with fast optimization methods to build a number of fundamental autonomy-enabling algorithms such as visual odometry (VO) [2], visual teach and repeat (VT&R) [3], and visual simultaneous localization and mapping (VSLAM) [4]. These techniques have demonstrated success in mapping and loop closing in real time, over tens of kilometres [5]. However, the Achilles heel for
This is a method that enables a robot to repeat any previously driven route fully autonomously, using only a visual sensor. VT&R has numerous applications, such as military munitions convoying [14], autonomous underground tramming [15], and sample and return scenarios [3], which are of primary focus for future missions to Mars. We have tested our techniques on real data collected from two different sensors: an Optech Ilris surveying lidar detection and ranging (lidar) sensor and a high-framerate Autonosys lidar video camera (see Fig. 1). Visual odometry results using a stereo camera are also reported for comparison and differential GPS was used for groundtruth. In addition, we present a light sensitivity experiment to analyze the stability of a representative keypoint detection and description algorithm on both camera images and laser intensity images collected over a 24 h period outside. In summary, this work represents an important step towards the development of robust, lighting invariant robotic vision techniques that combine the best of both worlds: the heritage of appearance-based computer vision methods with the lighting invariance of a laser scanner.

This paper is organized as follows. Section 2 begins with a review of previous works that have used laser scanners for navigation tasks, and concludes with an outline of our four main research objectives. In Section 3, we provide a brief overview of two different ranging principles that are used by our laser scanners, highlighting the relative advantages and disadvantages of each method. Section 4 describes the image formation process where we convert raw laser data into a stack of azimuth, elevation, range, and intensity images. Using this image stack, Section 5 describes how we generate keypoint measurements and perform keypoint matching and outlier rejection. In Section 6 we describe our VO algorithm and in Section 7 we provide a hardware description of the two laser scanners used in our experiments. Section 8 describes our first experiment, where we explore the stability of a representative keypoint detection and description algorithm on both camera images and laser intensity images collected over a 24 h period. We also include an estimation error analysis between images taken at different times of the day to demonstrate the consistent performance achievable with laser scanners. Section 9 presents our outdoor VO results using two different scanning lasers with stereo VO for comparison and DGPS for groundtruth. Section 10 presents our preliminary VT&R localization results, where the teach pass was conducted during the day and the repeat pass was conducted during the night. Lastly, Section 11 presents a discussion on the lessons learned and challenges that remain with using laser scanners for appearance-based motion estimation.
2. Related work

Laser scanners are used extensively for a number of scientific activities, such as archaeological site surveying [16], forest monitoring [17], traffic construction analysis [18], medical applications [19], aerial topographic mapping [20], and robotic mapping and localization tasks [21]. In most applications that use laser scanners, the goal is to align point cloud data, referred to as scan matching or scan registration, in order to build an accurate model or map of an environment. According to Zitova and Flusser [22], there exist many different scan matching methods that vary in how they accomplish feature detection, feature matching, transforming the model estimation, and image resampling and transformation. Perhaps the most common method for scan matching is the iterative closest point (ICP) algorithm [8], which uses the nearest neighbor assumption to establish point correspondences and minimizes a sum-of-squared-error objective function to compute the point cloud transformation. Other methods have used geometric primitives to establish point correspondences, such as surface normal vectors or local curvature of the scan points [23], or search for similar geometric shapes, such as cylinders and planes [24,25]. In addition to range data, laser scanners also provide intensity information, which has proven useful for the application of various feature detector/descriptors to accomplish data association.

Recognizing that laser intensity images provide a grayscale image of a scene is not a new idea. Kretschmer et al. [18] point out that in surveying, the reflectance images are often used by the surveyor to obtain a photo-realistic impression of the scanned area. In fact, most commercial surveyors use various reflective markers in the scene to act as tie points between different scan positions to make the data association problem much easier [26]. Bohm and Becker [27] developed an automated marker-free method for point cloud registration that uses point correspondences from the intensity images to estimate the rigid body transformations. SIFT features were extracted from the intensity images and Random Sample and Consensus (RANSAC) [28] was used for outlier detection. In order to dampen the areas of low and high reflectance, histogram equalization was used on all of the raw intensity images. Abymar et al. [29] developed a technique to fuse laser data with digital camera images to generate coloured 3D point clouds. SIFT features were used to obtain correspondences between laser and the camera intensity images and RANSAC was used for outlier detection. They compared their automatic marker-free approach with a reflective marker approach and showed that the errors were of the same order of magnitude.

In the mobile robotics literature, few have actually incorporated intensity information from a laser sensor for motion estimation. Neira et al. [30] developed a sensor fusion technique in planar environments using their variant of the Extended Kalman Filter (EKF), called the SPfilter, which incorporated both range and intensity data from a laser scanner to localize against a known map. Guivant et al. [31] described a SLAM system that used the intensity data from their laser scanner to identify reflective markers on landmarks in the environment, which simplified the data association problem. A similar strategy was used in the DARPA Urban Challenge, where several groups used intensity information from their laser sensors to detect lane markers and other cars [32–34]. Yoshitaka et al. [35,36] developed what they call intensity-ICP, which uses the laser intensity data to help establish point correspondences for their ICP algorithm. Similar to classical ICP, point correspondences are still established using the nearest neighbor method, but in addition to Euclidean distance, they also try to find points that have similar intensity values. More recently, Tong and Barfoot [37] developed a self-calibrating 3D groundtruth localization system that uses laser intensity information to detect stationary retroreflective landmarks.

Although the above-mentioned research used laser intensity information to aid in data association, they did not render the intensity data into an image to use local feature-based methods, making them very different from our work. To the best of our knowledge, only one other research group has used laser intensity images for motion estimation. May et al. [38], and later Ye and Bruch [39], developed 3D mapping and ego motion estimation techniques using a Swiss Ranger Time of Flight (TOF) camera. Unlike laser scanning devices, the Swiss Ranger uses an array of 24 LEDs to simultaneously illuminate a scene, offering the advantage of higher frame rates. However, TOF cameras often have a limited field of view, short maximum range, and are very sensitive to environmental noise [38]. Weigert et al. [40] were actually the first to use a Swiss Ranger for robotics applications; however, their method, as well as others that followed [41,42], only used range data from the sensor and not the intensity data. In contrast, May et al. [38] used laser intensity images to employ two feature-based methods for motion estimation: a KLT-tracker and frame-to-frame VO using SIFT features. Their results indicated that the SIFT approach yielded more accurate motion estimates than the KLT approach, but was less accurate than their ICP method, which used a network-based global relaxation algorithm [21]. Although May et al. demonstrated that frame-to-frame VO might be possible with a high-framerate TOF camera, the largest environment in which they tested was a 20 m long indoor hallway, with no groundtruth and no variation in ambient light. Furthermore, laser scanners are very different from TOF cameras in that they scan the scene with a single light source, introducing new problems such as image formation and image distortion caused by moving and scanning at the same time. Thus, a number of important questions need to be answered regarding the use of laser scanners for appearance-based motion estimation techniques. In particular, we set out to answer the following research questions:

1. How stable are descriptive features in laser intensity images under changes in ambient light?
2. Can we perform VO using a laser scanner in a stop-scan-go methodology with comparable results to stereo VO?
3. Can we perform VO using a laser scanner under continuous motion?

With regards to VT&R systems, those systems that have used scanning laser rangefinders as the primary sensor (e.g., Marshall et al. [15]) were restricted to indoor planar environments and did not utilize intensity information from their laser scanners. To the best of our knowledge, we are the first to explore the use a 3D laser scanner for VT&R in outdoor terrain and the first to apply appearance-based techniques to laser intensity images for this task. Since VT&R relies upon a robust VO pipeline, as an addendum to the third research question, we also sought to answer the following:

4. Can we perform VT&R localization using a laser scanner under continuous motion?

As this paper will show, it is indeed possible to use the intensity images from a laser scanner and apply appearance-based techniques that have been traditionally used with camera imagery. We show that the accuracy of our laser-based VO is comparable to stereo VO and include results from two different laser scanners. We also demonstrate that VT&R appears not just feasible, but in fact very promising, achieving an average localization error of just 0.36 m over a 170 m traverse.

This work is an extension of the work presented by McManus et al. [43], in the following ways: (i) we have extended our light sensitivity experiment to include an estimation error analysis between images taken at different times of the day, (ii) we have conducted field tests with a high-framerate Autonosys sensor in order to validate our approach on a moving platform in outdoor terrain, and (iii) we have included preliminary VT&R localization results using a high-framerate Autonosys sensor in outdoor terrain.
3. Ranging principles

This section will introduce the two most common methods for determining range and intensity in laser scanners, as it directly pertains to the devices used in this study. We consider two different light detection and ranging (lidar) sensors: one based on emitting individual pulses and the other based on emitting an amplitude modulated continuous wave. The basic theory behind the two different ranging principles will be discussed, as well as their respective advantages and disadvantages.

3.1. Pulsed time of flight (TOF) lidar

As the name implies, time of flight (TOF) lidar sensors compute range based on the time between the emission and return of a laser signal. In the case of a pulsed TOF lidar, a very precise time measurement is recorded from the moment a laser pulse is emitted to when it reflects back to the detector. Using the fact that the speed of light is a constant, one can directly compute the distance to an object, d, based on the time measurement, t, according to

\[ d = \frac{ct}{2} \]

Most lidar detectors use an avalanche photodiode (APD), which applies a high reverse bias voltage in order to increase the gain of the return signal [44]. For good ranging accuracy, pulsed TOF lidars generally increase the bias voltage over time in order to amplify weak signals that have returned from long distances. However, when the bias voltage is increased too high, the process can become nonlinear, making the recovery of the original shape and size of the pulse difficult. Since the shape and size of the pulse is related to the reflectivity, or intensity, pulsed TOF sensors inherently lose some intensity information through this amplification process, but have the benefit of long-distance ranging.

3.2. Amplitude modulated continuous wave (AMCW) lidar

Amplitude modulated continuous wave (AMCW) lidar uses the phase-shift principle to compute range, which measures the shift in phase of a reflected sinusoidally-modulated laser signal. Using this phase shift, \( \phi \), as well as the modulation frequency, \( \lambda \), one can compute the time of flight of the laser pulse, t, according to

\[ t = \frac{\phi}{2\pi\lambda} \]

which can then be used to compute range. Intensity information is determined by the difference in amplitude between the emitted and returned signal. The benefit of AMCW lidar versus pulsed TOF lidar is that it provides a larger dynamic range in intensity information, for the following reasons: (i) there is no need to increase the bias voltage of the detector since the detector is always on, and (ii) the detection process must remain linear in order to maintain the sinusoidal shape of the modulated signal. However, the disadvantage of AMCW lidar versus pulsed TOF lidar is their smaller maximum range, which results from an ambiguity in phase after half of a wavelength [45].

3.3. Beam steering

All laser scanners use rotating mirrors to steer the narrow laser beam across a raster pattern in azimuth and elevation to create range and intensity images. As this requires the beam to track a trajectory, there is always a small tracking error, which can often be measured by the mirror encoders. As will be described in the next section, we do not assume the image pixels occur regularly in azimuth and elevation but rather use the measured angles for each pixel.

4. Image formation

This section describes the image formation process where we convert raw laser data into a stack of azimuth, elevation, and intensity images. We then enhance the intensity image for use in our keypoint detector/descriptor, which is a GPU implementation of the SURF [46] algorithm. The first step in image formation is to develop a camera model. As both the Autonosys and the Iliris provide, approximately, equally spaced azimuth and elevation samples, we were able to use the raw data directly in a spherical camera model. Fig. 2 shows examples of a raw intensity image for both the Iliris and the Autonosys, as well as a camera image of the scene for comparison. As is immediately evident, the raw laser intensity images require some pre-processing in order to equalize the areas of high and low reflectance. Taking a similar approach as Bohm and Becker [27], we use an adaptive histogram equalization algorithm and also apply a Gaussian low-pass filter to achieve the desired enhancement (the justification for our pre-processing method will be described in more detail in Section 8). The associated azimuth, elevation, and range data is assembled into an array in the exact same order as the intensity image, forming what we call an image stack. This concept will prove useful in the next section when we introduce how to generate keypoint measurements and their associated uncertainties.

In addition to the above-mentioned preprocessing steps, filtering of the Autonosys data was also required. During our outdoor traverses, we noticed that the range data was noisy in some areas, which was a result of scattering inside the sensor that worsens when the sensor does not detect an object (e.g., the sky) or detects objects far in the distance (see Fig. 3 for an example intensity and range image gathered from outside). In order to reduce this noise in the outdoor Autonosys images, we performed the following filtering steps before applying adaptive histogram equalization: (i) remove all pixels beyond a range of 50 m and below a range of 1 m and (ii) remove all pixels with an intensity value below 150.

Although it is true that we are losing information by thresholding the images in this manner, we are also removing a great deal of noise, which can be troublesome for our feature detector and lead to erroneous keypoint measurements due to spurious range values. Fig. 3 shows the same image after applying the thresholds above, which shows a noticeable improvement in the range image, with very little effect in the intensity image.

It should be noted that there are a number of factors that influence laser intensity data, such as atmospheric attenuation, surface reflectance, local incidence angle, and range [47]. Applying squared range corrections to the intensity images seems quite common in the literature [47–49]. However, our preliminary analysis found that applying such a correction resulted in less detail in the foreground, which is not ideal for VO. We found that the adaptive-histogram approach described above worked well in practice, but this is certainly a topic open for future work.

5. Keypoint matching and outlier detection

This section describes how we form keypoint measurements and associated uncertainties using the image stack composed of azimuth, elevation, and intensity data. This measurement information is used both in our keypoint matching and outlier rejection step and forms the error terms used in visual odometry.

---

3 Multiple pulses are generally averaged in order to reduce noise [44].
5 Values based on recommendations from Autonosys Inc.
5.1. Keypoint matching

Before we outline our keypoint matching criteria, it is necessary to first introduce how we form keypoint measurements. The output of image formation is a stack of images, $I$—intensity ($I_\ell$), azimuth ($I_\theta$), elevation ($I_\phi$), and range ($I_r$)—derived from the raw lidar output and shown in Fig. 4. The intensity image, $I_\ell$, is used in our keypoint detector to detect interest points, and the azimuth, elevation, and range images, $\{I_\theta, I_\phi, I_r\}$, provide the metric information for each keypoint detection. An azimuth, elevation, and range measurement can be evaluated at any integer row, $r$, and column, $c$, as, $J_{rc}$, a $3 \times 1$ column,

$$J_{rc} := J(r, c) = \begin{bmatrix} I_{rc} \\ I_{\theta rc} \\ I_{\phi rc} \end{bmatrix},$$

where $I_{rc}$, $I_{\theta rc}$, and $I_{\phi rc}$ are the scalar azimuth, elevation, and range stored at this location in the image stack. We assume that the
Fig. 3. Illustrating the filtering process required for the outdoor Autonosys data. Unfiltered images are shown on the top row and their filtered versions are shown on the bottom row. All images are 480 × 360 with a 30° V × 90° H FOV. Most of the noise presents itself in the range images, as can be seen in 3(b). After we apply the thresholding, we lose some detail off in the distance, but it removes most of the noise in the sky.

Fig. 4. The image stack generated from the raw laser-rangefinder data. SURF keypoints are found in imagespace at sub-pixel locations and bilinear interpolation is used to find the azimuth, elevation, and range of the keypoint. Linearized error propagation from imagespace to azimuth/elevation/range is then used to determine the uncertainty of the measurement.

elements of each image are independent, identically-distributed samples such that

\[ \mathbf{I}_R = \mathbf{I}_{rc} + \delta \mathbf{I}_{rc} \sim \mathcal{N} (0, \mathbf{R}) , \quad \mathbf{R} := \text{diag} \left\{ \sigma_\theta^2, \sigma_\varphi^2, \sigma_r^2 \right\}, \]

where \( \mathbf{I}_{rc} \) is the true value, \( \delta \mathbf{I}_{rc} \) is zero-mean Gaussian noise, and the components of \( \mathbf{R} \) are based on the properties of the sensor (e.g., taken from the datasheet).

Keypoint detection returns a list of image locations, \( \mathbf{y}_i = [u \ v]^T \), with associated covariances, \( \mathbf{Y}_i \), where \( u \) and \( v \) are generally not integers. We use bilinear interpolation of \( \mathbf{I} \) to produce an azimuth/elevation/range measurement, \( \mathbf{z}_i \). Defining our bilinear function as \( \mathcal{B}(\cdot) \), and given the position of a keypoint in image space, \( \mathbf{y}_i \), we take the neighboring set of points in the azimuth, elevation, and range images, \( \delta_i := \{ \delta_\theta, \delta_\varphi, \delta_r \} \), and compute the interpolated azimuth, elevation, and range value, \( \mathbf{z}_i = \mathcal{B}(\mathbf{y}_i, \delta_i) \). The uncertainty, \( \mathbf{Q}_i \), associated with \( \mathbf{z}_i \), is produced by propagation of \( \mathbf{R} \) and \( \mathbf{Y} \) through the interpolation equations, such that \( \mathbf{Q}_i = \mathbf{J}_i \mathbf{Y}_i \mathbf{J}_i^T + \mathbf{R} \), where \( \mathbf{J}_i = \frac{\partial \mathcal{B}}{\partial \mathbf{y}} \bigg|_{\mathbf{y}_i} \).

One of the challenges with using laser scanners in non-convex environments is that slight changes in sensor orientation can result in large range deviations depending on the geometry of the scene (e.g., objects that have a high angle of incidence with respect to the laser beam or thin objects can display large range deviations). To illustrate this, we have included a pixel-wise standard-deviation range image in Fig. 5, which is comprised of 50 range images of the same scene, taken with the Ilris sensor. As can be seen, objects like the power line, the top of the dome and the chain-link fence represent regions were the range can deviate significantly.

As we use bilinear interpolation to compute keypoint measurements, if neighboring range values have a relatively large deviation (e.g., 20 cm) then the keypoint range uncertainty will also be large (see Fig. 6 for an illustration). To account for these large uncertainties, we use Mahalanobis error metrics [50] for hypothesis generation and outlier detection. This means that as long as the Mahalanobis distance remains below a certain threshold, our algorithm can track keypoints with large range uncertainty, provided that this range deviation is not the result of a jump discontinuity that occurs at a structure boundary. The reason we discard keypoints at structure boundaries is because they result in biased measurements since the bilinear interpolation inherently assumes that the local geometry is linearly continuous.

Now that we have introduced our method for generating keypoint measurements, we can outline our keypoint matching criteria. Defining the measurement of keypoint \( i \) in image \( n \) as \( \mathbf{z}_{n,i} \).
and the measurement of keypoint \( j \) in image \( m \) as \( z_{m,j} \), our keypoint selection criteria, in order, are:

1. Range values must agree to within some tolerance: \( \| r_{n,i} - r_{m,j} \| < \delta_r \).
2. Keypoint azimuth/elevation angles must be located within a local neighborhood: \( \left\| \Theta_{n,i} - \Theta_{m,j} \right\| < \delta_{\phi, \theta} \).
3. Keypoints must have the same Laplacian (i.e., a binary variable that distinguishes bright blobs on dark backgrounds and vice versa [46]).
4. Keypoint scales must agree to within some tolerance: \( |s_{n,i} - s_{m,j}| < \delta_s \).
5. The 64 element SURF descriptors must match to within some tolerance: \( 1 - \left\| d_{n,i}^q - d_{m,j}^q \right\| < \delta_d \).

As our sensor measurements are defined in a spherical coordinate system, we find the nearest neighbors in azimuth, elevation, and range, followed by the nearest neighbors in descriptor space. We use a generous range threshold (5 m) to prune keypoint matches at structure boundaries, which generally have significant range deviations and consequently act as outliers in the pose estimation. A local window around azimuth/elevation (10°) is used to reduce the search space for candidate features; otherwise, we would have to perform exhaustive matching across the whole image, which is too computationally expensive. If the candidate matches pass the first two tests, then we check that the scales of the SURF detections match within a certain threshold (0.9) and that the respective 64-element SURF descriptors agree, using a tolerance of 0.01.

5.2. Outlier rejection

RANSAC was used for outlier rejection and Horn’s 3-point method [51] was used for hypothesis generation. To account for measurement uncertainties and for robustness, we use Mahalanobis error metrics and the well-known German–McClure function [52] to compute the overall cost of each hypothesis. Recalling that each keypoint measurement, \( z_{k,j} \), corresponds to an observation of landmark \( j \) at time \( k \), the error term, \( e_{k,j} \), is simply

\[
e_{k,j} := z_{k,j} - g \left( x_{k-1,k}^j, P_{k-1}^j \right)
\]

where \( x_{k-1,k}^j \) is a \( 6 \times 1 \) parameterization of \( T_{k-1,k}^j \), which is the transformation from frame \( k \) to frame \( k-1 \), \( p_{k-1}^j \) is the vector from frame \( k-1 \) to landmark \( j \) expressed in frame \( k-1 \), and \( g(\cdot) \) builds the predicted azimuth/elevation/range value based on the current state estimate. Using the associated measurement uncertainty, \( Q_{k,j} \), the cost for each hypothesis is given by

\[
E_k := \sum_j w_j e_{k,j}^T Q_{k,j}^{-1} e_{k,j}, \quad w_j := \frac{1}{(e_{k,j}^T Q_{k,j}^{-1} e_{k,j} + \sigma)}.
\]

where \( \sigma \) is an \( M \)-estimator parameter. After we obtain our best hypothesis, we reject any matches where the Mahalanobis distance exceeds a certain threshold; i.e., if \( e_{k,j}^T Q_{k,j}^{-1} e_{k,j} > e_{\text{max}} \), reject.

6. Visual odometry

Visual odometry based on bundle adjustment is a core algorithm in the emerging VSLAM paradigm [4,5]. We have implemented the sliding window VO algorithm described by [53], and use the code to produce two motion estimates: one from stereo camera data, and the other from laser scans. Processing the laser data only requires differences in two areas: (i) keypoint formation, and (ii) error terms and associated Jacobians. The rest of the code blocks—keypoint tracking, hypothesis generation and outlier rejection using RANSAC, and sparse bundle adjustment—are identical to the traditional camera-based approach.

The standard error term used in VO systems is reprojection error—the difference between the observed keypoint location and the predicted keypoint location given the current state estimate. We use a similar error term based on the spherical camera model used to form the intensity images. Each measurement, \( z_{k,j} \), corresponds to an observation of landmark \( j \) at time \( k \). The error term, \( e_{k,j} \), is given by

\[
e_{k,j} := z_{k,j} - g \left( x_{0,k}, p_{0}^j \right),
\]

where \( x_{0,k} \) is a \( 6 \times 1 \) parameterization of \( T_{0,k} \), which is the transformation from frame \( k \) to the base frame, \( p_{0}^j \) is the vector from the base frame to landmark \( j \) and expressed in the base frame, and \( g(\cdot) \) is our sensor model that computes a predicted azimuth/elevation/range value based on the current state estimate. We assume that each measurement is corrupted by additive, zero-mean Gaussian noise, \( n_{k,j} \sim \mathcal{N}(0, Q_{k,j}) \), with covariance \( Q_{k,j} := J_{k,j} Y_{k,j} J_{k,j}^T + R_k \), where \( R_k \) is based on sensor specifications and \( J_{k,j} := \frac{\partial g}{\partial p} \bigg|_{p_{0}^j} \), where \( \mathbf{Z}(\cdot) \) is our bilinear interpolation function.

Given \( K \) poses and \( M \) measurements, we can write our system in matrix form as

\[
\begin{bmatrix}
  z_{1,1} \\
  \vdots \\
  z_{K,1} \\
  \vdots \\
  z_{K,M}
\end{bmatrix}, \quad
\begin{bmatrix}
  x_{0,1} \\
  \vdots \\
  x_{0,K}
\end{bmatrix}, \quad
\begin{bmatrix}
  p_{0}^1 \\
  \vdots \\
  p_{0}^M
\end{bmatrix}
\]

\[
= \mathbf{g}(x, p) := \begin{bmatrix}
  \mathbf{g}(x_{0,1}, p_{0}^1) \\
  \vdots \\
  \mathbf{g}(x_{0,K}, p_{0}^M)
\end{bmatrix}, \quad
\mathbf{Q} := \text{diag}(w_{1,1}, Q_{1,1}, \ldots, w_{K,1}, Q_{K,1}, w_{1,2}, Q_{1,2}, \ldots, w_{K,M}, Q_{K,M}),
\]

\[
\text{diag}(w_{1,1}, Q_{1,1}, \ldots, w_{K,1}, Q_{K,1}, w_{1,2}, Q_{1,2}, \ldots, w_{K,M}, Q_{K,M})
\]

\[
\text{diag}(w_{1,1}, Q_{1,1}, \ldots, w_{K,1}, Q_{K,1}, w_{1,2}, Q_{1,2}, \ldots, w_{K,M}, Q_{K,M})
\]
where we have included the Geman–McClure function,

$$w_{ij} := \frac{1}{(e_{ij}^T Q_{ij}^{-1} e_{ij} + \sigma)},$$

for robustness to outliers. Using the above quantities, we can then define the standard Mahalanobis objective function as

$$J(\mathbf{z}(x, p)) := \frac{1}{2} (\mathbf{z} - \mathbf{g}(x, p))^T \mathbf{Q}^{-1} (\mathbf{z} - \mathbf{g}(x, p)).$$

Performing a first-order linearization of the error terms about the current state estimate, \([\bar{x}, \bar{p}]\), and setting the derivative of \(J\) with respect to the perturbations of the state variables, \((\delta x, \delta p)\), to zero, yields the standard bundle adjustment system of equations \([54]\):

$$\begin{bmatrix} \mathbf{U} & \mathbf{W} \\ \mathbf{W}^T & \mathbf{V} \end{bmatrix} \begin{bmatrix} \delta x \\ \delta p \end{bmatrix} = \begin{bmatrix} e_x \\ e_p \end{bmatrix}.$$  

Through use of the Schur complement, one can manipulate the system of equations so as to exploit the sparse structure of the coefficient matrix, allowing for a much more computationally efficient solution—leading to the so-called sparse bundle adjustment formulation \([55]\):

$$\begin{bmatrix} \mathbf{U} - \mathbf{WV}^{-1}\mathbf{W}^T & 0 \\ \mathbf{W} & \mathbf{V} \end{bmatrix} \begin{bmatrix} \delta x \\ \delta p \end{bmatrix} = \begin{bmatrix} e_x - \mathbf{WV}^{-1}e_p \\ e_p \end{bmatrix}.$$  

As we method is based on the work by Konolige et al. \([53]\), within each window of nine poses, we omit six as design variables in \(\delta x\), but still optimize over all landmarks \(\delta p\). The resulting system of equations is then iteratively solved using the Levenberg–Marquardt algorithm \([56]\).

### 7. Hardware description

Our field robot was equipped with a Point Grey Research Bumblebee XB3 stereo camera, a Thales DG-16 Differential GPS unit, and two different laser scanners: an Optech Ilris 3D surveying sensor and the Autonosys LVC0702. Each of these sensors and their roles in this study will be described in this section.

#### 7.1. Point grey bumblebee stereo camera

We logged stereo images from our Bumblebee camera during all field tests in order to provide a comparison with laser-based VO. For the Autonosys field tests, we logged \(512 \times 384\) stereo images at 15 Hz for our Ilris field test we logged \(512 \times 384\) stereo images every half meter in a stop-and-go fashion as the Ilris is not a high-framerate sensor.

#### 7.2. Thales GPS

Two DG-16 Differential GPS units were used in this study: one mounted on the field robot and one set up as a static base station. For our field test conducted with the Ilris, Real Time Kinematic (RTK) corrections were used for improved GPS accuracy, however, due to interference issues, we encountered difficulties in maintaining the RTK mode for long periods of time. To remedy this for our Autonosys field tests, we used post-processed DGPS for groundtruth.

#### 7.3. Ilris

The Optech Ilris 3D is a surveying laser scanner that has a maximum horizontal and vertical field of view (FOV) of 40°. The Ilris is a pulse-based TOF lidar that can acquire 2500 points/s and offers a large dynamic scanning range from a minimum of 3 m to a maximum of over 1 km, with sub-centimetre accuracy up to 100 m (for our experiments, we restricted the maximum range to 500 m). In addition to range data, the sensor also provides 8-bit intensity information, which we use for image formation.

### 7.4. Autonosys

The Autonosys LVC0702 is a high-framerate phase-based lidar that uses a combined nodding and hexagonal mirror assembly to scan a 45°V/90°H FOV at much higher rates than traditional lidar sensors and can acquire up to 500,000 points/s. The basic premise behind the mirror design is to use a rotating hexagonal mirror for high-speed scanning in the horizontal direction and a nodding mirror to deflect the scan vertically where less angular speed is required. Nodding mirrors have the advantage of being able to collect all of the reflected light over their entire angular range, but the necessity to slow down and reverse direction of rotation at the edges of the FOV limits them to slow angular scan rates \([57]\). In contrast, hexagonal mirrors rotating at constant rate allow for fast angular scanning with no reversal in mirror motion. However, rotating polygonal mirrors have the disadvantage that some light collection efficiency is lost near the edges of their FOV. Thus, by combining both types of mirrors, the LVC0702 is able to achieve a compromise between speed and collection efficiency. The LVC0702 provides 15-bit intensity information, has a maximum range of approximately 53.5 m and can scan as fast as 10 Hz; however, increasing the frame rate results in lower image resolutions.

For our experiments, we narrowed the vertical field of view from 45° to 30° in order to capture 480 × 360 images at 2 Hz. The rational behind restricting the vertical field of view was to increase the angular resolution in the vertical direction while trying to maintain a feasible scanning rate. Higher resolution in the vertical field of view is ideal since the spacing between elevation points projected onto a flat plane increases with distance. Thus, the goal was to reduce spatial aliasing as much as possible (this is discussed in more detail in Section 11).

### 8. Stationary Ilris experiment

In this section, we answer the first of our four primary questions raised at the beginning of the paper: how stable are descriptive features in laser intensity image under changes in ambient light?

We begin by describing a light sensitivity experiment, which analyzed the stability of SURF keypoints on both camera and laser intensity images under a variety of lighting changes (see Fig. 7 for an example of the 100 strongest keypoints detected in a laser intensity image and camera intensity image). This experiment was conducted over a 24 h period and consisted of taking camera images and Ilris scans at half-hour intervals of an outdoor scene. Since the purpose of this experiment was to analyze the impact that changes in ambient lighting had on camera/laser intensity images, the Ilris remained stationary for the entire experiment.

To quantify the stability of SURF keypoints over the 24 h, we performed exhaustive pairwise matching of every image to every other image in the dataset. Given \(N_{ij}\) matches between image \(i\) and image \(j\), the similarity score, \(S_{ij}\), is

$$S_{ij} := \frac{N_{ij}}{N_{\text{max}}}$$

where \(N_{\text{max}}\) is the maximum number of possible matches, which was limited to 500 in this experiment as we found that detecting more keypoints did not significantly improve any of the matching scores (i.e., it was more costly with no benefit). For keypoint matching and outlier rejection, we used the methods described earlier in Section 5.

Fig. 8 shows the similarity matrices for the camera and laser over the entire 24 h period, where the light values represent a greater number of matches and the dark values represent fewer matches. We compared camera and laser similarity matrices for four different image processing options: (i) no image processing, (ii) adaptive histogram equalization, (iii) Gaussian low-pass filter,
and (iv) adaptive histogram equalization and a Gaussian low-pass filter (only the best two similarity matrices are included as figures). Of the four options, we determined that applying adaptive histogram equalization and a Gaussian low-pass filter was the best for the laser intensity images while the best camera similarity score was achieved using a Gaussian low-pass filter. Table 1 shows the mean similarity scores and standard deviation for the both camera and laser intensity images for the four different image processing options.

As expected, for the camera, the number of keypoint matches between daytime and nighttime images drops off significantly (in fact, goes to zero), which produces the dark boxes in the similarity matrix. Referring to the camera’s similarity matrix, it is evident that in general, the similarity score is quite low, illustrating the camera’s sensitivity to ambient light. The similarity score ranged from as high as 0.824 to as low as zero with a mean score of 0.132 ± 0.181.

It should be mentioned that, for the traditional camera, there are some off-diagonal values that appear to match as well as some of the values near the main diagonal, which may appear counterintuitive. However, this structure is the result of manually choosing among a limited set of exposure settings, as the camera we used does not have automatic gain control. Although this admittedly introduces some human error in the camera results, the important details to take note of are the sharp drop-offs between sunrise and sunset, which would still exist even if our camera had automatic gain control.

In the case of the laser intensity images, we found that the similarity score does drop off from day to night, indicating a potential sensitivity to ambient light. In fact, the similarity scores exhibited a larger deviation than expected, ranging from as low as 0.384 to as high as 0.772 throughout the entire 24 h period (the mean score was 0.592 ± 0.144). However, even for the most drastic changes in ambient light (i.e., from light to dark), we were still able to find at least 192 keypoint matches, which we found to be more than sufficient for motion estimation. To prove this, we also computed the pose estimation errors between every pair of images at all times. This was accomplished using the visual odometry algorithm described in Section 6, but limited to just a single pair of images. The state we estimate between image j and image i is defined as $x_{ij} = [r_{ij} \theta_{ij}]$, which is a 6 × 1 column of pose variables, where $r_{ij}$ is the translation and $\theta_{ij}$ are the Euler angles describing the rotation. Since the Ilris remained stationary the entire experiment, the translational and rotational errors were measured as $e_{r} = \|r_{ij} - 0\|$ and $e_{\theta} = \|\theta_{ij} - 0\|.$

Fig. 9 shows the translational estimation errors for the laser and camera in a similar fashion as the similarity matrices. For viewing purposes, we have adjusted the scales on these images by limiting the maximum error in the camera image to the maximum error in the laser image (which was 7.308 × 10⁻³ m) and normalizing the results. This way, one can better judge the relative performance between the camera and laser, since the camera errors were sometimes orders of magnitude larger than the laser errors, which corresponds to the sections with low similarity scores (i.e., very few to no matches). It is interesting to note the similar structure between these ‘error matrices’ and the similarity matrices, which was to be expected given that VO performance will directly relate to the number of inlying keypoint matches. A similar trend can also be observed with the rotational errors shown in Fig. 10.

Histograms of the errors are shown in Fig. 11, where we have again limited the maximum errors of the camera for scaling purposes. These plots confirm that even in the worst case, both the translational and rotational errors are very low when using the laser intensity images, confirming our hypothesis that 192 keypoint matches are sufficient for motion estimation.

It is difficult to say what portion of the results were impacted purely by changes in ambient light or by dynamic environmental

<table>
<thead>
<tr>
<th>Image processing</th>
<th>Camera</th>
<th>Ilris lidar</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) None</td>
<td>0.092 ± 0.150</td>
<td>0.329 ± 0.101</td>
</tr>
<tr>
<td>(ii) Hist. eq.</td>
<td>0.069 ± 0.135</td>
<td>0.443 ± 0.127</td>
</tr>
<tr>
<td>(iii) Gaussian filter</td>
<td>0.094 ± 0.151</td>
<td>0.375 ± 0.110</td>
</tr>
<tr>
<td>(iv) Hist. eq. + Gauss. filter</td>
<td>0.073 ± 0.150</td>
<td>0.502 ± 0.161</td>
</tr>
</tbody>
</table>

Fig. 7. Laser and camera intensity images taken at the same time of day. Blue circles represent light-on-dark blobs and red circles represent dark-on-light blobs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In the case of the laser intensity images, we found that the similarity score does drop off from day to night, indicating a potential sensitivity to ambient light. In fact, the similarity scores exhibited a larger deviation than expected, ranging from as low as 0.384 to as high as 0.772 throughout the entire 24 h period (the mean score was 0.592 ± 0.144). However, even for the most drastic changes in ambient light (i.e., from light to dark), we were still able to find at least 192 keypoint matches, which we found to be more than sufficient for motion estimation. To prove this, we also computed the pose estimation errors between every pair of images at all times. This was accomplished using the visual odometry algorithm described in Section 6, but limited to just a single pair of images. The state we estimate between image j and image i is defined as $x_{ij} = [r_{ij} \theta_{ij}]$, which is a 6 × 1 column of pose variables, where $r_{ij}$ is the translation and $\theta_{ij}$ are the Euler angles describing the rotation. Since the Ilris remained stationary the entire experiment, the translational and rotational errors were measured as $e_{r} = \|r_{ij} - 0\|$ and $e_{\theta} = \|\theta_{ij} - 0\|.$

Fig. 9 shows the translational estimation errors for the laser and camera in a similar fashion as the similarity matrices. For viewing purposes, we have adjusted the scales on these images by limiting the maximum error in the camera image to the maximum error in the laser image (which was 7.308 × 10⁻³ m) and normalizing the results. This way, one can better judge the relative performance between the camera and laser, since the camera errors were sometimes orders of magnitude larger than the laser errors, which corresponds to the sections with low similarity scores (i.e., very few to no matches). It is interesting to note the similar structure between these ‘error matrices’ and the similarity matrices, which was to be expected given that VO performance will directly relate to the number of inlying keypoint matches. A similar trend can also be observed with the rotational errors shown in Fig. 10.

Histograms of the errors are shown in Fig. 11, where we have again limited the maximum errors of the camera for scaling purposes. These plots confirm that even in the worst case, both the translational and rotational errors are very low when using the laser intensity images, confirming our hypothesis that 192 keypoint matches are sufficient for motion estimation.

It is difficult to say what portion of the results were impacted purely by changes in ambient light or by dynamic environmental
The similarity score ranged from 0 to 0.810 (i.e., 0–405 keypoint matches) with a mean score of 0.094 ± 0.151. Note the interesting structure in the similarity matrix, where there are sharp transitions from a non-zero score to a score of zero due to sunrise and sunset.

The similarity score ranged from 0.250 to 0.738 (i.e., 125–369 keypoint matches) with a mean score of 0.502 ± 0.161. Note that the lowest similarity scores occur between scans separated by approximately 12 h.

The translational error between image i and image j was measured according to $e_{ij} = \|r_{ij} - \mathbf{0}\|$. The maximum translational error for the camera was limited to the maximum translational error for the Ilris, which was $7.308 \times 10^{-3}$ m. We then normalized all of the errors to the interval [0, 1] for visualization purposes. Dark values represent errors closer to zero and white values represent errors equal to or greater than $7.308 \times 10^{-3}$ m. It is interesting to note the similar structure with the similarity matrices.

Factors. For instance, throughout the day the trees were being blown by the wind and the grass and plants in the foreground can be seen to move and track the direction of the sun; a phenomenon called heliotropism. Nonetheless, overall we found that a majority of the detected SURF keypoints were stable over the largest possible changes in ambient lighting. Not only does this experiment justify our choice of pre-processing method, but it also demonstrates the lighting-invariance of a laser scanner as compared to a traditional camera. In addition, these similarity matrices and estimation error matrices demonstrated that SURF keypoints extracted from laser intensity images are stable under the largest changes in ambient lighting, which was one of the main questions we sought to answer.

9. Visual odometry experiments

This section presents our visual odometry results using the algorithm outlined in Section 6. Two different 3D laser scanners were used: the low-framerate TOF Ilris lidar and the high-framerate phase-based Autonosys lidar. All experiments were conducted at the University of Toronto Institute for Aerospace Studies and include stereo visual odometry results for comparison.
(a) Rotational error matrix for the camera. Once again, we see that most of the low rotational errors occur along the main diagonal, as these images show the lowest variation in ambient light.

(b) Rotational error matrix for the laser. Similarly with the translational error matrix, we can see that the rotational errors are relatively consistent and quite low.

Fig. 10. Rotational error matrices over a 24 h period. The rotational error between image \(i\) and image \(j\) was measured according to 
\[
e_{\theta ij} = \| \theta_{ij} - 0 \|.
\]
The maximum rotational error for the camera was limited to the maximum rotational error for the Ilris, which was 0.1579°. We then normalized all of the errors to the interval \([0, 1]\) for visualization purposes. Dark values represent errors closer to zero and white values represent errors equal to or greater than 0.1579°.

(a) Translational error histogram. (b) Rotational error histogram.

Fig. 11. Translational and rotational error histograms for all matches over 24 h. The maximum camera errors in both translation and rotation have been limited to the respective maximum errors for the Ilris, which were \(7.308 \times 10^{-3}\) m in translation and 0.1579° in rotation. This was done for reasons of scale, since the translational and rotational errors experienced by the camera estimates were sometimes orders of magnitude larger than the Ilris, due to situations with an extremely low number of matched keypoints.

and DGPS for groundtruth. The VO pipeline used in these experiments was implemented in Matlab, meaning that our system did not run in real time. However, we currently have a system implemented in C++ that can operate at 20 Hz.

9.1. Ilris

As mentioned previously, the Ilris is not a high-framerate sensor, so we were unable to collect laser data while moving. Instead, we took laser scans and stereo images every half meter over a traverse of approximately 200 m. Our experiment began during a sunny afternoon and ended at night in pure darkness.\(^6\)

We note that this stop-and-go strategy is perfectly valid for some realistic exploration tasks, and was the strategy used with the Mars Exploration Rovers. This experiment represents the first step towards validating our visual odometry approach using laser intensity images and answers the second research question: can we perform VO using a laser scanner in a stop-scan-go methodology with comparable results to stereo VO?

Fig. 12 shows the two-dimensional view of the localization results with DGPS groundtruth. The dataset took approximately 13 h to collect, allowing for a wide range of lighting conditions over the course of the day. Figs. 13 and 14 shows the translational errors versus distance traveled in the regions where our DGPS was in RTK mode. As can be seen, the laser VO estimate experiences a drift along the z-axis, which was due to an accumulation of error in the platform’s orientation estimate. However, the RMS errors in the y and z axes were below 1% distance traveled and the orientation error in the z-axis could be significantly reduced with the aid of an inclinometer [58]. The stereo estimate also displayed good performance in two out of the three degrees of freedom, achieving low errors in the y-axis and z-axis, but higher errors in the x-axis.

Table 2 shows the RMS errors for both the laser and stereo in all three axes.

Fig. 15 shows the number of inlying keypoint matches for both the lidar and the stereo camera over the entire traverse. For most

\(^6\) A video of this run can be viewed on our youtube channel at http://www.youtube.com/user/utiasASRL, under the Appearance-Based Lidar playlist.
Table 2
Root-Mean-Squared translational errors given in percent of distance traveled for the Iliris test.

<table>
<thead>
<tr>
<th>Sensor</th>
<th>$E_x$</th>
<th>$E_y$</th>
<th>$E_z$</th>
<th>$E_{xyz}$</th>
<th>Avg. no. of keypoint matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iliris</td>
<td>0.68</td>
<td>0.62</td>
<td>2.48</td>
<td>2.64</td>
<td>121</td>
</tr>
<tr>
<td>Stereo camera</td>
<td>1.49</td>
<td>0.45</td>
<td>0.38</td>
<td>1.60</td>
<td>106</td>
</tr>
</tbody>
</table>

Overall, we found that the performance between stereo and laser VO was comparable, with each estimate showing a drift in error one in one of the degrees of freedom. This is likely due to the large inter-frame displacements on the order of 0.5–0.8 m, which made it difficult to track keypoints across multiple frames. At the very least, we have demonstrated that laser VO can perform just as well as stereo VO when in sunlight (i.e., under favorable conditions for stereo). Thus, we have answered the second question raised at the beginning of our paper, by demonstrating that VO can be accomplished using a laser scanner and compares well with stereo VO.

9.2. Autonosys

This section presents a preliminary field test result using a high-framerate Autonosys lidar, in order to answer our third research question: **can we perform VO using a laser scanner under continuous motion?**

We present a 100 m outdoor traverse with stereo VO for comparison and post-processed GPS for groundtruth. In this experiment we logged $512 \times 384$ stereo images at 15 Hz and $480 \times 360$ lidar images at 2 Hz while driving the robot at an average speed of approximately 0.5 m/s. The tests were conducted in the afternoon during a sunny day.

Fig. 16 shows a two-dimensional plot of our VO results, where we see that laser VO does not perform as well as stereo and displays significant drift in its estimate (Fig. 17 provides a clear illustration of the rapidly growing error in all three axes (see Figs. 18 and 19)). This drift in the VO estimate is the result of a bias that is inherently introduced from scanning and moving at the same time. In the case of the Autonosys, the sensor always scans from right to left due to its continuously spinning hexagonal mirror. Consequently, range values from one side of the image always arrive earlier than range values from the other side. For our experiments, the Autonosys was scanning at 2 Hz and the vehicle was moving at an average speed of 0.5 m/s, meaning that in one scan, the vehicle would have moved approximately 0.25 m. Since one of the implicit assumptions during image formation is that all pixel values are generated at the same time, this motion during scanning violates that assumption. Naturally, this means that for accurate VO some form of motion compensation must be used while scanning and moving at the same time. As this is not the main focus of the paper, we discuss possible solutions to this problem in Section 11, which are able to maintain tractability despite the immense data rate of the sensor. The much more interesting result will be shown in the next section, where we see that motion compensation is not actually needed for accurate visual-teach-and-repeat localization performance. Indeed, if one is purely interested in repeating a previously driven path of arbitrary distance, only locally consistent estimates are required (see Table 3).

10. Visual teach and repeat

This section presents our preliminary visual teach and repeat (VT&R) localization results using the Autonosys sensor. VT&R is a technique that allows a vehicle to repeat a previously driven route fully autonomously, using only a visual sensor for localization. Furgale and Barfoot [3] presented over 32 km of autonomous VT&R field test results in a planetary analogue environment, achieving...
Fig. 16. Visual odometry results with post-processed GPS used for ground truth.

Fig. 17. Translational errors versus distance traveled for laser VO. Note the rapid drift in error for all three axes.

Fig. 18. Translational error versus distance traveled for stereo VO.

Fig. 19. Number of inlying keypoint matches versus the distance traveled. On average, more matches were found between Autonosys images than stereo images: approximately 284 matches for the Autonosys and approximately 215 matches for stereo.

an autonomy rate of over 99% (i.e., their system was able to repeat almost all previously driven routes fully autonomously). However, the main limitation of their technique was the use of stereo cameras, as lighting changes would sometimes create difficulties in identifying a previously visited region. This lighting dependency issue was the motivation for our last research question: can we perform VT&R localization using a laser scanner under continuous motion?

As the name suggests, VT&R has two main stages: a teaching phase and a repeating phase. During the route teaching, the vehicle creates a map of the environment while it explores, which we represent as a topological network of connected keyframes (i.e., we use a purely relative map representation). During the route repetition phase, the system performs frame-to-frame VO for an incremental estimate update and then matches against the nearest keyframe along the network (i.e., localize against the map). This means that the system only cares about the relative estimate with respect to the map and not the global estimate. This subtle, yet important point is the main reason VT&R can enable such long-range autonomous repeating, despite the unbounded error growth inherent to relative localization. It should be noted that frame-to-frame VO and matching against the map are accomplished using the same bundle adjustment formulation shown in Section 6, with the following differences: (i) only two states are included in the window, \(\{x_0, x_{k-1}\}\), with the previous pose, \(x_{k-1}\), being held fixed and (ii) a no-motion prior is included to help to ensure stability in the estimate and bound it within a local neighborhood of its previous pose.

As these tests are preliminary, we present post-processed results, where we taught a 170 m route during the day and tried to manually repeat the same route during the night to see how well the VT&R localization engine could estimate its relative position to the teach pass. Defining the estimated relative displacement from the teach pass to the repeat pass as \(\hat{\rho}_t\), and the true relative position as \(\rho_t\) (measured using DGPS), the localization error was defined as

\[\varepsilon_t := \left\| \hat{\rho}_t - \rho_t \right\|\]

It is worth noting that stereo VT&R results could not be presented because of the drastic lighting changes, which is an important point to raise, as it supports the very motivation for this research.

Post-processed DGPS ground truth of the teach/repeat paths are shown in Fig. 20 and the localization error for the repeat pass is shown in Fig. 21. The system performed very well, with a mean error of 0.36 m, despite the lack of motion compensation. Fig. 22 shows the keypoint matches during the repeat pass for frame-to-frame VO and matching against the map. On average, there were a greater number of frame-to-frame VO matches then matches against the map. However, this was to be expected since the teach and repeat paths were not identical because we manually drove the route in the dark during the repeat pass, leading to slight viewpoint changes. Future work will look to extend this VT&R localization method online in a closed-loop autonomous driving system (i.e., similar to Furgale and Barfoot [3]) and test in a planetary analogue setting over multiple kilometers in a variety of lighting conditions.

11. Discussion and future work

One of the key assumptions we made during the image formation process was to assume that every pixel was generated at the exact same time. This assumption worked well for the Iliris, since there was no motion during each scan. However, for the Autonosys, this temporal distortion was sometimes very noticeable and resulted in warped images if there were any large motions during a scan (see Fig. 23(a) for an example). Although this image distortion may reduce the number of potential keypoint matches between frames, we found that our keypoint matching and outlier rejection method was robust enough to still detect a large number of inliers; in other words, image distortion did not appear to be a significant challenge for VO (see Fig. 23(c) for example keypoint tracks found between a severely warped image and a non-warped image). As mentioned earlier, the main challenge we encountered with the Autonosys data was a bias that produced a constant drift in our orientation estimates. We believe that this bias was inherently introduced by scanning and moving at the same time, since the range values to landmarks in one section of the image arrive earlier than others. Clearly, for accurate VO, motion compensation will be required. To address this issue, we are currently implementing the pose-interpolation method of Dong and Barfoot [59], where each individual laser pulse within an image is treated as a separate measurement with a
Finding ways to efficiently combine appearance representation. Finding ways to efficiently combine appearance represented in a different manner than the traditional point cloud formation, this is just another example of how laser data can be very different from ours as we use a camera model for image feature detector for data association. Although their range image viewed from above, which allowed them to use a Kanade–Tomasi kernel. This process forms an image that looks like a 2D map by projecting it into a 2D plane and convolving it with a Gaussian.

This concept, which uses continuous-time basis functions in order to reduce the dimensionality of the state-space for high data-rates, is not required; only a series of locally consistent map is needed to create a better local map to match against (i.e., better in a sense that more keypoints can be embedded within a given pose). By linearly interpolating between the first and last pose for all measurements in a laser image, the dimensionality of the problem is reduced and becomes tractable to solve. Furgale et al. [60] has recently introduced a more general formulation of this concept, which uses continuous-time basis functions in order to reduce the dimensionality of the state-space for high data-rate sensors, such as IMUs. Fortunately, for VT&R tasks, it appears that motion compensation is not required, which is not entirely surprising given that our VT&R system is based on a relative map representation. Thus, if the goal is to just repeat a previously driven path, an accurate global estimate of the map and vehicle trajectory is not required; only a series of locally consistent maps is needed (e.g., a submap or a keyframe).

Although we have used local appearance-based techniques solely on the intensity images, there is a great deal of work that could look at the application of these techniques to range images and how to fuse the information from both. Li and Olson [61] described a method that rasterizes three-dimensional lidar data by projecting it into a 2D plane and convolving it with a Gaussian kernel. This process forms an image that looks like a 2D map viewed from above, which allowed them to use a Kanade–Tomasi feature detector for data association. Although their range image is very different from ours as we use a camera model for image formation, this is just another example of how laser data can be represented in a different manner than the traditional point cloud representation. Finding ways to efficiently combine appearance and metric information available from laser data will undoubtedly prove useful for accurate motion estimation.

We also wish to present some preliminary findings using data collected from a Velodyne HDL-64E sensor. At present, we do not believe that the Velodyne is well suited for appearance-based techniques due to its low vertical resolution with only 64 vertical samples covering a field of view of 26.8°. The camera model for the Velodyne is much more complicated than either of the devices presented in this paper. The Velodyne comes with 64 individual sensors positioned at different vertical and horizontal locations within the sensor, each with a different elevation angle. To produce an image stack with this sensor, we perform the following steps: (i) compute the 3D position of each scan point in Euclidean space using all of the necessary offsets and calibration parameters, (ii) project all of the points back onto a cylinder resembling the outer shell of the sensor, and (iii) unwrap this cylinder to form an image. Due to the nonuniform distribution of the points, a triangle-based linear interpolation is used and the same preprocessing steps presented in this paper are applied. An example of a Velodyne intensity image is shown in Fig. 24. As can be seen, the level of detail in the vertical direction is simply too coarse to generate stable keypoints and there are a lot of artefacts present. Having said this, we stress that these are just preliminary findings, and there may be alternative image formation and preprocessing techniques that could be applied to enhance the images further.

It is also interesting to note that when sensing in nearly planar environments, the standard deviation of neighboring range values will naturally grow with distance as the spacing between points increases in the vertical direction (see Fig. 25). Since bilinear interpolation is used to generate measurements and measurement uncertainties, range uncertainty will also grow with distance, which is not unlike what is observed with stereo cameras. This of course is not the case when a distant landmark is perpendicular to the laser ray, but in general, the standard deviation of range values between neighboring points does increase with distance. It was for this reason that we decided to decrease the vertical field of view (from 45° to 30°) and thereby increase the vertical resolution.

Our aim for the future is to extend the VT&R localization system presented in this paper to work online and use it within an autonomous closed-loop driving system. We also plan to investigate how to incorporate surrounding keyframe information to create a better local map to match against (i.e., better in the sense that more keypoints can be embedded within a given keyframe to provide additional matches). We predict that local map improvements will be necessary since simple keyframe-to-keyframe matching is not robust to large motions or viewpoint changes. The motivation for developing a laser-based VT&R system comes from previous work [3]. One of the issues encountered with

---

Fig. 20. Post-processed DGPS groundtruth for the teach (day) and repeat (night) paths. Matching teach and repeat images are shown with their corresponding keypoint matches. Despite variations between day and night paths, as well as changes in the scene (e.g., cars present during the day but not at night), the system is able to effectively match the teach pass images to the repeat pass images.

Fig. 21. Repeat pass localization error versus distance traveled. Despite viewpoint changes due to non-identical paths between day and night, the average localization error was very low (approximately 0.36 m).

Fig. 22. Number of inlying keypoint matches versus distance traveled for the repeat pass, indicating both frame-to-frame VO matches and map matches. Due to viewpoint changes from non-identical day and night paths, the number of keypoint matches against the map are lower than the number of frame-to-frame VO matches, as one would expect.

---

In this image, we can see some warping in the checkerboard, which was due to rotational motion during scanning.

In the next frame, the checkerboard appears upright without any significant distortion.

Keypoint tracks on image $k+1$. Although we did not track any keypoints on the checkerboard, we were able to track more than 100 keypoints on the ground.

These images show some of the distortion resulting from scanning and moving at the same time. In this case, the vehicle was moving at approximately 0.5 m/s and the Autonosys was capturing at 2 Hz, meaning that each image was collected over approximately 0.25 m of travel. Fortunately, we found that we were still able to track enough keypoints between images, even in the worst of cases.

Processed Velodyne intensity image, 26.8°V × 360°H FOV, 64 × 717 pixels, 39 Hz framerate, static. Although some general structure in the Velodyne image is recognizable with the camera image, due to the low number of vertical samples, it is nearly impossible to extract any stable detail from the scene. As a consequence, we do not feel that the Velodyne is well suited for appearance-based techniques.

Standard deviation range image gathered with the Autonosys, where each pixel value represents the standard deviation of the neighboring range values. White values represent the least deviation and black represent the most. Looking at the ground plane, one can see that the range deviation increases with distance. This is not the case for the hills in the foreground and background, as they are almost perpendicular to the laser rays. The maximum range deviation was 26 m.

Fig. 23. These images show some of the distortion resulting from scanning and moving at the same time. In this case, the vehicle was moving at approximately 0.5 m/s and the Autonosys was capturing at 2 Hz, meaning that each image was collected over approximately 0.25 m of travel. Fortunately, we found that we were still able to track enough keypoints between images, even in the worst of cases.

1. Under the largest variations in lighting changes (i.e., from night to day), one can detect enough stable keypoints in laser intensity images to generate accurate motion estimates (at least for a series of stationary images, as demonstrated by our stationary Iliris experiment in Section 8),
2. Using a stop–scan–go methodology, one can perform VO using laser intensity images with comparable performance to stereo-based VO during the day (this was demonstrated by our Iliris VO experiment in Section 9.1),
3. For accurate VO using a scanning laser while under continuous motion, some form of motion compensation is required due to issues of motion distortion and temporal biases (as demonstrated by our Autonosys experiment in Section 9.2),
4. Preliminary VT&R localization results suggest that even without motion compensation, VT&R localization appears both feasible and accurate using a high-framerate laser scanner while under continuous motion (as demonstrated in Section 10).

This work was motivated by the need to develop efficient and accurate vision methods for motion estimation in the absence of consistent environmental lighting, which is an unavoidable issue in outdoor settings. Through four experiments we have demonstrated that it is entirely possible to use the rich heritage of appearance-based computer vision techniques with the accurate ranging and light invariance of a laser scanner. Ultimately, our goal is to implement a lidar-based VT&R system that uses appearance-based techniques for motion estimation and avoids the issue of lighting dependence in outdoor environments.

Acknowledgements

The authors would like to extend their deepest thanks to Dr. James O’Neill and Kevin Williams from Autonosys for providing us with the opportunity to use their Autonosys lidar and assisting us with the sensor integration and image processing. We would
also like to thank Art Sauer from CNSS Boards, Sensors and ADU Products for his assistance with the post-processed GPS software and Greg Broten from Defence Research and Development Canada (DRDC) Suffield for his hard work in collecting and delivering Velodyne datasets. In addition, we would also like to acknowledge Andrew Lambert from the Autonomous Space Robotics Lab (ASRL) for his help in installing the GPS payload, Chi Hay Tong from ASRL for his work on the GPU SURF algorithm, Goran Basic from ASRL for designing and assembling the Autonosys payload mount, and Hang Dong from ASRL for his help with the Autonosys field testing. Lastly, we also wish to thank NSF and the Canada Foundation for Innovation for providing us with the financial support necessary to conduct our experiments.

References

Colin McManus holds a B.A.Sc. (Hons., 2009) in Engineering Science (Aerospace Option) from the University of Toronto, a M.A.Sc. (2011) in Aerospace Engineering from the University of Toronto Institute for Aerospace Studies (UTIAS), and is currently pursuing a D.Phil. as a member of the Mobile Robotics Group at the University of Oxford.

Paul Furgale received his B.A.Sc. (2006) in Computer Science from the University of Manitoba and a Ph.D. (2011) in Aerospace Engineering from the University of Toronto Institute for Aerospace Studies (UTIAS). Paul is currently a Research Associate at the Autonomous Systems Lab at ETH Zürich. His research interests include computer vision and localization technology for planetary exploration rovers.

Timothy D. Barfoot holds a B.A.Sc. (Hons., 1997) in Engineering Science (Aerospace Option) from the University of Toronto and a Ph.D. (2002) in Aerospace Engineering from the University of Toronto Institute for Aerospace Studies (UTIAS). He is currently an Associate Professor at UTIAS, where he leads the Autonomous Space Robotics Laboratory. Before joining UTIAS, he worked at MDA Space Missions in the Controls and Analysis Group on applications of mobile robotics to space exploration and underground mining. Dr. Barfoot is a Canada Research Chair (Tier II) in Autonomous Space Robotics and a Professional Engineer of Ontario.