Article

The International Journal of Robotics Research
1–27
© The Author(s) 2014
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0278364914550133
ijr.sagepub.com

Visual mapping for natural gas pipe inspection

Peter Hansen[1], Hatem Alismail[2], Peter Rander[2] and Brett Browning[2]

Abstract

Validating the integrity of pipes is an important task for safe natural gas production and many other operations (e.g. refineries, sewers, etc.). Indeed, there is a growing industry of actuated, actively driven mobile robots that are used to inspect pipes. Many rely on a remote operator to inspect data from a fisheye camera to perform manual inspection and provide no localization or mapping capability. In this work, we introduce a visual odometry-based system using calibrated fisheye imagery and sparse structured lighting to produce high-resolution 3D textured surface models of the inner pipe wall. Our work extends state-of-the-art visual odometry and mapping for fisheye systems to incorporate weak geometric constraints based on prior knowledge of the pipe components into a sparse bundle adjustment framework. These constraints prove essential for obtaining high-accuracy solutions given the limited spatial resolution of the fisheye system and challenging raw imagery. We show that sub-millimeter resolution modeling is viable even in pipes which are 400 mm (16") in diameter, and that sparse range measurements from a structured lighting solution can be used to avoid the inevitable monocular scale drift. Our results show that practical, high-accuracy pipe mapping from a single fisheye camera is within reach.

Keywords

Visual mapping, visual odometry, pipe inspection, natural gas, field robotics, fisheye camera

1. Introduction

Pipe inspection is an important task for many industries ranging from natural gas production, chemical refineries, and gas distribution networks through to sewer maintenance. Moreover, as infrastructure ages, pipe inspection becomes more critical in order to avoid catastrophic failures of the system. Inspection of pipes is a significant challenge because pipes are often inaccessible or costly to access due to the pipe being buried, surrounding infrastructure, or being located in a high place which requires scaffolding or climbing equipment to reach it.

There have been a range of efforts across multiple industries to develop robot vehicles that can move inside such pipe networks. At one extreme are the pipeline inspection gauges (PIGs), which are passive data-collection vehicles used in pipeline inspection. Other examples include actively driven, articulating snake-like robots (Mirats Tur and Garthwaite, 2010; Schempf et al., 2010), variations on vehicles with actively driven wheels mounted on legs that can adjust to pipe diameter changes, and tracked vehicles. Many of these vehicles share common properties: they are tele-operated with a fisheye camera to provide visual context to the operator. Inspection may be performed using the fisheye imagery, or with secondary sensors, such as magnetic flux leakage (MFL) or eddy current sensors, and typically no pose system is present or it is of low quality.

Accurate pose estimates for pipe inspection robots can be used for automated mapping applications, registration against existing maps, and precise localization of detected pipe defects. In Lee et al. (2009) the authors develop a landmark-based pipe localization and mapping system for the MRINSPECT V robot (Roh et al., 2009). The landmarks are T-intersections and elbows whose type and pose relative to the robot are automatically resolved using shadows in fisheye imagery. Landmarks are connected with straight sections, and robot pose recovered using wheel odometry and two-axis gyro measurements. The landmark poses are matched to a given map for navigation. This system was extended in Lee et al. (2011) to include a rotating laser module for improved landmark recognition and pose

[1]Carnegie Mellon University Qatar, Doha, Qatar
[2]National Robotics Engineering Center, Robotics Institute, Carnegie Mellon University, PA, USA

Corresponding author:
Peter Hansen, Carnegie Mellon University Qatar, PO Box 24866 Doha, Qatar.
Email: phansen@qatar.cmu.edu

estimation, orientation sensing from a three-axis inertial measurement unit (IMU), and an automated loop closure system using a graph-based representation of the pipe network.

In this work, we focus on applications such as natural-gas production, where pipe surface structure changes at the scale of a millimeter are of concern. We develop an approach to using the fisheye imagery common across many actively driven pipe inspection vehicles, to not only estimate robot pose but also simultaneously build high-resolution 3D models of the inner pipe surface. Such models can be used for inspection, autonomy, or to provide improved situational awareness to the operator (e.g. in the same way it is used for ground robots; Kelly et al., 2011).

Visual odometry approaches deriving from multi-view geometry techniques for central projection cameras have proven successful in robotics applications (Triggs et al., 2000; Hartley and Zisserman, 2004; Nistér, 2004; Nistér et al., 2006). In prior work, we have shown that verged stereo imagery can be used to produce accurate, sub-millimeter models of pipes (Hansen et al., 2011). A limitation of this approach, which motivated the development of the fisheye system, is that many cameras are required to map all of the pipe because of the limited angle of view of standard stereo cameras. This presents difficulties in both constructing a multi-camera system due to space restrictions in small pipes, and calibrating and retaining accurate relative camera extrinsic pose estimates. Monocular fisheye imagery overcomes this limitation but at the cost of scale ambiguity or, correspondingly, scale drift if there is a scale initialization. Additionally, fisheye imagery with objects that are close to the lens present a challenge as the deviations from a central projection image formation model have much more impact than at longer range.

We present three contributions to address these challenges. First, we develop a novel calibration process that aims to minimize inaccuracies from our central projection assumption. Second, we leverage weak scene constraints incorporated into a bundle-adjustment framework to improve visual odometry and mapping accuracy. Third, we introduce a sparse structured lighting system enabling metric pose and mapping estimates to be obtained, and thereby overcome scale drift and ambiguity. We evaluate the performance of the approach in a fiberglass pipe network and demonstrate that highly accurate results can be obtained and sub-millimeter 3D models can be densely rendered.

In the next section (Section 2), we briefly present the hardware and datasets used in this work. Section 3 describes the calibration approach that we use to compensate for the non-central projection properties of the system and the sparse structured lighting for resolving scale ambiguity. We then describe the image processing pipeline (Section 4) and visual odometry algorithms (Section 5), the core of our contribution. In Section 6, we present a set of experiments and the performance results of our fisheye system. A discussion on camera selection is presented in Section 7, which highlights the advantage of the fisheye system over multi-camera approaches. For this, we present an overview of our earlier verged stereo pipe-mapping system and discuss challenges related to design, construction, and calibration. Finally, we conclude the work in Section 8.

2. Hardware and dataset collection

Referring to Figure 1, our test hardware includes a prototype pipe crawling robot, and a 400 mm (16”) internal diameter fiberglass pipe network containing both straight sections and T-intersections.

The robot is equipped with a forward facing *fisheye camera*: a 190° angle-of-view Fujinon fisheye lens mounted to a 1280 × 960 resolution RGB CCD FireWire camera. Grayscale images (8-bit or 16-bit) are logged to the robot’s onboard computer at up to 15 frames per second (fps) during wireless tele-operation through the pipe network. All lighting within the pipe network is provided by eight high-intensity (3 W each) LEDs mounted on a square frame surrounding the fisheye camera (see Figure 1(a)). A sample dataset image logged in a straight section and another logged in a T-intersection are provided in Figure 2. As is evident, there are variations in lighting between the images. The large circular patterns (most noticeable in the straight section image) result from the manufacturing process of fiberglass pipes in which a single fixed-width length of fiberglass sheet is wrapped around a cylindrical core in a helix creating a continuous overlapping seam. The same pattern does not appear on the surface of solid metal pipes.

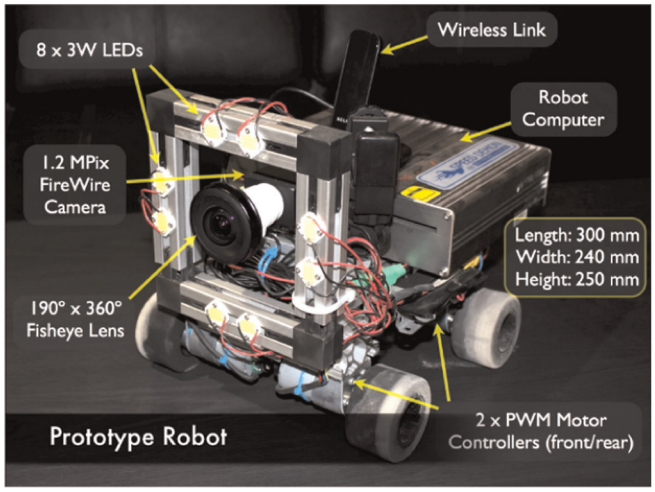
Although not visible in Figure 1(a), a structured lighting system consisting of two laser diode modules (LDMs) was retrofitted to the robot. (Edmund Optics micro LDM #57-100: 0.9 mW output power, 655 nm wavelength, focus range 50 mm–∞.) The lasers are mounted on either side of the camera and during operation each produces a visible red spot on the interior surface of the pipe within the field of view of the camera. The structured lighting is used to obtain metric pose and structure estimates.

3. Image formation and calibration

A critical component of the pipe-mapping algorithm is accurate sensor modeling and calibration. This includes the selection and calibration of the intrinsic image formation model for the fisheye camera, and the relative extrinsic poses of each laser relative to the camera.

3.1 Fisheye model

Image formation is modeled using a central projection polynomial mapping, which is a common selection for fisheye cameras (e.g. Xiong and Turkowski, 1997; Kannala and Brandt, 2006). A scene point coordinate $\mathbf{X}_i$ (X, Y, Z) defined relative to the single effective viewpoint $(0,0,0)^{\mathrm{T}}$ of the camera can be parameterized as



(a) Prototype robot with forward-facing fisheye camera.

(b) The 400 mm (16") internal diameter fiberglass pipe network, and the robot position inside a section of pipe.

Fig. 1. The prototype robot and fiberglass pipe network used for testing. Image datasets are collected as the robot is tele-operated through the pipe network.

$$\mathbf{X}_i(X, Y, Z) = \mathbf{X}_i(\theta, \phi, l) = l \begin{bmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{bmatrix} \quad (1)$$

where the angles θ and φ are, respectively, colatitude and longitude. The projected fisheye coordinate $\mathbf{u}'_i(u', v')$ is

$$\mathbf{u}'_i(u', v') = \left(\sum_{i=1}^{5} k_i \theta^i \right) \begin{bmatrix} \cos\phi \\ \sin\phi \end{bmatrix} + \mathbf{u_0} \quad (2)$$

where $\mathbf{u_0}$ (u_0, v_0) is the principal point. The polynomial function models the extreme radial distortion in the fisheye imagery. Notably, (2) can be written as a linear function of the intrinsic parameters $\boldsymbol{\omega}$:

$$\omega = (k_1, k_2, k_3, k_4, k_5, u_0, v_0)^{\mathrm{T}} \quad (3)$$

For convenience we use the shorthand $\mathbf{u}'_i = \mathcal{F}(\omega, \mathbf{X}_i)$ to denote the projection of a scene point to a fisheye image coordinate using (2).

A fisheye lens is inherently non-central, having a locus of depth-dependent viewpoints analogous to a shifting entrance pupil along its principal axis (Gennery, 2006). The magnitude of this entrance pupil shift is generally minimal when viewing distal scenes. It is, therefore, a common and often sufficiently accurate assumption to use central projection fisheye models for outdoor visual odometry and mapping applications where imaged scene points are far from the camera. For example, Royer et al. (2007) achieved accurate localization and mapping in outdoor environments using a 130° angle-of-view fisheye camera modeled using a central projection fifth-order polynomial. The advantages of using a central projection model are simplified calibration and the ability to utilize well established multiple-view geometry techniques for pose estimation and mapping (e.g. Hartley and Zisserman, 2004).

We use a central projection model for the advantages discussed. A novel calibration method is used to limit inaccuracies from this central projection assumption, which is

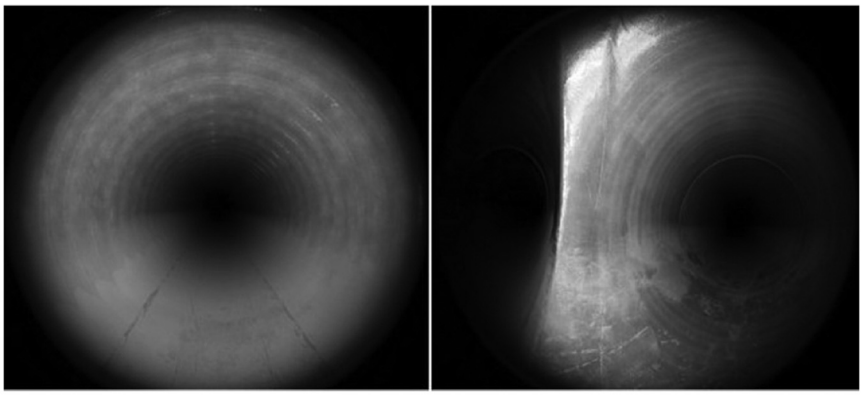


Fig. 2. Sample fisheye images logged in a straight section of the pipe network (left), and while turning in a T-intersection (right).

necessary for our pipe-mapping application as the imaged pipe surface is close to the fisheye lens. The calibration method supplements 'traditional' calibration data (i.e. images of a planar checkerboard target) with visual feature tracks found in a straight section of the pipe network. A discussion of the image processing techniques used to find the feature tracks is reserved for Section 4.

Using the visual feature tracks ensures the calibration data includes scene points imaged at operating distances very similar to those in the pipe datasets logged. This provides the opportunity to calibrate the central projection intrinsic model parameters that most accurately describe image formation at these distances. The checkerboard data ensures there is a distribution of calibration data points over the full angle of view of the fisheye camera.

3.2. Fisheye calibration

Figure 3 shows a subset of the 18 checkerboard images and the visual feature tracks, or 'pipe data' for short, used to calibrate the fisheye camera. The images of the planar checkerboard target were collected indoors prior to pipe data collection. The pipe data feature tracks were found from 80 selected keyframes logged while traversing a straight section of our pipe network. The length of the straight pipe spanned by these keyframes was close to 1 m. An initial approximate calibration of the fisheye camera using only checkerboard data is sufficiently accurate for finding the feature tracks. The checkerboard and pipe data are combined and used together to calibrate the final estimate of the camera intrinsic parameters $\boldsymbol{\omega}$ in (3).

It is evident in Figure 3 that the visual feature tracks extend toward the periphery of the camera field of view, but do not appear near the image center. As mentioned, the checkerboard data ensures there is a distribution of calibration data over the full field of view, and in particular near the image center where visual feature tracks are found when turning in T-intersections. This highlights the need to calibrate a single central projection camera model which most accurately describes image formation over the camera's full angle of view.

For all proceeding discussions, the pose $P(R,\mathbf{t})$ of a camera in a reference coordinate frame is

$$P(R,\mathbf{t}) = \begin{bmatrix} R & \mathbf{t} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix} \tag{4}$$

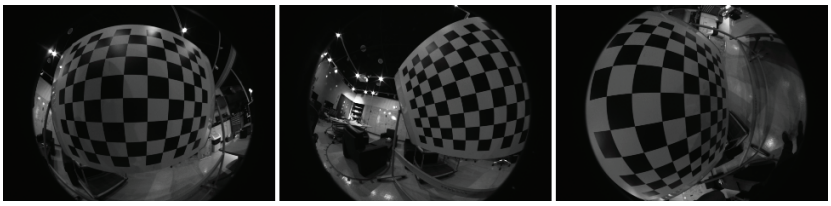
where $R \in SO(3)$ is a rotation and $\mathbf{t} \in \mathbb{R}^{3\times 1}$ a translation. A scene point with homogeneous coordinate $\mathbf{X}(X, Y, Z, 1)$ in the reference coordinate frame projects to the coordinate $\widehat{\mathbf{X}} = P\,\mathbf{X}$ in the camera coordinate frame. For convenience, homogeneous and inhomogeneous scene coordinates will be used interchangeably.

3.2.1. Checkerboard error. Each of the $n_k = 18$ images of the planar checkerboard target were collected by the camera at a pose P_k relative to the checkerboard target coordinate frame. The known Euclidean coordinates $\mathbf{X}$ of the checkerboard grid-points project to coordinates $\widehat{\mathbf{X}}_k = P_k\mathbf{X}$ in camera frame k, from which their reprojected fisheye image coordinates $\widehat{\mathbf{u}}_k = \mathcal{F}(\omega, \widehat{\mathbf{X}}_k)$ can be found. The reprojection errors are the Euclidean image distances between the detected grid-point image coordinates $\mathbf{u}_k$ and their reprojected coordinates $\widehat{\mathbf{u}}_k$. The checkerboard calibration error ϵ_g is the sum of squared grid-point image reprojection errors over all n_k images,

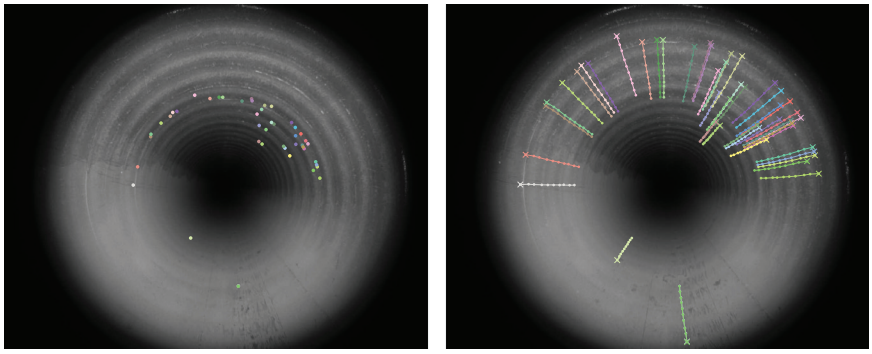
$$\epsilon_g = \sum_{k=1}^{n_k}\sum_{i=1}^{n_i} ||\mathbf{u}_k^i - \hat{\mathbf{u}}_k^i||^2 \tag{5}$$

where n_i is the number of checkerboard grid-points.

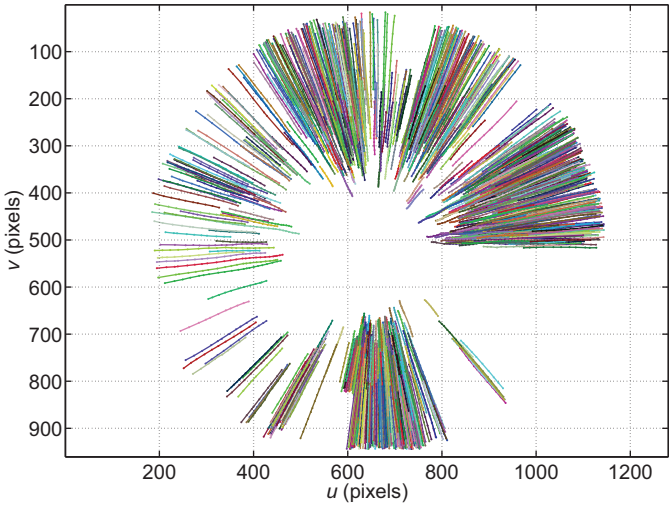
3.2.2. Pipe data error. The pipe data contains visual features tracks in $n_p = 80$ image keyframes: each scene point is only tracked across a subset of the keyframes. The pose P_p of each camera frame is defined relative to a straight cylinder with arbitrary fixed radius r. The coordinates $\widetilde{\mathbf{X}}$ of all observed scene points in the cylinder coordinate frame are constrained to lie on the cylinder, and are parameterized using cylindrical coordinates



(a) Subset of the 18 checkerboard images.



(b) Sample pipe data features detected in keyframe 1 (left), and the same features marked with crosses in keyframe 10 (right). The solid lines show the paths of the features through keyframes 1 to 10.



(c) All pipe data feature tracks. The solid lines show the paths of the features (scene points) tracked over multiple keyframes.

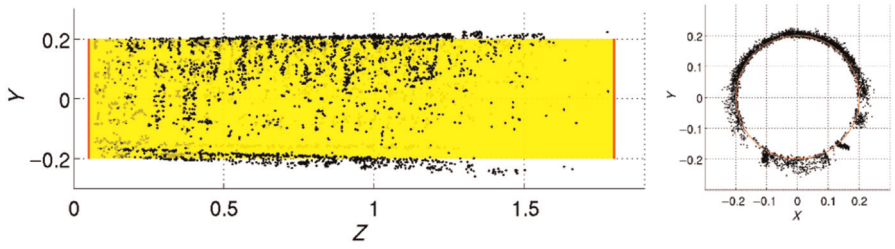
Fig. 3. Overview of the camera calibration data: (a) a subset of the checkerboard calibration images (18 total), (b) exampletracks example feature tracks between pipe data keyframes, and (c) the set of all pipe data feature tracks collected for the 80 keyframes.

$$\widetilde{\mathbf{X}}_i(\tilde{\phi}, \tilde{l}, r) = \begin{bmatrix} r\cos\tilde{\phi} \\ r\sin\tilde{\phi} \\ \tilde{l} \end{bmatrix} \tag{6}$$

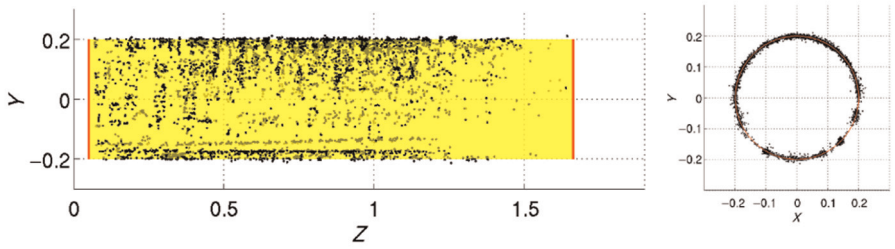
where $\tilde{\phi}$ is the orientation about the cylinder axis, and l is a distance along the cylinder axis.

The scene points observed in a given camera frame p project to the coordinates $\widehat{\mathbf{X}}_p = P_p\widetilde{\mathbf{X}}$ in the camera frame, and then to the reprojected fisheye image coordinates $\widehat{\mathbf{u}}_p = \mathcal{F}(\omega, \widehat{\mathbf{X}}_p)$. The pipe data calibration error $\boldsymbol{\epsilon}_p$ is the sum of squared image coordinate errors between the scene point observations and their reprojections,

$$\boldsymbol{\epsilon}_p = \sum_{p=1}^{n_p}\sum_{i=1}^{n_{i(p)}} ||\mathbf{u}_p^i - \widehat{\mathbf{u}}_p^i||^2 \tag{7}$$



(a) SBA result using calibration from only checkerboard data. The tapering in the sparse reconstruction does not correspond to the cylindrical structure of the pipe.



(b) SBA result using calibration from checkerboard and pipe data.

Fig. 4. SBA without scene constraints for a small straight section of pipe using: (a) calibration from only checkerboard data and (b) calibration from checkerboard and pipe data. A cylinder has been fitted to each set of SBA optimized scene points for reference. Including pipe data in the calibration improved the accuracy of sparse scene point reconstruction.

where $n_{i(p)}$ is the number of scene point observations in image p. The index i is not a global index to the set of all scene points. It is an index only to the observations in the image.

3.2.3. Parameter estimation. The calibration is a single non-linear optimization of the camera intrinsic parameters $\boldsymbol{\omega}$ in (3), each checkerboard camera pose P_k, each pipe data camera pose P_p, and the constrained scene point coordinates $\widetilde{\mathbf{X}}$ for the pipe data. This optimization minimizes the combined sum of squared reprojection errors $\boldsymbol{\epsilon}_g$ and $\boldsymbol{\epsilon}_p$.

For each iteration, each checkerboard camera pose P_k is updated, and the new scene point coordinates $\widehat{\mathbf{X}}_k$ in the frame computed. The constrained pipe data scene point coordinates $\widetilde{\mathbf{X}}$ are also updated, as well as each pipe data camera pose P_p, allowing the new observed scene point coordinates $\widehat{\mathbf{X}}_p$ in each frame to be found. Using the updated scene point coordinates in the checkerboard and pipe data camera frames, a linear estimate of the camera intrinsic parameters $\boldsymbol{\omega}$ is found which minimizes the combined checkerboard and pipe errors in (5) and (7).

The radius r for the pipe data in (6) is fixed to $r = 1$ and not optimized during calibration. Changing the radius simply rescales the magnitude of the change in position between the camera poses for the pipe data (i.e. a gauge freedom).

3.2.4. Results and discussion. The standard deviations $\boldsymbol{\sigma}(\sigma_u, \sigma_v)$ for the checkerboard reprojection errors were (0.630, 0.641) pixels, and (0.404, 0.436) pixels for the pipe reprojection errors.

Figure 4 illustrates the effectiveness of using the pipe data for calibration. It shows sparse bundle adjustment (SBA) results for an alternate set of visual feature tracks in a straight section of the pipe network using two sets of intrinsic parameters $\boldsymbol{\omega}$. In both cases, no cylindrical scene constraints were enforced when optimizing the scene point coordinates. The first set was calibrated using only the checkerboard data (Figure 4(a)), and the second set calibrated using both the checkerboard and pipe data (Figure 4(b)). A cylinder has been fitted to the SBA optimized scene point coordinates for reference. The tapering of the scene points for the checkerboard-only calibration in Figure 4(a) is the result of scale drift and highlights the limitations of calibrating the camera using only checkerboard data. A direction for future work is the implementation and evaluation of calibration using a non-central projection model such as CAHVOR (Gennery, 2006).

3.2.5. Changing pipe diameter. The central projection intrinsic parameters $\boldsymbol{\omega}$ were optimized using checkerboard data and visual feature tracks in a 400 mm internal diameter pipe, which is the size used in all experiments presented in Section 6. The same calibration procedure could be used in other pipe networks assuming that visual feature tracks could be obtained in a straight section with a fixed internal diameter. In some respects, this would be a form of

camera auto-calibration, also referred to as self- or online calibration.

Auto-calibration is the ability to calibrate intrinsic camera parameters from visual correspondences across multiple views of unstructured scenes, that is, without specially constructed calibration targets (Faugeras et al., 1992). Techniques have been developed for rotation-only viewpoint constraints (Hartley, 1994; Xiong and Turkowski, 1997), and for unconstrained camera viewpoint changes (Zhang, 1996; Kang, 2000; Fitzgibbon, 2001; Mičušík and Pajdla, 2003; Hartley and Kang, 2005; Li and Hartley, 2005; Thirthala and Pollefeys, 2005; Li and Hartley, 2006). The latter methods operating without viewpoint constraints are the most flexible for real-world applications. However, restrictions are often placed on the intrinsic model form (Fitzgibbon, 2001; Mičušík and Pajdla, 2003), and as detailed by Fitzgibbon (2001), they are often suitable only for finding an 'approximate' calibration whose accuracy is limited compared to more traditional offline calibration using calibration targets.

Although prior scene constraints are used for the pipe data, no special calibration target is constructed. Again, it is necessary only that visual feature tracks can be found in a small straight section of pipe. The main requirement, however, is for the calibration to achieve more than only an 'approximate' calibration. We demonstrated that when compared to the more traditional checkerboard-only calibration, the accuracy of calibration improves when the pipe data is included. We conclude from this that the calibration procedure could potentially be used for auto-calibration in alternate pipe networks with different pipe diameters.

3.3. Structured lighting extrinsic calibration

The two laser diodes are mounted either side of the camera, as shown in Figure 5. Each laser emits a ray which intersects the pipe's interior surface, creating a visible laser spot in each fisheye image. Calibrating the relative pose between the camera and a laser ray enables the Euclidean coordinate of the ray/scene point of intersection to be calculated from the fisheye image coordinate of the laser spot. When combined with the cylindrical scene priors detailed in Section 5, this property enables a metric pipe radius and metric camera position estimates to be estimated.

The extrinsic pose $L_i(R_i,\mathbf{t}_i)$ of laser i is parameterized relative to the camera coordinate frame using the minimum four degrees of freedom (see Figure 5):

$$L_i = \begin{bmatrix} R_i & \mathbf{t}_i \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix} \tag{8}$$

$$R = (R_X(\alpha)R_Y(\beta))^{\mathrm{T}}, \quad \mathbf{t} = -R(t_X, t_Y, 0)^{\mathrm{T}} \tag{9}$$

Each laser extrinsic pose L_i is calibrated using images of a planar checkerboard target with the imaged laser spot visible at some point on the checkerboard. After extracting the checkerboard grid coordinates $\mathbf{u}_p$ in each image, the pose P_p of the camera relative to the checkerboard target is estimated using a non-linear minimization of the sum of squared grid-point reprojection errors.

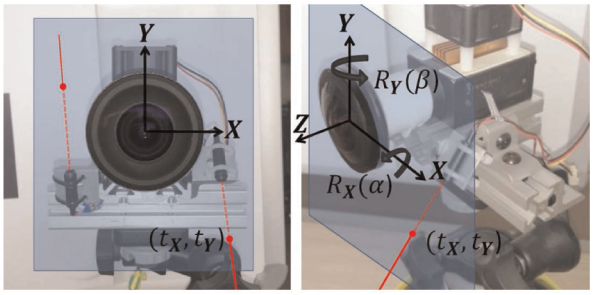


Fig. 5. The structured lighting system containing two laser diodes rigidly mounted either side of the camera. The extrinsic pose of each laser relative to the camera is given in (8), and can be fully described using four degrees of freedom.

For an estimate of the laser extrinsic pose L_i, the Euclidean point of intersection between the laser ray and checkerboard for frame p can be computed using the pose P_p, and then projected to the fisheye image coordinate $\widehat{\mathbf{l}}_p^i$. The laser reprojection error is the image distance between this projected coordinate and the detected coordinate $\mathbf{l}_p^i$. A non-linear estimate of the laser extrinsic pose L_i is found which minimizes the sum of squared laser reprojection errors $\epsilon_{L(i)}$ in all $n_{p(i)}$ images,

$$\epsilon_{L(i)} = \sum_{p=1}^{n_{p(i)}} ||\mathbf{l}_p^i - \widehat{\mathbf{l}}_p^i||^2 \tag{10}$$

The standard deviations $\boldsymbol{\sigma}(\sigma_u, \sigma_v)$ of the image reprojection errors for the top laser were (0.114, 0.096) pixels, and (0.120,0.108) pixels for the bottom laser. Figure 6 shows the epipolar line for each laser in a sample fisheye image. The epipolar lines are used to constrain the search space during the automatic laser spot detection described in Section 4.

4. Image processing

The image processing steps described in this section provide the necessary data for the pose estimation and mapping, and are all implemented offline after dataset collection. These steps include finding visual feature correspondences between selected image keyframes, identification/grouping of keyframes in straight sections and T-intersections of the pipe network, and the extraction of the laser spot image coordinates from the structured lighting system. The fisheye camera can log 8-bit or 16-bit grayscale images, and the image processing steps are capable of processing either.

4.1. Feature correspondences and keyframing

An efficient region-based Harris detector (Harris and Stephens, 1988) based on the implementation in Nistér

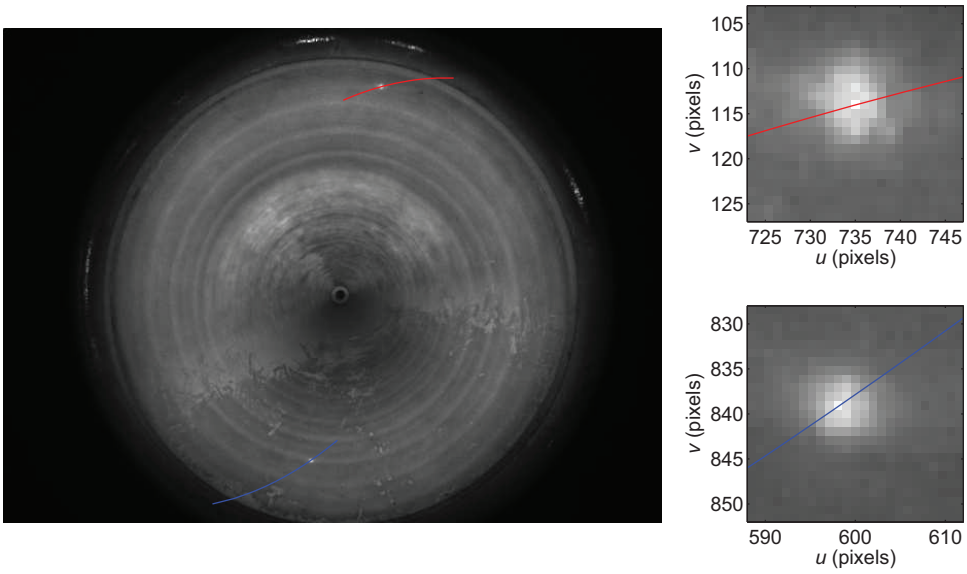


Fig. 6. The epipolar lines for the two laser diodes. Zoomed-in versions with the imaged laser spots visible are shown in the right column.

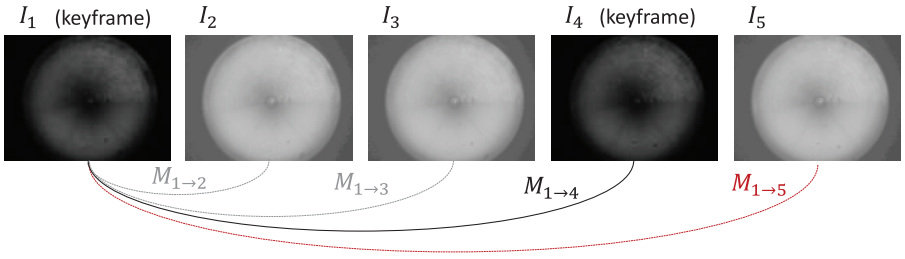


Fig. 7. Image keyframing is used to increase the baseline separation between frames. When the number of correspondences $M_{i\to j}$ falls below a threshold ($M_{1\to5}$ in the example), image I_{j-1} is selected as the next keyframe (I_4 in the example).

et al. (2006) is used to extract features in each image. Each image is divided into 2×2 regions, and the strongest $N = 200$ features per region retained based on their 'cornerness' score.

Given a current image I_i, correspondences $M_{i\to j}$ are found between image I_i and a proceeding image I_j using cosine similarity matching of 11×11 grayscale template descriptors evaluated for the detected features. Each of these 11×11 template descriptors is interpolated from a 31×31 region surrounding the feature. Five-point relative pose (Nistér, 2004) and RANSAC (Fischler and Bolles, 1981) are used to remove outliers and provide an initial estimate of the essential matrix E (Hartley and Zisserman, 2004). Using a cosine similarity metric achieves some invariance to lighting variations between images.

For all unmatched features in image I_i, a guided zero-mean normalized cross-correlation (ZNCC) is applied to find their pixel coordinate in image I_j. Here, 'guided' refers to a search within an epipolar *region* in image I_j. Since we implement ZNCC in the original fisheye imagery, we back-project each integer pixel coordinate to a spherical coordinate $\boldsymbol{\eta}$, and use the estimate of the essential matrix E to constrain the epipolar search regions using $|\boldsymbol{\eta}_j^{\mathrm{T}} E \boldsymbol{\eta}_i| < thresh$, where the subscripts denote the image. This guided ZNCC significantly improves the percentage of feature correspondences found between images.

Image keyframing is implemented when finding the visual correspondences. The goal of keyframing is to increase the baseline separation between the frames used during pose estimation and mapping. This reduces computational expense, and typically improves the accuracy of the estimates. However, the latter is conditional on maintaining a suitable number of visual correspondences between the keyframes. Referring to Figure 7, we continue to find correspondences $M_{i\to j}$ between the current image I_i and proceeding images I_j. When the number of correspondences falls below a heuristic threshold, the previous image I_{j-1} is selected as the keyframe. The correspondences are always found directly between the keyframes to avoid unnecessary integration of feature localization and ZNCC errors. As keyframe selection is based solely on the number of correspondences that can be found, the number of

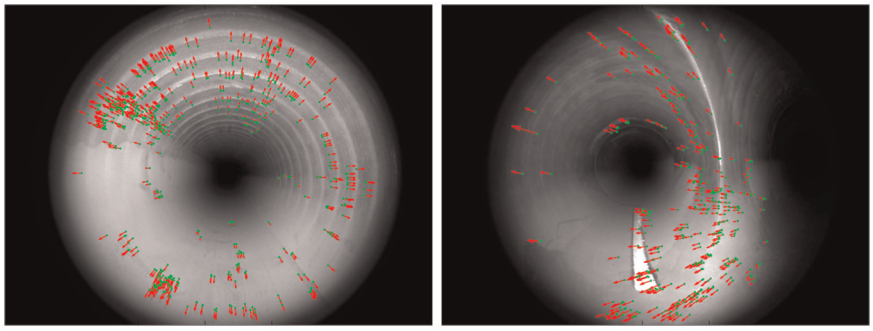


Fig. 8. Sparse optical flow vectors in a straight section (left) and T-intersection (right) obtained using a combination of Harris feature matching and epipolar guided ZNCC.

images between selected keyframes can change depending on the camera frame-rate, robot speed, and image appearance itself (e.g. texture), which influences the ease of feature matching.

Figure 8 shows subsets of the sparse optical flow vectors between keyframes selected in both a straight section and a T-intersection of our test pipe network: only subsets are shown for display purposes. The vectors (arrows) show the change in image coordinates of the feature correspondences between the keyframes. Features are 'tracked' across multiple keyframes by recursive matching using the method described.

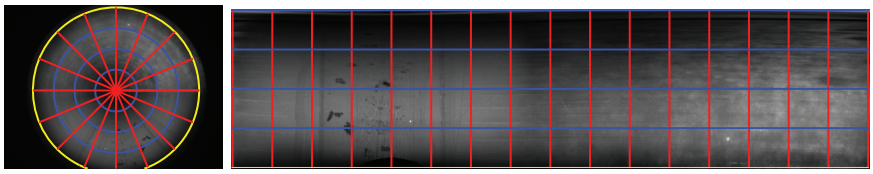
Using a region-based version of the Harris detector improves the uniformity of feature distributions in the images in the presence of strong lighting variations. Consequently, a more uniform distribution of visual correspondences (sparse optical flow) is found which covers much of the camera's full wide-angle field of view. This is advantageous during pose estimation and mapping for two reasons. Firstly, the correspondences produce more distinct motion patterns than narrow-field-of-view cameras. This, and the fact that the focus of expansion/contraction is more frequently visible in wide-angle cameras, improves the ability to decouple rotational and translational motion and achieve more accurate pose estimates (Nelson and Aloimonos, 1988; Gluckman and Nayar, 1998; Neumann et al., 2002; Strelow and Singh, 2004). Secondly, a more uniform distribution of sparse reconstructed scene points is found which is beneficial when enforcing the scene fitting constraints detailed in Section 5.

4.1.1. Feature detector selection. Designing a robust feature tracking system for the imagery in the pipe network is challenging for numerous reasons, not all mutually exclusive. They include: the limited image texture in sections of the pipes; strong projective deformations due to the close proximity/position of the internal pipe surface relative to the camera; and the camera's strong radial distortion.

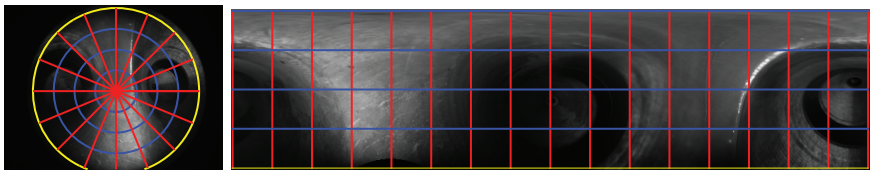
Images captured by the fisheye camera at similar poses can vary significantly in appearance due to both radial distortion and projective changes. Two-dimensional discrete scale space (Lindeberg, 1990, 1994) detectors/descriptors are often used when attempting to match correspondences between images with large appearance changes. State-of-the-art examples include SIFT (Lowe, 2004) and SURF (Bay et al., 2008). However, they are primarily invariant only to fronto-parallel scale change (i.e. varying resolution of imaged objects), and in-image rotations. In our application, the projective changes between views extend beyond simple front-parallel scale change and in-image rotations as the inner pipe surface is parallel to the direction of motion down the pipe. Moreover, SIFT and SURF offer little invariance to appearance changes resulting from radial distortion. Our evaluations using SIFT and SURF demonstrated no significant improvements in matching performance over the Harris detector and template descriptor method selected. This matching performance was based on the number and percentage of correspondences correctly matched.

A potentially more suitable alternative for our application are variants of the SIFT detector/descriptor that have been designed for central projection wide-angle fisheye and catadioptric (Nayar, 1997) image processing. The image processing steps are formulated as operations on the unit view sphere of a camera, a concept popularized in Daniilidis et al. (2002), which provides invariance to scale, camera rotations $R \in SO(3)$, and radial distortion. Using Bülow's derivation for scale-space on the sphere in Bülow (2004), two variants of SIFT were developed for wide-angle images in Hansen et al. (2007, 2010), the latter named pSIFT. However, both require an image re-sampling operation to perform the necessary spherical diffusion operations. This introduces unwanted image interpolation artifacts. More recently sRD-SIFT was developed (Lourenço et al., 2012) which avoids image re-sampling and has similar matching performance to pSIFT (number and percentage of matched features), but with improved keypoint position accuracy. We tested pSIFT and again found only marginal performance improvements for the number and percentage of correct correspondences over the selected Harris detector and template descriptor. Again, the large projective changes extending beyond just scale change and in-image rotations limited performance.

A final option considered was to process log-polar versions of the fisheye images. Example log-polar images in a



(a) Straight section



(b) T-intersection

Fig. 9. Original fisheye images, and log-polar version of the images. Converting the fisheye images to log-polar images introduces large interpolation artifacts which limit feature localization accuracy.

straight section and a T-intersection are provided in Figure 9 for reference. For the straight section, the log-polar mapping provides improved invariance to both the radial distortion in the fisheye imagery and the projective changes between images. However, the opposite is true for the T-intersection. Another major disadvantage of using log-polar images is the need for extensive image interpolation. The image artifacts created by interpolation limit feature localization accuracy.

For all feature detector/descriptors tested, the repeatability of feature detection and the ability to robustly match feature descriptors between views is limited primarily by large projective changes. As mentioned, these projective changes are not limited to simple fronto-parallel scale change and in-image rotations. This limits the repeatability of feature detection and matching performance of the scale-invariant detectors SIFT, SURF, and pSIFT. The matching results using the region-based Harris detector and template descriptors proved similar to these tested alternatives, but with the advantage of being the most computationally inexpensive. The epipolar guided ZNCC step was included as a secondary step and greatly improves the final percentage of feature correspondences found between images.

4.2. Straight section detection

The image keyframes must be divided into straight sections and T-intersections. This is necessary so that the straight section and T-intersection scene constraints can be correctly applied during pose estimation and mapping.

Techniques for classifying different pipe segments have previously been developed. In Lee et al. (2009), binary thresholding, morphology, and shape analysis of in-pipe imagery from a fisheye camera were used to classify straight sections, elbows, and T-intersections for landmark-based navigation. Illumination was provided by a carefully designed 64-element high flux LED array. Using the same robot, an alternate technique was presented in Lee et al. (2011) using images of visible laser beam projections provided by a rotating laser module. Our technique is most similar to the former and uses only the fisheye imagery.

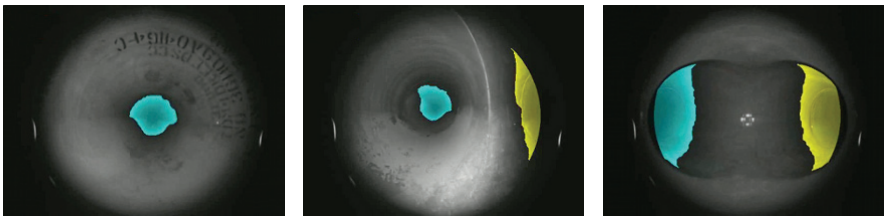
To classify each image as belonging to a straight section or T-intersection, the image resolution is first reduced by sub-sampling pixels from every second row and column. A binary thresholding is applied to extract dark blobs within the field of view of the camera, followed by binary erosion and clustering of the blobs. The largest blob is selected and the second moments of area L and L_0 computed about the blob centroid and principal point, respectively. An image is classified as straight if the ratio L_0/L is less than an empirical threshold; we expect to see a large round blob near the center of images in straight sections.

Figure 10(a) to (c) shows the blobs extracted in three sample images, and initial classification of each image. After initial classification, a temporal filtering is used to correct misclassification, as illustrated in Figure 10(d). This filtering enforces a minimum straight/T-intersection cluster size. In the example in Figure 10(d) there are five straight-section groups and four T-intersection groups.

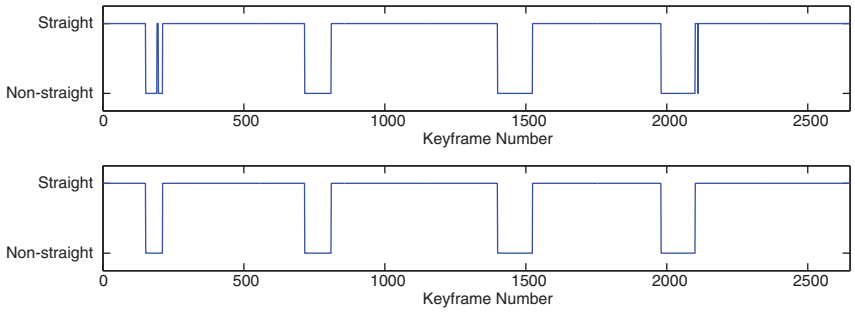
Referring to Figure 10(d), the shape of the dark spot and extracted blob may be non-circular. This is the result of some degree of non-uniform lighting provided by the LEDs, and variations in the reflective properties of the interior pipe surface caused by varying surface texture and degrees of dust build-up. Despite this, the straight section detector provides robust performance.

4.3. Structured lighting

The image coordinates of both laser diode spots are detected automatically in each keyframe using the epipolar lines to constrain the search space (see Figure 6). For each laser spot, the initial coordinate estimate is taken as the pixel with the largest weighted intensity value. The



(a) Straight section. (b) T-intersection. (c) T-intersection.



(d) Initial classification (top), and after applying temporal filtering (bottom). Each of the four T-intersection clusters is a unique T-intersection in the pipe network.

Fig. 10. Straight section and T-intersection image classification. Example initial classifications (a), (b), (c), and the classification of all keyframes before and after temporal filtering (d).

weighting for each pixel is its Gaussian distance from the nearest point on the epipolar line. Using this epipolar weighting constraint provides reliable detection of the initial integer laser-spot coordinates. All pixels with intensity values within 75% of the maximum are selected, within a five-pixel radius of its position, and the final coordinate $\mathbf{l}_i(u,v)$ computed as the weighted average over all the selected pixels $\{u,v\}$,

$$\mathbf{l}_i = \frac{1}{\sum_{\{u,v\}} I(u,v)} \left(\sum_{\{u,v\}} u\, I(u,v), \sum_{\{u,v\}} v\, I(u,v) \right)^{\mathrm{T}} \qquad (11)$$

where $I(u,v)$ is the grayscale intensity value at pixel position u,v.

The laser spot pattern is often dispersed due to the reflective nature of the interior surface of the fiberglass pipes used during testing. Using the weighted average in (11) improves the accuracy of the laser spot detection in these cases. Examples of detected laser spot coordinates in sample images from one of the logged datasets are shown in Figure 11.

5. Pose estimation and mapping

The image processing steps described in Section 4 provide the data necessary for the pose estimation and mapping. This includes visual correspondences between keyframes, the image coordinates of the structured lighting laser spots, and the grouping of the image keyframes into straight sections and T-intersections. Using this data, the pose estimation and mapping consists of three main steps:

1. Section 5.1: Obtain the pose estimation and mapping results for each straight section using a sliding-window SBA and a localized straight cylinder fitting constraint. When available, the structured lighting is used to resolve monocular scale ambiguity.
2. Section 5.2: Obtain the pose estimation and mapping results for each T-intersection using a two-cylinder intersection model fitting constraint. This step effectively joins straight sections processed in the first step.
3. Section 5.3: When possible, perform loop closure correction using additional visual correspondences found between image keyframes in overlapping sections of the pipe network.

Each step is described in detail in the relevant section. The justification for and effectiveness of including scene constraints are presented in the results in Section 6.

5.1. Straight sections

For each new keyframe in a straight section, the feature correspondences are used to estimate the new camera pose

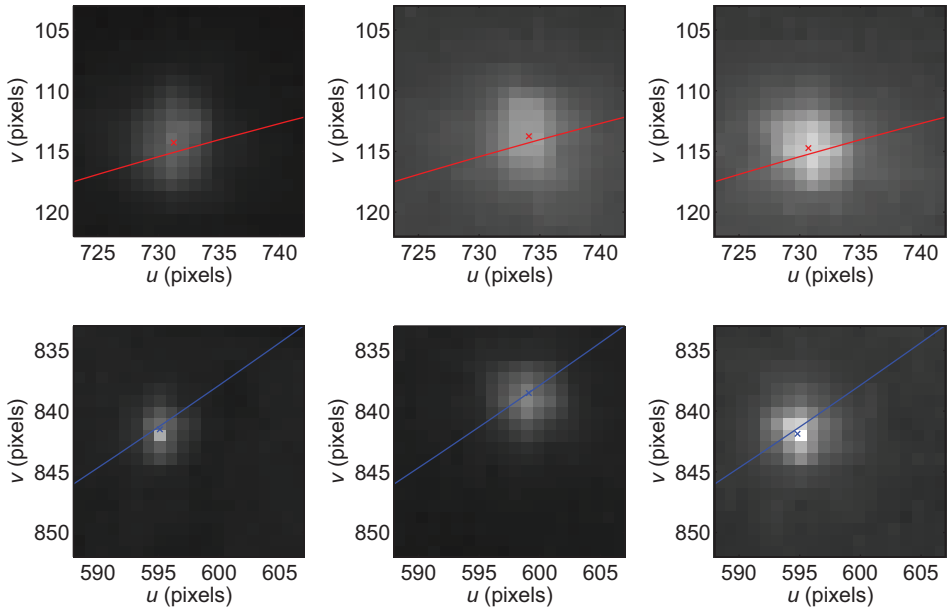


Fig. 11. The laser spot positions (crosses) automatically detected in three sample dataset images. The top and bottom rows correspond to the top and bottom lasers, respectively. The line in each image patch is the epipolar line of the laser which is used to guide the laser spot detection.

and sparse scene point coordinates. This is performed using Nister's five-point relative pose algorithm (Nistér, 2004) to obtain an initial unit-magnitude pose change estimate, optimal triangulation to reconstruct the scene points (Hartley and Zisserman, 2004), and prior reconstructed scene coordinates in the straight section to resolve relative scale.

After every 50 keyframes, a modified sliding-window SBA is implemented which includes a localized straight cylinder fitting. A 100 keyframe window size is used. Although the keyframing procedure described in Section 4.1 does not enforce a fixed number of keyframes to be found per given length of pipe, for our datasets a 100 keyframe window typically spans a segment of pipe approximately 1 m in length. The SBA is a multi-objective least squares minimization of image reprojection errors $\boldsymbol{\epsilon}_{\mathbf{I}}$, scene point regularization errors $\boldsymbol{\epsilon}_{\mathbf{X}}$, and when available the structured lighting errors $\boldsymbol{\epsilon}_{\mathbf{L}}$. (The structured lighting system was not implemented on all datasets evaluated.) The scene point and structured lighting errors are both functions of the fitted cylinder whose parameterization will be provided. The SBA attempts to find the non-linear estimate of the camera poses P in the window, observed scene point coordinates $\mathbf{X}$ in the window, and the fitted cylinder C which minimize the combined error $\boldsymbol{\epsilon}$:

$$\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_{\mathbf{I}} + \tau\,\boldsymbol{\epsilon}_{\mathbf{X}} + \kappa\,\boldsymbol{\epsilon}_{\mathbf{L}} \tag{12}$$

The parameter τ is a scalar weighting with units of pixels^2/m^2, and controls the trade-off between the competing error terms $\boldsymbol{\epsilon}_{\mathbf{I}}$ and $\boldsymbol{\epsilon}_{\mathbf{X}}$. The parameter κ is a unit-less scalar weighting of the structured lighting error.

5.1.1. Image reprojection error $\boldsymbol{\epsilon}_I$. For image I_p in the sliding window, with pose $P_p(R_p, \mathbf{t}_p)$, the reprojected coordinates $\hat{\mathbf{u}}_p$ of the subset $\mathbf{X}_p$ of scene points observed in the image are

$$\hat{\mathbf{u}}_p = \mathcal{F}(\omega, P_p\mathbf{X}_p) \tag{13}$$

The image reprojection error $\boldsymbol{\epsilon}_{\mathbf{I}}$ is the sum of squared differences between all image feature observations $\mathbf{u}$ in the sliding window, and their reprojected scene point coordinates $\hat{\mathbf{u}}$,

$$\boldsymbol{\epsilon}_{\mathbf{I}} = \sum_{p=1}^{n_p}\sum_{i=1}^{n_{i(p)}} ||\mathbf{u}_p^i - \hat{\mathbf{u}}_p^i||^2 \tag{14}$$

where $n_p = 100$ is the number of frames in the sliding window, and $n_{i(p)}$ is the number of scene point observations in image I_p.

5.1.2 Scene point error $\boldsymbol{\epsilon}_X$. The scene point error $\boldsymbol{\epsilon}_{\mathbf{X}}$ is the sum of squared errors between the straight cylinder and the optimized scene point coordinates $\mathbf{X}$ observed within the sliding window. The cylinder pose C is defined relative to the first camera keyframe pose $P_m(R_m, \mathbf{t}_m)$ in the sliding window as the origin. It is parameterized using four degrees of freedom,

$$C(\tilde{R}, \tilde{\mathbf{t}}) = \begin{bmatrix} R_X(\gamma)\,R_Y(\beta) & (t_X, t_Y, 0)^{\mathrm{T}} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix} \tag{15}$$

where R_A denotes a rotation about the axis A, and t_A denotes a translation in the axis A. Using this parameterization, all observed scene points in the sliding window project to the coordinates $\widetilde{\mathbf{X}}$ in the cylinder coordinate frame via the transform

$$\widetilde{\mathbf{X}} = C\, P_m\, \mathbf{X} \tag{16}$$

The error for each of the $n_{\mathbf{X}}$ points in the sliding window is defined as the Euclidean distance between its coordinate $\widetilde{\mathbf{X}}^i$ and the nearest point on the cylinder. The scene point error $\boldsymbol{\epsilon}_{\mathbf{X}}$ is the summation of these squared Euclidean distances over all observed scene points,

$$\boldsymbol{\epsilon}_{\mathbf{X}} = \sum_{i=1}^{n_{\mathbf{X}}} \left(\sqrt{(\tilde{X}^i)^2 + (\tilde{Y}^i)^2} - r \right)^2 = \sum_i (\delta \tilde{r}^{\,i})^2 \tag{17}$$

where r is the pipe radius. Referring to (12), we use an empirically selected value $\tau = 500^2$ (pixels^2/m^2). Using this value, a one-pixel image reprojection error, $||\mathbf{u}^i - \hat{\mathbf{u}}^i|| = 1$, and a 2 mm scene point error, $\delta \tilde{r}^i = 0.002$, contribute equally to the overall error term $\boldsymbol{\epsilon}$ in (12). Increasing the value τ would further increase the strictness of the scene fitting constraints.

The scene point error in (17) is a function of the fitted cylinder pose C and its radius r. Using a correct metric value for the radius r enables metric visual odometry and scene reconstruction results to be obtained, that is, it resolves monocular scale ambiguity. The structured lighting error discussed next is used to find a metric estimate of the radius r. In the experiments presented in Section 6, the structured lighting system was not implemented for several datasets, and a fixed a priori measured radius of $r = 200$ mm was provided.

5.1.3 Structured lighting error $\boldsymbol{\epsilon}_L$. The structured lighting error allows a metric radius r to be estimated during each sliding-window SBA, significantly improving the flexibility of the system for field applications. This radius r is optimized in the sliding-window SBA simultaneously with all camera poses, observed scene points, and the cylinder pose C.

To find the structured lighting error for an image frame p in the sliding window, the point of intersection $\mathbf{W}(X, Y, Z)$ between the fitted cylinder and each laser i is first found. Recall that the fitted cylinder pose C is defined relative to the first camera pose P_m in the sliding window. The pose Q of the cylinder relative to laser i, with extrinsic pose L_i relative to the camera, is therefore

$$Q = C\, P_m\, P_p^{-1}\, L_i^{-1} \tag{18}$$

The point of intersection $\mathbf{W}$ in the cylinder coordinate frame is parameterized as

$$\mathbf{W} = \mathbf{V}_0 + s(\mathbf{V}_1 - \mathbf{V}_0) \tag{19}$$

where s is a scalar, and

$$\mathbf{V}_0(X_0, Y_0, Z_0) = Q\,(0, 0, 0, 1)^{\mathrm{T}} \tag{20}$$

$$\mathbf{V}_1(X_1, Y_1, Z_1) = Q\,(0, 0, 1, 1)^{\mathrm{T}} \tag{21}$$

Using the constraint $X^2 + Y^2 = r^2$, and the parameterization in (19), we have

$$(X_0 + s(X_1 - X_0))^2 + (Y_0 + s(Y_1 - Y_0))^2 = r^2 \tag{22}$$

which is a quadratic in the unknown s. The largest signed solution for s is substituted into (19) to find the point of intersection $\mathbf{W}$. This point is projected back to the camera coordinate frame,

$$\widehat{\mathbf{W}} = P_p\, P_m^{-1}\, C^{-1}\, \mathbf{W} \tag{23}$$

and then to the coordinate $\widehat{\mathbf{l}}_p^i = \mathcal{F}(\omega, \widehat{\mathbf{W}})$ in the fisheye image. The structured lighting error is the sum of squared laser reprojection errors for both lasers over the set of all $n_p = 100$ images in the sliding window,

$$\boldsymbol{\epsilon}_{\mathbf{L}} = \sum_{p=1}^{n_p} \sum_{i=1}^{2} ||\mathbf{l}_p^i - \widehat{\mathbf{l}}_p^i||^2 \tag{24}$$

where $\mathbf{l}_p^i$ is the detected laser spot coordinate for laser i in image frame p.

The structured lighting error metric could be used with any number of lasers. The only requirements are that the extrinsic pose of each laser relative to the camera is known, and that the cylinder/laser intersection spot is visible and detectable in the fisheye images.

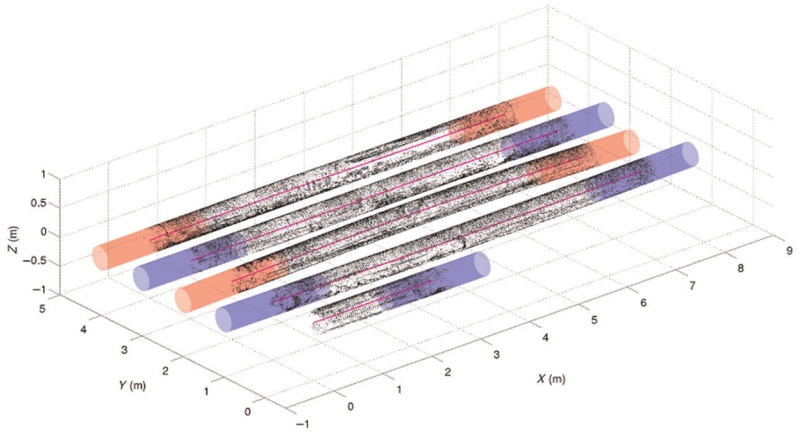
5.1.4 Huber weighting and outlier removal. Due to the challenging raw imagery, outliers are often present during pose estimation and mapping. Limited image texture and a low signal-to-noise ratio in some sections of the pipe network result in feature correspondence outliers, and specular reflections can result in laser spot detection outliers.

To minimize the influence of outliers, a Huber weighting is applied to all scene point reprojection errors ($\mathbf{u} - \hat{\mathbf{u}}$), scene fitting errors ($\delta \tilde{r}$), and laser reprojection errors ($\mathbf{l} - \widehat{\mathbf{l}}$) before computing the final sum of squared errors $\boldsymbol{\epsilon}_{\mathbf{I}}$, $\boldsymbol{\epsilon}_{\mathbf{X}}$, and $\boldsymbol{\epsilon}_{\mathbf{L}}$.

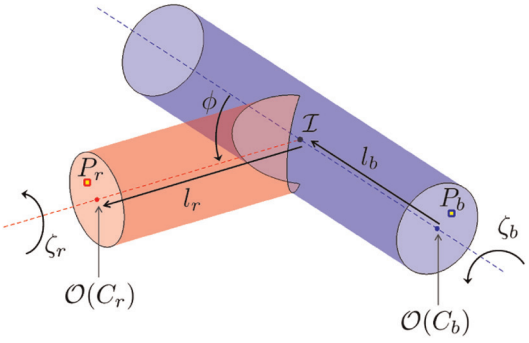
As a secondary step, outliers are removed at multiple stages (iteration steps) of each sliding-window SBA. This includes removing detected scene point outliers based on image reprojection errors, scene point outliers based on scene fit errors, and detected laser spot outliers based on laser reprojection errors. For each outlier type, the outlier removal threshold *thresh* is computed using the median absolute deviation (MAD) of the signed errors, *thresh* = ρ MAD (**errs**). For example, given all the scene fitting errors $\mathbf{errs} = \delta\widetilde{\mathbf{r}}$ in the sliding window, the threshold is

$$thresh = \rho\, \mathrm{MAD}(\mathbf{errs}) \tag{25}$$

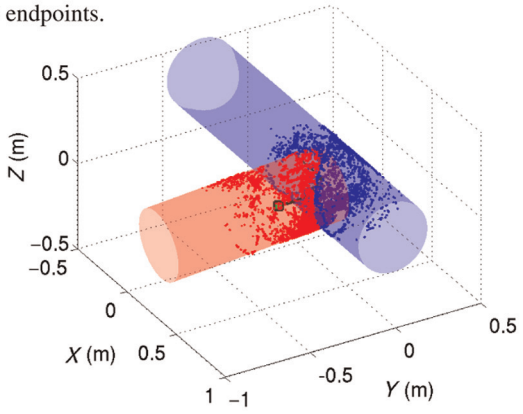
$$= \rho\, \mathrm{median}\{|\mathbf{errs} - \mathrm{median}\{\mathbf{errs}\}|\} \tag{26}$$



(a) Straight sections with cylinders fitted to endpoints.



(b) The two-cylinder T-intersection model parameters (refer to text for detailed description). The lighter 'red' (vertical part of 'T') cylinder axis is constrained to intersect the darker 'blue' (horizontal part of 'T') cylinder axis at a unique point $\mathcal{I}$.



(c) Visual odometry and scene reconstruction result using the T-intersection model. The scene points have been automatically assigned to each section allowing cylinder fit regularization terms to be used within the SBA framework.

Fig. 12. A T-intersection is modeled as the intersection of cylinders fitted to the straight sections of the pipe. Respectively, the darker 'blue' and lighter 'red' shades distinguish the horizontal and vertical sections of the 'T', as illustrated in (b). The figure is best viewed in color in the electronic version.

where ρ is a scalar and median $\{\mathbf{a}\}$ is the median of the values in the set $\mathbf{a}$. We use the same value $\rho = 5.2$ for all outlier types. Computing the outlier removal thresholds using MAD was also used for robust outlier removal for vision-based localization in Scaramuzza and Siegwart (2008). This secondary outlier removal step is particularly useful when the percentage of outliers is large.

5.2. T-intersections

The general procedure for processing the T-intersections is illustrated in Figure 12. After processing each straight section, straight cylinders are fitted to the scene points in the first and last 1 m segments (Figure 12(a)). As illustrated in Figure 12(b), a T-intersection is modeled as two intersecting cylinders: the lighter 'red' (vertical part of 'T') cylinder axis intersects the darker 'blue' (horizontal part of 'T') cylinder axis at a unique point $\mathcal{I}$.

The start and end cylinder poses are each parameterized using (15), using the first and last camera poses in the straight section as origins, respectively. The optimal cylinder fit minimizes the sum of squared Euclidean distances between the reconstructed scene points and their nearest point on the cylinder. A fixed a priori measured radius $r = 200$ mm is used for datasets not utilizing the structured lighting system. For the others, the radius r is also optimized during cylinder fitting.

The T-intersection model is parameterized as follows. Let P_r be the first/last camera pose in a vertical red section, and C_r be the cylinder fitted with respect to this camera as the origin. Similarly, let P_b be the last/first camera pose in a horizontal blue section, and C_b be the cylinder fitted with respect to this camera as the origin. The parameters ζ_r and ζ_b are rotations about the axis of the red and blue cylinders, and l_r and l_b are the signed distances of the cylinder origins $\mathcal{O}(C_r)$ and $\mathcal{O}(C_b)$ from the intersection point $\mathcal{I}$. Finally, ϕ is the angle of intersection between the two cylinder axes

in the range $0° \leq \phi < 180°$. These parameters fully define the change in pose Q between P_b and P_r, and ensure that the two cylinder axes intersect at a single unique point $\mathcal{I}$. Letting

$$D = \begin{bmatrix} & & 0 \\ R_Z(\zeta_r) & & 0 \\ & & l_r \\ 0^{\mathrm{T}} & & 1 \end{bmatrix} \begin{bmatrix} & & 0 \\ R_Z(\zeta_b)R_Y(\phi) & & 0 \\ & & l_b \\ 0^{\mathrm{T}} & & 1 \end{bmatrix} \tag{27}$$

where R_A is a rotation about axis A, the change in pose Q between P_b and P_r is

$$Q = C_r^{-1} D C_b \tag{28}$$

SBA is used to optimize all camera poses P in the T-intersection between P_r and P_b, as well as all new scene points $\mathbf{X}$ in the T-intersection, and the T-intersection model parameters $\Phi(\zeta_r, l_r, \zeta_b, l_b, \phi)$. The objective function minimized is the same form as (12), except only the image reprojection error (14) and scene fitting error (17) are minimized. The same value $\tau = 500^2$, robust cost function, and outlier removal scheme are also used.

Special care needs to be taken when computing the scene fitting error $\boldsymbol{\epsilon}_{\mathbf{X}}$ in (17) as there are two cylinders in the T-intersection model. Specifically, we need to assign each scene point to only one of the cylinders, and compute the individual error terms in (17) with respect to this cylinder. This cylinder assignment is performed for each scene point by finding the distance to each of the cylinder surfaces, and selecting the cylinder for which the absolute distance is a minimum. Figure 12(c) shows the results for one of the T-intersections after SBA has converged. The shading of the scene points (dots) represent their cylinder assignments.

5.3 Loop closure

For multi-loop datasets, loop closure correction is used to reduce integrated pose estimate inaccuracies. At present we use the graph-based optimization system g^{2}o (Kümmerle et al., 2011), although alternate graph-based techniques with comparable performance and improved efficiency could be used (e.g. Dubbelman et al., 2012).

Using g^{2}o, the graph vertices are the set of all keyframe poses described by their Euclidean coordinates and orientations (quaternions). The graph edges connect temporally adjacent keyframes, and the loop closures connecting keyframes in the same section of the pipe network visited at different times. Each edge has an associated relative pose estimate between the vertices it connects, and a covariance of the estimate.

Loop closures in image datasets are frequently identified using visual place recognition techniques which compare visual image content to evaluate similarity. Popular algorithms such as FABMAP (Cummins and Newman, 2008, 2011) have evolved from the visual 'bag of words' (BoW) method presented in Sivic and Zisserman (2003). An alternate approach to loop closure detection used for navigation in pipe networks was presented in Lee et al. (2011). The pipe network is represented as a set of nodes connected by edges. Each node is a T-intersection or elbow joint, and each edge is a straight section of pipe. When the robot detects a node, the pose of the node is estimated and then compared to all previously mapped nodes. A thresholding is applied to determine if a correct loop closure has been found.

Fig. 13. The pipe network used in experiments with the four T-intersections T1 through T4 labeled.

While BoW-based algorithms could be explored for operation in large-scale pipe networks, we take an approach more similar to Lee et al. (2011) and use the known overall structure of our constrained test network for loop closure detection. Given the known straight section of pipe where the robot begins, and counting left/right turns, we identify when the robot returns to a straight section visited previously: see Figure 13. Let $P_{\mathbf{i}}$ and $P_{\mathbf{j}}$ be the keyframe poses in the same straight section of pipe visited at different times. We compute the Euclidean distances $l_{\mathbf{i}}$ and $l_{\mathbf{j}}$ to the T-intersection centroid $\mathcal{I}$ at the end of the straight section. Poses P_i are P_j are selected as a candidate pair if $l_i \approx l_j$.

The relative pose estimate between each candidate loop pair $P_i \leftrightarrow P_j$ is refined using image correspondences found by matching SIFT descriptors for the previously detected Harris corners. SIFT descriptors are used at this stage to achieve improved rotation invariance as the relative orientation of the robot about the cylinder axis may have changed. Prior knowledge of the surrounding scene structure (fitted cylinders) is used to resolve the monocular scale ambiguity of the relative pose estimates obtained using the five-point algorithm and RANSAC, followed by non-linear refinement.

For our implementation, the relative uncertainties between all graph edges (temporally adjacent keyframes and loop closure keyframes) are evaluated numerically.

6. Experiments and results

Three separate datasets (A, B, C) were logged during teleoperation of the robot through the pipe network. A summary of the datasets is provided in Table 1, including the robot's sequence through the network (refer to Figure 13),

Table 1. A summary of the three datasets (A, B, C) logged in the pipe network using the prototype robot described in Section 2. The sequence of the robot through the pipe network is provided. Loop closure was performed for datasets B and C in the straight section between T-intersections T1 and T4. All images logged were grayscale.

Dataset	A	B	C
Sequence	start–T1–T2–T3–T4–end	start–T4–T3–T2–T1–T4–end	start–T4–T1–T2–T3–T4–T1–end
Distance (m)	≈ 38.0	≈ 45.0	≈ 48.0
Bit depth	8	16	16
fps	7.5	15.0	15.0
Resolution	1280 × 960	1280 × 960	1280 × 960
Number of images	26,000	24,000	32,000
Number of keyframes	2760	4170	4040
Loop closure	–	T1–T4	T1–T4
Structured lighting	no	no	Yes

total distance traveled, image resolution and bit depth of the grayscale images, camera frame rate, and the number of images and selected keyframes. The details of loop closure and structured lighting are also provided in the table. Loop closure correction was performed for datasets B and C in the straight section of pipe between T-intersections T1 and T4. The structured lighting system was only available for dataset C.

The visual odometry and sparse scene reconstruction results were found for each dataset. These results are presented and discussed in Section 6.1. Using the visual odometry results, a dense scene reconstruction for dataset A was generated. The details for generating this reconstruction and results are presented in Section 6.2.

6.1. Visual odometry and sparse scene reconstruction

The visual odometry and sparse scene reconstruction results for each dataset are shown in Figures 14 (dataset A), 15 (dataset B), and 16 (dataset C). The labels T1 through T4 in the figures correspond to those displayed in Figure 13. As the structured lighting system was not available for datasets A and B, a constant measurement of the pipe radius $r = 200\,\text{mm}$ was used during cylinder fitting and computation of the scene fit errors ϵ_X. For dataset C, the radius r was estimated during each sliding-window SBA using the structured lighting error ϵ_L in (24).

Loop closure for datasets B and C was performed using the procedure described in Section 5.3. For each dataset, 15 loop closure poses in the straight section of pipe connecting T-intersections T1 and T4 were used. All results presented for these datasets are post-loop-closure. A zoomed-in view of T-intersection T1 for dataset C is provided in Figure 17 after loop closure. It demonstrates the consistency of the two reconstructions of the T-intersection and the loop closure performance.

An ideal performance metric for our system is direct comparison of the visual odometry (camera pose) and scene structure results to accurate ground truth. However, obtaining suitable ground truth is particularly challenging due to the unavailability of GPS in a pipe, the limited accuracy of standard-grade IMUs, and the inability to accurately survey ground truth of the internal pipe structure.

Our current *approximate* ground truth is hand-held laser distance measurements estimating the distance between all T-intersection centroids $\mathcal{I}'$ (see Figure 12(b)). These measurements are compared to the T-intersection centroid $\mathcal{I}$ distance results from the pose estimation and mapping in Table 2. The laser ground truth measurements are only approximations as, for practical reasons, the measurements were taken between the centers of the upper exterior surfaces of each T-intersection. These reference points do not correlate directly to the 'true' centroids $\mathcal{I}$. Moreover, there were minor alterations to the pipe network in the period between collecting the datasets (sections of pipe temporarily removed then placed back). As as result, the errors reported in Table 2 are themselves only approximations, and may over- or underestimate the true errors.

The absolute percentage errors for all distance measurements are less than 1%, with the largest absolute error measured being 0.097 m (0.85%) between T1 and T3 for dataset C. This was the dataset which used the structured lighting system to estimate the internal diameter. Overall, similar accuracy was achieved for all datasets.

The limitation of any visual odometry system is the integration of incremental errors which can result in non-linear pose error growth. As a result, over longer transits the percentage errors reported in Table 2 are likely to increase. Loop closure, which was used effectively for two of our datasets, is one technique for correcting/minimizing long-range visual odometry errors when the robot returns to a previously visited location. As demonstrated in Lee et al. (2011), using a known graph-based structure of pipe networks nodes (e.g. T-intersections and elbow joints) can facilitate loop closure. If available, any prior metric information of the pipe network structure, such as the location of network nodes, could also be used to further minimize long-range visual odometry errors.

6.1.1. Effectiveness of scene constraints. Including scene constraints significantly improved the accuracy of pose

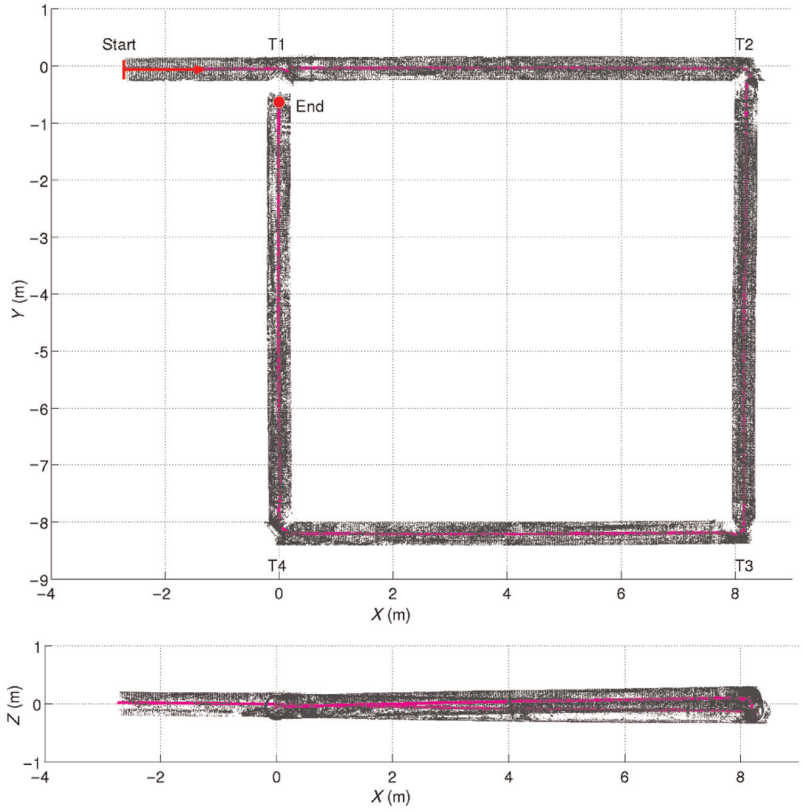


Fig. 14. Dataset A: Visual odometry (solid line) and sparse scene reconstruction (dots). The robot moved in the sequence start–T1–T2–T3–T4–end. No loop closure was implemented.

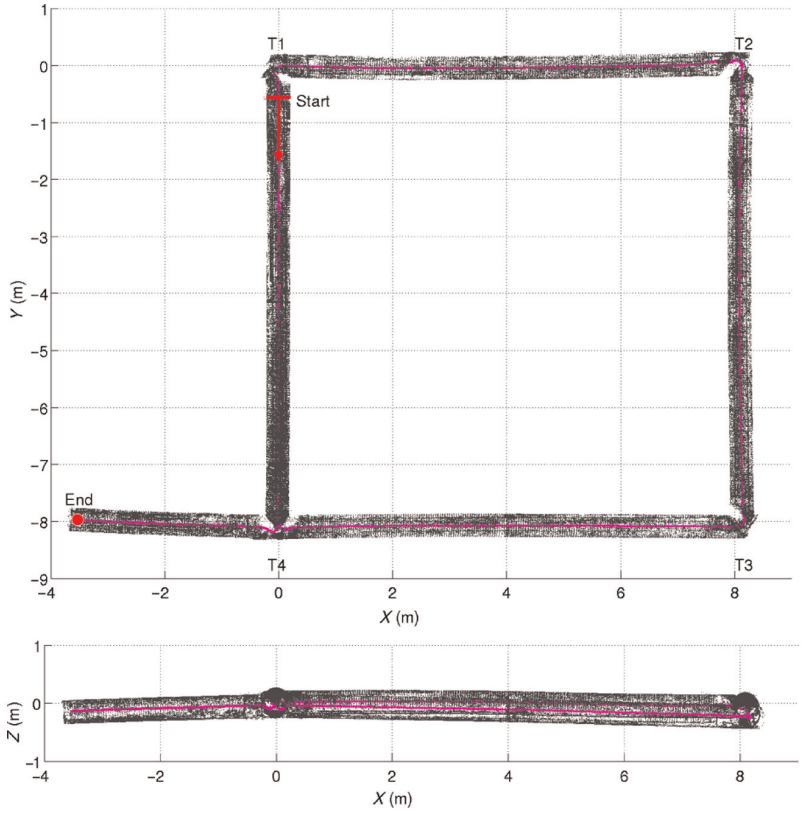


Fig. 15. Dataset B: Visual odometry (solid line) and sparse scene reconstruction (dots). The robot moved in the sequence start–T4–T3–T2–T1–T4–end. Loop closure was implemented between T1 and T4.

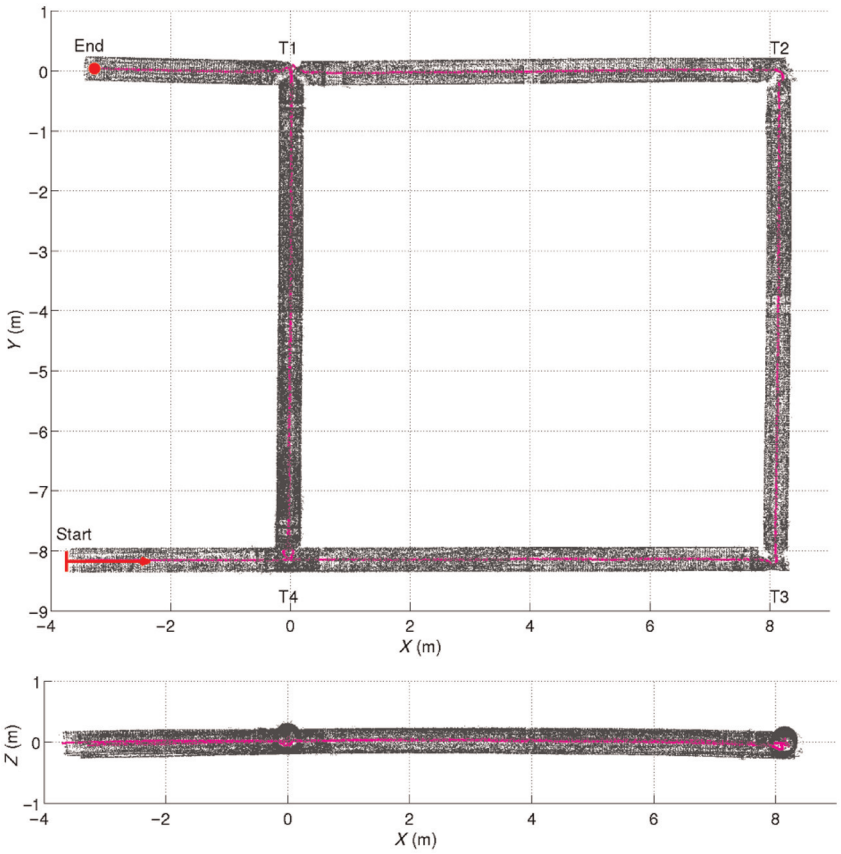


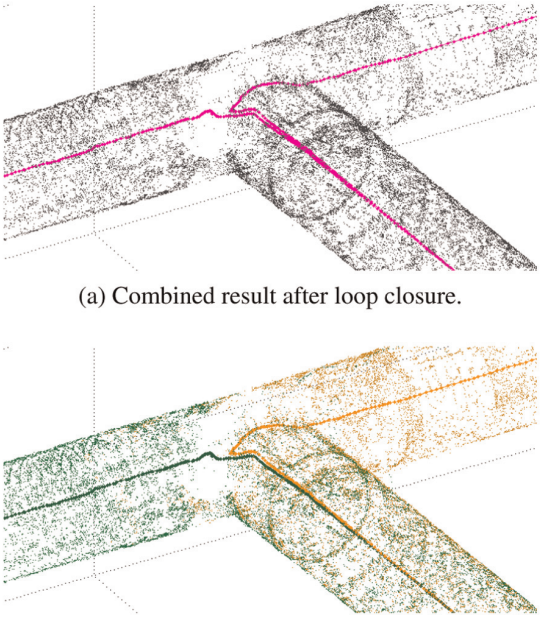
Fig. 16. Dataset C: Visual odometry (solid line) and sparse scene reconstruction (dots). The robot moved in the sequence start–T4–T1–T2–T3–T4–T1–end. Loop closure was implemented between T1 and T4.

Table 2. The distances between the T-intersection centroids I measured using a hand-held laser range finder (approximate ground truth: see text for detailed description), and those found from the visual odometry and mapping ('VO' in table). The signed distance errors and signed percentage errors are evaluated relative to the laser measurements.

Distance	T1–T2	T2–T3	T3–T4	T4–T1	T1–T3	T2–T4
Laser (m)	8.150	8.174	8.159	8.110	11.468	11.493
A: VO (m)	8.184	8.131	8.138	8.161	11.514	11.543
A: Error (m)	0.034	−0.043	−0.021	0.051	0.046	0.050
A: Error (%)	0.42	−0.53	−0.26	0.63	0.41	0.44
B: VO (m)	8.114	8.118	8.142	8.105	11.473	11.492
B: Error (m)	−0.036	−0.056	−0.017	−0.005	0.005	−0.001
B: Error (%)	−0.44	−0.69	−0.21	−0.07	0.04	−0.01
C: VO (m)	8.127	8.156	8.133	8.115	11.565	11.437
C: Error (m)	−0.023	−0.018	−0.026	0.005	0.097	−0.056
C: Error (%)	−0.28	−0.22	−0.31	0.06	0.85	−0.49

estimation and mapping. To illustrate, Figure 18 provides a qualitative comparison of the results for T-intersection T3 in dataset C with and without the use of the T-intersection model constraints. For the latter, a generic SBA is applied to all frames in the T-intersection minimizing only image reprojection errors $\epsilon_{\mathbf{I}}$. The relatively poor accuracy without scene constraints is due to several limitations of the fisheye camera system.

When compared to perspective cameras with similar pixel resolution, fisheye cameras have a larger angle of view, but decreased spatial resolution. This decreased spatial resolution limits the overall ability to reliably detect and measure pixel position changes of an imaged scene point across small changes in camera pose, especially in the presence of both feature detection position uncertainty and ZNCC matching noise. The spatial resolution of our



(a) Combined result after loop closure.

(b) Combined result after loop closure. The scene points reconstructed in the first and second transit through the T-intersection are displayed with different intensities.

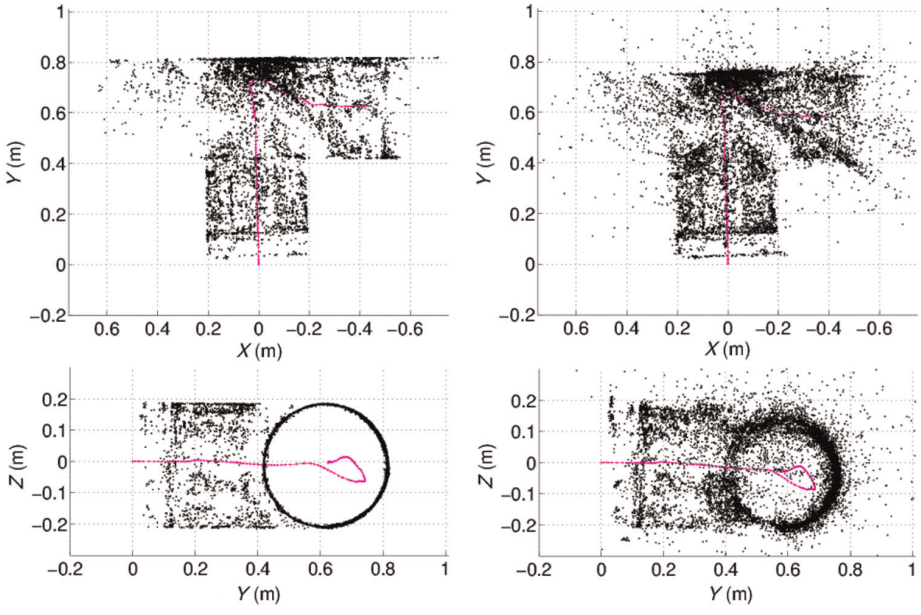
Fig. 17. The visual odometry and sparse reconstruction for T-intersection T1 in dataset C. The robot entered the T-intersection from the same direction two different times. The top image (a) shows the combined visual odometry and sparse scene points. The bottom image (a) uses separate intensities for the scene points reconstructed during the first and second transits through the T-intersection (orange and green colors respectively in the electronic version).

fisheye system could easily be improved by selecting a higher-resolution camera.

The second limitation relates specifically to operation in a pipe. As discussed in Section 4.1.1, the motion of the robot and the close proximity of the pipe surface creates strong projective changes between views which limits the number of frames that features can be tracked. This limits the maximum effective baseline used to optimize scene point coordinates, which is the Euclidean distance between the first and last camera pose each scene point is tracked. As is the case with stereo systems, scene reconstruction accuracy improves with increasing baseline, assuming of course the relative pose between the views is accurately estimated/calibrated.

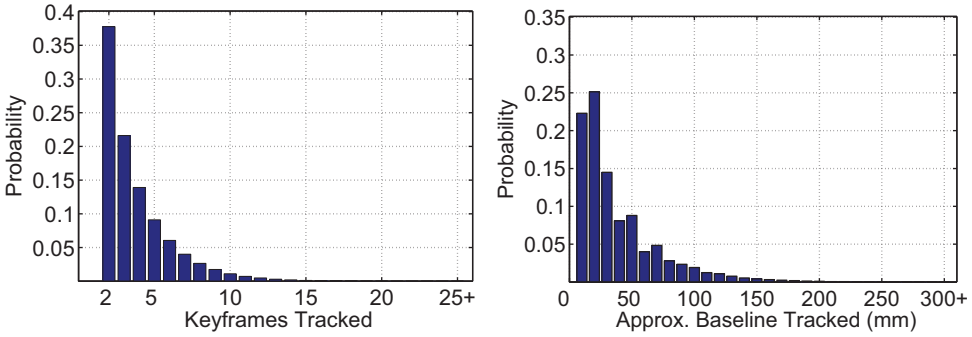
Figure 19 shows, for each dataset, probability distribution functions for the number of frames each scene point is tracked and the maximum effective baseline. It is clear from the figure that the effective baselines for scene points can be small. This is particularly true for tracked scene points in the T-intersection where camera motion between keyframes is predominately rotational and not translational. The scene constraints are particularly effective when processing the T-intersections for this reason, as demonstrated in Figure 18(b). Again, increasing the camera resolution would improve scene reconstruction accuracy across small baselines by increasing spatial resolution.

Although the scene constraints used were very effective for improving pose estimating and mapping performance, some considerations had to be taken into account when processing the straight sections of the pipe network. In practice, modeling each long straight section of our pipe network as a perfect straight cylinder is too restrictive.

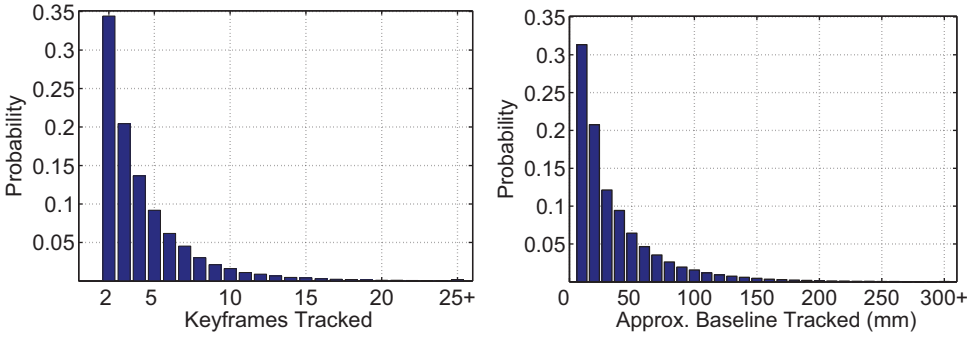


(a) T3 with model constraints. (b) T3 without model constraints.

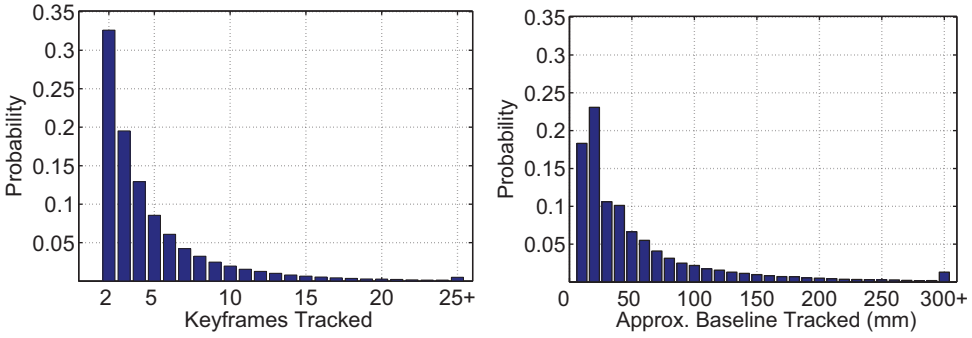
Fig. 18. Top and side views of the SBA result for T-intersection T3 in dataset C with (a) and without (b) the T-intersection model scene constraints. The results are shown in a local coordinate frame. The solid line is the camera path, and the points the sparse 3D reconstruction. Including scene constraints significantly improves the accuracy of the results, particularly the sparse scene reconstruction.



(a) Dataset A.



(b) Dataset B.



(c) Dataset C.

Fig. 19. Probability distributions for the number of keyframes each observed scene point is tracked, and the approximate effective baseline of the tracked scene points. This effective baseline is the Euclidean change in position between the first and last camera poses a scene point is tracked.

Firstly, each individual pipe segment contains some degree of curvature/sag. Secondly, the individual segments used to construct the long straight sections of pipe (see Figure 13) are not precisely aligned. It is for this reason that we only perform straight-cylinder-fitting locally within the 100 keyframe sliding-window SBA, which typically spans a 1 m length of pipe, and limit the maximum value for τ in (12). Doing so permits some gradual pipe curvature to be achieved in the results, as visible in Figure 12(a). For more severe pipe misalignments or deformations from excessive sag, modifications to the straight cylinder model may be more suitable. For example, a cubic spline modeling of the cylinder axis may be appropriate, despite the significant increase in computational expense when computing the scene point regularization errors.

As our pipe network contained only straight sections and T-intersections, only scene model constraints for these were derived and implemented. We anticipate that the same general procedure could be used in more complex pipe networks containing alternate configurations and connections. An example is curved elbow joints which appear frequently in pipe networks, and for which the current straight cylinder fitting model would be unsuitable. By extending the classification system in Section 4.2 to include elbow joints

Table 3. The pipe radius estimates for the cylinders fitted to re-constructed scene points at the start and end of each straight section for dataset C: refer to Figure 16. The scale of the reconstructed scene points is estimated using the structured lighting. There are two values for T1–T4 as the robot moved through the straight section connecting T1–T4 twice.

Section	start–T4	T1–T4	T1–T2	T2–T3	T3–T4	T1–end
r_{start} (mm)	199.30	199.54 199.95	199.32	199.83	197.83	197.29
r_{end} (mm)	196.66	199.26 199.32	197.53	200.01	200.10	199.29

(e.g. Lee et al., 2009), it would then be necessary only to define a suitable prior scene constraint, compute a scene fit error metric, and add the scene fit error as a regularization term during SBA. The regularization term can be weighted by any scalar multiplier to change its contribution to the overall error.

6.1.2 Accuracy of structured lighting. The structured lighting system improves the flexibility of the pipe-mapping system by providing a means for estimating the internal pipe radius. The structured lighting system was used while processing dataset C, and the accuracy of the results in Table 2 are similar to those for datasets A and B which both used a fixed measured pipe radius r. This demonstrates the suitability of the system for estimating the internal radius r.

Referring to Figure 12, straight cylinders are fitted to the endpoints of each straight section before processing the T-intersections. As a more quantitative measure of the structured lighting accuracy, the radii of these cylinders fitted for dataset C are give in Table 3. There are two measurements for the start/end cylinders for the straight section between T1 and T4 as the robot traveled through this section twice. There is a high repeatability between the corresponding start/end radii for this section. Additionally, with the exception of r_{end} for start–T4, all fitted radii are within a few millimeters of the expected value $r = 200$ mm.

Figure 20 shows the sensitivity of the top and bottom lasers over a range of pipe radii. This sensitivity is the change in fisheye pixel coordinate du of the laser spot along the epipolar line for a change in pipe radius dr. The values were computed using the fisheye intrinsic and laser extrinsic calibration, and assuming the camera was centered in the pipe with the principal axis aligned with the pipe axis (i.e. looking directly down the pipe). Increasing the sensitivity of each laser, and the number of lasers, could improve the overall accuracy of the structured lighting system. Both could be achieved through hardware design. For the former, increased sensitivity could be achieved by increasing the camera resolution, the baseline between the camera and each laser, or both. For reference, the current baseline between the camera and each laser is only approximately 50 mm. For the latter, a ring of laser diodes could be constructed producing a visible laser spot field in the images.

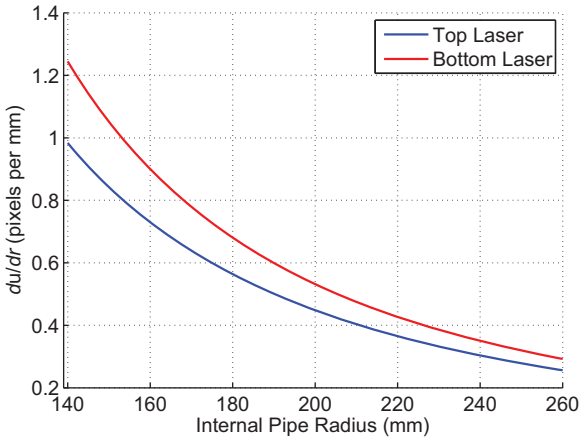
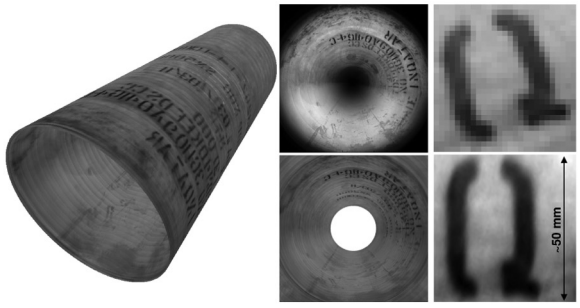


Fig. 20. The estimated sensitivity $d\mathbf{u}/dr$ for each laser over a range of pipe radii. All values were computed assuming the camera is centered in a cylindrical pipe, with its principal axis aligned with the cylinder axis (i.e. looking directly down the pipe).

6.2. Dense rendering

Once the camera pose and structure estimates have been obtained, an appearance map of the interior surface of the pipe network can be produced. This map can be used as input for automated corrosion detection algorithms and direct visualization in rendering engines. Figure 21 shows the appearance map of the pipe network produced using dataset A. The figure includes a zoomed-in view of a small straight section (Figure 21(a)) and T-intersection (Figure 21(b)) to highlight the detail. Logging color images would allow a full RGB model to be produced.

The appearance map shown is a dense 3D grayscale point cloud which could be extended to a full facet model. The Euclidean point cloud coordinates were set using the cylinder-fitting results for both the straight sections and T-intersections. The grayscale intensity value for each point was obtained by mapping it into all valid fisheye image keyframes, and then taking the average sampled image intensity value over all the keyframes. Here, valid is defined as having a projected angle of colatitude $50° < \theta < 90°$ (i.e. near the periphery of the fisheye images where spatial resolution is a maximum). Sampling the intensity values from multiple images enables a high-resolution appearance map to be produced without excessive blurring. This is



(a) Small straight section 1 m in length. The middle column shows an original fisheye image (top), and dense reconstruction near the same location in the pipe (bottom). The right column is a small section of pipe cropped from the original image (top), and the dense reconstruction (bottom).

(b) T-intersection.

(c) Full dataset.

Fig. 21. Dense 3D grayscale appearance map of the pipe network; (a) and (b) are zoomed-in sections of (c).

demonstrated in Figure 21(a). It shows a sample letter printed on the interior surface of the pipe in an original fisheye image, and the high-resolution dense appearance map for the same section of pipe with intensity values sampled from multiple image keyframes. The consistency of this lettering further highlights the accuracy of the camera pose estimates obtained.

Additional techniques could be employed to enhance the dense appearance maps. They include polarization options to reduce specular reflections in the pipe, and gain-mask correction to correct for variations in ambient lighting from the LEDs.

7. Camera selection discussion

Our pipe-mapping system was required to produce accurate full coverage metric maps of a pipe network. As demonstrated in Section 6, the fisheye system proved suitable for this application.

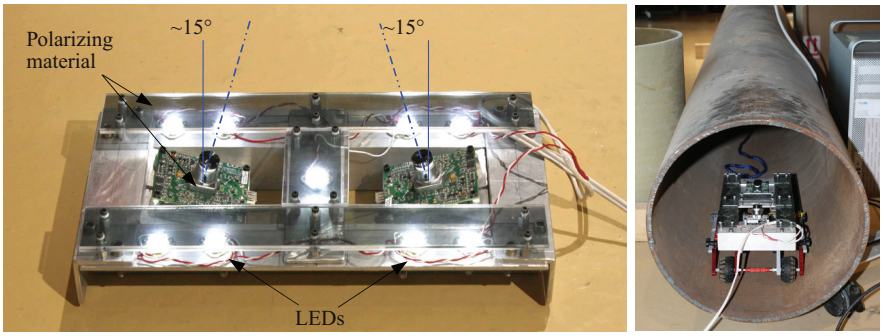
A range of alternate camera options was considered for pipe mapping, however, and as mentioned in Section 1, we previously developed a verged perspective stereo system (Hansen et al., 2011). Stereo cameras are frequently used for robot localization and mapping applications and enable metric scene reconstruction from a single stereo pair. The greatest advantage of our verged stereo system, which is summarized in this section, was the ability to generate high-resolution appearance maps of local patches of the interior surface of a pipe directly from dense stereo matching. However, unlike the fisheye system, there were numerous hardware-related concerns for deployed pipe-mapping operations which will be discussed. These concerns motivated the development of the fisheye pipe-mapping system.

7.1 Verged stereo system overview

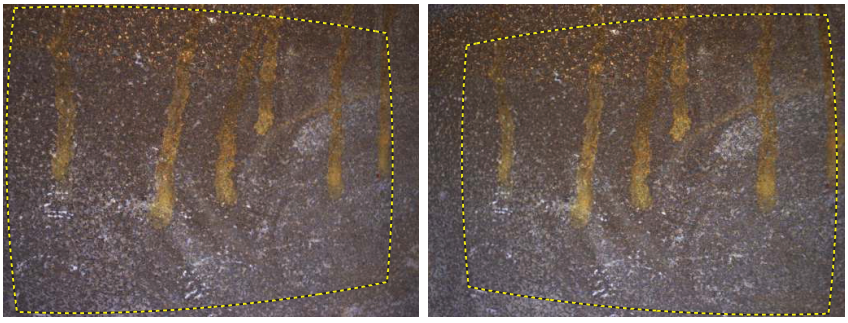
The stereo perspective camera is shown in Figure 22. It consists of two 1024×768-pixel-resolution FireWire cameras fitted with 4.0 mm focal length S-mount lenses. The stereo baseline is 150 mm, and each camera is verged inwards by 15° to achieve large overlapping fields of view on the interior pipe surface. The cameras are mounted on a large aluminium frame, as well as nine LEDs (the same as used for the fisheye system) which provide ambient lighting in the pipe. Polarizing material was placed above the LEDs and camera CCD sensors, oriented in orthogonal directions, to minimize specular reflections in the imagery.

Multiple datasets were collected in a 4.0 m long straight section of steel pipe with an internal diameter of 400 mm (16”). Accurate camera pose estimates were achieved using the stereo system, with errors well below 1% for distance traveled. Additionally, high-resolution dense appearance maps were produced from dense stereo matching. A more detailed description of the image processing techniques and results are presented in Hansen et al. (2011).

Referring to Figure 23, the accuracy of the results was improved after correcting for a changing extrinsic pose between the left and right cameras, which we attribute to thermal variations in the metal housing: the LEDs and cameras produce heat during operation. For this, we used our online stereo continuous extrinsic recalibration (OSCER) algorithm presented in Hansen et al. (2012). It estimates a unique five-degree-of-freedom extrinsic stereo pose for each frame, derived only from sparse stereo correspondences, and refined using a Kalman filter. Despite the significantly improved results shown in Figure 23, the



(a) The verged perspective stereo camera system, and position of camera inside the 400 mm internal diameter steel pipe used during testing.



(b) Sample left and right raw images (not stereo-rectified). The dashed lines indicate the overlapping fields of view of the individual cameras.

Fig. 22. The stereo camera system (a) and sample imagery (b). The cameras were verged inwards to achieve large overlapping fields of view.

recalibration is unable to estimate a changing stereo baseline (i.e. gauge freedom). Periodic and time-consuming offline calibration would be required for this.

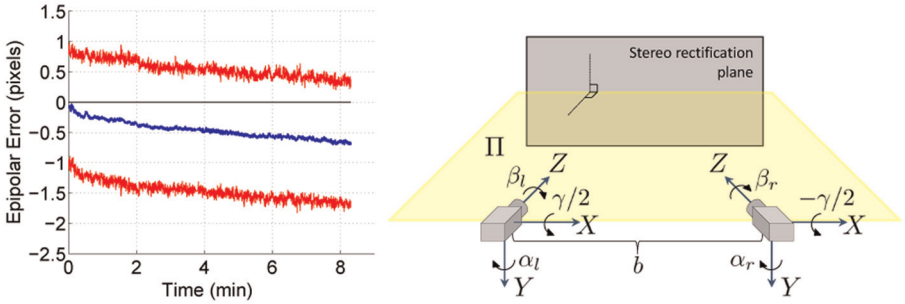
In addition to the recalibration issue discussed, there are other hardware challenges to consider when using stereo for pipe mapping. The first relates to operation in pipe networks with changing pipe diameters, and the second relates to the requirement for imaging the entire inner circumference of the pipe (i.e. full pipe coverage).

7.1.1. Changing pipe diameter. The first concern for any stereo camera system is the ability to maintain large overlapping left/right fields of view when operating in pipes with changing diameters. The vergence angle of our stereo camera was selected to maximize image overlap in the 400 mm internal diameter test pipe used in experiments. Unfortunately, large changes to the vergence angle would be required to ensure maximum image overlap as the pipe diameter changes, as illustrated in Figure 24. An active system would be required to change vergence angle in deployed operation, requiring additional hardware, and creating added relative stereo extrinsic calibration challenges.

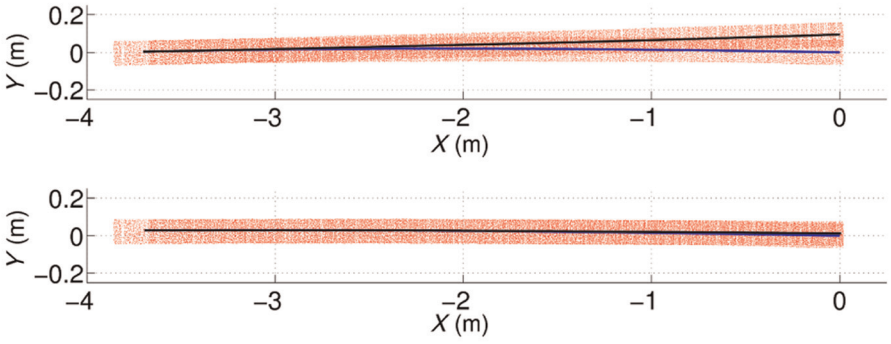
The second concern for a stereo camera operating in pipes with changing diameters is the limited focal range of many perspective lenses at short viewing distances. For our stereo camera, variations in the imaged scene point depths caused by the pipe curvature and camera vergence angle were greater than the focal range of the lenses at close viewing distance. The result was some out-of-focus blurring in regions of the images. Changing the overall pipe diameter, which changes the distance of the camera from the imaged pipe surface, would increase out-of-focus blurring. One potential solution to this problem would be the use of an active varifocal lens to re-focus when needed. However, the overall size and weight of the camera system would be increased, and online intrinsic recalibration of both cameras would be required.

In contrast, using a monocular fisheye system avoids the issue of maintaining left/right stereo overlap. Moreover, the current short focal length fisheye lens used has a considerably larger focal range than perspective cameras at short viewing distances. This gives it the ability to capture in-focus images over a large range of pipe diameters.

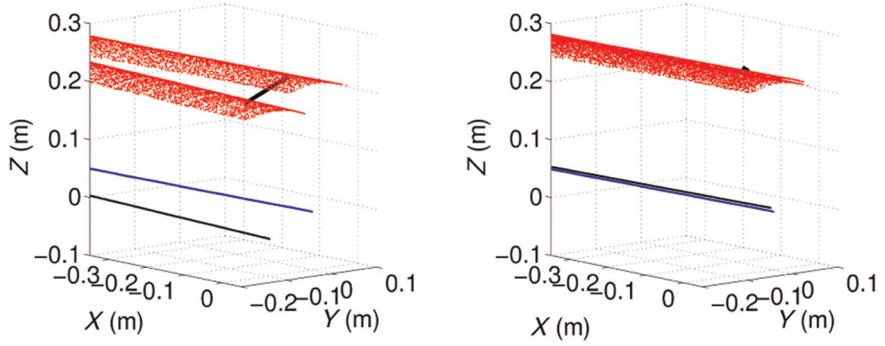
7.1.2. Full pipe coverage. An important requirement for pipe mapping is the ability to image and map the entire inner surface of the pipe network. In Li et al. (2012) and Yu et al. (2014), this requirement was formulated as a region-



(a) The temporal variation of the stereo-rectified epipolar errors before recalibration of the five-degree-of-freedom relative extrinsic pose with parameters $\Phi = (\alpha_l, \beta_l, \alpha_r, \beta_r, \gamma)$. The mean (middle line) and $\pm 3\sigma$ error bars of the epipolar errors are displayed.



(b) Visual mapping of a 4 m section of pipe without (top) and with (bottom) online stereo recalibration.

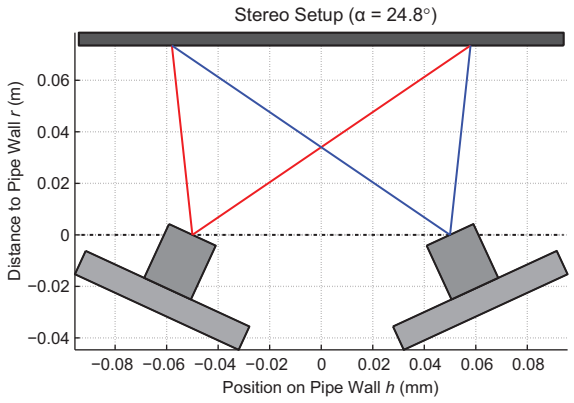


(c) Reconstruction at the start and end without (left) and with (right) online recalibration. The same scene point mapped at the start and the end is connected by a solid line.

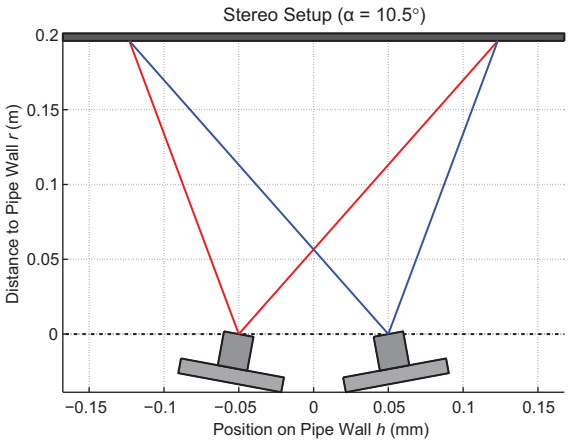
Fig. 23. Visual odometry and pipe-mapping results for a straight 4.0 m long section of 400 mm internal diameter pipe using the perspective stereo system.

guarding problem for robotic pipe inspection. In this context, the region-guarding problem was to find a minimum set of inspection spots that could be used to image/inspect the entire inner surface of a pipe network. Here we restrict our discussion to the simple requirement for the camera system to image the entire circumference of the inner pipe surface in the local neighborhood of the robot.

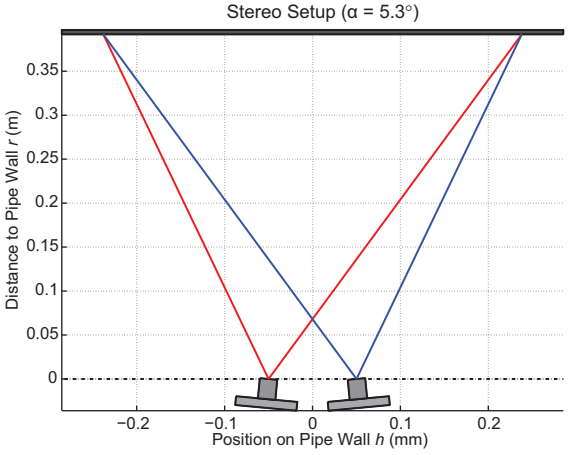
For a perspective stereo camera, which has a limited angle of view, it would be necessary to construct a ring of cameras, as illustrated in Figure 25. This presents space concerns for a stereo system, particularly in small-diameter pipe networks. Furthermore, the accuracy of the pipe mapping would be dependent on accurate intrinsic and extrinsic calibration of many cameras. As mentioned, even for our



(a) 150 mm (≈ 6”) diameter: vergence angle α = 24.8°.



(b) 400 mm (≈ 16”) diameter: vergence angle α = 10.5°.



(c) 810 mm (≈ 32”) diameter: vergence angle α = 5.3°.

Fig. 24. Stereo camera vergence angles α required to achieve maximum image overlap in pipes with internal diameters of (a) 150 mm, (b) 400 mm, and (c) 810 mm. The vergence angle α of each camera is provided in the captions. The camera parameters used are: 1/3” format image sensor, 4.0 mm focal length lens, 100 mm stereo baseline. The lines represent camera angle of view.

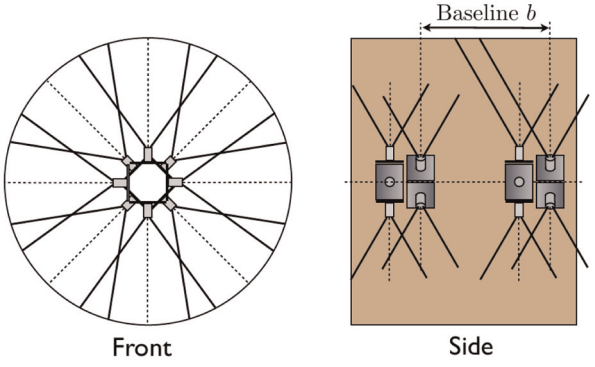


Fig. 25. An example multiple stereo camera system capable of imaging the entire inner surface of a section of pipe. Multiple cameras are required to ensure overlapping fields of view.

single stereo camera system, an online extrinsic recalibration of the relative left/right camera poses was necessary to achieve accurate results.

The advantage of the wide-angle fisheye system presented is its ability to achieve full pipe coverage using a single camera, as demonstrated in the results in Section 6. Relative camera extrinsic calibration challenges are avoided, and the overall size of the camera hardware remains compact.

8. Conclusions

A visual odometry and mapping system was presented for non-destructive in-pipe inspection. A target domain for the system is inspection of natural gas pipes where the

detection of structure changes, for example corrosion, is critically important. Camera pose estimates and sparse scene reconstruction results are generated from fisheye imagery calibrated using a novel combination of checkerboard data and visual feature tracks in a straight section of pipe. To find the pose and sparse structure estimates, a traditional SBA is extended to include weak geometric priors of the pipe network components. More specifically, straight-section and T-intersection scene fitting errors are included as regularization terms in the SBA framework. A structured lighting system was developed and incorporated into the SBA framework with the scene fitting constraints, enabling metric estimates for camera position and scene structure to be obtained without requiring a full stereo solution.

Visual odometry and scene reconstruction results were presented for three datasets logged in a 400 mm (16”) internal diameter fiberglass pipe network. The accuracy of the pipe network reconstruction was evaluated by comparing the distance between all T-intersections in the network. All distance measurements obtained were well within $\pm 1\%$ of ground truth laser distance estimates, with the dataset using the structured lighting showing comparable accuracy to those using a fixed a priori estimate of the metric pipe structure (i.e. pipe radius r). Moreover, these results were used to demonstrate the effectiveness of scene constraints for reducing SBA inaccuracies caused by the limited spatial resolution of the fisheye imagery and challenges associated with tracking features in the challenging raw imagery.

A sample dense 3D appearance map of the internal pipe structure was produced for one of the datasets using the visual odometry and scene reconstruction results. This demonstrated the ability to produce a consistent appearance map which could be used as input for automated inspection algorithms, or imported into rendering engines for visualization and direct metric measurement. This improves on a current industry practice where a remote operator manually inspects imagery from an in-pipe inspection robot without localization or structure information.

Funding

This publication was made possible by the NPRP (grant number 08-589-2-245) from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

Bay H, Ess A, Tuytelaars T, et al. (2008) Speeded-up robust features (SURF). *Computer Vision and Image Understanding* 110(3): 346–359.

Bülow T (2004) Spherical diffusion for 3D surface smoothing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(12): 1650–1654.

Cummins M and Newman P (2008) FAB-MAP: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research* 27(6): 647–665.

Cummins M and Newman P (2011) Appearance-only SLAM at large scale with FAB-MAP 2.0. *The International Journal of Robotics Research* 30(9): 1100–1123.

Daniilidis K, Makadia A and Bülow T (2002) Image processing in catadioptric planes: Spatiotemporal derivatives and optical flow computation. In: *Proceedings of the third workshop on omnidirectional vision*, pp. 3–10.

Dubbelman G, Hansen P, Browning B, et al. (2012) Orientation only loop-closing with closed-form trajectory bending. In: *International conference on robotics and automation*, pp. 815–821.

Faugeras OD, Luong QT and Maybank SJ (1992) Camera self-calibration: Theory and experiments. In: Sandini G (ed.) *Computer Vision ECCV'92* (Lecture Notes in Computer Science, vol. 588). Berlin/Heidelberg: Springer, pp. 321–334.

Fischler MA and Bolles RC (1981) Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* 24(6): 381–395.

Fitzgibbon A (2001) Simultaneous linear estimation of multiple view geometry and lens distortion. In: *International conference on computer vision and pattern recognition*.

Gennery D (2006) Generalized camera calibration including fish-eye lenses. *International Journal of Computer Vision* 68(3): 239–266.

Gluckman J and Nayar S (1998) Ego-motion and omnidirectional cameras. In: *Sixth international conference on computer vision*, pp. 999–1005.

Hansen P, Alismail H, Browning B, et al. (2011) Stereo visual odometry for pipe mapping. In: *International conference on intelligent robots and systems*, pp. 4020–4025.

Hansen P, Alismail H, Rander P, et al. (2012) Online continuous stereo extrinsic parameter estimation. In: *IEEE conference on computer vision and pattern recognition*, pp. 1059–1066.

Hansen P, Corke P and Boles W (2010) Wide-angle visual feature matching for outdoor localization. *The International Journal of Robotics Research* 29(2–3): 267–297.

Hansen P, Corke P, Boles W, et al. (2007) Scale invariant feature matching with wide angle images. In: *International conference on intelligent robots and systems*, pp. 1689–1694.

Harris C and Stephens M (1988) A combined corner and edge detector. In: *Proceedings of the fourth Alvey vision conference*, pp. 147–151.

Hartley R (1994) Self-calibration from multiple views with a rotating camera. In: *Proceedings of the third European conference on computer vision*, pp. 471–478.

Hartley R and Kang SB (2005) Parameter-free radial distortion correction with centre of distortion estimation. In: *International conference on computer vision*, pp. 1834–1841.

Hartley R and Zisserman A (2004) *Multiple View Geometry in Computer Vision*. 2nd edn. Cambridge: Cambridge University Press.

Kang SB (2000) Catadioptric self-calibration. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 201–207.

Kannala J and Brandt S (2006) A generic camera model and calibration method for conventional, wide-angle and fish-eye lenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(8): 1335–1340.

Kelly A, Chan N, Herman H, et al. (2011) Real-time photorealistic virtualized reality interface for remote mobile robot control. *The International Journal of Robotics Research* 30(3): 384–404.

Kümmerle R, Grisetti G, Strasdat H, et al. (2011) g^2o: A general framework for graph optimization. In: *International conference on robotics and automation*, pp. 3607–3613.

Lee D, Moon H and Choi H (2011) Autonomous navigation of in-pipe working robot in unknown pipeline environment. In: *IEEE international conference on robotics and automation*, pp. 1559–1564.

Lee J, Roh S, Kim D, et al. (2009) In-pipe robot navigation based on the landmark recognition system using shadow images. In: *IEEE international conference on robotics and automation*, pp. 1857–1862.

Li H and Hartley R (2005) A non-iterative method for correcting lens distortion from nine point correspondences. In: *Proceedings of OmniVision'05*.

Li H and Hartley R (2006) Plane-based calibration and auto-calibration of a fish-eye camera. In: *Computer Vision – ACCV 2006* (Lecture Notes in Computer Science, vol. 3851). Berlin/-Heidelberg: Springer, pp. 21–30.

Lindeberg T (1990) Scale-space for discrete signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12(3): 234–254.

Lindeberg T (1994) *Scale-Space Theory in Computer Vision*. Norwell, MA: Kluwer Academic Publishers.

Li X, Yu W, Lin X, et al. (2012) On optimizing autonomous pipeline inspection. *IEEE Transactions on Robotics* 28(1): 223–233.

Lourenço M, Barreto J and Vasconcelos F (2012) sRD-SIFT: Keypoint detection and matching in images with radial distortion. *IEEE Transactions on Robotics* 28(3): 752–760.

Lowe D (2004) Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2): 91–110.

Mičušík B and Pajdla T (2003) Estimation of omnidirectional camera model from epipolar geometry. In: *Proceedings of the IEEE Computer Society conference on computer vision and pattern recognition*, pp. 485–490.

Mirats Tur JM and Garthwaite W (2010) Robotic devices for water main in-pipe inspection: A survey. *Journal of Field Robotics* 27(4): 491–508.

Nayar S (1997) Catadioptric omnidirectional camera. In: *Proceedings of the IEEE Computer Society conference on computer vision and pattern recognition*, pp. 482–488.

Nelson CR and Aloimonos J (1988) Finding motion parameters from spherical motion fields (or the advantages of having eyes in the back of your head). *Biological Cybernetics* 58: 261–273.

Neumann J, Fermüller C and Aloimonos Y (2002) Eyes from eyes: New cameras for structure from motion. In: *Proceedings of the third workshop on omnidirectional vision*, pp. 19–26.

Nistér D (2004) An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(6): 756–770.

Nistér D, Naroditsky O and Bergen J (2006) Visual odometry for ground vehicle applications. *Journal of Field Robotics* 23(1): 3–20.

Roh S, Kim D, Lee J, et al. (2009) In-pipe robot based on selective drive mechanism. *International Journal of Control, Automation, and Systems* 7(1): 105–112.

Royer E, Lhuillier M, Dhome M, et al. (2007) Monocular vision for mobile robot localization and autonomous navigation. *International Journal of Computer Vision* 74(3): 237–260.

Scaramuzza D and Siegwart R (2008) Appearance-guided monocular omnidirectional visual odometry for outdoor ground vehicles. *IEEE Transactions on Robotics* 24(5): 1015–1026.

Schempf H, Mutschler E, Gavaert A, et al. (2010) Visual and nondestructive evaluation inspection of live gas mains using the Explorer^{TM} family of pipe robots. *Journal of Field Robotics* 27(3): 217–249.

Sivic J and Zisserman A (2003) Video Google: A text retrieval approach to object matching in videos. In: *International conference on computer vision*, pp. 1470–1477.

Strelow D and Singh S (2004) Motion estimation from image and inertial measurements. *The International Journal of Robotics Research* 23(12): 1157–1195.

Thirthala S and Pollefeys M (2005) The radial trifocal tensor: A tool for calibrating the radial distortion of wide-angle cameras. In: *IEEE Computer Society conference on computer vision and pattern recognition*, pp. 321–328.

Triggs B, McLauchlan PF, Hartley RI, et al. (2000) Bundle adjustment – A modern synthesis. In: *Proceedings of the international workshop on vision algorithms: Theory and practice*, pp. 298–372.

Xiong Y and Turkowski K (1997) Creating image-based VR using a self-calibrating fisheye lens. In: *International conference on computer vision and pattern recognition*, pp. 237–243.

Yu W, Li M and Li X (2014) Optimizing pyramid visibility coverage for autonomous robots in 3D environment. *Control and Intelligent Systems* 42(1): 9–16.

Zhang Z (1996) On the epipolar geometry between two images with lens distortion. In: *Proceedings of the international conference on pattern recognition*, pp. 407–411.