Stereo Visual Odometry for Pipe Mapping

Peter Hansen, Hatem Alismail, Brett Browning and Peter Rander

Abstract— **Pipe inspection is a critical activity in gas production facilities and many other industries. In this paper, we contribute a stereo visual odometry system for creating high resolution, sub-millimeter maps of pipe surfaces. Such maps provide both 3D structure and appearance information that can be used for visualization, cross registration with other sensor data, inspection and corrosion detection tasks. We present a range of optical configuration and visual odometry techniques that we use to achieve high accuracy while minimizing specular reflections. We show empirical results from a range of datasets to demonstrate the performance of our approach.**

I. INTRODUCTION

Stereo vision has proven to be a useful tool for mobile robot localization and mapping in indoor, and outdoor extraterrestrial and underwater environments [1], [2], [3], [4], [5]. If a calibrated stereo rig is used, then metric pose estimates and sparse world structure (maps) can be obtained which can then be upgraded using dense stereo reconstruction methods. In this paper, we present a stereo visual odometry system for building high resolution appearance and 3D structure maps for pipes such as those used in Liquified Natural Gas (LNG) production facilities.

Pipe inspection is a critical task in many industries, particularly those involved in natural gas production where corrosion can be a critical safety hazard. Regular inspection of such pipes is therefore necessary to avoid potentially catastrophic failures. Current practice, though, relies on manual visual inspection (e.g. [6]), which can be difficult for operators due to fatigue, lack of scale, or suitable visualization tools. Non-vision alternatives such as magnetic flux leakage (MFL) have other limiting factors such as accuracy, false positives, and poor visualization tools. In either case, precise localization of the vehicle in the pipe and localization of a feature on the pipe surface may be challenging. We aim to address these issues through an automated visual mapping system that can produce high resolution, sub-millimeter 3D appearance maps of the pipe surface. Such maps can be used for direct metric measurements, for visualization in a 3D rendering engine, or as input to automatic corrosion detection algorithms.

In prior work [7], we presented an automated vision approach to this problem where we developed an accurate monocular visual odometry (VO) algorithm to map the inside surface of the pipe. Our results showed high accuracy was achievable, but that restrictive assumptions about pipe geometry were required to resolve the monocular scale ambiguity. We introduced a novel incorporation of these supplied constraints into the Bundle Adjustment optimization process that produced the final polished result.

In this work, we develop a new stereo visual odometry algorithm applicable for operations in pipe environments. Stereo visual odometry is a well studied problem, with work on feature detection and tracking (e.g. [8]), pose estimation (e.g. [9]), and non-linear, least squares bundle adjustment for polishing solutions [10], [11]. Despite a number of successful applications to several specific domains [2], [3], [4] (albeit not in pipes to our knowledge), it remains non-trivial to develop a stereo visual odometry solution for domains where visual structure and appearance is very different from the above scenarios. As such, we present two contributions in this paper. First, we show how stereo visual odometry can be extended to operate with pipe environments and what is required in terms of features, tracking, stereo configuration, calibration, and bundle adjustment to achieve reliable results. In particular, the restricted confines of pipes mean that most lensing configurations are not able to achieve full focus over the pipe surface. This creates challenges for feature localization and therefore visual odometry accuracy. Our second contribution is a physical implementation that addresses the challenges of stereo, minimizing specular reflections, and minimizing blur. The result is a compact verged stereo vision system that no longer requires restrictive assumptions on pipe geometry and through empirical tests, we show that it is able to produce high accuracy, sub-millimeter maps. We evaluate the performance of our system on a number of datasets collected specifically for the task.

In the next section, we describe our stereo test rig and the lighting and filtering approach required to minimize the impact of specular reflections inside metal pipes. In section III, we describe the method we use to establish stereo correspondences. In section IV we describe the core details of our algorithm including mechanisms to increase the number and quality of stereo correspondences, robust tracking and pose estimation. In section V, we show the accuracy of our system on several datasets before concluding the paper in section VI.

II. STEREO CAMERA SYSTEM

The stereo camera is required to image a scene (interior pipe surface) at a distance of approximately 200mm from the camera. The overlapping field of view of many commercially available fronto-parallel stereo cameras is very limited, or non-existent, at this working distance. Furthermore, their lenses Minimum Operating Distance (MOD) often exceeds 200mm, which would limit their ability to produce focused images. For these reasons we have designed and assembled a custom stereo camera. We detail the hardware and calibration procedure in this section.

This paper was made possible by the support of an NPRP grant from the Qatar National Research Fund. The statements made herein are solely the responsibility of the authors.

Alismail, Rander, and Browning are with the Robotics Institute/NREC, Carnegie Mellon University, Pittsburgh PA, USA, {halismai,rander,brettb}@cs.cmu.edu. Hansen and Browning are with the Qri8 lab, Carnegie Mellon University, Doha, Qatar phansen@qatar.cmu.edu

A. Hardware

Referring to figure 1, the stereo camera consists of two 1024×768 RBG color Firewire cameras mounted on a rigid aluminum frame. Each camera has a 1/3" format CCD, and is fitted with a 6.0mm focal length S-mount lens with a 150mm MOD. The baseline separation between camera centers is approximately 140mm, and each camera is verged (rotated) inwards by approximately 15 degrees. This verging of the cameras is critical for ensuring sufficient stereo image overlap, as illustrated in figure 2. Nine 3.5 Watt Light Emitting Diodes (LEDs) are mounted on the aluminum frame, and provide the only light source during dataset collection. To minimize specularities, we positioned polarizing material above the LEDs, and the same polarizing material in the orthogonal direction within the camera mounts (i.e. polarized lenses). During datasets collection, we log synchronous, time-stamped RGB images from the stereo camera at 7.5 frames per second (fps). The exposure time and gain of the individual cameras were configured manually.

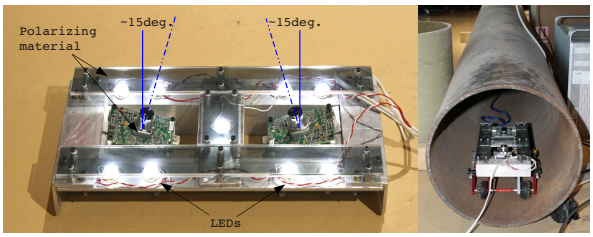
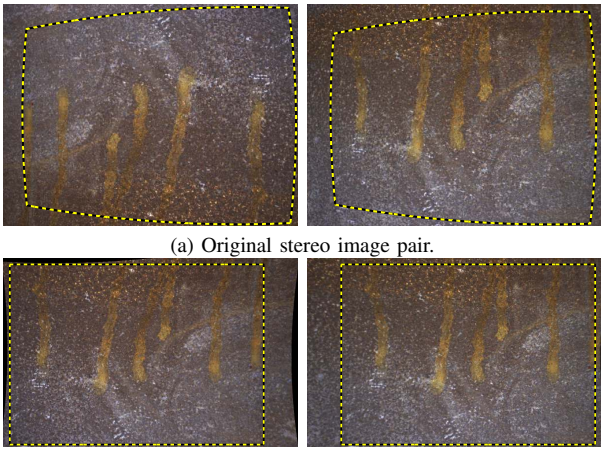


Fig. 1: The prototype verged stereo camera, and its position inside the 16" pipe used in the experiments in section V. This is a typical pipe diameter used in LNG processing facilities.



(a) Original stereo image pair.

(b) Stereo rectified image pair.

Fig. 2: The dashed lines enclose the *approximate* overlapping regions of the left and right cameras in the (a) original stereo image pair, and (b) stereo rectified image pair. These images were taken from the first pipe dataset in section V.

B. Calibration

For each camera (left and right), the intrinsic image formation model and parameters were obtained using the Matlab Calibration Toolbox[1].

[1] `http://www.vision.caltech.edu/bouguetj/calib_doc/index.html`.

Extrinsic calibration is accomplished by optimizing reprojection errors of calibration targets as follows. The left and right camera poses, P_l and P_r respectively, are the 3×4 matrices

$$P_l(R_l, \mathbf{t}_l) = [R_l \,|\, \mathbf{t}_l] \quad P_r(R_r, \mathbf{t}_r) = [R_r \,|\, \mathbf{t}_r], \tag{1}$$

where R_l, R_r are 3×3 rotation matrices, and $\mathbf{t}_l, \mathbf{t}_r$ are 3×1 translation vectors. These camera poses describe the mapping of the coordinate $\mathbf{X}_g = (X_g, Y_g, Z_g)^T$ of a scene point in the global coordinate frame, to the coordinate $\mathbf{X}_l$ in the left camera frame, and $\mathbf{X}_r$ in the right frame:

$$\mathbf{X}_l = R_l \, \mathbf{X}_g + \mathbf{t}_l \tag{2}$$

$$\mathbf{X}_r = R_r \, \mathbf{X}_g + \mathbf{t}_r. \tag{3}$$

The pose S of the right camera with respect to the left, referred to as the stereo extrinsic pose, is

$$S(R, \mathbf{t}) = [R \,|\, \mathbf{t}] \tag{4}$$

$$= [R_r \, R_l^T \mid -R_r \, R_l^T \, \mathbf{t}_l + \mathbf{t}_r], \tag{5}$$

and defines the mapping $\mathbf{X}_r = R \, \mathbf{X}_l + \mathbf{t}$.

Before collecting each of the datasets presented in section V, we estimate S using a set of 50 stereo images of a checkerboard pattern with known geometry in a global coordinate frame. For a given estimate of P_l and S, the position of the checkerboard corners in the left and right images of a stereo image pair can be obtained using the camera intrinsic parameters. The error between these reprojected positions and their observed (detected) position is the reprojection error. We optimize P_l for each stereo pair, as well as the single extrinsic pose S which applies to all stereo image pairs, by minimizing the sum of squared reprojection errors in all stereo image pairs. This non-linear optimization is implemented using Levenberg-Marquardt and a quaternion parameterization of all rotation matrices.

III. STEREO CORRESPONDENCES

A. Stereo Rectified Images

Grayscale stereo rectified images are used to find stereo scene point correspondences. Using the intrinsic calibration parameters for the left and right cameras, the normalized pinhole (ray-based) coordinates $\mathbf{x}_l(x_l, y_l, 1)$ and $\mathbf{x}_r(x_r, y_r, 1)$ of any pixel in the left and right image, respectively, can be derived. The stereo rectified image coordinates $\mathbf{u}_l$ and $\mathbf{u}_r$ are produced by rotating the rays about the camera centers, and then applying a pinhole projection using the left and right camera matrices K_l and K_r:

$$\mathbf{u}_l = K_l \, \tilde{R}_l \, \mathbf{x}_l \tag{6}$$

$$\mathbf{u}_r = K_r \, \tilde{R}_r \, \mathbf{x}_r. \tag{7}$$

The rotations used in the rectification, $\tilde{R}_l$ and $\tilde{R}_r$, rotate the cameras principal axes so that they are orthogonal to the vector joining the camera centers (i.e. the baseline), and the epipoles in the rectified images are horizontally aligned[2].

The camera matrices have the form

$$K_l = \begin{bmatrix} f & 0 & u_{0_l} \\ 0 & f & v_{0_l} \\ 0 & 0 & 1 \end{bmatrix} \quad K_r = \begin{bmatrix} f & 0 & u_{0_r} \\ 0 & f & v_{0_r} \\ 0 & 0 & 1 \end{bmatrix}, \tag{8}$$

[2] See [12] for a more detailed description of stereo rectification.

where f is the focal length, u_{0_l} and v_{0_l} is the coordinate of the principal point in the left image, and u_{0_r} and $v_{0_r} = v_{0_l}$ is the coordinate of the principal point in the right image. An example pair of color stereo rectified images was shown in figure 2b. They are 1199×768 pixels in size, and have a focal length of $f = 1200$ pixels.

B. Initial Stereo Correspondences

Correspondences between the left and right rectified stereo images are found using a combination of sparse feature detection/matching and Zero-mean Normalized Cross Correlation (ZNCC).

A sparse set of Harris corners [13] are detected in the left and right images. To enforce a uniform distribution of features in the image, a region-based scheme is used [7]; the image is divided into 6×8 regions, and the 30 features in each region with the largest Harris 'cornerness' score after non-maxima suppression (3×3 region) are retained. A quadratic interpolation of the cornerness score is used to achieve sub-pixel accuracy. The initial set of feature correspondences is obtained by thresholding the cosine similarity between the SIFT descriptors [14] assigned to each feature. Since the stereo extrinsic pose S has been estimated, a guided matching along epipolar lines is used, whereby a feature in the left image can only be matched to a feature in the right if the v pixel coordinates satisfy $|v_l - v_r| < 5$ pixels. To refine the accuracy of the right image feature coordinates, ZNCC is used within a small 11×11 region surrounding the Harris feature position in the right image. If there is a local minima in the ZNCC score, the feature is retained, and the sub-pixel position is calculated using a quadratic interpolation of the ZNCC score.

Many of the features detected in the left image are not matched during the first step described. For all the unmatched features, their *estimated* position $\mathbf{u}_r$ in the right image is obtained by finding the difference $\delta u = u_l - u_r$ of the nearest 5 matched features in the left image, and setting $\mathbf{u}_r = (u_l - \delta u, v_l)^T$. ZNCC is then used to refine this position within an 11×11 window surrounding the estimated position. Again, a correspondence is only found if there is a local minima in the ZNCC score, and a quadratic interpolation of this score is used to achieve sub-pixel accuracy. This step significantly improves the percentage of feature correspondences and exploits the smooth structure of the pipe surface.

The final step is outlier rejection. We make no assumptions regarding the scene structure, and reject outliers using a robust cost function. This cost function is the Median Absolute Deviation (MAD) of the errors $e_v = v_l - v_r$:

$$\text{MAD} = \text{median}\left(|e_{v_i} - \text{median}(e_v)|\right). \tag{9}$$

A correspondence $\mathbf{u}_{l_i} \leftrightarrow \mathbf{u}_{r_i}$ is retained only if $e_{v_i} < \gamma\,\text{MAD}$. A value of $\gamma = 4.0$ is used.

C. Stereo Triangulation

Given the set of the stereo correspondences $\mathbf{u}_l \leftrightarrow \mathbf{u}_r$, the triangulated scene point coordinates $\tilde{\mathbf{X}}_l$ in the left rectified frame are

$$\tilde{\mathbf{X}}_l = \frac{b}{d} \begin{pmatrix} u_l - u_{0_l} \\ \frac{1}{2}(v_l + v_r - 2v_{0_l}) \\ f \end{pmatrix}, \tag{10}$$

where $b = || - R^T \mathbf{t}||$ is the stereo baseline, and

$$d = (u_l - u_{0_l}) - (u_r - u_{0_r}) \tag{11}$$

is the disparity.

IV. VISUAL ODOMETRY

A. Temporal Correspondences

Given two pairs of stereo rectified images captured at different times, the temporal correspondences between the pairs are found as follows:

1) Find the stereo correspondences $\mathbf{u}_l \leftrightarrow \mathbf{u}_r$ and the scene point coordinates $\tilde{\mathbf{X}}_l$ in the first stereo pair.
2) Find the stereo correspondences $\mathbf{u}'_l \leftrightarrow \mathbf{u}'_r$ and scene point coordinates $\tilde{\mathbf{X}}'_l$ in the second stereo pair.
3) Find corresponding features in the left images by thresholding the ambiguity of their SIFT descriptors [14].
4) Use RANSAC and the Efficient Perspective-n-Points (EPnP) algorithm [15] to remove outliers and obtain an initial estimate of the change in pose $Q(\delta R, \delta \mathbf{t})$ between the left rectified cameras (see section IV-B).
5) For all features in the left image that were not matched, use their scene coordinates $\tilde{\mathbf{X}}_l$, and the estimated change in pose $Q = [\delta R | \delta \mathbf{t}]$, to find their estimated coordinate $\mathbf{u}'_l$ in the second left rectified image:
$$\mathbf{u}'_l = K_l \left(\delta R\, \tilde{\mathbf{X}}_l + \delta \mathbf{t}\right). \tag{12}$$
6) Use ZNCC to find refine the position $\mathbf{u}'_l$ within a 5×5 window surrounding the estimate position[3].
7) Estimate the coordinate $\mathbf{u}'_r$ in the second right rectified image, and use ZNCC to refine the position[3]. Using (10) and (11), the estimated coordinate is
$$\mathbf{u}'_r = \begin{bmatrix} u_l - u_{0_l} + u_{0_r} - \frac{b\,f}{\tilde{Z}_l} \\ v'_l \end{bmatrix} \tag{13}$$
8) If a new feature observation in the second pair was found using steps 5–7, assign this feature the same SIFT descriptor (i.e. do not recompute the SIFT descriptor using the second left image).

Rather than keep correspondences between every adjacent pair of stereo images in our datasets, we select only key-frames based on the method in [16]. The key-frames selected are the stereo image pairs separated by the largest median sparse optical flow magnitude below 50 pixels. This median is evaluated at step 4 to avoid unnecessary computations.

Step 2 attempts to find a fixed number of stereo pair correspondences using the region-based Harris detector. However, steps 5–8 can add additional features tracked from the previous frame. Of this final set of all features, some are discarded to try and maintain a near constant number — this constant number is (6×8) regions $\times\, 30 = 1440$. All of the features tracked from previous frames are retained since we want to maximize the number of key-frames we observe a given scene point. A subset of the stereo correspondences found in step 2 are removed. A strategic selection is used,

[3]The position is only found if there is a local minima in the ZNCC score. If the position is found, a quadratic interpolation of the score is used to further improve accuracy.

whereby the pose estimate Q is used to identify and remove those that we expect to leave the camera field of view first. These are the features nearest to the focus of contraction.

The algorithm described enables features to be *tracked* across many key-frames — a global index is assigned to each scene point. To illustrate, figure 2 shows the rectified stereo images in one of our datasets taken inside a pipe. The images move from left to right as the robot moves forward through the pipe. Figure 3 shows the probability distribution of the number of pixels each feature was tracked (rectified images are 1199×768 pixels in size). The results shown are for approximately 500 key-frames over a distance traveled of nearly 4 meters.

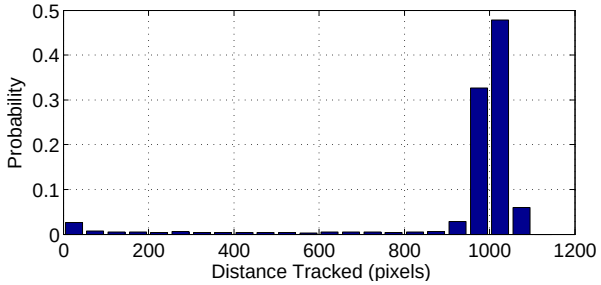


Fig. 3: Probability density function of the absolute distance each feature is tracked in the right stereo rectified camera. The width of the bins is 50 pixels.

By tracking points across many frames, an observation of its coordinate in each of the images is obtained. This is useful for sparse bundle adjustment [12] — see section IV-C. After each 25 key-frames, a sparse bundle adjustment of the previous 50 frames is implemented, which optimizes all left camera poses and scene point coordinates. Finally, for the pipe datasets described in section V, the final processing step is a sparse bundle adjustment of all camera poses and scene point coordinates.

B. Initial change in pose estimate between frames

The change in pose Q between the left rectified key-frames is obtained using the Efficient Perspective-n-Points (EPnP) algorithm [15]. It solves the change in pose Q using a set of scene point coordinates, $\tilde{\mathbf{X}}_l$ in the first left verged frame, and their corresponding homogeneous coordinates $\mathbf{u}'_l$ in the second left rectified frame, and are related by (12). EPnP is implemented within a RANSAC framework [17], using subsets of 25 randomly sampled correspondences.

Ideally, both the left and right observations $\mathbf{u}'_l$ and $\mathbf{u}'_r$ should be used to estimate the change in pose Q. Therefore, a non-linear refinement of the change in pose Q using both the left and right observations in the second frame is used. Using (12) and (13), the estimated coordinates $\hat{\mathbf{u}}'_l$ and $\hat{\mathbf{u}}'_r$ in the second stereo pair can be obtained for a given scene point $\tilde{\mathbf{X}}_l$ and estimate of Q. We seek the non-linear estimate of $Q = [\delta R|\delta \mathbf{t}]$ which minimizes the error

$$\epsilon_Q = \sum_n D(\hat{\mathbf{u}}'_l - \mathbf{u}'_l)^2 + D(\hat{\mathbf{u}}'_r - \mathbf{u}'_r)^2, \qquad (14)$$

where the summation is taken over all n correspondences, and where D is the geometric distance between homogeneous coordinates.

C. Sparse Bundle Adjustment

As mentioned, Sparse Bundle Adjustment (SBA) [12] is used in an attempt to find an optimal estimate of all the left rectified camera poses and scene point coordinates $\mathbf{X}$ defined in a global coordinate frame.

The reprojected coordinate of a scene point $\mathbf{X}_g$ in the left and right rectified images of camera k can be found. Denote these coordinates $\hat{\mathbf{u}}_{l_{kg}}$ and $\hat{\mathbf{u}}_{r_{kg}}$. Since we also have their observed coordinates $\mathbf{u}_{l_{kg}}$ and $\mathbf{u}_{r_{kg}}$, the reprojection errors can be measured. If the error uncertainty of the observed positions is isotropic Gaussian, the maximum likelihood estimate of the camera poses and scene coordinates are the ones which minimize the sum of squared reprojection errors

$$\sum_k \sum_g D(\hat{\mathbf{u}}_{l_{kg}} - \mathbf{u}_{l_{kg}})^2 + D(\hat{\mathbf{u}}_{r_{kg}} - \mathbf{u}_{r_{kg}})^2, \qquad (15)$$

where the inner summation is taken over all scene points g observed in camera k. SBA is the process of minimizing this sum of squared reprojection errors. Since the process is non-linear, we use Levenberg-Marquardt, and cannot guarantee a globally optimal solution.

V. EXPERIMENTS, RESULTS & DISCUSSION

A. Carbon Steel Pipe

The stereo system has been developed for mapping carbon steel pipes used in the LNG industry. We obtained two datasets, run 1 and run 2, each including approximately 4000 stereo images pairs captured by the stereo camera as it traversed through a 4 meter long, 16 inch diameter carbon steel pipe. Figure 1 shows the positioning of the stereo camera within the pipe. For both run 1 and run 2, the camera first moved forwards down the length of the pipe, and then backwards to the same starting position. There were 971 key-frames for run 1, and 956 key-frames for run 2.

The results for both run 1 and run 2 are shown in figure 4. The blue lines show the path of the camera when moving forward, the red lines the path when moving in reverse, and the green dots the reconstructed 3D scene points. The black crosses are the reconstructed scene points associated with manually augmented marks on the pipe. These marks are located at the start and end of the pipe, on the uppermost surface, and whose image coordinates in the images were manually selected.

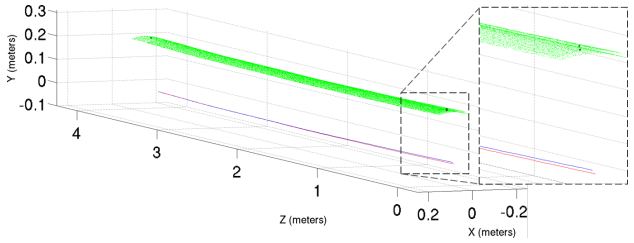
To make a quantitative assessment of accuracy, the reconstructed coordinates of the manually augmented marks in the pipe were used. At the start position, the reconstructed coordinate of the first mark is X_a. When the camera reaches the end of the pipe, the reconstructed coordinate of the second mark is X_b. After moving backwards to the start of the pipe, the second estimate of the first mark is X_c. Note that all marks are defined in the global coordinate frame (the first left camera is at the origin of this frame). The distances between these points are

$$d_{fwd} = ||X_a - X_b|| \qquad (16)$$

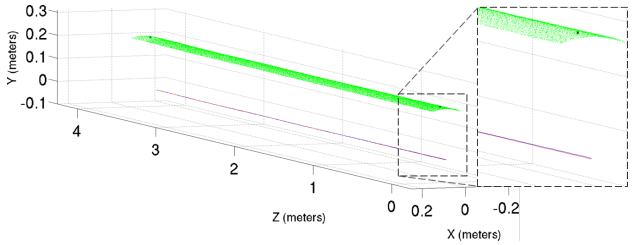
$$d_{rev} = ||X_b - X_c|| \qquad (17)$$

$$d_{end} = ||X_a - X_c||. \qquad (18)$$

The ground truth distance $d_{fwd} = d_{rev}$ is the precisely measured distance between the marks, and the ground truth



(a) Run 1.



(b) Run 2.

Fig. 4: Visual odometry results for the pipe datasets. The blue and red lines are, respectively, the camera path moving forward and in reverse. The green dots are the reconstructed scene points, and the black marks are the reconstructed coordinates of manually augmented marks on the pipe which are used as ground truth. All results have been rotated into a coordinate system similar to that used in [7].

distance d_{end} is zero. Table I compares the ground truth distance to those obtained using the visual odometry estimates. The absolute percentage errors reported for d_{end} are calculated as the error d_{end} divided by the total absolute distance traveled, $d_{fwd} + d_{rev}$. The results show that the absolute percentage errors are all below 0.1%, and are relatively repeatable for both runs.

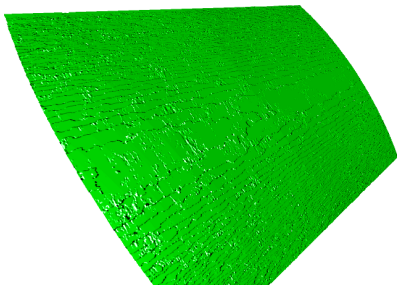
TABLE I: Results for the two pipe datasets. All distances have units of millimeters.

Dataset	Distance	Ground Truth	Visual Odometry	Abs. Error
Run 1	d_{fwd}	3689.83	3690.035	0.205 (0.0056%)
	d_{rev}	3689.83	3690.096	0.266 (0.0072%)
	d_{end}	0.000	6.566	6.566 (0.0890%)
Run 2	d_{fwd}	3689.83	3686.319	3.511 (0.0952%)
	d_{rev}	3689.83	3686.341	3.489 (0.0946%)
	d_{end}	0.00	1.717	1.717 (0.0233%)

As outlined in the introduction, our goal is to produce appearance maps of the internal surface of LNG pipes. However, we only achieve a very sparse scene reconstruction using the visual odometry algorithm — there are approximately 50,000 reconstructed scene points for each pipe dataset. The number of points could be increased by detecting and tracking more image features. This significantly increases computational cost, especially when implementing sparse bundle adjustment. As an alternative, we use our initial camera pose estimates, and a dense stereo reconstruction of all key-frame pairs, to produce a dense reconstruction of the internal pipe surface.

Figure 5 illustrates the appearance map generated for a small segment of the forward run of the first pipe dataset. The appearance map contains a dense 3D point cloud, and a color associated with each of these points. At present we use a block matching dense stereo algorithm which searches for dense stereo correspondences, strictly along epipolar lines, using sum of absolute differences. We have found this method to provide results comparable to more sophisticated algorithms such as graph cuts, but is much faster.

(a) Colorized point cloud.



(b) A closeup view of the dense 3D structure.

Fig. 5: A small segment of the appearance map produced for the forward trajectory of the first pipe dataset. The appearance map includes the 3D structure of the internal pipe surface, and a color associated with each scene point. The camera poses were found using the visual odometry algorithm described in sections III and IV.

B. Road

The target application of our system is the appearance mapping of the large pipe networks found in LNG processing facilities. As we currently only have access to a 4 meter length of pipe, we have collected a supplementary dataset to test the long-range accuracy of our system.

Referring to figure 6, the stereo camera was mounted downward facing on the side of a mobile robot. Over 10,000 stereo image pairs were logged as the robot moved in excess of 30 meters in a near straight line. Some example rectified stereo image pairs are shown in figure 7. The rectified pair in the bottom row of this figure show the brick pavers viewed when the robot traveled over a large speed bump towards the end of the dataset. The visual odometry results for a small segment of dataset are illustrated in figure 8. The change in elevation (z axis) occurred when the robot moved over the large speed bump. A full sparse bundle adjustment of all camera poses and scene points was not implemented due to the size of the dataset.

Ground truth measurements of differential distance traveled were measured using a laser distance sensor mounted on a tripod – the reported sensor accuracy is ± 1.5 millimeters. These measurements are compared to the visual odometry estimates in table II. The visual odometry estimates are the Euclidean distance between camera centers associated with the key-frames when the sensor readings were taken.

Fig. 6: The stereo camera mounted on a mobile robot (left), and the length of road used for dataset collection (right).

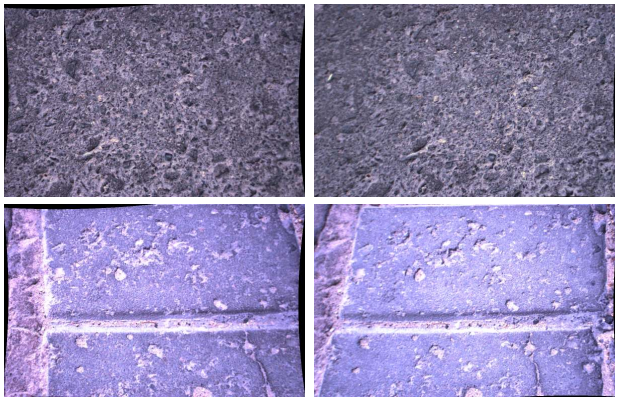


Fig. 7: Two sample stereo rectified images from the road dataset. The top row shows the road surface, and the bottom row shows the brick pavers on the speed bump.

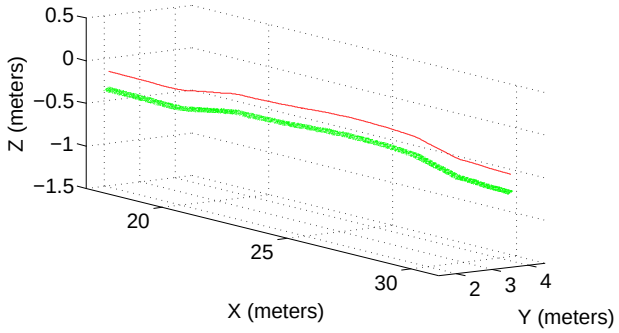


Fig. 8: The visual odometry results for a small section of the road dataset. The red line is the path of the left camera center, and the green dots are the sparse reconstructed scene points. The change in elevation (z axis) occurred when the robot moved over a large speed bump.

TABLE II: Results for the road dataset using the optimized calibration. All distances have units of millimeters.

Measurement	Ground Truth	Visual Odometry	Abs. Error	
d_1	10588.682	10551.729	36.953	(0.349%)
d_2	18317.680	18226.751	90.929	(0.496%)
d_3	30483.033	30226.617	256.416	(0.841%)

Although the percentage errors for distance traveled are larger than those for the pipe datasets, they are still less than 1%. The road dataset presented additional challenges which were not encountered in the pipe datasets, and this may explain the increased percentage errors. They include greater lighting variations and specularites resulting from a non-polarized natural light source, and greater depth discontinuities in the scene resulting in occlusions in the images.

VI. CONCLUSIONS AND FUTURE WORKS

In this work we presented a stereo-based visual odometry algorithm for pipe inspection. The algorithm was evaluated and tested on real datasets inside a pipe, and further validated on a long outdoor run under a very closed range. Our results show that stereo vision can be a very suitable sensor for the task of pipe inspection. The ability to acquire detailed 3D maps of the interior surface of the pipe is of high value in such inspection tasks. In the future, we plan to further validate the algorithm in complex pipe networks, with varying pipe diameters and sharp turns. Furthermore, we would like to investigate more accurate and efficient methods to represent the internal surface of the pipe, with possibly making use of some known "rough" topology of the surface.

VII. ACKNOWLEDGMENTS

The authors would like to thank Samitha Ekanayake and Mohamed Mustafa for their support in building and maintaining the data collection vehicle used for the results described in this paper.

REFERENCES

[1] G. Sibley, C. Mei, I. Reid, and P. Newman, "Adaptive relative bundle adjustment," in *Robotics Science and Systems Conference*, 2009.
[2] D. Nistér, O. Naroditsky, and J. Bergend, "Visual odometry for ground vehicle applications," *JFR*, vol. 23, no. 1, pp. 3–20, January 2006.
[3] M. Maimone, Y. Cheng, and L. Matthies, "Two years of visual odometry on the mars exploration rovers," *JFR*, vol. 24, no. 3, pp. 169–186, March 2007.
[4] P. Corke, C. Detweiler, M. Dunbabin, M. Hamilton, D. Rus, and I. Vasilescu, "Experiments with underwater robot localization and tracking," in *ICRA*, 2007.
[5] M. Agrawal and K. Konolige, "Rough terrain visual odometry," in *International Conference on Advanced Robotics (ICAR)*, August 2007.
[6] H. Schempf, E. Mutschler, A. Gavaert, G. Skoptsov, and W. Crowley, "Visual and nondestructive evaluation inspection of live gas mains using the Explorer family of pipe robots," *Journal of Field Robotics*, vol. 27, no. 3, pp. 217–249, 2010.
[7] P. Hansen, H. Alismail, P. Rander, and B. Browning, "Monocular visual odometry for robot localization in LNG pipes," in *International Conference on Robotics and Automation*, 2011.
[8] A. Schmidt, M. Kraft, and A. Kasiński, "An evaluation of image feature detectors and descriptors for robot navigation," in *ICCVG'10*, Berlin, Heidelberg, 2010, pp. 251–259.
[9] H. Alismail, B. Browning, and M. B. Dias, "Evaluating pose estimation methods for stereo visual odometry on robots," in *In proceedings of the 11th International Conference on Intelligent Autonomous Systems (IAS-11)*, 2010.
[10] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment - a modern synthesis," ser. ICCV '99. London, UK: Springer-Verlag, 2000, pp. 298–372.
[11] R. Hartley and F. Kahl, "Optimal algorithms in multiview geometry," in *Proceedings of the 8th Asian conference on Computer vision - Volume Part I*, ser. ACCV'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 13–34.
[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2003.
[13] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings Fourth Alvey Vision Conference*, 1988, pp. 147–151.
[14] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.
[15] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An Accurate O(n) Solution to the PnP Problem," *IJCV*, vol. 81, no. 2, pp. 155–166, 2008.
[16] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd, "Real time localization and 3D reconstruction," in *CVPR*, 2006.
[17] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comms. of the ACM*, pp. 381–395, 1981.