Bundle Adjustment in Large-Scale 3D Reconstructions based on Underwater Robotic Surveys

Chris Beall*, Frank Dellaert*, Ian Mahon†, Stefan B. Williams†
*College of Computing, Georgia Institute of Technology, Atlanta, GA 30332
†Australian Centre for Field Robotics, The University of Sydney, 2006 NSW
cbeall3@gatech.edu, dellaert@cc.gatech.edu, {i.mahon,s.williams}@cas.edu.au

***Abstract*—In this paper we present a technique to generate highly accurate reconstructions of underwater structures by employing bundle adjustment on visual features, rather than relying on a filtering approach using navigational sensor data alone. This system improves upon previous work where an extended information filter was used to estimate the vehicle trajectory. This filtering technique, while very efficient, suffers from the shortcoming that linearization errors are irreversibly incorporated into the vehicle trajectory estimate.**

This drawback is overcome by applying smoothing and mapping to the full problem. In contrast to the filtering approach, smoothing and mapping techniques solve for the entire vehicle trajectory and landmark positions at once by performing bundle adjustment on all the visual measurements taken at each frame. We formulate a large nonlinear least-squares problem where we minimize the pixel projection error of each of the landmark measurements.

The technique is demonstrated on a large-scale underwater dataset, and it is also shown that superior results are achieved with smoothing and mapping as compared to the filtering approach.

I. Introduction

Simultaneous Localization and Mapping (SLAM) has enjoyed widespread application in large scale reconstruction problems. City-scale reconstructions using images have been achieved in recent years [1], [2], [3], and a lot of progress has already been made in developing efficient algorithms to solve these large-scale reconstruction problems.

Motivated by continuing deterioration of underwater ecosystems, there has been a growing interest in adapting large-scale SLAM and smoothing and mapping (SAM) techniques to work in underwater environments. Translating standard SLAM techniques to the underwater reconstruction domain presents various difficulties, in part due to the challenging conditions and the limited types of sensors that can be used underwater. Furthermore, visibility is often limited, while poor illumination begins to play a significant role at greater depths.

Early techniques for underwater mapping were introduced in [4] which proposed a complete framework for sparse 3D mapping of the seafloor. Problems such as incremental position estimation, recursive global alignment of the final trajectory, and 3D reconstruction of the topographical map were tackled. In [5] the vehicle positions are estimated within a visual-based delayed state SLAM framework. The vehicle position estimation incorporates relative pose constraints from image correspondences. The result is an efficient filtering method that works on large trajectories. This approach was validated experimentally using monocular imagery collected of the RMS Titanic. Along the same line, the method proposed in [6] was used within scientific expedition surveys of submerged coral reefs. The result was a composite 3D mesh representation which allowed marine scientists to interact with the data gathered during the mission.

Pizarro et al. devoted close attention to low level image processing algorithms, from feature extraction to relative pose transformation between cameras [7]. The result was an enriched structure from motion (SfM) technique for sparse 3D reconstruction, where the steps were adapted to suit specific underwater conditions. The method was validated within controlled water tank conditions by comparing image based reconstruction to accurate laser scan measurements.

Dense reconstructions of submerged structures have been obtained in [8]. However, this was a very small-scale method where the stereo system was mounted on a controlled manipulator arm, so that the camera rotation and translation were known. Dense reconstruction has also been proposed as a second stage after sparse SfM [9]. Piecewise planar models were constructed by tracking landmarks in the scene in [10], with new landmarks being added as they became visible.

A complete system capable of large-scale underwater 3D reconstructions was presented in [11]. State estimates are recovered using an extended information filter (EIF), which takes advantage of the sparseness of the information matrix [5]. The efficiency of the EIF was further improved by updating the Cholesky factor directly, rather than recomputing the entire Cholesky factorization at each update step [12]. This is possible because the information matrix remains sparse, and only the last columns have to be recomputed as more vehicle states are added. Nonetheless, EKF and EIF filtering approaches still suffer from the permanent incorporation of linearization errors.

SAM was successfully applied to underwater stereo sequences in [13], and while the reconstruction had a large number of landmarks, the area covered by the 3D recon-

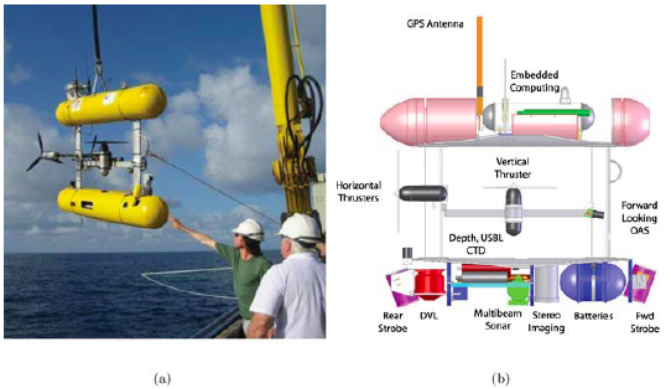


Figure 1: (a) The AUV Sirius being retrieved after a mission aboard the R/V Southern Surveyor. (b) AUV system diagram showing two thrusters, stereo camera pair and lights, as well as navigation sensors.

Sensor	Specification
Attitude + heading	Tilt ±0.5 deg, compass ±2 deg
Depth	Digiquartz pressure sensor, 0.01%
Velocity	RDI 1,200-kHz Navigator DVL ±2 mm/s
Altitude	RDI navigator four-beam average
USBL	TrackLink 1,500 HA (0.2-m range, 0.25 deg)
GPS receiver	uBlox TIM-4S
Camera	Prosilica 12-bit, 1360 × 1024 charge coupled device stereo pair

Table I: Summary of the Sirius AUV specifications

struction was not large-scale. In this paper we build on the work by [5], [11], [12], and show that more consistent results can be obtained by applying SAM to the full problem. We demonstrate our approach on a large-scale underwater dataset collected at the Scott Reef off the coast of Australia, and show that the solution is qualitatively superior to the filtering based solution that has been obtained previously. We perform full bundle adjustment over all cameras and a subset of landmarks detected in the images collected by the Autonomous Underwater Vehicle (AUV) . By increasing the reconstruction accuracy, we are able to provide oceanographers with better tools to study the ocean floor.

The remainder of this paper is organized as follows: Section II explains the data collection process, and in section III the initial filtering approach upon which we improve is outlined. This is followed by a discussion of the Smoothing and Mapping solution and our results in sections IVand V, respectively.

II. Data Collection

The data used for this work was collected using the Autonomous Underwater Vehicle *Sirius*, operated by the Australian Centre for Field Robotics, University of Sydney, Australia. A picture and system diagram of *Sirius* are shown in Fig. 1. The AUV is equipped with a suite of oceanography sensors, as well as a stereo-camera, multi-beam sonar, doppler-velocity log, and a GPS that can be used to geo-reference underwater surveys when the vehicle surfaces. The AUV is

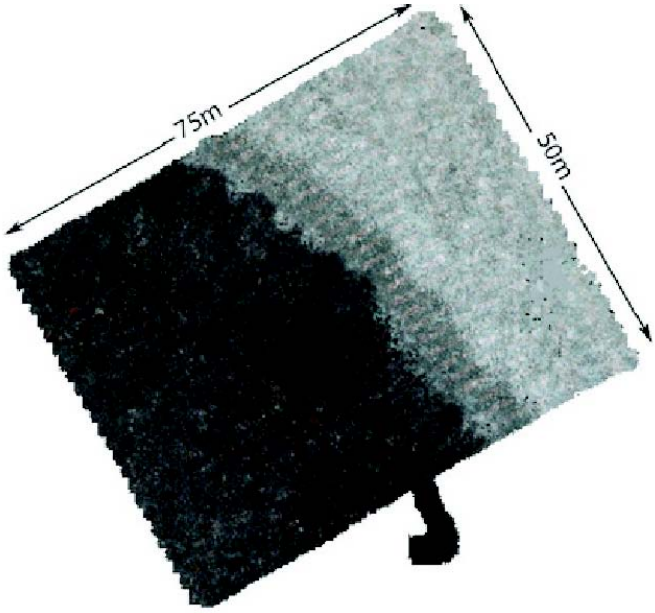


Figure 2: Top view of Scott Reef 3D reconstruction off western Australia covering an area of $50 \times 75m$ with 9831 stereo image pairs.

passively stable in pitch and roll, meaning that the stereo rig mounted in the bottom compartment is always imaging the sea floor. The stereo camera pair has a baseline of approximately 7 cm and a resolution of 1360×1024, with a field of view of 42×34 degrees. The sensors relevant to the work in this paper are listed in Table I.

Data was collected at the South Scott Reef, which is located 300 km northwest of Cape Leveque, Western Australia. The survey covered an area of $50m \times 75m$, and consists of 9831 stereo image pairs, taken at a frequency of about 1Hz, with the AUV programmed to maintain an altitude of 2m.

The AUV was programmed to follow a predetermined trajectory in the survey area. The rectangular area was densely covered with back-and-forth transects. The spatial overlap in between temporally consecutive images generally is around 50%. It is important to note that the AUV's navigation sensors (not the camera) were used to travel along the pre-planned path, and the sensors' measurements were used within the context of an EIF to reconstruct the trajectory that was actually followed.

Images collected by the stereo camera are preprocessed to compensate for lighting and wavelength-dependent color absorption of the water.

III. Trajectory Estimation

The vehicle state and trajectory are estimated using the data collected by the navigation sensors. This is done using an improved version of the viewpoint augmented navigation (VAN) framework [5], [12]. Reconstruction results obtained using the VAN framework for the dataset used in this paper are shown in Fig. 2.

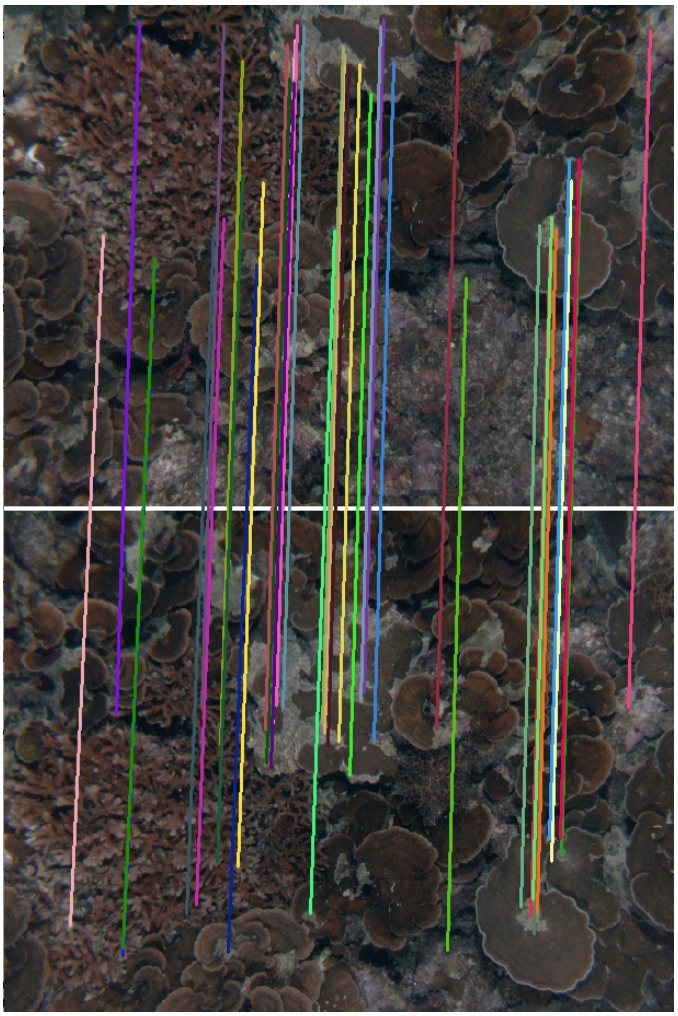


Figure 3: Features matched between two images taken at different AUV poses by the left camera.

The stereo imagery collected with the AUV is only used to introduce loop closures to the solution. This is a necessary step in large-scale reconstructions to ensure self-consistency of the resulting map. Creating loop closure hypotheses involves extracting features in both images of each stereo pair, finding valid stereo correspondences, and triangulating the corresponding 3D points using the midpoint method [14]. Features from images that are likely to have come from the same specific location in the survey area are then compared to establish correspondences. Once made, these are then used to robustly compute the rotation and translation between the two AUV poses at which the images were taken, and if successful, this information is added to the EIF. The loop closure search is carried out in multiple passes. In entirety, 48002 loop closure hypotheses were made, and 7951 of these were actually accepted and added to the filter. For more details on this work, please see [11]. An example of features matched between two images taken at different AUV poses by the left camera is shown in Fig. 3.

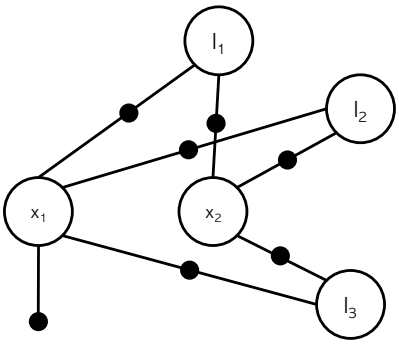


Figure 4: Factor Graph of two camera poses x and three landmarks l. The black nodes are the factors, which represent the landmark measurements taken by the respective cameras. The factor graph is used as the underlying computing paradigm for our SAM algorithm.

IV. Smoothing and Mapping

Smoothing and Mapping refers to the process of simultaneously estimating the full set of robot poses and landmark locations in the scene [15]. When this technique is applied to visual image features and camera poses, it is called Structure from Motion [16]. The structure of the scene is inferred from visual measurements taken by the camera from different vantage points. Features are extracted and matched between all the images [17], [18], and camera poses and landmark locations are optimized to minimize a cost function. The non-linear minimization process to estimate the best possible scene configuration is also referred to as bundle adjustment in the literature [19].

Bundle adjustment has been applied to create highly accurate, city-scale reconstructions from large photo-collections [2], [3]. Applying SAM to data collected by *Sirius* is an excellent way to create a highly accurate model of the survey area, making these models more useful to oceanographers. Whereas the VAN framework discussed in section IIIonly makes use of image correspondences for loop closure generation, our SAM approach differs substantially as we also use correspondences between temporally consecutive image pairs. In other words, in addition to the feature matches introduced by the 7951 loop closures, we also add all the feature correspondences that can be found between the 9831 camera poses along the trajectory.

Factor graphs offer a natural representation for the SAM problem. A factor graph is a bipartite graph containing only two types of nodes: state variables and factors. In our case, the unknown camera poses $X = \{x_i | \, i \in 1...M\}$ and landmarks $L = \{l_j | \, j \in 1...N\}$ correspond to the set of state variables. The factors in the graph represent the landmark measurements $Z = \{z_k | \, k \in 1...K\}$ that are made by the cameras. An example of a factor graph is shown in Fig. 4.

We minimize the non-linear cost function

$$\sum_{k=1}^{K} \| h_k(x_{i_k}, l_{j_k}) - z_k \|^2_{\Sigma_k} \tag{1}$$

where $h_k(.)$ is the measurement function of landmark l_j from camera x_i, and the notation $\|.\|^2_\Sigma$ represents the squared Mahalanobis distance with covariance Σ. We assume that we have normally distributed Gaussian measurement noise. Assuming a rectified image pair, we have individual landmark measurements (the features detected in the images) given by $z = (u_L, u_R, v)$, where u_L and u_R are the horizontal pixel coordinates of the left and right camera, respectively, and v is the vertical coordinate. For a properly calibrated stereo rig, v may be assumed to be the same for left and right images after rectification. The cameras used as part of the stereo rig are modeled using a pinhole camera model, and the standard projection equations apply. In practice one considers a linearized version of the problem, and the terms in equation 1 can be linearized as

$$\begin{aligned} & h_k(x_{i_k}, l_{j_k}) - z_k \\ \approx \quad & \left\{ h_k(x^0_{i_k}, l^0_{j_k}) + H^{i_k}_k \delta x_{i_k} + J^{j_k}_k \delta l_{j_k} \right\} - z_k \end{aligned} \tag{2}$$

where $H^{i_k}_k, J^{j_k}_k$ are the Jacobians of $h_k(.)$ with respect to x_{i_k}, l_{j_k} evaluated at $(x^0_{i_k}, l^0_{j_k})$.

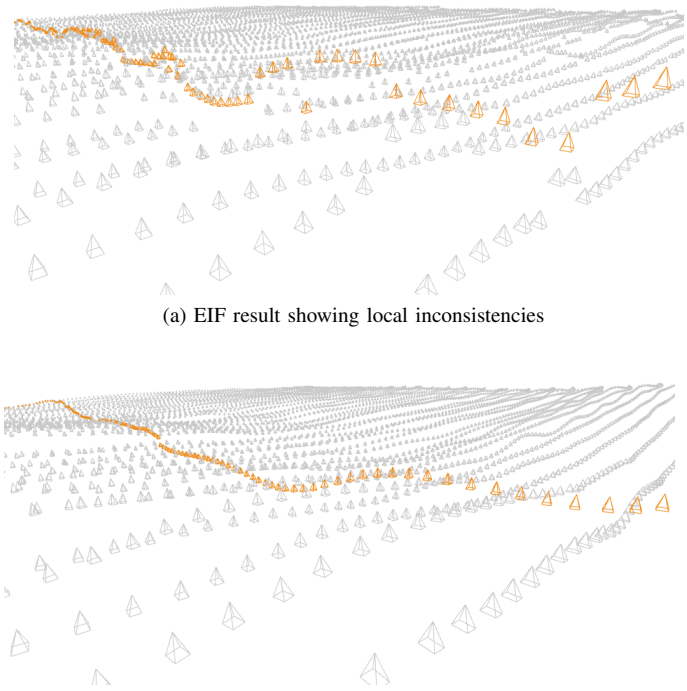
During optimization the ordering in which variables are elimiated is crucial for performance. This also applies to matrix factorization methods used in the EIF, and we use an AMD ordering [20]. For more details on the SAM optimization process, we refer the interested reader to [15].

V. SAM Results

For the experiments in this paper, image features are matched between consecutive stereo rig poses, as well as for loop closure observations. These feature matches are then used to construct the factor graph for our SAM problem. In all, the factor graph contains 9831 camera poses, 185261 landmarks, and 350988 factors. The camera poses computed by the EIF algorithm are used to initialize the camera poses for the SAM algorithm. Due to problematic lighting conditions, feature matching and relative pose estimation was unsuccessful in about 1% of the data, and odometry constraints from the EIF solution are used in their place. In other words, the factor graph contains many thousands of factors for landmark measurements, and a very small number of odometry factors. The odometry factors are added to ensure a contiguous trajectory without gaps.

Visual inspection of the results shows noticeable inconsistencies in the EIF camera trajectory, while the SAM solution is much smoother. Fig. 5 shows an area near the beginning of the trajectory where the differences are particularly significant. The point cloud and camera trajectory that is output from the SAM algorithm is shown in Fig. 6.

Ground truth is not available for this dataset. However, assuming a high quality camera calibration, the pixel projection errors $h_k(x_{i_k}, l_{j_k}) - z_k$ from eq. 1 provide a meaningful metric. This error allows for the direct comparison of map consistency resulting from the EIF and SAM solutions. The root mean square errors (RMSE) for EIF and SAM optimzation are 8.32 and 0.26, respectively. The root mean square error for the EIF solution is significantly larger than that of the SAM solutions, because relative visual odometry from consecutive poses was not used to estimate the trajectory in the EIF, but was only used to introduce loop closure constraints. Projection RMSE are shown in Fig. 7. The runtime for the SAM optimization is 287 seconds on an Intel Core 2 Duo 2.53Ghz MacBook Pro.



(a) EIF result showing local inconsistencies

(b) SAM result showing a very smooth trajectory

Figure 5: Partial view showing camera poses along the AUV trajectory. The first leg north is highlighted for clarity. The SAM result in (b) is notably smoother than the EIF result (a).

VI. Conclusions

In this paper we showed that smoothing and mapping leads to 3D maps that are more consistent than those resulting from employing a filtering approach. SAM provides a globally more optimal result due to optimizing over all cameras and landmarks, and does not suffer from the incorporation of linearization errors as do filters, since the current state estimate is always available to relinearize around. Full bundle adjustment on this dataset took just under 5 minutes. Future work includes applying more efficient SAM algorithms, such as Tectonic SAM [21], which hierarchically split up the problem for faster optimization, and consequently reduce the processing time required.

Acknowledgements

We thank Ian Mahon and Stefan B. Williams of the Australian Centre for Field Robotics for sharing some of their underwater datasets with us.

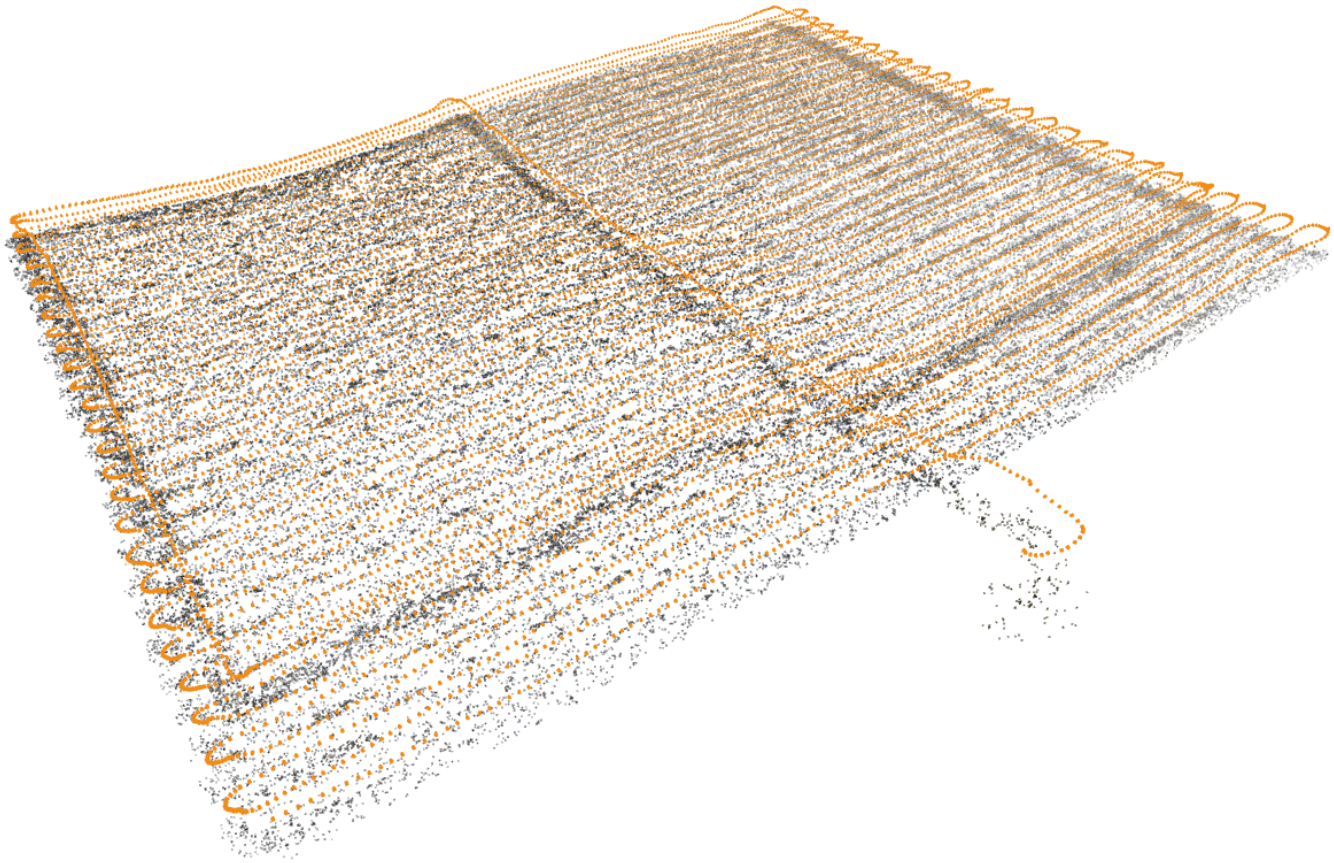


Figure 6: Camera trajectory and ocean floor point cloud with point colors taken from the corresponding images.

REFERENCES

[1] Y. Jeong, D. Nister, D. Steedly, R. Szeliski, and I. Kweon, "Pushing the envelope of modern methods for bundle adjustment," in *CVPR*, 2010.

[2] N. Snavely, S. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," in *SIGGRAPH*, 2006, pp. 835–846.

[3] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building rome in a day," 2009.

[4] S. Negahdaripour and H. Madjidi, "Stereovision imaging on submersible platforms for 3-d mapping of benthic habitats and sea-floor structures," *Oceanic Engineering, IEEE Journal of*, vol. 28, no. 4, pp. 625–650, Oct. 2003.

[5] R. Eustice, H. Singh, and J. Leonard, "Exactly sparse delayed-state filters for view-based slam," *IEEE Trans. Robotics*, vol. 22, no. 6, pp. 1100–1114, Dec 2006.

[6] S. Williams, O. Pizarro, M. Johnson-Roberson, I. Mahon, J. Webster, R. Beaman, and T. Bridge, "Auv-assisted surveying of relic reef sites," in *OCEANS 2008. Proceedings of IEEE*, Kobe, Japan, 2008, pp. 1–7.

[7] O. Pizarro, R. Eustice, and H. Singh, "Large area 3-d reconstructions from underwater optical surveys," *Oceanic Engineering, IEEE Journal of*, vol. 34, no. 2, pp. 150–169, April 2009.

[8] V. Brandou, A. Allais, M. Perrier, E. Malis, P. Rives, J. Sarrazin, and P. Sarradin, "3D reconstruction of natural underwater scenes using the stereovision system iris," in *OCEANS 2007. Proceedings of IEEE*, Aberdeen, Scotland, 2007, pp. 1–6.

[9] A. Sedlazeck, C. Albrechts, K. Koser, and R. Koch, "3D reconstruction based on underwater video from ROV KIEL 6000 considering underwater imaging conditions," in *OCEANS 2009. Proceedings of IEEE*, Bremen, Germany, May 2009.

[10] T. Nicosevici and R. Garcia, "Online robust 3D mapping using structure from motion cues," in *OCEANS 2008 - MTS/IEEE Kobe Techno-Ocean*, April 2008, pp. 1–7.

[11] M. Johnson-Roberson, O. Pizarro, S. Williams, and I. Mahon, "Generation and visualization of large-scale three-dimensional reconstructions from underwater robotic surveys," *Journal of Field Robotics*, vol. 27, no. 1, pp. 21–51, 2010.

[12] I. Mahon, S. Williams, O. Pizarro, and M. Johnson-Roberson, "Efficient view-based SLAM using visual loop closures," *IEEE Trans. Robotics*, vol. 24, no. 5, pp. 1002–1014, Oct 2008.

[13] C. Beall, B. Lawrence, V. Ila, and F. Dellaert, "3D Reconstruction of Underwater Structures," 2010.

[14] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[15] F. Dellaert, "Square Root SAM: Simultaneous location and mapping via square root information smoothing," in *Robotics: Science and Systems (RSS)*, 2005.

[16] S. Ullman, *The interpretation of visual motion*. The MIT press, Cambridge, MA, 1979.

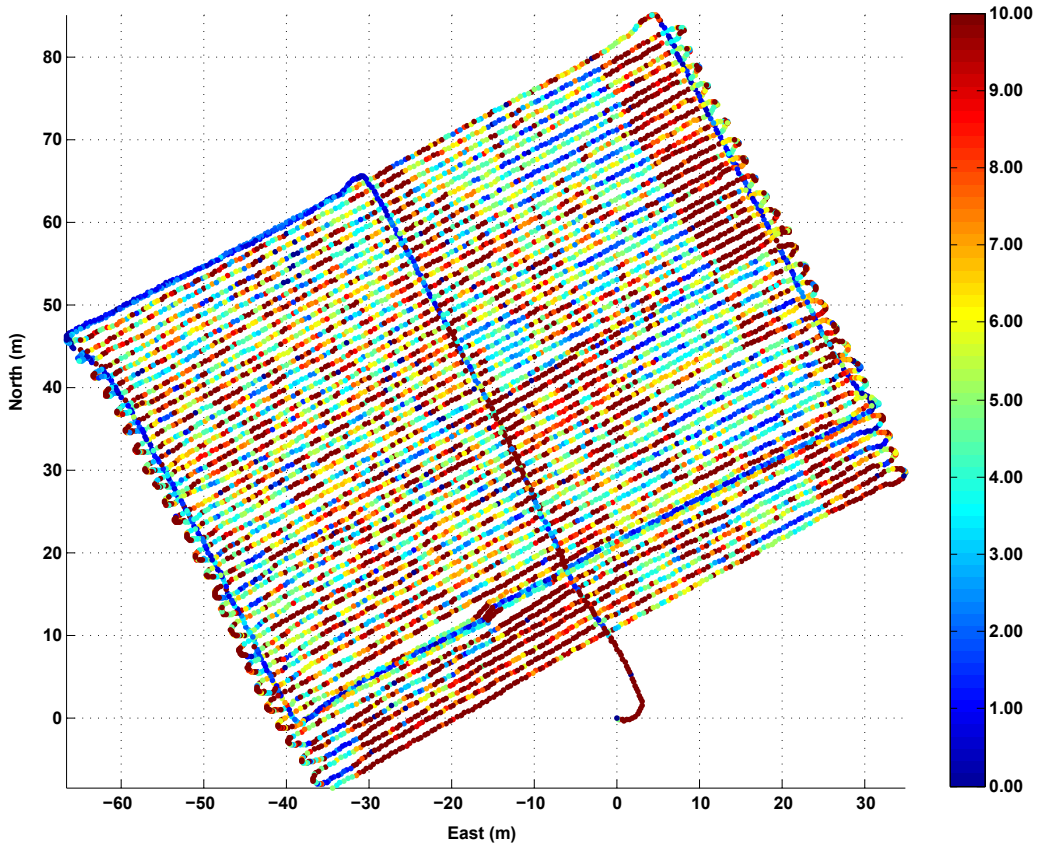
[17] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. J. of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[18] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: speeded up robust features," in *Eur. Conf. on Computer Vision (ECCV)*, 2006.

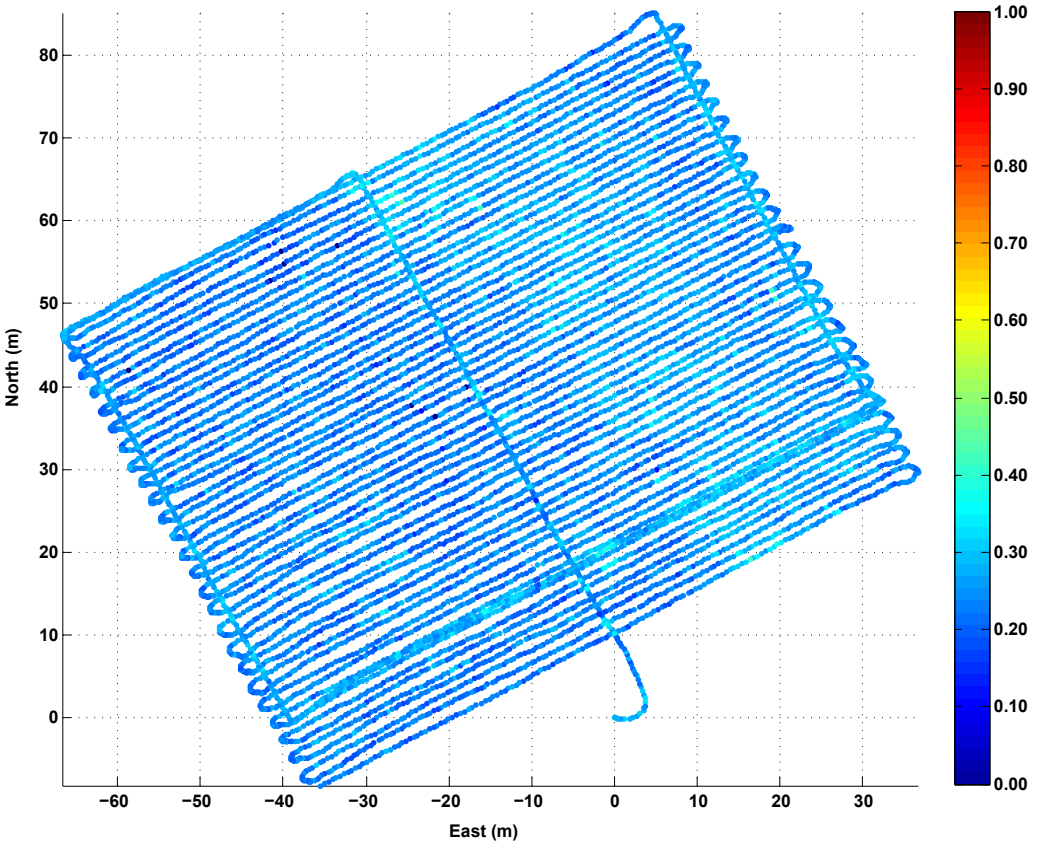
[19] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle adjustment – a modern synthesis," in *Vision Algorithms: Theory and Practice*, ser. LNCS, W. Triggs, A. Zisserman, and R. Szeliski, Eds. Springer Verlag, Sep 1999, pp. 298–375.

[20] P. Amestoy, T. Davis, and I. Duff, "An approximate minimum degree ordering algorithm," *SIAM Journal on Matrix Analysis and Applications*, vol. 17, no. 4, pp. 886–905, 1996.

[21] K. Ni and F. Dellaert, "Multi-level submap based slam using nested dissection," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2010. [Online]. Available: http://frank.dellaert.com/pubs/Ni10iros.pdf



(a) Extended Information Filter result



(b) Smoothing and Mapping result

Figure 7: Root mean square projection errors per camera, with camera locations plotted using the respective optimization results. The error is shown in color. Note the difference in error scales.