Bi-objective Bundle Adjustment: Towards Large-Scale Visual Sea-floor Mapping with a Minimal Sensor Suite

Michael Warren*, Peter Corke*, Oscar Pizarro†, Stefan Williams† and Ben Upcroft*
*Faculty of Science and Engineering,
Queensland University of Technology, Brisbane, Qld, Australia
†Australian Centre for Field Robotics (ACFR),
University of Sydney, Sydney, NSW, Australia

I. Introduction

Visual sea-floor mapping is a rapidly growing application for Autonomous Underwater Vehicles (AUVs). AUVs are well-suited to the task as they remove humans from a potentially dangerous environment, can reach depths human divers cannot, and are capable of long-term operation in adverse conditions. The output of sea-floor maps generated by AUVs has a number of applications in scientific monitoring: from classifying coral in high biological value sites [10] to surveying sea sponges to evaluate marine environment health [4].

In order to generate self consistent visual maps with accurately geo-referenced imagery over large swathes, accurate localization of the AUV is a strict requirement. While localization is relatively easy for surface vehicles due to GPS access, subsurface vehicles are either dependent on beacon based infrastructure (analogous to GPS localization) or Simultaneous Localization and Mapping (SLAM) using on-board sensors. In many subsurface environments of interest, beacon based infrastructure is unavailable or extremely sparse, meaning that SLAM is the only viable option for accurate localization.

In many AUV based substrate monitoring applications, an Information or Delayed state filtered SLAM solution [4, 1] is the standard method to integrate a large number of sensors and achieve an adequate pose solution. For visual mapping, using a set of downward facing cameras and active light strobes, imagery is taken at regular intervals and geo-referenced from the SLAM solution to generate 2D mosaics and 3D reconstructions of the sea floor [2].

Many land and airborne robots utilize Visual Odometry (VO) to estimate vehicle pose from sequential monocular or stereo frames [8, 9], in addition to some underwater scenarios [7]. Visual odometry has been demonstrated to perform well as a single estimator for estimating pose (used in combination with loop closure detection to constrain error growth), but also has the potential to be used in combination with other sensors in a filtered framework. By tracking visual features on the sea-floor it has distinct advantage as a passive pose estimator with a rich information output, and is capable of rivaling much more expensive inertial sensors in generating motion and orientation updates.

In contrast to other vision-based sensing scenarios, the imagery from the Sirius AUV [2] presents some difficulties when performing ‘traditional’ VO. In order to conserve energy used for strobing, imagery captured by Sirius is of very low frequency and low overlap ($\sim 30\%$), meaning that feature observations are fleeting and difficult to triangulate accurately. This adversely affects estimated pose using techniques suited to high overlap imagery. Such limited visual information manifests itself in rapid pose estimate degeneration using standard 6-DOF VO techniques. However, it is possible to take advantage of the constrained motion of the AUV (see Sec. II) and include additional readings from a minimal set of other sensors to constrain the VO and produce accurate pose over large trajectories in this specific scenario. Applications of this research may assist future development in two key ways: deployment of future vehicles at lower cost and increased operation time due to a reduced sensor suite, and capability improvement to existing vehicles by adding additional sensor information to the filtered solution.

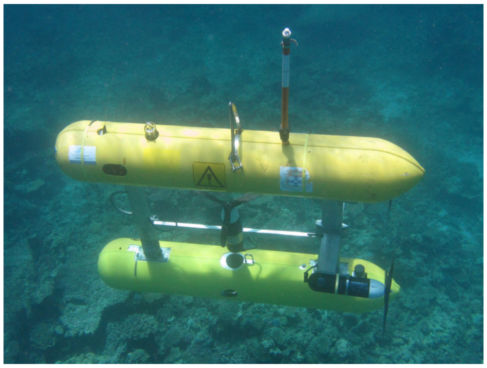

Fig. 1. The Sirius AUV on deployment in Scott Reef, WA, Australia

For this workshop, this paper presents active research into performing high accuracy sea-floor mapping using only stereo vision with extremely low overlap imagery and magnetometer input from the Sirius AUV. By taking advantage of the constrained motion of the AUV and integrating magnetometer data to correct yaw drift, accurate pose estimation is achieved using a minimal set of sensors. A brief introduction to the methodology, including a novel 2-point pose estimator and modified bundle adjustment are presented, and preliminary results on a 300m trajectory are shown. As a qualitative assessment of the trajectory estimation, 3D reconstructions of the observed scene are performed using the image data and pose estimates.

II. The Sirius AUV

The Sirius AUV (Fig. 1) is a modified version of the SEABed AUV, a mid-size underwater robotic vehicle primarily designed for large-scale sea-floor mapping for marine science and reef health monitoring. The AUV is equipped with a large set of oceanographic instruments including a magnetometer and a high-resolution (1360×1024) downward facing stereo camera pair ($\sim 7.5cm$ baseline) with strobes for imagery. The vehicle typically captures imagery at $1Hz$ from a height of $2m$ above the sea-floor while maintaining a forward velocity of approximately $0.5m/s$. Key to the development of theory presented here, this AUV design is passively stable in pitch and roll, meaning its motion is effectively constrained to only four degrees of freedom. Typically, roll and pitch of the vehicle rarely exceeds $1°$, particularly in the still water environments in which the AUV operates, actively avoiding impacts from strong currents and wave motion nearer the surface.

III. Methodological Approach

Here we present both the proposed visual odometry pipeline, and separately address 3D mesh generation and texturing from the final pose output.

The standard visual odometry pipeline follows three main repeating steps for each captured stereo pair:

- Structure triangulation
- Camera pose update
- Bundle adjustment

Our modified visual odometry algorithm follows the same basic pipeline, but with two major differences:

- A novel 2-point pose estimator that assumes a zero or negligible roll and pitch in the solution
- The inclusion of an additional objective in the bundle adjustment stage, assisting to minimise angular drift in the final pose estimate.

The proposed pipeline is shown in Figure 2. We emphasize here that the only input to the proposed pipeline is stereo images and temporally registered magnetometer data, no additional sensors are included.

A. Structure Triangulation

From the initial image pair at the start of the sequence, 3D structure is initialized via stereo triangulation. It is assumed that the homogeneous transform between the stereo camera pair is known and fixed, constraining scale of the scene and cameras to a known metric value. After any new pose update previously unseen structure is triangulated from the stereo pair. Additionally, new observations of known 3D structure are used to optimally re-triangulate the points via least squares optimization.

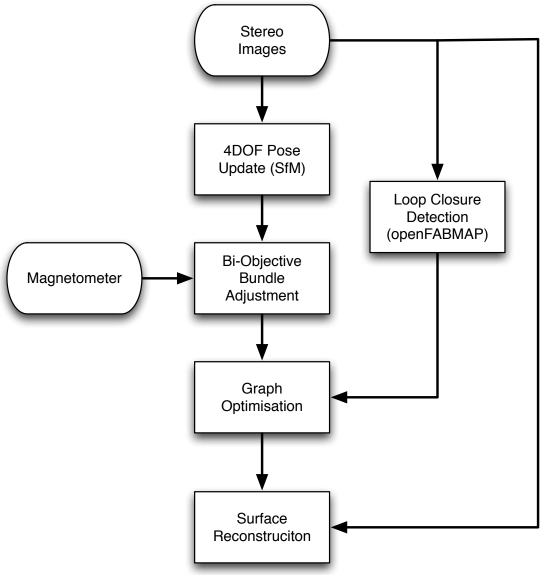

Fig. 2. The proposed Visual SLAM Pipeline

B. Camera Pose Update

Following the optimal triangulation of observed structure, the pose of the next camera pair (specifically the left camera, from which the right hand camera is derived) is extracted via SURF based feature matching to recover the 2D projections of the 3D points. This is achieved by solving a linear system of equations including the observed scene points $\mathbf{X} = \begin{bmatrix} X & Y & Z & 1 \end{bmatrix}^T$ and their projections $\mathbf{x} = \begin{bmatrix} u & v & 1 \end{bmatrix}^T$ into the image to find the elements of the matrix encoding the camera pose: $\mathbf{M} = [\mathbf{R}|\mathbf{t}]$ via the projection equation:

$$\mathbf{x} = \mathbf{P}\mathbf{X}, \mathbf{P} = \mathbf{K}\mathbf{M}$$

where $\mathbf{P}$ is termed the camera matrix and $\mathbf{K}$ is the camera intrinsics matrix.

In the standard 6-DOF case, a minimum of 3 points is required to extract the elements which define the pose: $x, y, z, \gamma, \phi, \theta$. However, if it is assumed that the roll γ and pitch ϕ movement in adjacent poses is negligible (*i.e.* zero) a 4-DOF pose estimate can be generated from the observation of only two points. This concept is similar to the absolute camera pose problem with known vertical direction given by an IMU [3]. Here, the rotation matrix $\mathbf{R}$ is simplified to the following case (we parameterize yaw, θ, in terms of variable q, where $\cos\theta = \frac{1-q^2}{1+q^2}$ and $\sin\theta = \frac{2q}{1+q^2}$):

$$\mathbf{R} = \begin{bmatrix} \frac{1-q^2}{1+q^2} & \frac{-2q}{1+q^2} & 0 \\ \frac{2q}{1+q^2} & \frac{1-q^2}{1+q^2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and

$$\mathbf{t} = \begin{bmatrix} t_x & t_y & t_z \end{bmatrix}^T$$

Hence, the required solution for $\mathbf{M}$, with some mathematical manipulation, becomes:

$$\begin{bmatrix} 0 & -1 & v \\ 1 & 0 & -u \\ -v & u & 0 \end{bmatrix} \mathbf{K} \begin{bmatrix} \frac{1-q^2}{1+q^2} & \frac{-2q}{1+q^2} & 0 & t_x \\ \frac{2q}{1+q^2} & \frac{1-q^2}{1+q^2} & 0 & t_y \\ 0 & 0 & 0 & t_z \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = 0$$

Analytically solving this linear system of equations given two scene points $\mathbf{X}_1, \mathbf{X}_2$ and their projections $\mathbf{x}_1, \mathbf{x}_2$ gives two closed form solutions for q, from which can be extracted four potential values of theta: $\theta_1, -\theta_1, \theta_2, -\theta_2$. By checking the residual of the projections two values are immediately rejected, and the residual of a third point is used to find the correct θ. It is then possible to substitute the value for q and recover the other three degrees of freedom. This 2-point pose estimator is placed in a MLESAC-based iterative estimator to achieve robustness in the presence of outliers.

C. Bi-Objective Bundle Adjustment

Following a pose update, a sliding window of the most recent camera positions $\hat{\mathbf{P}}$ and observed structure $\hat{\mathbf{X}}$ are optimized via bundle adjustment by minimizing the residual error in the projection of each estimated 3D point $\hat{\mathbf{X}}_j$ into each camera $\hat{\mathbf{P}}_i$: $\epsilon_{ij(c)} = \mathbf{x}_{ij} - \hat{\mathbf{x}}_{ij}$, where $\mathbf{x}_{ij}$ is the projection of scene point $\mathbf{X}_j$ into camera $\mathbf{P}_i$, and $\hat{\mathbf{x}}_{ij}$ is the projection of the corresponding estimate. The convergence of the algorithm is quantified by the reduction in the residual cost function over the estimated camera poses and scene structure:

$$\varepsilon_c^2 = \frac{1}{nm} \Sigma_i^n \Sigma_j^m \parallel \epsilon_{ij(c)} \parallel^2$$

However, even with bundle adjustment to optimise the pose and scene structure, drift is still present in the trajectory. This is most obvious in yaw, where global camera orientation can drift by up to $40°$ over $500m$. A solution to this problem involves the understanding that bundle adjustment is a special case of nonlinear least-squares solving. Using this, it is possible to introduce additional objectives and optimize not only based on the image re-projection error but additional constraints provided by other sensors [6].

By introducing a rotational cost term, ε_r, it is possible to optimize camera pose using both re-projection error and readings from an IMU or magnetometer by way of a rotational residual: $\epsilon_{i(r)} = \mathbf{r}_i - \hat{\mathbf{r}}_i$, where $\mathbf{r}_i$ is the orientation estimate provided by the additional sensor and $\hat{\mathbf{r}}_i$ is the corresponding estimate from visual odometry:

$$\varepsilon_r^2 = \frac{1}{n} \Sigma_i^n \parallel \epsilon_{i(r)} \parallel^2$$

Here, we parameterize the orientation in the form of a Rodriguez vector: $\mathbf{r} = \begin{bmatrix} \gamma & \phi & \theta \end{bmatrix}^{\mathbf{T}}$ and assume the difference $\epsilon_{i(r)}$ is small. In the case of our constrained motion estimate, and because of the parameterization of the rotation, it is possible to introduce a cost dependent only on one dimension, yaw, and use a magnetometer to provide the additional data. Since a magnetometer provides a global orientation it is possible to correct the orientation of the vehicle globally to maintain straight trajectories over large distances. The error in both the re-projection and orientation can be considered independent and Gaussian, hence weighted by a covariance, and the costs can be added to give a bi-objective cost:

$$E(\mathbf{x}, \mathbf{r}) = \frac{1}{(\sigma_x)^2 mn} \Sigma_i^n \Sigma_j^m \parallel \epsilon_{ij(c)} \parallel^2 + \frac{1}{(\sigma_r)^2 n_i} \Sigma_i^n \parallel \epsilon_{i(r)} \parallel^2$$

$$= \varepsilon_c^2 + \lambda^2 \varepsilon_k^2$$

where $\lambda = \frac{\sigma_x}{\sigma_r}$, indicating the ratio of the two covariances. Implementing this bi-objective bundle adjustment using magnetometer data to constrain the yaw motion will reduce angular drift and give a better pose estimate.

D. 3D Meshing and Texturing

Following a refined estimate of camera poses over the entire trajectory based on the 2-point pose estimator and bi-objective bundle adjustment, textured surface reconstructions are generated from the imagery for further analysis. For each stereo pair, dense feature matching with a number of consistency checks and smoothing operations [5] is performed on the imagery to gain dense depth maps.

Following a consistent depth map from each pair, a dense set of 3D oriented points is generated and a Poisson mesh fitted to the points. Each stereo mesh is arranged into a common reference frame denoted by the stereo poses, and a second Poisson surface fitted to 10 pairs in a windowed fashion. This process preserves local mesh quality to a high degree while smoothing any poorly reconstructed sections. Texture is added by projecting each vertex in the mesh back into the estimated camera poses and extracting the color of the associated image pixel. These surfaces can be stitched together and visualized in 3D to assist further research such as estimating individual coral growth and reef complexity.

IV. Preliminary Results

Initial results from both the modified visual odometry algorithm and surface reconstruction are presented. The VO pipeline was run on a dataset gathered by Sirius during a trip to Scott Reef, North of Western Australia during a Field Trip in 2011. Over 900 poses, the algorithm was evaluated using both the 4-DOF pose estimator with only standard bundle adjustment, and again with the 4-DOF pose estimator with a bi-objective bundle adjustment that includes yaw data from the on-board magnetometer. These are compared to the output of an Information filter using a number of alternative sensors as a 'ground truth'. The results are graphed in Figure. 3. It is immediately observable that over the 300m trajectory, the unconstrained solution (blue) drifts over time while the additional constraint from the yaw sensor prevents large scale drift from the correct heading over time. In Figures 4 and 5 examples of the 3D meshing and texturing pipeline based on a subsection of the poses in Figure 3 are presented.

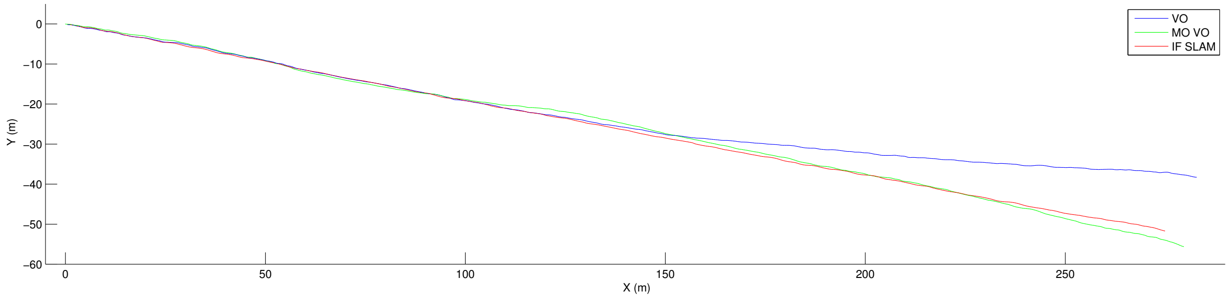

Fig. 3. A plot of 4-DOF VO with standard bundle adjustment (blue) and 4-DOF VO with bi-objective BA (green) over a 300 metre trajectory compared to an Information Filter based SLAM solution as ground truth (red)

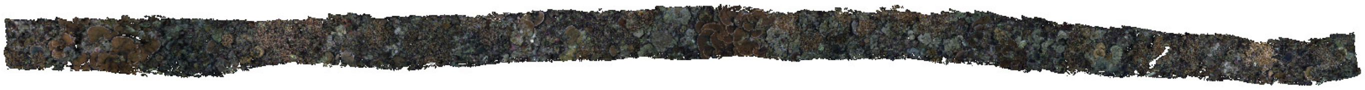

Fig. 4. A high resolution mesh generated by the reconstruction pipeline from 100 camera poses. (see Sec. IV)

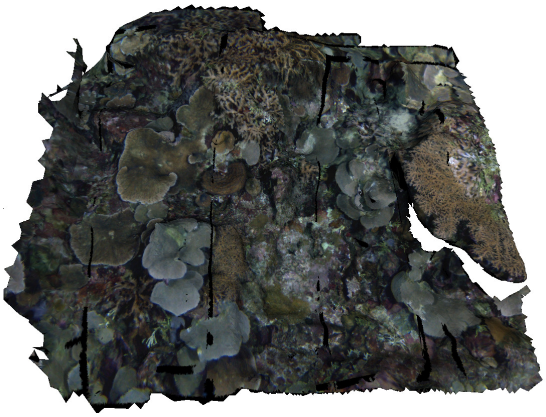

Fig. 5. A close up view of a sample of the reconstructed mesh indicating the quality of reconstruction (see Sec. IV)

V. Conclusion

A technique for performing accurate visual pose estimation using only low-overlap stereo images and yaw data has been presented. Quantitative results are shown from the constrained visual odometry technique over a 300m trajectory and reconstructions generated from this pose estimate show qualitative accuracy. This research will potentially enable future sub-sea mapping of high interest locations with increased accuracy on existing AUVs and enable methods of producing sea-floor maps with lower cost AUV hardware. Future work will involve demonstrating the technique on a full mission of the Sirius AUV, utilizing loop closure via openFABMAP and graph relaxation to constrain VO drift over the entire mission, and the development large scale environment reconstructions from the data.

References

[1] R M Eustice. Large-area visually augmented navigation for autonomous underwater vehicles. *Ph.D. dissertation, Massachusetts Inst. Technol. Woods Hole Oceanogr. Inst, Woods Hole, MA*, 2005.

[2] M Johnson-Roberson, Oscar Pizarro, S.B. Williams, and I. Mahon. Generation and Visualization of Large-Scale Three-Dimensional Reconstructions from Underwater Robotic Surveys. *Journal of Field Robotics*, 27 (1):21–51, 2010.

[3] Zuzana Kukelova, Martin Bujnak, and Tomas Pajdla. Closed-form solutions to the minimal absolute pose problems with known vertical direction.

[4] I. Mahon, S.B. Williams, O. Pizarro, and M. Johnson-Roberson. Efficient View-Based SLAM Using Visual Loop Closures. *IEEE Transactions on Robotics*, 24(5): 1002–1014, October 2008.

[5] David McKinnon, Ryan N Smith, and Ben Upcroft. A Semi-Local Method for Iterative Depth-Map Refinement. In *International Conference on Robotics and Automation (ICRA)*, 2012.

[6] J Michot and A Bartoli. Bi-objective bundle adjustment with application to multi-sensor slam. *3DPVT'10*, 2010.

[7] O. Pizarro, R. Eustice, and H. Singh. Large area 3d reconstructions from underwater surveys. *Oceans '04 MTS/IEEE Techno-Ocean '04 (IEEE Cat. No.04CH37600)*, 2:678–687.

[8] Michael Warren, D. McKinnon, H. He, and Ben Upcroft. Unaided stereo vision based pose estimation. In *Australasian Conference on Robotics and Automation*, Brisbane, 2010. Australian Robotics and Automation Association.

[9] Michael Warren, David Mckinnon, Hu He, Arren Glover, and Michael Shiel. Large Scale Monocular Vision-only Mapping from a Fixed-Wing sUAS. In *Field and Service Robotics*, pages 1–14, 2012.

[10] S Williams, Oscar Pizarro, and Michael Jakuba. AUV benthic habitat mapping in South Eastern Tasmania. *Field and Service Robotics*, pages 1–10, 2010.