2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Macau, China, November 4-8, 2019

Contour based Reconstruction of Underwater Structures Using Sonar, Visual, Inertial, and Depth Sensor

Sharmin Rahman[1], Alberto Quattrini Li[2], and Ioannis Rekleitis[1]

***Abstract*— This paper presents a systematic approach on real-time reconstruction of an underwater environment using Sonar, Visual, Inertial, and Depth data. In particular, low lighting conditions, or even complete absence of natural light inside caves, results in strong lighting variations, e.g., the cone of the artificial video light intersecting underwater structures, and the shadow contours. The proposed method utilizes the well defined edges between well lit areas and darkness to provide additional features, resulting into a denser 3D point cloud than the usual point clouds from a visual odometry system. Experimental results in an underwater cave at Ginnie Springs, FL, with a custom-made underwater sensor suite demonstrate the performance of our system. This will enable more robust navigation of autonomous underwater vehicles using the denser 3D point cloud to detect obstacles and achieve higher resolution reconstructions.**

I. Introduction

Underwater cave exploration is one of the most extreme adventures pursued by humans [1]. It is a dangerous activity with more than 600 fatalities, since the beginning of underwater cave exploration, that currently attracts many divers. Generating models of the connectivity between different underwater cave systems together with data on the depth, distribution, and size of the underwater chambers is extremely important for fresh water management [2], environmental protection, and resource utilization [3]. In addition, caves provide valuable historical evidence as they present an undisturbed time capsule [4], and information about geological processes [5].

Before venturing beyond the light zone with autonomous robots, it is crucial to ensure that robust localization and mapping abilities have been developed. Constructing a map of an underwater cave presents many challenges. First of all, vision underwater is plagued by limited visibility, color absorption, hazing, and lighting variations. Furthermore, the absence of natural light inside underwater caves presents additional challenges; however, the use of an artificial light can be used to infer the surrounding structures [6]. The most common underwater mapping sensor is based on sonar, which, when mounted on a moving platform, requires a secondary sensor to provide a common frame of reference for the range measurements collected over time. Furthermore, the majority of sonar sensors generate multiple returns in enclosed spaces making mapping caves extremely difficult.

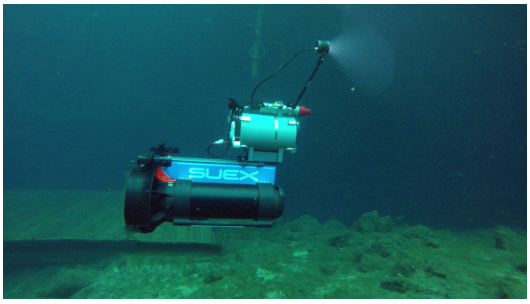


Fig. 1. Stereo, inertial, depth, and acoustic sensor and video-light mounted on a dual diver propulsion vehicle (DPV), at the Blue Grotto cavern.

In our earlier work, the cone of light perceived by a stereo camera was used to reconstruct offline the boundaries of a cave in Mexico [6]. No other sensor was available and a stereo-baseline of $0.03\,\mathrm{m}$ limited the accuracy of the reconstruction for objects further than a couple of meters. More recently, augmenting the visual-inertial state estimation package OKVIS [7], we fused visual and inertial data together with acoustic range measurements from a pencil beam sonar, which provide more reliable distance estimate of features. This allows a more robust and reliable state estimation [8], [9]. One of the limitations is the granularity of the resulting 3D point cloud: only few keypoints are typically tracked, resulting in very sparse 3D point cloud, which cannot be directly used, for example, by an Autonomous Underwater Vehicle (AUV) to navigate and avoid obstacles. Applying a direct-based method, such as LSD-SLAM [10], is not straightforward, given the sharp changes in illumination in the underwater scene. A fundamental difference with most vision based estimation approaches is that in a cave environment, the light source is constantly moving thus generating shadows that are also moving. Consequently the majority of the strong features cannot be used for estimating the pose of the camera.

In this paper, we propose a novel system that is able to track the state estimate and at the same time improve the 3-D reconstruction from visual-edge based information in the cave boundaries. In particular, the proposed approach for real-time reconstruction of the cave environment with medium density is based on an underwater visual odometry system that combines acoustic (sonar range), visual (stereo camera), inertial (linear accelerations and angular velocities), and depth data to estimate the trajectory of the employed sensor suite. The inspiration for a denser point cloud comes from the following observation: visual features on the boundaries

[1]S. Rahman and I. Rekleitis are with the Computer Science and Engineering Department, University of South Carolina, Columbia, SC, USA srahman@email.sc.edu, yiannisr@cse.sc.edu

[2]A. Quattrini Li is with the Department of Computer Science, Dartmouth College, Hanover, NH, USA alberto.quattrini.li@dartmouth.edu

The authors would like to thank the National Science Foundation for its support (NSF 1513203, 1637876). The authors would like to acknowledge the help of the Woodville Karst Plain Project (WKPP).

created by shadows, occlusion edges, and the boundaries of the artificial illumination (video light) – see Fig. 1 – are all located at the floor, ceiling, and walls of the cave. The point cloud resulting from such edges is then optimized in a local bundle adjustment process, and can be used for providing a denser reconstruction, enabling the deployment of AUVs like Aqua2 [11] with agile swimming gaits [12], navigating around obstacles without disturbing the sediment at the bottom. Experiments in caverns and caves validate the proposed approach.

The paper is structured as follows. In the next section, we present related work, specifically focusing on state estimation and 3D reconstruction. Section III describes the proposed method. Experimental results are presented in Section IV. Section V concludes the paper.

II. Related Work

Robotic exploration of underwater caves is at its infancy. One of the first attempts was to explore a Cenote, a vertical shaft filled with water [13], by the vehicle DEPTHX (DEep Phreatic THermal eXplorer) [14] designed by Stone Aerospace [15], equipped with LIDAR and sonar. More recently, Mallios *et al*. demonstrated the first results of an AUV performing limited penetration, inside a cave [16]. The main sensor used for SLAM is a horizontally mounted scanning sonar. A robotic fish was proposed for discovering underwater cave entrances based on vision performing visual servoing, with experiments restricted to a swimming pool [17]. More recently, Sunfish [18] – an underwater SLAM system using a multibeam sonar, an underwater dead-reckoning system based on a fiber-optic gyroscope (FOG) IMU, acoustic DVL, and pressure-depth sensors – has been developed for autonomous cave exploration. The design of the sensor suite we use is driven by portability requirements that divers have [19], not permitting the use of some sensors, such as multibeam sonar or DVL.

Corke *et al*. [20] compared acoustic and visual methods for underwater localization showing the viability of using visual methods underwater in some scenarios. In recent years, many vision-based state estimation algorithms – PTAM [21], ORB-SLAM [22], LSD-SLAM [10], DSO [23], COLMAP [24] – have been developed using monocular, stereo, or multi-camera systems mostly for indoor and outdoor environments. Vision is often combined with IMU for more accurate estimation of pose, for example, MSCKF [25], OKVIS [26], Visual-Inertial ORB-SLAM [27], and VINS-Mono [28]. Our comprehensive comparison of state-of-the-art open-source visual SLAM packages for underwater [29] shows improvement of performance for visual-inertial odometry (VIO) systems over pure visual odometry (VO) systems; at the same time, many challenges are still present, including track loss.

Structure-from-Motion (SfM) from unstructured collections of photographs to build the 3-D model of the scene has been addressed in different solutions, e.g., *Bundler* [30] and *VisualSFM* [31]. They provided an algorithmic analysis to improve the computational complexity and the performance accuracy. *COLMAP* [24] proposes an SfM algorithm to improve on the state-of-the-art incremental SfM methods for 3D reconstruction from non-ordered image collections. They provide scene graph augmentation, a next best view selection mechanism, and an efficient triangulation and Bundle Adjustment (BA) technique. COLMAP outperforms state-of-the-art SfM system on benchmark datasets with a large number of photos from the Internet with varying camera density and distributed over a large area. Multiview Stereo (MVS) is another well known method for reconstruction. Merrell *et al*. [32] presented a *viewpoint-based* approach to fuse multiple stereo depth maps for reconstructing 3-D shape from video. By decoupling the processing into two stages, they are able to run large-scale reconstructions in real-time using a GPU implementation for efficient computation. The computational power available on board of AUVs is very limited, making the deployment of bundle adjustment based methods not feasible.

Recently, direct methods (e.g., LSD-SLAM [10], DSO [23]) and the semi-direct method (SVO [33]) based SLAM systems show promising performance in 3-D reconstruction of large-scale maps in real time, as well as accurate pose estimation based on direct image alignment. However, theses methods are sensitive to *brightness consistency* an assumption which limits the baseline of the matches and in low visibility with small contrast environments like underwater, often results into tracking loss [29]. For good reconstruction, direct methods require perfect photometric calibration for modeling the gain and exposure. DSO [23] shows an improvement in performance providing a full photometric calibration that accounts for lens attenuation, gamma correction, and known exposure times. In purely monocular vision based direct SLAM, like DSO, the initialization is slow and requires minimal rotational changes.

In our previous work [6], we proposed an offline system based on a stereo camera, ORB-SLAM, and artificial light to have a good reconstruction of the cave. In our current work, we overcome some of the limitations of our previous work, including making the system operating in real-time, augmenting outlier-rejection scheme, and integrating it with a more robust visual inertial odometry system that includes also acoustic-range and depth measurements.

III. Technical Approach

The proposed approach augments [8], [9] to generate real-time a denser reconstruction of underwater structures exploiting the boundaries of the structure and the cone-of-light. This proposed system is depicted in Fig. 2. For completeness, we briefly introduce the system hardware and visual inertial method that includes acoustic and depth measurements – see [19], [26], [8], [9] for more details. Then, we describe the proposed 3D reconstruction based on contour matching and the local optimization of the point cloud.

A. System Overview

The sensor suite is composed of stereo camera, mechanical scanning profiling Sonar, IMU, pressure sensor, and an on-board computer. This custom-made sensor suite can be

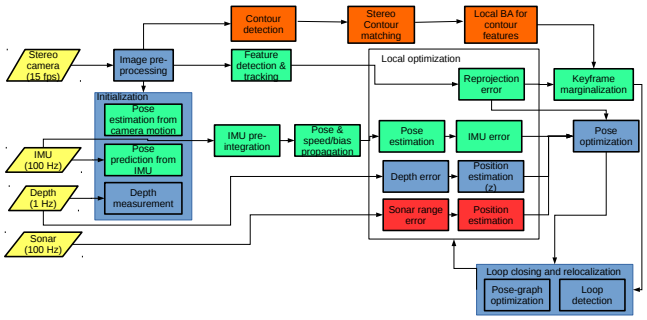


Fig. 2. Block diagram of the proposed system; in yellow the sensor input with frequency from the custom-made sensor suite, in green the components from OKVIS, in red and blue the contribution from our previous works [8] and [9], and in orange the new contributions in this paper.

deployed by divers as well as mounted on a single or dual Diver Propulsion Vehicle (DPV) [19]. The hardware was designed with cave mapping as the target application. As such, the sonar scanning plane is parallel to the image plane which provides data at a maximum of $6\,\mathrm{m}$ range, scanning in a plane over $360°$, with angular resolution of $0.9°$.

B. Notations and States

The reference frames associated to each sensor and the world are denoted as C for Camera, I for IMU, S for Sonar, D for Depth, and W for World. Let us denote ${}_X\mathbf{T}_Y = [{}_X\mathbf{R}_Y|{}_X\mathbf{p}_Y]$ the homogeneous transformation matrix between two arbitrary coordinate frames X and Y, where ${}_X\mathbf{R}_Y$ represents the rotation matrix with corresponding quaternion ${}_X\mathbf{q}_Y$ and ${}_X\mathbf{p}_Y$ denotes the position vector.

The state of the robot R is denoted as $\mathbf{x}_R$:

$$\mathbf{x}_R = [{}_W\mathbf{p}_I^T, {}_W\mathbf{q}_I^T, {}_W\mathbf{v}_I^T, {\mathbf{b}_g}^T, {\mathbf{b}_a}^T]^T \quad (1)$$

It contains the position ${}_W\mathbf{p}_I$, the quaternion ${}_W\mathbf{q}_I$, the linear velocity ${}_W\mathbf{v}_I$. All of them are in the IMU reference frame I with respect to the world reference frame W. In addition, the gyroscopes and accelerometers bias $\mathbf{b}_g$ and $\mathbf{b}_a$ are also estimated and stored in the state vector.

The corresponding error-state vector is defined in minimal coordinates, while the perturbation for the optimization problem defined next, takes place in the tangent space:

$$\delta\boldsymbol{\chi}_R = [\delta\mathbf{p}^T, \delta\mathbf{q}^T, \delta\mathbf{v}^T, \delta{\mathbf{b}_g}^T, \delta{\mathbf{b}_a}^T]^T \quad (2)$$

C. Tightly-coupled Non-Linear Optimization Problem

The cost function $J(\mathbf{x})$ for the tightly-coupled non-linear optimization includes the reprojection error $\mathbf{e}_r$, the IMU error $\mathbf{e}_s$, sonar error $\mathbf{e}_t$, and the depth error e_u:

$$J(\mathbf{x}) = \sum_{i=1}^{2}\sum_{k=1}^{K}\sum_{j\in\mathcal{J}(i,k)} {\mathbf{e}_r^{i,j,k}}^T \mathbf{P}_r^k \mathbf{e}_r^{i,j,k} + \sum_{k=1}^{K-1} {\mathbf{e}_s^k}^T \mathbf{P}_s^k \mathbf{e}_s^k + \sum_{k=1}^{K-1} {\mathbf{e}_t^k}^T \mathbf{P}_t^k \mathbf{e}_t^k + \sum_{k=1}^{K-1} {e_u^k}^T P_u^k e_u^k \quad (3)$$

with i denoting the camera index – $i = 1$ for left, $i = 2$ for right camera in a stereo camera – and landmark index j observed in the k^{th} camera frame. $\mathbf{P}_r^k$, $\mathbf{P}_s^k$, $\mathbf{P}_t^k$, and P_u^k denote the information matrix of visual landmarks, IMU, sonar range, and depth measurement for the k^{th} frame respectively.

The reprojection error $\mathbf{e}_r$ describes the difference between a keypoint measurement in camera coordinate frame C and the corresponding landmark projection according to the stereo projection model. The IMU error term $\mathbf{e}_s$ combines all accelerometer and gyroscope measurements by *IMU pre-integration* [34] between successive camera measurements and represents the *pose*, *speed and bias* error between the prediction based on previous and current states. Both reprojection error and IMU error term follow the formulation by Leutenegger *et al.* [26].

The sonar range error $\mathbf{e}_t$, introduced in our previous work [8], represents the difference between the 3D point that can be derived from the range measurement and a corresponding visual feature in 3D.

The depth error term e_u can be calculated as the difference between the rig position along the z direction and the water depth measurement provided by a pressure sensor. Depth values are extracted along the *gravity* direction which is aligned with the z of the world W – observable due to the tightly coupled IMU integration. This can correct the position of the robot along the z axis.

The *Ceres Solver* nonlinear optimization framework [35] optimizes $J(\mathbf{x})$ to estimate the state of the robot in Eq. (1).

D. Feature Selection and 3D Reconstruction from Stereo Contour Matching

To ensure that the VIO system and the 3D reconstruction can be run in real-time in parallel, we replaced the OKVIS feature detection method with the one described in [36], which provides a short list of the most prominent features based on the *corner response* function in the images. This reduces the computation in the *frontend* tracking and, as shown in the results, retains the same accuracy with less computational requirements.

Fig. 3. Image in a cave and the detected contours.

A real-time stereo contour matching algorithm is utilized followed by an outlier rejection mechanism to produce the point-cloud on the contour created by the moving light; see Fig. 5(c) for an example of all the edge-features detected. The approach of Weidner *et al.* [6] has been adapted for the contours from the intersection of the cone of light with the cave wall; see Fig. 3 for the extracted contours from an underwater cave. In particular, adaptive thresholding the images based on the light and dark areas ensures that the illuminated areas are clearly defined. In our current work, we also found that sampling from pixels which have rich gradients, e.g., edges, provide better and denser point-cloud reconstructions. As such, both types of edges – the ones marking the boundaries between the light and dark areas and the others from visible cave walls – are used to

reconstruct the 3-D map of the cave. The overview of the augmenting Stereo Contour Matching method in our tightly-coupled Sonar-Visual-Inertial-Depth optimization framework is as follows.

For every *frame* in the local optimization window, a noisy edge map is created from the edges described above. This is followed by a filtering process to discard short contours by calculating their corresponding bounding boxes and only keeping the largest third percentile. This method retains the highly defined continuous contours of the surroundings while eliminating spurious false edges, thus allowing to use the pixels on them as good features to be used in the reconstruction. In a stereo frame, for every image point on the contour of the left image a BRISK feature descriptor is calculated and matched against the right image searching along the epipolar line. Then a sub-pixel accurate localization of the matching disparity is performed. Another layer of filtering is done based on the grouping of the edge detector, i.e., keeping only the consecutive points belonging to the same contour in a stereo pair. These stereo contour matched features along with depth estimation is projected into 3-D and then projected back for checking the reprojection error consistency resulting into a point-cloud with very low reprojection error.

The reason behind choosing stereo matched contour features rather than tracking them using a semi-direct method or using a *contour tracking* [37] method is to avoid any spurious edge detection due to lighting variation in consecutive images, which could lead to erroneous estimation or even tracking failure. The performance of SVO [33], an open-source state-of-the-art semi-direct method, in underwater datasets [38], [29] validates the above statement. In addition, though indirect feature extractors and descriptors are invariant to photometric variations to some extent, using a large number of features for tracking and thus using them for reconstruction is unrealistic due to the computational complexity of maintaining them.

E. Local Bundle Adjustment (BA) for Contour Features

In the current optimization window, a local BA is performed for all newly detected stereo contour matched features and the keyframes they are observed in, to achieve an optimal reconstruction. A joint non-linear optimization is performed for refining k^{th} keyframe pose ${}_W\mathbf{T}_{C_i}{}^k$ and homogeneous *landmark* j in world coordinate W, ${}_W\boldsymbol{l}^j = [{l_x}^j, {l_y}^j, {l_z}^j, {l_w}^j]$ minimizing the cost function:

$$J(\mathbf{x}) \quad = \quad \sum_{j,k} \rho(\mathbf{e}^{j,k^T} \mathbf{P}^{j,k} \mathbf{e}^{j,k}) \tag{4}$$

Hereby $\mathbf{P}^{j,k}$ denotes the information matrix of associated landmark measurement, ρ is the Huber loss function to *downweigh* outliers. The reprojection error, $\mathbf{e}^{j,k}$ for landmark j with matched keypoint measurement $\mathbf{z}_{j,k}$ in image coordinate in the respective camera i is defined as:

$$\mathbf{e}^{j,k} \quad = \quad \mathbf{z}^{j,k} - \mathbf{h}_i({}_W\mathbf{T}_{C_i}{}^k, {}_W\boldsymbol{l}^j) \tag{5}$$

with camera projection model $\mathbf{h}_i$. We used Levenberg-Marquardt to solve the local BA problem.

IV. Experimental Result

The experimental data were collected using a custom made sensor suite [19] consisting of a stereo camera, an IMU, a depth sensor and a mechanical scanning Sonar, as described in Section III-A. More specifically, two USB-3 uEye cameras in a stereo configuration provide data at 15 Hz, an IMAGENEX 831L mechanical scanning Sonar sensor acquires a full 360° scan every four seconds; the Bluerobotics Bar30 pressure sensor provides depth data at 1 Hz; a MicroStrain 3DM-GX4-15 IMU generates inertial data at 100 Hz; and an Intel NUC running Linux and ROS consolidates all the data. A video light is attached to the unit to provide artificial illumination of the scene. The Sonar is mounted on top of the main unit which contains the remaining electronics. In Fig. 1 the unit can be seen deployed mounted on a dual Diver Propulsion Vehicle (DPV); please note, the system is neutrally buoyant and stable. The experiments were run on a computer with an Intel i7-7700 CPU @ 3.60GHz, 32 GB RAM, running Ubuntu 16.04 and ROS Kinetic and on an Intel NUC with the same configuration.

The data is from the ballroom at Ginnie Springs, FL, a cavern open to divers with no cave-diving training. It provides a safe locale to collect data in an underwater cave environment. From entering the cavern at a depth of seven meters, the sensor was taken down to fifteen meters, and then a closed loop trajectory was traversed three times.

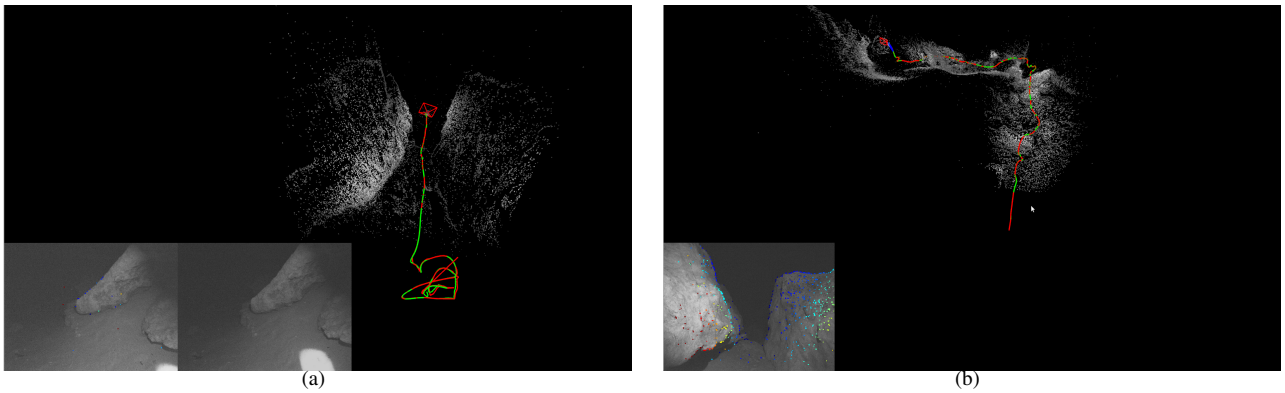
In the following, we present, first, preliminary experiments with DSO [23] showing the problem with photometric consistency. Second, as there is no ground truth available underwater, such as a motion capture system, we qualitatively validate our approach from the information collected by the divers during the data collection procedure.

A. Comparison with DSO

DSO is one of the best performing state-of-the-art direct VO method which uses a sparse set of high intensity gradient pixels. Josh *et al*. [29] show few cases where DSO generates very good 3-D reconstructions in challenging underwater environments. Fig. 4 shows the result of DSO in the underwater cave dataset in two different runs, Fig. 4(a) and Fig. 4(b). DSO did not track for the full length of cave; instead it was able to keep track just for a small segment due to the variation of the light and hence violating the *photometric consistency* assumption of a direct method. Also, the *initialization* method is critical as it requires mainly translational movement and a very small rotational change due to the fact that it is a pure monocular visual SLAM. We ran DSO using different starting points of the dataset to have a better initialization, the best one we got in Fig. 4(b) – eventually failed too due to the poor lighting conditions.

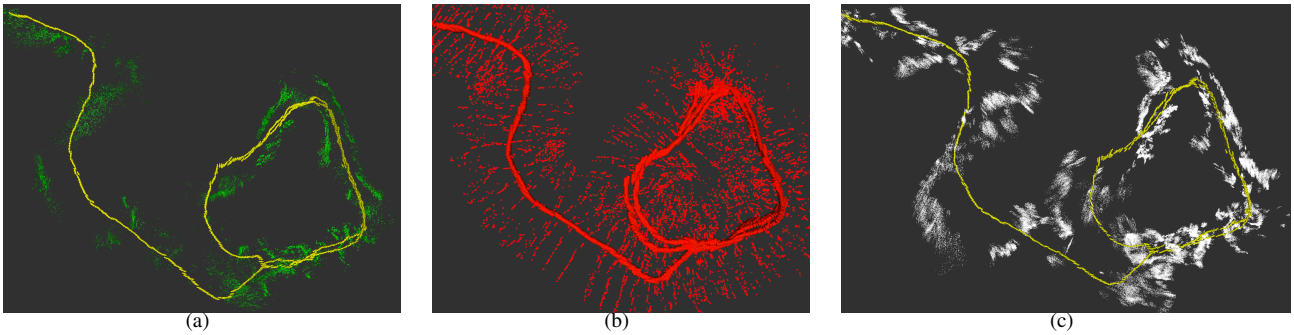
B. Odometry and 3D Cave-Wall Reconstruction

The length of the trajectory produced by our method is 87 meters, consistent with the measures from the divers. Fig. 5 shows the whole trajectory with the different point clouds generated by the features used for tracking, Sonar data, and stereo contour matching. Keeping a small set of features for



(a) (b)

Fig. 4. Partial trajectories generated by DSO. Fig. 4(a) Incorrect odometry and failing to track just after a few seconds and Fig. 4(b) longer trajectory after starting at a place with better illumination which also fails later on.



(a) (b) (c)

Fig. 5. (a) Odometry using only a few strong features (green) for tracking. (b) Scanning Sonar measurements (red) aligned along the trajectory. (c) Reconstruction of the cave using the edges detected in the stereo contour points (gray).

only tracking helps to run the proposed approach in real-time, without any dropped sensor data, on the tested computers. As shown in the figure, Sonar provides a set of sparse but robust points using *range* and *head_position* information. Finally, the stereo contour matched point generates a denser point-cloud to represent the cave environment.

Fig. 6 highlights some specific sections of the cavern, with the image and the corresponding reconstruction – in gray, the points from the contours; in red the points from the Sonar. As it can be observed, our proposed method enhances the reconstruction with a dense point cloud; for example rocks and valleys are clearly visible in Fig. 6.

V. Discussion

The proposed system improves the point cloud reconstruction and is able to perform in real time even with additional processing requirements. One of the lessons learned during the experimental activities is that the light placement affects also the quality of the reconstruction. In the next version of the sensor suite, we plan to mount the dive light in a fixed position so that the cone of light can be predicted according to the characteristics of the dive light. Furthermore, setting the maximum distance of the Sonar according to the target environment improves the range measurements.

While this work presents the first initiative towards real-time semi-dense reconstruction of challenging environments with lighting variations, there are several scopes for improvements. One future work of interest is to combine a *direct* method and an *indirect* method, similar to [33], but instead of relying on the direct method for tracking, we would rely on the robust Sonar-Visual-Inertial-Depth estimate. Thus we will achieve a denser 3-D reconstruction by jointly minimizing the *reprojection* and *photometric* error followed by a robust tracking method. We also plan to acquire ground truth trajectories [39] by placing *AprilTags* along each trajectory for quantitative analysis. By deploying the sensor suite on a dual DPV more accurate results due to the greater stability are expected – see Fig. 1 for preliminary tests.

References

[1] S. Exley, *Basic Cave Diving: A Blueprint for Survival*. National Speleological Society Cave Diving Section, 1977.

[2] "Climate Change and Sea-Level Rise in Florida: An Update of "The Effects of Climate Change on Florida's Ocean and Coastal Resources."," Florida Ocean and Coastal Council, Tallahasee, FL, Tech. Rep., 2010.

[3] Z. Xu, S. W. Bassett, B. Hu, and S. B. Dyer, "Long distance seawater intrusion through a karst conduit network in the Woodville Karst Plain, Florida," *Scientific Reports*, vol. 6, pp. 1–10, Aug 2016.

[4] A. Abbott, "Mexican skeleton gives clue to American ancestry," *Nature News*, May 2014.

[5] N. Kresic and A. Mikszewski, *Hydrogeological Conceptual Site Models: Data Analysis and Visualization*. CRC Press, 2013.

[6] N. Weidner, S. Rahman, A. Quattrini Li, and I. Rekleitis, "Underwater Cave Mapping using Stereo Vision," in *Proc. ICRA*, 2017, pp. 5709–5715.

[7] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015.

[8] S. Rahman, A. Quattrini Li, and I. Rekleitis, "Sonar Visual Inertial SLAM of underwater structures," in *Proc. ICRA*, 2018, pp. 5190–5196.

[9] ——, "SVIn2: An Underwater SLAM System using Sonar, Visual, Inertial, and Depth Sensor," in *Proc. IROS*, 2019.

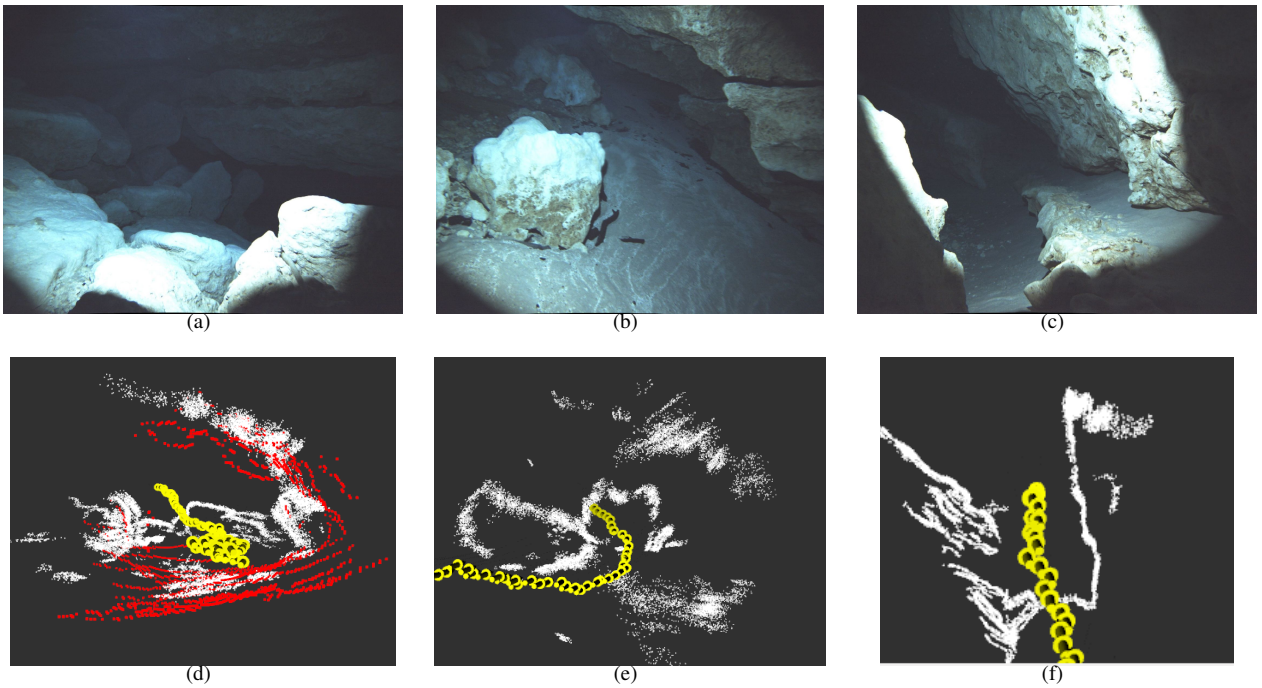


Fig. 6. Stereo contour reconstruction results in (d)-(f) and the corresponding images in (a)-(c) respectively.

[10] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-Scale Direct Monocular SLAM," in *Proc. ECCV*. Springer Int. Pub., 2014, vol. 8690, pp. 834–849.

[11] G. Dudek *et al.*, "A visually guided swimming robot," in *Proc. IROS*, 2005, pp. 1749–1754.

[12] D. Meger, J. C. G. Higuera, A. Xu, P. Giguere, and G. Dudek, "Learning legged swimming gaits from experience," in *Proc. ICRA*, 2015, pp. 2332–2338.

[13] M. Gary, N. Fairfield, W. C. Stone, D. Wettergreen, G. Kantor, and J. M. Sharp Jr, "3D mapping and characterization of sistema Zacatón from DEPTHX (DEep Phreatic THermal eXplorer)," in *Proc. of KARST: Sinkhole Conference ASCE*, 2008.

[14] W. C. Stone, "Design and Deployment of a 3-D Autonomous Subterranean Submarine Exploration Vehicle," in *Int. Symp. on Unmanned Untethered Submersible Technologies (UUST)*, no. 512, 2007.

[15] Stone Aerospace , "Digital Wall Mapper," URL:http://stoneaerospace.com/digital-wall-mapper/, Apr. 2015.

[16] A. Mallios *et al.*, "Toward autonomous exploration in confined underwater environments," *J. Field Robot.*, vol. 33, pp. 994–1012, 2016.

[17] S.-F. Chen and J.-Z. Yu, "Underwater cave search and entry using a robotic fish with embedded vision," in *Chinese Control Conference (CCC)*, 2014, pp. 8335–8340.

[18] K. Richmond, C. Flesher, L. Lindzey, N. Tanner, and W. C. Stone, "SUNFISH®: A human-portable exploration AUV for complex 3D environments," in *MTS/IEEE OCEANS Charleston*, 2018, pp. 1–9.

[19] S. Rahman, A. Quattrini Li, and I. Rekleitis, "A modular sensor suite for underwater reconstruction," in *MTS/IEEE Oceans Charleston*, 2018, pp. 1–6.

[20] P. Corke, C. Detweiler, M. Dunbabin, M. Hamilton, D. Rus, and I. Vasilescu, "Experiments with underwater robot localization and tracking," in *Proc. ICRA*, 2007, pp. 4556–4561.

[21] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *IEEE and ACM Int. Symp. on Mixed and Augmented Reality*, 2007, pp. 225–234.

[22] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, 2015.

[23] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, 2018.

[24] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. CVPR*, 2016, pp. 4104–4113.

[25] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in *Proc. ICRA*, 2007, pp. 3565–3572.

[26] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual-inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015.

[27] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular SLAM with map reuse," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 796–803, 2017.

[28] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, 2018.

[29] B. Joshi, S. Rahman, M. Kalaitzakis, B. Cain, J. Johnson, M. Xanthidis, N. Karapetyan, A. Hernandez, A. Quattrini Li, N. Vitzilaios, and I. Rekleitis, "Experimental Comparison of Open Source Visual-Inertial-Based State Estimation Algorithms in the Underwater Domain," in *Proc. IROS*, 2019.

[30] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3d," in *ACM transactions on graphics (TOG)*, vol. 25, no. 3, 2006, pp. 835–846.

[31] C. Wu, "Towards linear-time incremental structure from motion," in *IEEE Int. Conf. on 3D Vision-3DV*, 2013, pp. 127–134.

[32] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-M. Frahm, R. Yang, D. Nistér, and M. Pollefeys, "Real-time visibility-based fusion of depth maps," in *Proc. ICCV*, 2007, pp. 1–8.

[33] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "SVO: Semidirect Visual Odometry for Monocular and Multicamera Systems," *IEEE Trans. Robot.*, vol. 33, no. 2, 2017.

[34] C. Forster *et al.*, "On-Manifold Preintegration for Real-Time Visual–Inertial Odometry," *IEEE Trans. Robot.*, vol. 33, no. 1, pp. 1–21, 2017.

[35] S. Agarwal, K. Mierle, and Others, "Ceres Solver," http://ceres-solver.org, 2015.

[36] J. Shi *et al.*, "Good features to track," in *Proc. CVPR*, 1994, pp. 593–600.

[37] J. J. Tarrio and S. Pedre, "Realtime edge based visual inertial odometry for MAV teleoperation in indoor environments," *J. Intell. Robot. Syst.*, pp. 235–252, 2017.

[38] A. Quattrini Li, A. Coskun, S. M. Doherty, S. Ghasemlou, A. S. Jagtap, M. Modasshir, S. Rahman, A. Singh, M. Xanthidis, J. M. O'Kane, and I. Rekleitis, "Experimental comparison of open source vision based state estimation algorithms," in *Proc. ISER*, 2016.

[39] E. Westman and M. Kaess, "Underwater AprilTag SLAM and calibration for high precision robot localization," Carnegie Mellon University, Tech. Rep. CMU-RI-TR-18-43, 2018.