Underwater Monocular Image Depth Estimation using Single-beam Echosounder

Monika Roznere and Alberto Quattrini Li

***Abstract*— This paper proposes a methodology for real-time depth estimation of underwater monocular camera images, fusing measurements from a single-beam echosounder. Our system exploits the echosounder's detection cone to match its measurements with the detected feature points from a monocular SLAM system. Such measurements are integrated in a monocular SLAM system to adjust the visible map points and the scale. We also provide a novel calibration process to determine the extrinsic between camera and echosounder to have reliable matching. Our proposed approach is implemented within ORB-SLAM2 and evaluated in a swimming pool and in the ocean to validate image depth estimation improvement. In addition, we demonstrate its applicability for improved underwater color correction. Overall, the proposed sensor fusion system enables inexpensive underwater robots with a monocular camera and echosounder to correct the depth estimation and scale in visual SLAM, leading to interesting future applications, such as underwater exploration and mapping.**

I. Introduction

Exploration is fundamental for many underwater work, from archaeological preservation [1] to ecological surveys [2], and it will continue to advance with the technological progress of autonomous underwater robotic systems. Thus far, one of the main challenges is in *visual underwater perception*, notably in Simultaneous Localization and Mapping (SLAM) [3], which, if solved, can enhance the situational awareness of the robots and enable autonomy. SLAM is particularly difficult for low-cost Remotely Operated Vehicles (ROVs) and Autonomous Underwater Vehicles (AUVs), often configured with low-end sensors, such as inexpensive Inertial Measurement Unit (IMU), compass, pressure sensor, single-beam echosounder, and monocular camera.

Many state-of-the-art real-time visual SLAM systems are feature-based methods, which use raw images to extract features, track them over subsequent frames, and finally estimate poses and 3-D points [4]. While high accuracy was demonstrated with stereo cameras and IMUs – typically high-end in the underwater domain – low-cost vehicles are far from being robust enough to enable autonomous operation. In cases when the IMU is unreliable and stereo camera is unavailable, low-cost vehicles must rely on purely-visual monocular SLAM systems, which suffer from ambiguous depth scale and drift [5].

This paper addresses the problem of estimating image depth from a monocular camera on an inexpensive commercially available ROV, by integrating distance measurements from a low-cost single-beam echosounder – see Fig. 1 for a depiction of the problem in focus. Distance measurements from the echosounder are matched with estimated 3-D points from the monocular visual SLAM system, and a scale correction is applied to the estimated 3-D points or camera pose. To ensure proper matching, we devise a calibration method to determine the extrinsic between camera and echosounder that minimizes the matching error of measurements from the two sensors of a known object. From our previous work [6], [7], this paper provides the following contributions:

- A calibration algorithm based on cone fitting that utilizes a simple sphere. This allows for recovery of extrinsic between camera and echosounder.
- A method for projecting the echosounder measurement cone onto the monocular camera image frames and matching its readings to the extracted feature points from a monocular SLAM system.
- A real-time sensor fusion approach to integrate echosounder measurements into a monocular SLAM system, thus improving the depth estimate and scale.
- An implementation with ORB-SLAM2 [8] and analysis of pool and sea experiments that highlight the feasibility of our approach for image depth correction.
- An example application of underwater color correction given the improved estimated image depth of the scene.

This work represents a first effort towards inexpensive solutions for underwater perception to make low-cost underwater vehicles more autonomous and accessible to the scientific and industrial communities. The promising results provide

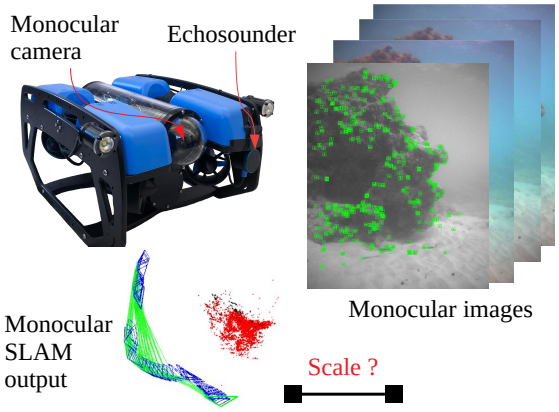


Fig. 1: Given a monocular camera and an echosounder mounted on a low-cost underwater robot (BlueROV2), how can scale be corrected for a monocular SLAM system?

The authors are with Department of Computer Science, Dartmouth College, Hanover, NH USA {monika.roznere.gr, alberto.quattrini.li}@dartmouth.edu

 DOI: 10.1109/IROS45743.2020.9341105

insights for future directions.

This paper is structured as follows: the next section presents background work on SLAM and sensor fusion. Section III describes the calibration approach and the depth fusion in a monocular SLAM system. Sections IV and V analyze the experimental results and discuss extensions. Finally, Section VI concludes the paper.

II. Background

State-of-the-art visual odometry and SLAM systems employ two main classes of methods for estimating camera motion. (1) Direct methods minimize the alignment error based on intensity values between images (e.g., LSD-SLAM [9] and DSO [10]). (2) Indirect methods minimize reprojection errors of tracked features (e.g., ORB-SLAM2 [8] and work of Lim et al. [11]). Hybrid methods combine both methodologies, e.g., SVO [12].

The basic sensor configuration of such methods is composed of a monocular camera, which suffers from scale ambiguity and drift [13], resulting in an incorrectly scaled map point cloud, negatively affecting the situational awareness of the robot, especially for control and planning.

To increase the quality of the estimates calculated by the state estimation algorithms, it is common to fuse data from other sensors, for example multi-calibrated cameras [14], [15] or IMU [16]–[22]. In the underwater realm, SLAM systems are mainly based on sonar – an exteroceptive sensor, whose measurements will not be affected by drift, as seen with low-end IMUs. Folkesson et al. [23] proposed the use of a blazed array sonar for real-time feature tracking. More recently, Richmond et al. [24] described an underwater SLAM system for autonomous cave exploration that uses a multi-beam sonar, an underwater dead-reckoning system based on fiber-optic gyroscope (FOG) IMU, an acoustic Doppler Velocity Log (DVL), and pressure-depth sensors. Similarly, SVIn2 [25] system fused measurements from a profiling scanning sonar together with IMU, stereo camera images, and pressure sensor. To reliably apply many of these systems, it is important to undergo multi-sensor calibration, such as camera to multi-beam sonar [26] and camera to imaging sonar [27], [28].

Instead of fusing multiple sensors, recent work integrates fiducial markers into the environment to act as ground truth parameters for the SLAM system, either depending solely on the markers (SPM-SLAM [29]) or using a fusion of keyframes and markers (UcoSLAM [30]). Other methods [31], [32] explicitly address changes in the scene, e.g., in illumination, by preprocessing the images. The image enhancement methods do not typically depend on information from the environment or do require high-end sensors (DVL).

In our work, we consider low-cost ROVs and AUVs not equipped with costly devices, but usually installed with a monocular camera and a single-beam echosounder, e.g., BlueROV2. Our method uses a monocular SLAM system – NO reliable high-frequency IMU is installed on the robot. We address the problem of "how to alleviate the issue of scale ambiguity affecting monocular SLAM with measurements from a single-beam echosounder?"

III. Approach

Our proposed system – see Fig. 2 – will correct the depth scale for a monocular SLAM system, given a camera and an echosounder with overlapping field of view. We define the echosounder and camera model (Section III-A) to then enable real-time projection of the sound cone onto the image frame. We propose a calibration and optimization procedure to ensure proper projection (Section III-B), and we describe the method to fuse the echosounder measurement with the SLAM system in Section III-C.

A. Echosounder and Camera Model

To measure the distance to an object, a single-beam echosounder emits an acoustic pulse i and listens to the reflected pulses, recording the time of flight. The time of flight t_i of the strongest reflected pulse is used to calculate the distance measurement $m_i = v \cdot (t_i/2)$, where v is the sound velocity in water. Note that sound beams propagate approximately in a cone – see Fig. 2 (right).

The echosounder returns valid measurements m_i when the distance to the object is between d_0 – the known blanking distance or the minimum distance that the sensor can reliably detect at – and $d_{\max}$ – the maximum distance for the sensor. In general, an invalid measurement occurs when the following inequality does not hold:

$$d_0 \leq m_i \leq d_{\max} \tag{1}$$

Furthermore, if the echosounder's position ${}_C\mathbf{t}_E$ and direction unit vector $\vec{\mathbf{v}}$ with respect to the camera reference frame $\{C\}$ are known, its sound cone can be projected onto the image frame for every new distance reading m_i. The sound cone can be approximated as a circle with the center ${}_C\mathbf{c}_i$ in the camera reference frame as:

$${}_C\mathbf{c}_i = {}_C\mathbf{t}_E + m_i \cdot \vec{\mathbf{v}} \tag{2}$$

The radius of the circles is calculated as

$$r_i = m_i \cdot \tan(a), \tag{3}$$

where a is the sonar's known cone angle.

Then by applying a camera projection model, we can obtain the pixel value for the center – see Fig. 2 (right). For a pinhole camera model with camera intrinsics K, the pixel coordinates u and v are:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \cdot {}_C\mathbf{c}_i \tag{4}$$

While a more complex model could be used, we approximate the projection of the cone slice on the image with a circle. In this way, we only need to compute another point on the cone slice using the radius r_i, by projecting it on to the image and determining the circle in the image plane passing through that point, with center (u, v).

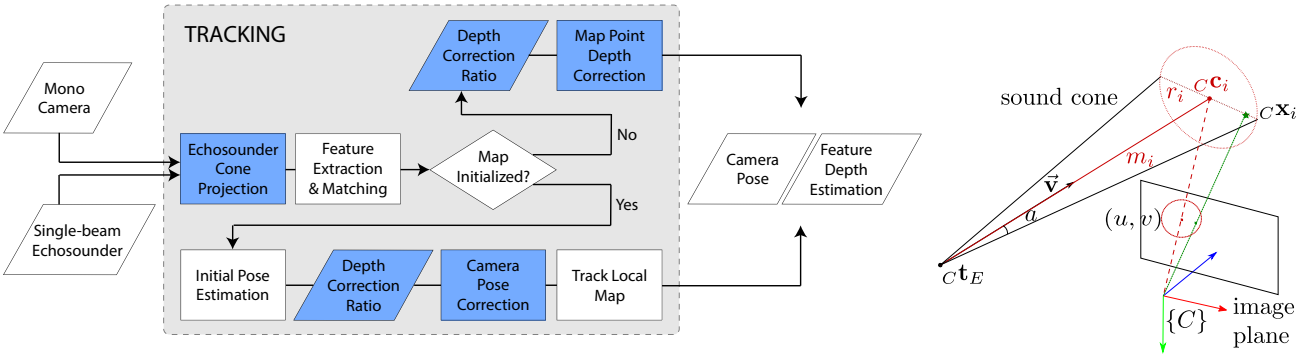


Fig. 2: Overview of the integration of the single-beam echosounder and monocular camera, with an indirect-based monocular SLAM system (left). Echosounder-camera model (right). If a point ${}_C\mathbf{x}_i$ is detected by the echosounder and is also visible by the camera, the echosounder detection cone can be projected onto the image frame, encircling the point's feature f.

B. Calibration

The echosounder's position ${}_C\mathbf{t}_E$ and direction vector $\vec{\mathbf{v}}$ are calculated by calibrating over a collected dataset $\mathbf{X}$ of ${}_C\mathbf{x}_i$ 3-D points of a known target object – detected by the sonar in the camera reference frame – and the corresponding echosounder readings m_i.

We use a known-sized sphere as the target. Spheres are typically used in acoustic calibration due to the guarantee that some sound will reflect back to the sensor [33]. By synchronizing the image input and echosounder readings, the robot can move around and observe the loss of detection moments. If the sphere is visible in the image frame, its approximate 3-D center point is calculated by the circular blob shape detection and by solving the Perspective-n-Point problem. When the echosounder produces a valid reading, the distance measurement and the corresponding current 3-D point is saved. This set $\mathbf{X}$ of n 3-D points and measurements describe the shape of the echosounder's coverage cone.

Note, the 3-D data points are approximations of the true points of acoustic reflections detected by the echosounder. The relative error is correlated to the size of the sphere, the distance from the sphere to the robot, and the distance from the echosounder to the camera. We assume the error to be minimal, and despite the slight inaccuracy the 3-D points are useful as estimations for the calibration procedure.

The calibration algorithm is implemented as an optimization process. The goal is to find the best camera-echosounder extrinsic ${}_C\mathbf{t}_E$ and $\vec{\mathbf{v}}$ that minimize the error between the ${}_C\mathbf{x}_i$ 3-D points in $\mathbf{X}$ and their corresponding measurements m_i. The solution should satisfy the constraint that all points lie within the sound cone. More formally:

$$\begin{aligned} &\arg\min_{{}_C\mathbf{t}_E,\vec{\mathbf{v}}} \sum_{i=1}^{n} (\|{}_C\mathbf{x}_i - {}_C\mathbf{t}_E\| - m_i)^2 \\ &\text{s.t.} \quad \forall\, {}_C\mathbf{x}_i \in \mathbf{X}, \quad o({}_C\mathbf{x}_i) \leq r({}_C\mathbf{x}_i) \end{aligned} \tag{5}$$

where $o({}_C\mathbf{x}_i) = \|({}_C\mathbf{x}_i - {}_C\mathbf{t}_E) \times \vec{\mathbf{v}}\|$ is the shortest (orthogonal) distance between ${}_C\mathbf{x}_i$ and the cone axis – direction vector $\vec{\mathbf{v}}$ – calculated with the cross product, and $r({}_C\mathbf{x}_i) = (({}_C\mathbf{x}_i - {}_C\mathbf{t}_E) \cdot \vec{\mathbf{v}}) \cdot \tan(a)$ is the radius of the circle, a slice of the cone, that ${}_C\mathbf{x}_i$ lies on.

The echosounder position ${}_C\mathbf{t}_E$ and direction vector $\vec{\mathbf{v}}$ may be initialized with hand-measured values to minimize the chance of falling into a local minimum. Additional constraints on the extrinsic can be added in the optimization to reflect the mounting position of the echosounder, e.g., if the echosounder is mounted on the left side of the camera, then the x component of ${}_C\mathbf{t}_E$ can only be negative.

The extrinsic parameters ${}_C\mathbf{t}_E$ and $\vec{\mathbf{v}}$ resulting from this optimization process are used for properly fusing the feature points from the images and echosounder readings m_i.

C. SLAM Depth Fusion

While absolute scale cannot be recovered from a monocular SLAM system, sonar readings can correct this ambiguity. We assume that the SLAM system is feature-based, because compared to direct-based method, indirect-based ones have shown to track longer underwater [3]; in underwater scenarios, illumination changes frequently, resulting in loss of localization for methods tracking pixel intensities. The main steps of a SLAM system include: an initialization phase to start the tracking with features visible from different points of view; a tracking phase, for comparing new frames with the current map using feature points; and a map updating phase, for optimizing the 3-D points using a set of keyframes and for performing loop closure [4].

Echosounder integration and depth scale correction occur in the tracking phase, more specifically in the map initialization and per image frame tracking. By adjusting the map points during map initialization, the SLAM system may begin its process with a more accurate initial scale. Likewise, per image frame tracking, particularly when estimating the initial camera pose, requires the camera pose to be adjusted with the correct depth scale to account for any error in the motion model or in the sudden changes in view.

Algorithm 1 shows how to calculate the depth correction ratio used for adjusting the map points or the camera pose. First, iterate through all of the features whose pixel points lie within the projected sound cone and take the closest point to the camera (Lines 1-8). That point is the one that according

Algorithm 1 Depth ratio calculation

Input: list of current visual feature points $\boldsymbol{f}_j$ and corresponding 3-D map points $\mathcal{F}_v$, echosounder measurement m_i, current camera pose ${}_C T_W$ in the world $\{W\}$ reference frame, camera-echosounder extrinsic ${}_C\mathbf{t}_E$, $\vec{\mathbf{v}}$
Output: Depth ratio d_i

/*Find the closest feature point to the cone*/
1: ${}_W\mathbf{x}_s = \infty$
2: **for** (every $\boldsymbol{f}_j, {}_W\mathbf{x}_j$ in $\mathcal{F}_v$) **do**
3: **if** (in_projected_cone($\boldsymbol{f}_j, m_i, {}_C\mathbf{t}_E, \vec{\mathbf{v}}$)) **then**
4: **if** (dist_to_cam(${}_W\mathbf{x}_j, {}_C T_W$) < dist_to_cam(${}_W\mathbf{x}_s, {}_C T_W$)) **then**
5: ${}_W\mathbf{x}_s = {}_W\mathbf{x}_j$
6: **end if**
7: **end if**
8: **end for**
/*Find new depth estimate of the visual point matching echosounder reading*/
9: $d_m = \arg\min_{{}_W\mathbf{x}_s} (\| {}_C T_W \cdot [{}_W\mathbf{x}_s^T \ 1]^T - [{}_C\mathbf{t}_E^T \ 1]^T \| - m_i)^2$
10: **return** $d_i = d_m / \| {}_W\mathbf{x}_s - {}_W\mathbf{t}_C \|$

to the echosounder model should be corresponding to the measurement. The new depth estimate d_m of the found map point ${}_W\mathbf{x}_s$ is calculated by optimizing its position along the line of camera view to fit with the echosounder's reading (Line 9) and the ratio is calculated according to the current map point depth (Line 10).

IV. Evaluation and Application

In this section, we illustrate the steps for collecting and applying echosounder measurements. First, our ESCalibr[1] application is operated to help collect echosounder data for calculating its position and direction vector with respect to the camera. This is followed by details on integrating echosounder extrinsic and measurements into a SLAM system for image frame projection and depth scale correction. Finally, we will display how the echosounder's measurements and the image feature points can be used with our image color enhancement method [6].

While our methodology for fusing echosounder measurements can be applied to any indirect based monocular SLAM system, we modified monocular ORB-SLAM2 [8], a real-time keyframe-based SLAM system that has shown good performance in underwater datasets [3], [5]. The optimized extrinsic from the calibration step are used to match feature points detected from ORB-SLAM2 and to adjust the corresponding depth values with the echosounder readings.

All experiments and data collection were performed in a swimming pool or in the Caribbean Sea. We used the BlueROV2, its installed Ping echosounder[2], and either the Sony IMX322LQJ-C camera[3] (included in the BlueROV2) or the Sony IMX273 camera[4] (installed for separate performance evaluation). The former camera – used in the pool – has a resolution of 5 MP, a horizontal field of view (FOV) of 80°, and a vertical FOV of 64°. While, the latter camera – used in the sea – has a resolution of 1.6 MP, a horizontal FOV of 96°, and a vertical FOV of 72°. The echosounder has a maximum range of 30 m and a cone angle a of 30°.

[1] https://github.com/dartmouthrobotics/escalibr
[2] https://bluerobotics.com/learn/ping-sonar-technical-guide/
[3] https://www.bluerobotics.com/store/sensors-sonars-cameras/cameras/cam-usb-low-light-r1/
[4] https://www.flir.com/products/blackfly-s-usb3/?model=BFS-U3-16S2C-CS

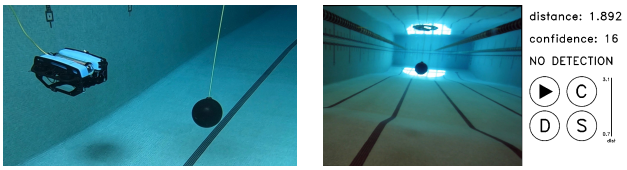


Fig. 3: Black glass sphere (left) and GUI (right).

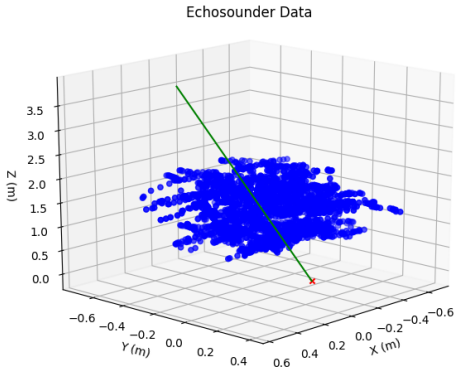


Fig. 4: The echosounder's position ${}_C\mathbf{t}_E$ (red x) and direction vector $\vec{\mathbf{v}}$ (green line) are calculated via the optimization process described in Equation (5), based on the measurements m_i and the 3-D points ${}_C\mathbf{x}_i$ of the detected sphere (blue dots).

A. ESCalibr: Echosounder and Camera Calibration

For experimental setup, we suspended a black glass sphere with a diameter of 25.4 cm into the water at an arbitrary depth. We observed that the thin rope, which holds the sphere in water, is undetectable by the echosounder. Fig. 3 (left) presents the simple setup needed for data collection.

We use our ESCalibr application to help us visualize what the robot sees, the current echosounder reading and confidence level (if applicable), and the amount of data points collected so far at different distances. The GUI, snapshot seen in Fig. 3 (right), also allows us to see in real time the sphere detection and results from 3D point calculation. After a period of time, the user can end the application and save the data of collected detection points.

Fig. 4 displays 4000 data points collected over 3 runs that were detected with high confidence level. The echosounder's hand-measured position is $(-0.17\,\text{cm},\ 0.08\,\text{cm},\ 0.09\,\text{cm})$. After calibration, the position ${}_C\mathbf{t}_E$ became $(-0.166\,\text{cm},\ 0.101\,\text{cm},\ -0.049\,\text{cm})$, shown as a red x, with a direction vector $\vec{\mathbf{v}}$ of $\langle 0.080, -0.146, 0.963 \rangle$, shown as a green line.

B. Depth Extraction and Adjustment

To validate the calibration extrinsic and the depth scale correction accuracy, we set up a pyramid of boxes with fiducial markers – acting as ground truth targets – and move the BlueROV2 to different view points. See Table I for the results of 10 runs, half with the hand-measured (**b**) values and half with the calibrated (**c**) extrinsic values. The

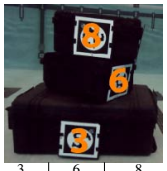
	View 1			View 2				View 3		
										
Tag ID	3	6	8	1	2	5	6	0	5	7
a	0.461	0.469	0.435	0.846	1.025	0.705	0.719	0.777	0.825	0.777
b	0.363	0.334	0.380	0.329	0.464	0.363	0.311	0.400	0.934	0.240
c	0.169	0.174	0.161	0.645	0.742	0.537	0.578	0.158	0.158	0.161

TABLE I: Depth scale error in Root Mean Square Error (m). **a:** Regular Monocular ORB-SLAM2. **b:** Adjusted with hand-measured echosounder extrinsic. **c.** Adjusted with calibrated echosounder extrinsic.

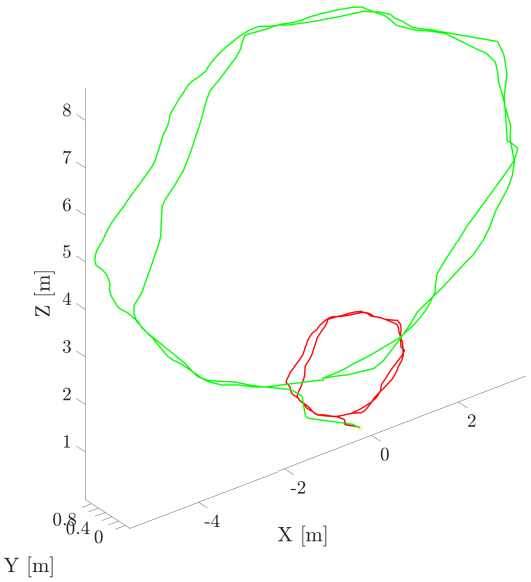


Fig. 5: Trajectory results of the robot circling a reef rock. Red: Monocular ORB-SLAM2 [8] implementation. Green: Monocular ORB-SLAM2 and echosounder integration.

calibrated values provided the best results, except for **View 2**. The fault most likely occurred while calculating the depth scale ratio. If a bad map point – e.g., a new corner appeared and assumed to be near – is chosen, then the effect ripples through the rest of the map points. Otherwise, hand-measured parameter values provide decent results as well.

We also conducted an experiment to evaluate the results after loop-closing. Here, the robot circled around a reef rock, identical to what is depicted in Fig. 6. As illustrated in Fig. 5, the SLAM and echosounder integration results in a trajectory of the same form as the regular SLAM implementation, but its scale is much larger, and corresponding to the actual size of the reef rock. This heavily implies that without the echosounder integration, the robot "thinks" it is closer to the (actually larger) rock than it is in reality.

C. Application: Image Enhancement

Our proposed method can be applied to robotic vision-based applications, such as our image enhancement method [6] (see the paper for further details). This method depends on the availability of image depth information, or distance values between the camera and the objects of interest in the scene. One distance value is not enough, as it will not accurately color correct parts of the image, especially when foreground objects are shaped uniquely or are at different locations in the scene. In this case, ORB-SLAM2 feature points with adjusted depth values can provide additional needed data.

Fig. 6 shows the steps to apply depth values to our image enhancement process [6] and results: **(a)** is the raw undistorted image seen by the robot. In parallel, ORB-SLAM2 detects features in the scene, as in **(b)**. Here, we estimate the depth values in the regions between the feature points by applying the Voronoi diagram. With monocular ORB-SLAM2, the system may randomly set the scene with low **(c)** or high **(d)** depth scale estimates, which leads to under- or over-enhancement, respectively. On the other hand, our approach **(e)** with SLAM and echosounder integration shows the best results, with more detail and no over-correction.

Image enhancement is one possible application for our system. Other underwater robotic operations include obstacle avoidance, exploration, and scene reconstruction.

V. Discussion and Future Steps

The jointly calibrated system of single-beam echosounder and monocular camera yields much potential to underwater tasks, especially when integrated with SLAM algorithms. While the proposed method was tested with ORB-SLAM2 [8], it will be beneficial to analyse it with other SLAM systems. Other extensions include system integration with a more suitable IMU or stereo camera.

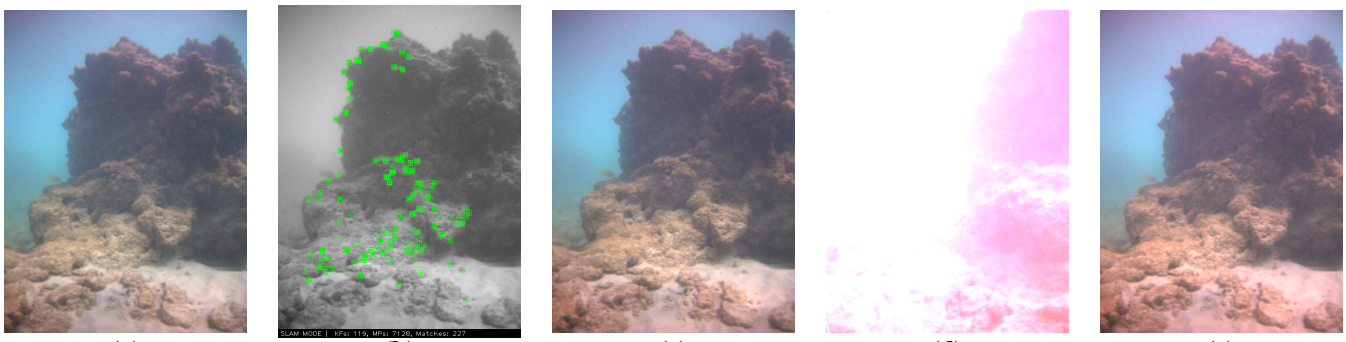
Currently, the sonar's reading is matched with the closest map point in its sound cone, which is misleading if the chosen point is on a parallel plane, like a wall or floor, not detectable by the sonar. To account for these false positives, one could add measurement uncertainty to the map points.

Furthermore, while the echosounder was shown to improve the depth scale during SLAM operation, we would like to also extend its capabilities to mitigate drift. We plan to integrate the echosounder readings into the map optimization phase to ensure that adjustments in keyframes also take into account of the sonar values.

While the proposed system was applied to image enhancement, it would be interesting to extend it to other underwater robotic tasks, like autonomous object avoidance or tracking.

VI. Conclusion

We presented a new method for integrating a low-cost single-beam echosounder and monocular camera together to improve SLAM and underwater robotic tasks, such as image enhancement. This paper provides analyses on experiments in a pool and in the sea to show the feasibility of this new design, as well as a discussion on accuracy improvements



(a) **(b)** **(c)** **(d)** **(e)**

Fig. 6: Image enhancement [6] with SLAM depth estimates. **(a)** Raw. **(b)** ORB-SLAM2 output. **(c)** Enhanced with low SLAM depth estimates. **(d)** Enhanced with high SLAM depth estimates. **(e)** Enhanced by proposed method.

and future steps. In broad sense, mounting inexpensive sensors on low-cost ROVs and AUVs will effectively augment their autonomy, increasing their applicability in many fields.

ACKNOWLEDGMENT

The authors thank the members of Dartmouth RLab and the Bellairs Research Institute of Barbados for experimental support. This work is supported in part by the Dartmouth Burke Research Initiation Award and NSF CNS-1919647.

REFERENCES

[1] "The world's underwater cultural heritage," http://www.unesco.org/new/en/culture/themes/underwater-cultural-heritage/underwater-cultural-heritage/, Accessed 02/20/2020 2020.

[2] O. Hoegh-Guldberg and J. F. Bruno, "The impact of climate change on the world's marine ecosystems," *Science*, vol. 328, no. 5985, 2010.

[3] 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.

[4] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," *IEEE Robot. Autom. Mag.*, vol. 18, no. 4, pp. 80–92, 2011.

[5] 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.

[6] M. Roznere and A. Quattrini Li, "Real-time model-based image color correction for underwater robots," in *Proc. IROS*, 2019.

[7] ——, "On the mutual relation between SLAM and image enhancement in underwater environments," *ICRA Underwater Robotics Perception Workshop*, 2019, (best paper award).

[8] 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.

[9] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. ECCV*, 2014.

[10] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, 2017.

[11] H. Lim, J. Lim, and H. J. Kim, "Real-time 6-DOF monocular visual SLAM in a large-scale environment," in *Proc. ICRA*, 2014.

[12] 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.

[13] H. Strasdat, J. M. M. Montiel, and A. J. Davison, "Scale drift-aware large scale monocular slam," in *Proc. RSS*, 2010, pp. 73–80.

[14] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, 2017.

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

[16] 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.

[17] 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.

[18] J. Salvi, Y. Petillo, S. Thomas, and J. Aulinas, "Visual SLAM for underwater vehicles using video velocity log and natural landmarks," in *MTS/IEEE OCEANS*, 2008, pp. 1–6.

[19] C. Beall, F. Dellaert, I. Mahon, and S. B. Williams, "Bundle adjustment in large-scale 3d reconstructions based on underwater robotic surveys," in *Proc. OCEANS*, 2011, pp. 1–6.

[20] F. Shkurti, I. Rekleitis, M. Scaccia, and G. Dudek, "State estimation of an underwater robot using visual and inertial information," in *Proc. IROS*, 2011, pp. 5054–5060.

[21] G. Loianno, C. Brunner, G. McGrath, and V. Kumar, "Estimation, control, and planning for aggressive flight with a small quadrotor with a single camera and imu," *IEEE J. Robot. Autom.*, vol. 2, no. 2, pp. 404–411, 2016.

[22] Y. Zhang, J. Tan, Z. Zeng, W. Liang, and Y. Xia, "Monocular camera and imu integration for indoor position estimation," in *EMBS*, 2014.

[23] J. Folkesson, J. Leonard, J. Leederkerken, and R. Williams, "Feature tracking for underwater navigation using sonar," in *Proc. IROS*. IEEE, 2007, pp. 3678–3684.

[24] 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.

[25] S. Rahman, A. Quattrini Li, and I. Rekleitis, "SVIn2: An Underwater SLAM System using Sonar, Visual, Inertial, and Depth Sensor," in *Proc. IROS*, 2019, pp. 1861–1868.

[26] N. Hurtós, X. Cufí, and J. Salvi, "Calibration of optical camera coupled to acoustic multibeam for underwater 3d scene reconstruction," in *Proc. OCEANS*. IEEE, 2010, pp. 1–7.

[27] S. Negahdaripour, H. Sekkati, and H. Pirsiavash, "Opti-acoustic stereo imaging: On system calibration and 3-d target reconstruction," *IEEE Trans. Image Process.*, vol. 18, no. 6, pp. 1203–1214, 2009.

[28] A. Lagudi, G. Bianco, M. Muzzupappa, and F. Bruno, "An alignment method for the integration of underwater 3d data captured by a stereovision system and an acoustic camera," *Sensors*, 2016.

[29] R. Muñoz Salinas, M. J. Marín-Jimenez, and R. Medina-Carnicer, "SPM-SLAM: Simultaneous localization and mapping with squared planar markers," *Pattern Recognition*, vol. 86, pp. 156–171, 2019.

[30] R. Muñoz Salinas and R. Medina-Carnicer, "UcoSLAM: Simultaneous localization and mapping by fusion of keypoints and squared planar markers," *Pattern Recognition*, vol. 101, 2020.

[31] R. Schettini and S. Corchs, "Underwater image processing: state of the art of restoration and image enhancement methods," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, p. 14, 2010.

[32] Y. Cho and A. Kim, "Visibility enhancement for underwater visual slam based on underwater light scattering model," in *Proc. ICRA*. IEEE, 2017, pp. 710–717.

[33] D. A. Demer, L. Berger, M. Bernasconi, E. Bethke, K. Boswell, D. Chu, R. Domokos, A. Dunford, S. Fassler, S. Gauthier, *et al.*, "Calibration of acoustic instruments," 2015.