IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 October 2023, accepted 30 October 2023, date of publication 3 November 2023, date of current version 8 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3329834

RESEARCH ARTICLE

Camera-Sonar Combination for Improved Underwater Localization and Mapping

ALEXANDRE CARDAILLAC AND MARTIN LUDVIGSEN, (Member, IEEE)
Department of Marine Technology, Norwegian University of Science and Technology, 7491 Trondheim, Norway

Corresponding author: Alexandre Cardaillac (alexandre.cardaillac@ntnu.no)

This work was supported by the BugWright2 European Union's Horizon 2020 project under grant agreement No 871260.

ABSTRACT Taking advantage of the complimentary properties of sonars and cameras can improve underwater visual odometry and point cloud generation. However, this task remains difficult as the image generation concepts are different, giving challenges to direct acoustic and optic feature matching. Solving this problem can improve applications such as underwater navigation and mapping. A camera-sonar combination is proposed for real time scale estimation using underwater monocular image features combined with a multibeam forward looking sonar. The detected features from a monocular SLAM framework are matched with the acoustic features based on the relative distances in instrument reference frame calculated using the two data streams, and used to estimate a depth ratio. The ratio is optimised over a large sample set to ensure scale stability. The sensor combination enables real time scale estimation of the trajectory and the mapped environment, which is a requirement for autonomous systems. The proposed approach is experimentally demonstrated for two underwater environments and scenarios, a subsea module mapping and a ship hull inspection. The results demonstrate the efficiency and applicability of the proposed solution. In addition to correctly restoring the scale, it significantly improves the localization and outperforms the tested dead reckoning and visual inertial SLAM methods.

INDEX TERMS Imaging sonar, visual SLAM, underwater perception, 3D reconstruction.

I. INTRODUCTION

Situational awareness of robots is fundamental to enable their autonomy. Simultaneous Localisation And Mapping (SLAM) [1] methods can significantly contribute to the autonomy as they improve the knowledge and understanding of the environment where the robots are operating in real time. However, when the method depends on a monocular camera only, the scale information of the resulting map and calculated vehicle path is lost or ambiguous. The scale information is important for both localization and mapping to enable the autonomy of the robotic platforms.

Underwater scenes can be difficult to observe and understand with an optical camera because of the light conditions in underwater environments. The light refraction makes the object appear blurry or distorted and the produced image becomes dimmer as the depth increases, resulting in loss of color and contrast perception. The turbidity of the water and the floating particles have an negative impact on the visual range by scattering and absorbing light. Using artificial light, movements of the light source and receiver may cause challenging light and shadow patterns. The advantage of optical imaging is the high resolution and rich information content in the data. Acoustic signals do not depend on seawater turbidity and allow larger observation ranges for underwater structures and objects. Compared to optical data, the resolution of acoustical data is considerably lower. For underwater vehicles, multibeam forward-looking sonars (MB-FLS) is often used to provide accurate observations of the surroundings to enable collision avoidance and safe paths.

In spite of the challenges to optical underwater imaging, features can be detected using computer vision and learning methods through processing to adjust and compensate for the effects present underwater. For MB-FLS imagery, the 3D information of the features is not available, only the relative

The associate editor coordinating the review of this manuscript and approving it for publication was Chengpeng Hao.

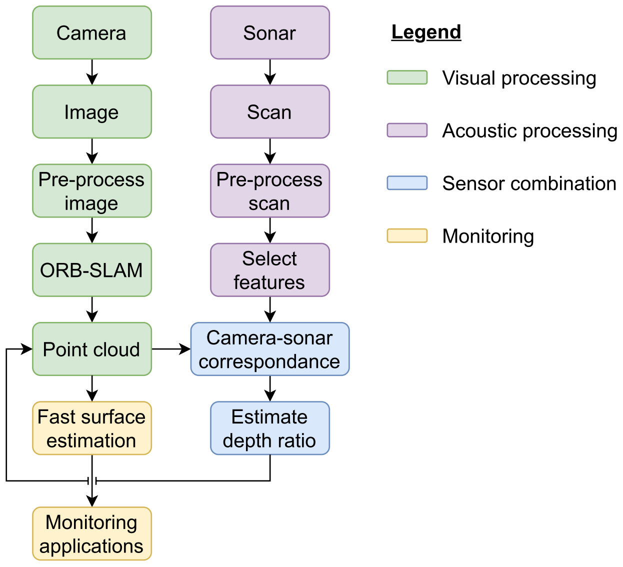


FIGURE 1. Overview of the proposed approach to perform camera-sonar combination by matching the sensors respective features.

azimuths and distances are computed, leaving the elevation of the data points ambiguous.

Visual SLAM (VSLAM) is a an active research field where the algorithms and methods are set up to match the available sensors and their configuration to detect and identify features in the environment. Both single and multiple camera systems are used, but the former is most common [2], [3], [4]. The sensors are easy to use and deploy, but for single camera systems in particular, the navigation solution experiences drift. The resulting scale ambiguity represents a particular challenge for SLAM based methods. To improve the results a second camera, a depth camera or an IMU can be added [5], [6], [7]. Monocular SLAM methods have also been developed or augmented specifically for the underwater environment [8], [9]. Sonars can be used for range detection by time of flight measurements. In [10], a FLS is employed with a feature based approach using detection of well-constrained landmarks to accurately estimate 3D points for mapping purposes. A filter-based approach is adopted in [11], the registered scans are processed together with an IMU and a DVL to create online a 2D grid map of the environment. In [12], a method to combine an IMU, a stereo camera and a mechanical scanning profiling sonar is proposed. The camera and sonar are combined over multiple samples. Patches based on the visual features are created and used to determine if they correspond to the features previously observed by the sonar. The complementary properties of optic and acoustic sensors represents a promising solution [13], [14], [15]. In [16], particle filter is used as the data association technique to calibrate the camera and sonar to obtain an accurate transformation matrix. The 3D camera features are then projected onto the sonar scan using the sonar coordinate system. A VSLAM framework is augmented in [17] where the camera is combined with a single beam echosounder to dynamically restore the scale of the SLAM estimate. To this end, the acoustic cone is modelled and matched to the best corresponding visual feature. The depth ratio is then calculated and applied to all the 3D points and the estimated trajectory.

The efforts made to enable optical and acoustic data combination depend strongly on the setup and application, and very often with low level feature matching mechanisms, i.e., using the main image characteristics such as shape and texture. Direct combination at the feature level for improved navigation and mapping requires more advanced inter-sensor calibration and methods, and is not well studied, and requires specific sensors and calibration routines. Using a sonar in addition to a camera provides a robust, drift free, and consistent solution, together with a basic sensor suite with implementation that are convenient to operate.

This paper aims at combining a monocular camera and a MB-FLS for improved underwater localisation and mapping independent of inertial or gyro data, making it suitable also in areas where inexpensive magnetometer based gyros are not feasible. The optical images are processed in a VSLAM framework to obtain a trajectory and point cloud over time. The sonar measurements are first used to rescale the SLAM estimates by finding correspondences between the sonar features and camera features. A depth ratio is estimated during the initialisation and updated online using the Maximum Likelihood Estimation (MLE). The depth ratio is a single value describing the factor to correct the depth scale of the VSLAM framework. It allows to convert the SLAM's distance unit to meters. The correspondences between the two sensors can be done thanks to the prior knowledge of intrinsic and extrinsic calibration details for the camera. The overlapping acoustic and visual areas can then be estimated, and a sonar feature can be represented as a segment in the camera image. Feature matching is performed based on the relative distances, which also helps removing outliers from the set of visual features such as particles. The matched points enable estimation of the depth ratio used later for the entire set of poses and visual points from the VSLAM framework. The optical-acoustic data combination is performed within the SLAM framework itself, which enables a verification step for the visual features. This provides improved localisation and mapping accuracy together with scale correction. Finally, 3D surface estimation based on the generated re-scaled point cloud is performed, using an adapted Poisson surface reconstruction approach.

To achieve real time underwater SLAM for low cost ROVs used for ship hull inspection is the main objective of this work, and the solutions are applicable to any underwater vehicle equipped with camera and MB-FLS. The camera and sonar sensor models are first presented in Section II including explanations of the inter-sensor correspondence mechanisms. Section III describes the depth ration estimation obtained using the inter-sensor feature matching method. The experiments and results are then presented and discussed in Sections IV and V. Finally, conclusive remarks are

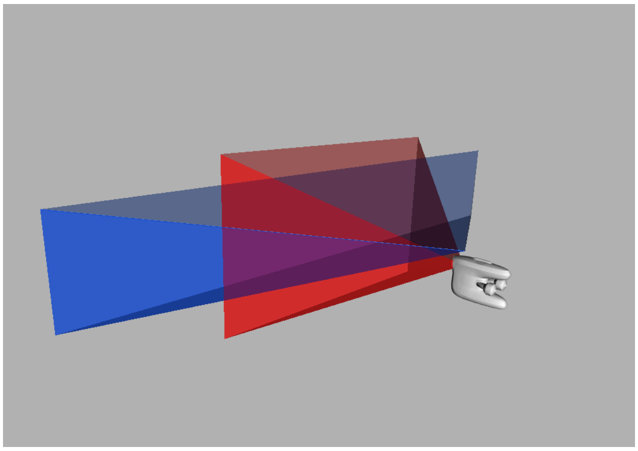


FIGURE 2. The footprints of the perception sensors are represented while the ROV is facing a wall. The camera field of view is in red, and the field of view for the forward looking sonar is shown in blue.

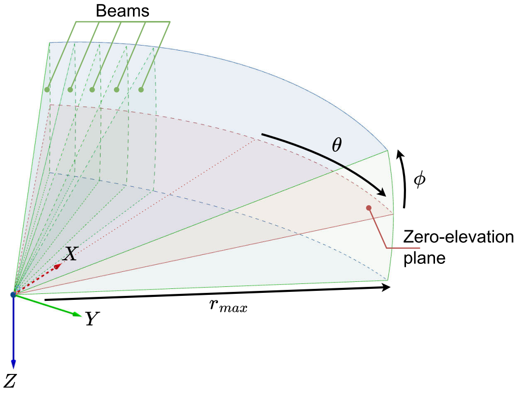


FIGURE 3. The footprint of the forward looking sonar is represented with the corresponding geometry. The minimum and maximum elevation planes are represented, as well as the zero-elevation plane, in which all the planes are merged to after the processing of the measurements. A beam ***i*** is also represented, going through all the elevation planes.

formulated in Section VI. An overview of the components and their interactions in the proposed approach is presented in Figure 1.

II. SENSOR MODELS

To understand how correspondences between the features in the sonar data and the optical camera imagery can be created, both sensor models are described. The field of view for the sensors have a large overlap and are horizontally aligned. However, there is a vertical offset because the sonar is mechanically mounted above the camera. The footprints of both sensors are represented in Figure 2. In this paper, we refer to the camera measurements as images, and to sonar measurements as scans, where a scan is defined as the data from all acoustic beams for a single acoustic ping.

A. SONAR MODEL

The sonar emits sound pulses referred to as pings and using the wave properties of acoustics, the multi-element transmitter and receiver array enables directionality for both signal transmission and reception providing acoustic beams. These beams have a vertical and horizontal opening and direction defined by the transducer element array and the signal transceiver. The pings propagate to a target before they are reflected back to the sonar receiver. The target range is estimated based on the signal travel time and the bearing is calculated using the phase difference measured using the transducer array. Most MB-FLS have a one dimensional transducer element array resulting in undefined depression angles for the echos and the sonar can therefore not derive the vertical position of the targets. This means that each point on the sonar imagery is a point on a 3D arc going from the minimum elevation to the maximum allowed by the sensor.

The sonar employed in the experiments was a Blueprint Oculus 750/1200 kHz with horizontal aperture of 130° and 20° vertically. It has 512 beams uniformly distributed and with angular width $\sigma_h = 0.25°$. The ping rate is controlled and configured to 10Hz. For the experiment used in this work, the sonar was configured in high frequency mode, corresponding to 1200 kHz, with low gain. A fixed maximum range of 4 meters was set to correlate to the visible range of the camera.

A sonar scan represents a 2D acoustic intensity array representing the features in polar coordinates $[\theta, r]^\top$, where θ is the azimuth angle and r is the range. The scan is formed as a polar image. The 3D geometry of the beams are represented in Figure 3 where the 3D acoustic features are expressed in spherical coordinates $[\theta, \phi, r]^\top$ with the elevation angle ϕ. When the acoustic ping reflections have returned to the sonar, the measurements are merged into a polar grid without elevation information. This grid can be projected onto the zero-elevation plane and placed in the relevant 3D reference frame. To represent the acoustic points in a 3D Cartesian world, given that the elevation angle is known or estimated, the coordinates need to be converted. The spherical-Cartesian coordinates conversion is formulated as

$$P = \begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix} = r \begin{bmatrix} \cos\phi \cos\theta \\ \cos\phi \sin\theta \\ \sin\phi \end{bmatrix}, \tag{1}$$

where P is the 3D point in Cartesian coordinates. The inverse conversion is also possible, and given by

$$r = \sqrt{P_x^2 + P_y^2 + P_z^2}, \tag{2}$$

$$\theta = \tan^{-1}\left(\frac{P_y}{P_x}\right), \tag{3}$$

$$\phi = \tan^{-1}\left(\frac{P_z}{\sqrt{P_x^2 + P_y^2}}\right). \tag{4}$$

B. CAMERA MODEL

The camera used in the presented experiments has an imaging frequency of 25Hz and a resolution of 1280 × 720px. It has a vertical and horizontal Field Of Views (FOV) underwater of ~48° and ~77° respectively. The image was calibrated underwater using a checkerboard and follows the pinhole model which formulates the 2D-3D correspondence as

$$p = \frac{P}{P_z} K, \tag{5}$$

converting the 3D point P in the world to the 2D point p in pixels in the image, using the intrinsic matrix of the camera K defined as

$$K = \begin{pmatrix} f_x & 0 & c_u \\ 0 & f_y & c_v \\ 0 & 0 & 1 \end{pmatrix}. \tag{6}$$

The focal length is described by (f_x, f_y), and (c_u, c_v) are the pixel coordinates of the optical centre of the camera.

C. CAMERA-SONAR CORRESPONDENCE

To combine both sensors, correspondences and mapping functions must be setup. They are defined based on both models to formulate the features of the first sensor in the second's sensor frame. For the correspondences, we consider the camera to be the origin of the local reference frame and use the features detected by the SLAM framework. Since the sonar is aligned with the camera with offset only vertically, the transformation matrix is simplified and constitutes an identity matrix for the rotation and a translation vector $[0, 0, t_z]^\top$ describing the vertical offset t_z. This removes the need of computing 6-DoFs sensor transformations. However, this comes with the risk of calibration imprecision which can significantly impact the results. Given that the main objective of this work is to inspect underwater structures, the imprecision is negligible since the operation will be performed with a close range to the objects.

Because the elevation of the sonar features, ϕ, is ambiguous, the exact corresponding points on the camera image cannot be known from the sonar data directly. Instead, the potential locations of an interest point can be represented by a moving vertical segment for each beam, where the beam and the image plane coincide. For an ideal setup where the camera is perfectly calibrated and the mounting offset between the camera and sonar is exactly compensated, each beam corresponds to a segment of the pixel column in the optical image. Because both sensors have different vertical field of views and have a vertical offset, the intersecting beam segment does not include the entire pixel column, and its length varies with the distance to the target. u_i represents the corresponding pixel column for a sonar beam of azimuth θ_i,

$$u_i = f_x \tan(\theta_i) + c_x. \tag{7}$$

However, since a beam has an angular width σ_h wider than the pixel width, there are multiple corresponding pixel columns for each beam in the optical image. The first and last columns must be calculated with $\theta_i \pm \sigma_h$ where σ_h represents the angle between the beam's central axis and its boundary. To obtain the list of possible vertical pixels for a given beam, the pinhole formula is not sufficient, the sonar range r_i and the vertical offset t_z must be included in the estimation of the vertical pixel locations v_i,

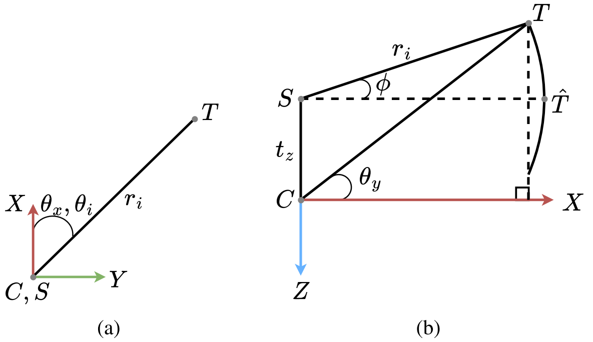


FIGURE 4. The geometry involved to obtain the pixel position on the optical image of a sonar feature T is represented. (a) is a top-down view, in the Oxy reference plane, and shows the parameters used to obtain the horizontal position u_i of the pixel with the azimuth θ_i and range r_i of the sonar beam i. (b) presents a side view, in the Oxz plane, with a possible sonar beam elevation ϕ. C and S respectively correspond to the camera and sonar positions.

$$v_i = f_y \tan(\theta_y) + c_y. \tag{8}$$

The elevation angle for the light ray from the target to the camera is represented by θ_y, and is obtained given a target with location T seen by the sonar at a distance r_i. Its estimated position $\hat{T}$ is initially placed on the sonar's zero-elevation plane and moved along the elevation circle arc constrained by the acoustic beam vertical width, for all $\phi \in [\phi_{min}, \phi_{max}]$. Given that the camera is at the origin of the reference frame, the coordinates of T are enough to obtain the angle θ_y, such that

$$\theta_y = \text{atan2}(T_z, T_x). \tag{9}$$

Equation (8) can be simplified to avoid multiple operations with tangents,

$$v_i = \begin{cases} f_y \dfrac{T_z}{T_x} + c_y, & \text{if } T_x \neq 0 \\ c_y, & \text{otherwise.} \end{cases} \tag{10}$$

This problem can be solved in a 2D environment, in the Oxz plane, since the camera and the sonar are horizontally aligned. Therefore, the possible positions of T are computed as follows:

$$T = \begin{bmatrix} t_x \\ t_z \end{bmatrix} + r_i \begin{bmatrix} \cos\phi \\ \sin\phi \end{bmatrix}. \tag{11}$$

A visual representation of the parameters is displayed in Figure 4 with a top-down and side views. The angles are defined relative to a target T and then used together with the Pinhole definitions to obtain a list of pixel candidates on the optical image. The top-down views shows the alignment of the sensors and the horizontal angles from both sensor are the same and remain constant regardless of the vertical angle. The side view shows how are the vertical angles related, given the vertical offset of the sonar.

The mapping from the sonar scan to the camera image is now established with (7) and (10), considering only the overlapping areas. However, a sufficient number of points, which are distributed in the image, are matched to ensure that robust results are obtained in the following sections and can be applied to the not-overlapping areas.

Each visual feature on a sonar line should correspond to a sonar feature. Matching the two features enables the estimation of a depth ratio.

III. DEPTH RATIO ESTIMATION

The feature matching mechanism was developed and executed in three steps followed by the estimation of the ratio.

1) All features are detected in the corresponding image and scan. In the case of the optical camera, 2D features are required as well as triangulated 3D points. To this end, a monocular V-SLAM framework is utilized, ORB-SLAM [18]. For the sonar scans, only 2D features are sampled.
2) All the features are filtered, and only the closest visual and acoustic points are kept.
3) The features are matched based on their respective relative distances and constrained by the possible locations on the image plane.
4) Each detected correspondence is processed to obtain a depth ratio and the MLE is applied over all matches to obtain a unique and consistent depth ratio.

In the following section III-A, the first two steps are covered, the selection and filtering of the good features to match. The last two steps are presented in section III-B, the actual matching of visual and acoustic features and how they are used to obtain scale information.

A. CAMERA-SONAR FEATURE MATCHING

The ORB-SLAM framework for monocular image data provides a trajectory and 3D point cloud over time. It has real-time performance and can work in large environments. It performs feature detection and matching for each image and builds a pose graph over time which enables loop closure and camera relocalisation capabilities. It is a popular lightweight framework that has proven to be very efficient in many applications [19], [20], including in underwater environments [21], [22] in spite of the challenges related to the ORB descriptor applied to image features and characteristics common underwater. Scenarios close to the water surface will often suffer from non-uniform ambient lightning conditions, and in deeper water the motion of the camera and light carrying robot may cause dynamic light and shadow patterns. The Contrast Limited Adaptive Histogram Equalisation (CLAHE) has proven to be an efficient method to compensate for non-uniform lighting environment to highlight the present features before the images are passed through a marine snow filter [23]. The ORB-SLAM's embedded parallax mechanism is made more flexible to enable continuous triangulation of points with a slow speed manoeuvring ROV and high camera frame rate. This mechanism also accounts for unwanted features coming from dynamic objects by computing the local median disparity of the tracked features and keeping only those below a threshold. This also results in a fast initialisation process.

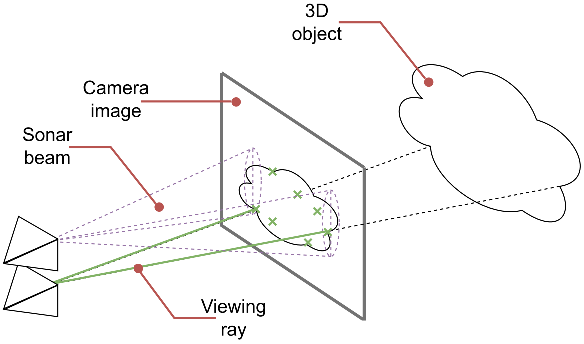


FIGURE 5. A 3D scene is represented with the camera-sonar correspondence and matching mechanisms. The green rays come from the camera, and the purple beams from the sonar.

Only the polar features of the sonar scans are required for the matching mechanism as they already hold positioning information related to the vehicle's reference frame, i.e., the distance to the object and its horizontal angle relative to the ROV. However, depending on the surrounding structures, the scans might include a significant amount of noise. They are therefore preprocessed to remove the noise and to highlight areas with structural information. A combination of a Gaussian filter and CLAHE is used for that purpose, significantly diminishing the noise while at the same time increasing the intensity values of the structures in sight. Furthermore, this approach allows uniform intensity over the scan sequence.

One feature per beam is selected, the closest with a reflectance above a high reflectance threshold. They correspond to the features with the highest chances of being visually detected as they should also be the closest to the camera.

Three sets of data are now available: the closest sonar points, the 3D point cloud, and the corresponding 2D features on the current image. For each sonar line on the image, based on the sonar feature information and (7) and (8), the closest 3D point that has its 2D correspondence lying on the line is matched. The 3D representation of the data types is displayed in Figure 5 with the sonar beams and camera reprojections.

B. MLE OF THE DEPTH RATIO

Because the ROV is continuously moving, timing is essential. The sonar processing is tightly integrated in the SLAM framework and the time difference between the sonar scans and optical images is monitored to make sure they are synchronised. If the latency is below a threshold, defined as a percentage of the rate difference, the depth ratio is computed for this sonar measurement. This latency check is applied to ensure both sensors are observing the same scene. With a high

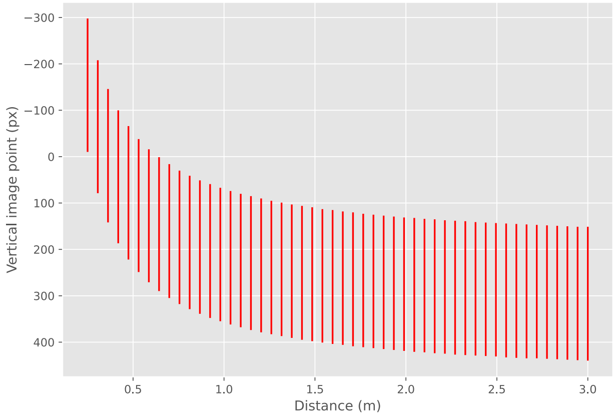


FIGURE 6. The visual table of the camera-sonar correspondences is computed with acoustic distances ranging from 0.25m to 3.0m. Each red line corresponds to a possible projected sonar beam on an image, i.e., its vertical pixel coverage.

latency, it is very likely that the sensors captured the scene from different locations.

Each sonar point, corresponding to a vertical line in the image, is now matched to a visual feature if such exists. For each match, a distance ratio is computed using the visual distance and acoustic distance, such that

$$d_i^c = ||\eta - P_i||, \tag{12}$$

$$_c d_i^s = \frac{r_i}{d_i^c}. \tag{13}$$

where η is the ROV position, P_i the 3D visual point used to obtain the visual distance d_i^c to the camera, r_i the sonar range, and $_c d_i^s$, the distance ratio for the match i. Once performed on each match, a new set of values is obtained. The MLE is employed to extract the final depth ratio, as it robustly find a consistent estimate. More data is accumulated over time, which makes the MLE adapt and estimate a value closer to the optimal one. The set of distance ratios is assumed to be following a Normal distribution $\mathcal{N}(\mu, \sigma^2)$ with mean μ and standard deviation σ. Its probability density function (pdf) is defined as follows,

$$P(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{14}$$

for the observation x. For simplification, the log likelihood is applied by taking the natural logarithm of the expression. This is possible because the natural logarithm is a monotonically increasing function. The equation (14) becomes

$$\ln\left(P(x; \mu, \sigma)\right) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(x-\mu)^2}{2\sigma^2}. \tag{15}$$

Iteratively maximising the above equation, or minimising its negative equivalent, results in optimised estimated values $\hat{\mu}$ and $\hat{\sigma}$ for the current data collected. The depth ratio λ is then assigned to the mean of the estimated normal distribution such that

$$\lambda = \hat{\mu}. \tag{16}$$

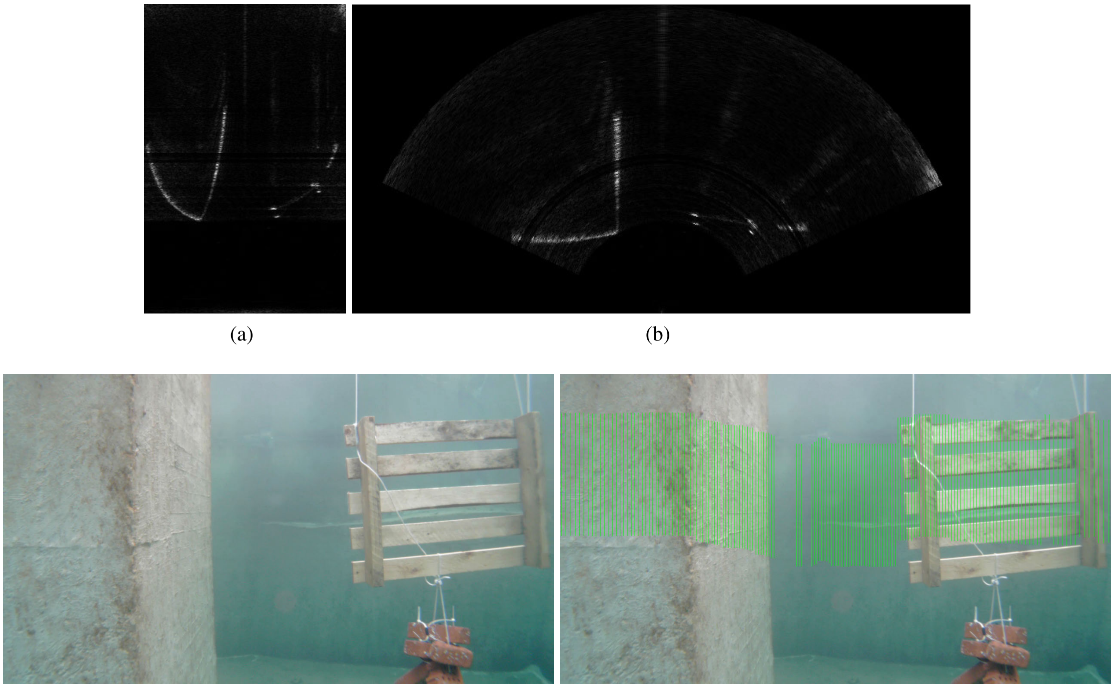
The continuous scale correction using the MLE enables a stable correction of the trajectory and point cloud over time, which results in improved localisation performance. Additionally, the normal distribution is used for outlier rejection and correction of the visual 3D points. The points that are more that $2\hat{\sigma}$ away from $\hat{\mu}$, corresponding approximately to the 95% confidence interval, are considered as outliers. The points inside this interval are updated, i.e. displaced further away or closer to match the predicted depth ratio. The elevation and azimuth angles of the updated points remain the same. This verification step is possible because the sonar is accurate and reliable, and has no error growth over time. Therefore, a 3D point with an irregular individual depth ratio can be detected and rejected to prevent the SLAM system from using it for future estimates.

IV. EVALUATION

In this section, the proposed approach is tested and quantitatively compared with three alternative SLAM navigation approaches and the ground truth. The camera-sonar correspondence model is first validated. The list of camera-sonar correspondence possibilities was computed geometrically and plotted in Figure 6. The red bars represent the possible intersections of the sonar beams with the image plane. Given the geometrical configuration, the beams with lower acoustic ranges intersect the higher parts of the image plane because the sonar is placed above the camera. Beams with larger acoustic ranges converge towards the center of the image.

To experimentally validate the setup, objects with known positions were placed in a pool and the ROV, equipped with the camera and sonar, positioned in front of them. The ROV was equipped with a GNSS receiver mounted on a pole, enabling the computation of its position and the distance between the ROV and the objects. The visual results of the first test scene are displayed in Figure 7. The main objects of the scene were detected by the sonar and were easily recognisable because they included high acoustic intensity values. Figure 7d is visually correct since the projected beams on the close objects in the camera image are higher than the rest, because the detected acoustic features corresponding to the objects reported close distances. When observing the stone pillar in both the sonar scan (Figure 7b) and camera image (Figure 7d), it is possible to understand how the sonar beams intersect the image plane at different sections. The edge of the pillar being the closest part of the pillar to the camera and sonar, the intersection segment is naturally higher, closer to the sonar's depth. And the further away the points are from the edge, the further away they are from the camera and sonar, gradually moving the intersection segment towards the center of the image, close to the camera's depth.

The numerical errors were estimated using GNSS as ground truth and are reported in Table 1. Here, the setup was tested in an additional scene, where the vehicle was facing the corner of the pool. While the second column shows a measure of the sonar accuracy, the third reports how well the sonar



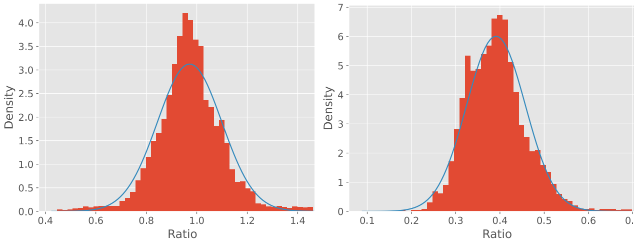
(a) (b)

(c) (d)

FIGURE 7. (a) is the original sonar scan as a polar image with rows as ranges and columns as bearings. It is converted to cartesian corrdinates in (b). The contours of the objects in the scene are recognisable as they present high intensity values. (c) is the original camera image. After applying the sonar features on the image, (d) is obtained. Each green line corresponds to an intersection of a sonar beam with the camera image plane. They highlight the possible locations of the acoustic features on the image plane.

TABLE 1. Camera-sonar correspondence results.

Scene	Distance error	Horizontal angle error
1	$0.04m \pm 0.08m$	$0.22° \pm 0.5°$
2	$0.02m \pm 0.07m$	$0.15° \pm 0.5°$



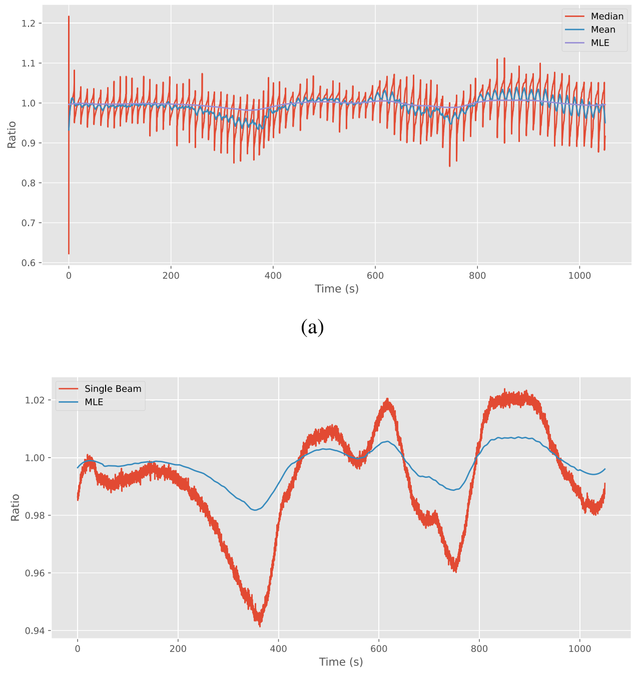
(a) (b)

FIGURE 8. Distance ratios were collected and accumulated over a sequence in two different scenes. Histograms (a) and (b) corresponds to the two scenes and Normal probability distributions were fitted and their density functions displayed on top of the histograms.

beams are corresponding to the image features, i.e., the angle difference.

Experimentally, the distance ratios acquired during field trials were found to fit a Normal distribution. The previous two scenes were expanded to include a few minutes of image and scan sequence of the surroundings to compute and create a set of ratios over time and with a changing scene. The corresponding histograms are displayed in Figure 8 with the density function on top. This validated the choice of the Normal distribution for the MLE. In both scenes, the distance and angle errors are very low, with centimeter level accuracy for the distances, 0.04m and 0.02m, and decimal level accuracy for the angles, 0.22° and 0.15°. For the ship hull mapping and inspection application considered in this work, theses errors are acceptable since the operations are performed close to the structures. For example, given the results from Table 1, at three meters distance, the maximum expected total error of the point correspondence is $\sim 0.15m$, and $\sim 0.05m$ on average.

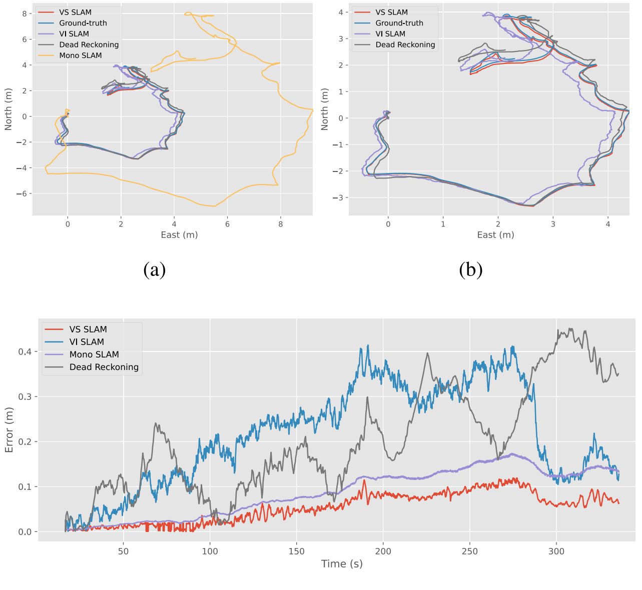
Ideally, the estimated depth ratio should converge towards 1, meaning the scale does not need to be re-updated. Our approach using the MLE is compared to three alternative approaches, including using simply the median or the mean, and using a single central beam. In this scenario, for each method, the depth ratio is estimated and applied every time there is a new keyframe created in the SLAM framework. The convergence rate of each method can be observed in Figure 9. While all methods converge rapidly, only the proposed one is continuously stable once it has converged. This is especially important for real-time operations as scale errors can quickly propagate to the depending systems. Peaks can appear when there is a sudden change of geometry in the scene, or when the ROV is turning, but they are immediately corrected. The method using the median showed high variations because it



(a)

(b)

FIGURE 9. **The depth ratio evolution over time is displayed. The proposed method using the MLE is compared to the mean and median in (a) and to a single central beam estimation in (b).**



(a)

(b)

(c)

FIGURE 10. **The 2D trajectories of all the methods used for comparisons are plotted in (a) and only the scaled ones in (b). (c) displays the position error over time of all the trajectories including the manually rescaled monocular SLAM trajectory.**

TABLE 2. **Performance metrics for trajectory evaluation.**

Method	ATE	RPE	Init. time	Completeness
Mono SLAM	0.091m	0.088m	1.2s	92%
VI SLAM	0.229m	0.203m	10.9s	84%
Dead Reckoning	0.201m	0.168m	0.0s	100%
VS SLAM	0.053m	0.049m	1.2s	98%

was heavily influenced by the new values added to the set and therefore by the shifts in sonar range. In comparison, using the mean was more stable, but still contained oscillatory results. When a single beam was used, the results improved. However, this method was more prone to noise, which can then destabilize the future SLAM estimates. Using the proposed method based on the MLE brings scale stability and consistency over longer periods of time, and once it has converged, it remains stable with low variance around the convergence point.

To show how the proposed method performs and improves the VSLAM framework, the trajectory estimate was compared to the default monocular SLAM from ORB-SLAM, visual-inertial SLAM, and dead reckoning using an IMU and a DVL. Additionally, a trajectory interpolating GNSS fixes and visual markers was computed and used as the ground truth.

The estimated trajectory of each method is displayed in Figure 10. They were all manually aligned. This visual comparison enables a first assessment of the method's performance and of the rescaled trajectory from the proposed method. The trajectory of the monocular SLAM (Mono SLAM), although correct, is off scale and can not be used for robotic applications. However, it was manually rescaled for the purpose of comparison. The rescaled version of the trajectory using the sonar (VS SLAM) appears close to the ground-truth compared to the other solutions. The visual-inertial SLAM (VI SLAM) was also able correctly rescale the trajectory, however, as the scale factor is estimated during the initialisation, if it is incorrectly estimated, it will lead to acceleration bias errors which can quickly propagate to the position estimates. Also, the noise of the low cost IMU influenced the plotted trajectory negatively. The dead reckoning solution performed well but showed apparent drift over time that made the trajectory end at a different location. The numerical results are highlighted in Table 2, with for each method, the Absolute Trajectory Error (ATE) computed with the Root Mean Square Error (RMSE), and the Relative Position Error (RPE). They were calculated over the whole trajectory. The initialisation time is also included, corresponding to how much time the framework needed to converge to an initial position estimate. Finally, the completeness represents how much of the dataset is successfully covered by the method, i.e., how many estimates were provided over time compared to the data available. Typically, long initialisation processes and visual tracking losses will result in significant loss of coverage.

The dead reckoning method was quickly initialising and always provided an estimate. However, it drifted quickly. The VI-SLAM showed the largest drift and lowest accuracy of the candidates, but thanks to the ORB-SLAM capabilities, the

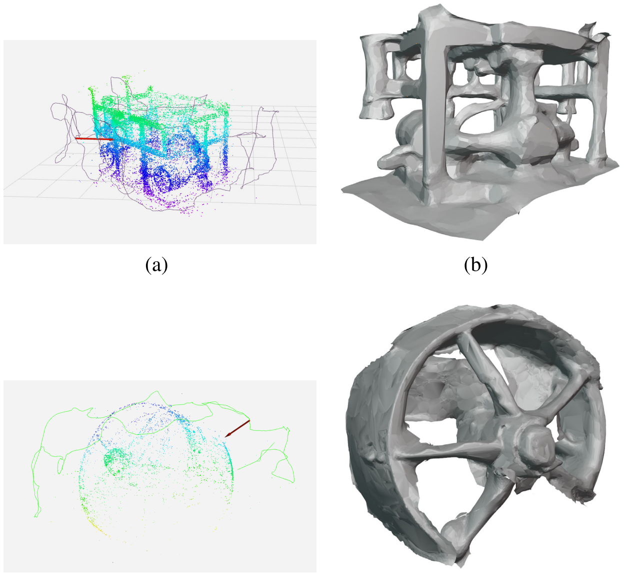

(a) (b) (c) (d)

FIGURE 11. (a) and (c) are the rescaled point clouds and trajectories from the proposed pipeline, and (b) and (c), their corresponding 3D models from the online Poisson surface estimation.

trajectory still ended close to the ground-truth. In comparison, the proposed method had very low ATE and RPE, and also ended close to the ground-truth. However, this method lost accuracy during turns.

V. APPLICATIONS

The proposed method can be used for image enhancement using depth prior, robotic navigation, or 3D reconstruction. The latter will be explored in this section in two independent inspection scenarios. The first one consists of an inspection of a subsea module, and the second, of an inspection of a ship propeller. Underwater inspections are important to assess the structure integrity. In the case of remote inspections, the operation is typically overviewed by an inspector monitoring the inspection through a transmitted visual stream. Establishing scale to the scene allows additional and automatically processed data enabling better inspection condition and assessment of the structure.

The 3D reconstruction is based on the online generated point cloud from the proposed approach, combining the camera and the sonar within the ORB-SLAM framework, ensuring the estimation of a correctly scaled model. It is performed in real-time using the inactive rescaled visual point from the modified SLAM framework. The inactive points represent the SLAM 3D points not being tracked or modified. The Poisson surface estimation was applied to obtain a set of 3D faces displayed to the operator for monitoring purposes. This method is particularly efficient for the mapping application since it can work with noisy data and misregistered points while estimating the surface fast. The 3D surface was estimated in real time to facilitate object and place recognition for the inspector enabling real time updates of the mission plan based on the findings. The results for the two scenarios are displayed in Figure 11, with both the prior point cloud and the resulting estimated surface. The geometry of the 3D objects is not exact, but provides a representative presentation with the correct scale. In the case of inspection missions, the generated model can be exported along with annotations from the inspector, making the inspection process more efficient, repeatable and accurate.

VI. CONCLUSION

A new approach to monocular SLAM estimates rescaling is presented, using a MB-FLS for scale estimation. The proposed camera-sonar combination includes estimation of individual sonar beam coverage in the optical image, enabling visual-acoustic feature matching. This allows depth ratio estimation after the application of the maximum likelihood estimation, providing continuous rescaling, and stability. The proposed pipeline was experimentally tested and the results show improvements in stability and robustness compared to known methods.

However, it was observed during the experiments that the position error tends to increase when the vehicle turns. This is likely due to the camera-sonar calibration imprecision. A calibration step in the processing pipeline would improve these errors and will be studied in the future. Also, the method would benefit from a tighter integration of the sonar in the SLAM framework, including the addition of parameters such as the speed of sound and the reprojection errors of the acoustic features. This would also enable the sonar to keep estimating the pose of the vehicle during periods of camera outages, for example when images become too blurry to keep tracking the visual features. For monitoring applications, especially for inspection missions, semantic SLAM can significantly improve scene understanding and therefore, inspection results. This will be studied in the case of ship hull inspection, using the previously developed LIACi dataset [24].

ACKNOWLEDGMENT

The authors would like to thank the members of the Applied Underwater Robotic Laboratory (AUR-Laboratory) for their experimental support.

REFERENCES

[1] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I," *IEEE Robot. Autom. Mag.*, vol. 13, no. 2, pp. 99–110, Jun. 2006.

[2] 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, Oct. 2015.

[3] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Computer Vision—ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 834–849.

[4] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. 6th IEEE ACM Int. Symp. Mixed Augmented Reality*, Nov. 2007, pp. 225–234.

[5] 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, Aug. 2018.

[6] 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, Oct. 2017.

[7] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 573–580.

[8] M. Ferrera, J. Moras, P. Trouvé-Peloux, and V. Creuze, "Real-time monocular visual odometry for turbid and dynamic underwater environments," *Sensors*, vol. 19, no. 3, p. 687, Feb. 2019.

[9] S. Xu, T. Luczynski, J. S. Willners, Z. Hong, K. Zhang, Y. R. Petillot, and S. Wang, "Underwater visual acoustic SLAM with extrinsic calibration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 7647–7652.

[10] E. Westman, A. Hinduja, and M. Kaess, "Feature-based SLAM for imaging sonar with under-constrained landmarks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 3629–3636.

[11] C. Cheng, C. Wang, D. Yang, W. Liu, and F. Zhang, "Underwater localization and mapping based on multi-beam forward looking sonar," *Frontiers Neurorobotics*, vol. 15, Jan. 2022, Art. no. 801956.

[12] S. Rahman, A. Q. Li, and I. Rekleitis, "Sonar visual inertial SLAM of underwater structures," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 5190–5196.

[13] 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, Jun. 2009.

[14] S. Bejarano, P. J. Mumby, J. D. Hedley, and I. Sotheran, "Combining optical and acoustic data to enhance the detection of Caribbean forereef habitats," *Remote Sens. Environ.*, vol. 114, no. 11, pp. 2768–2778, Nov. 2010.

[15] A. Spears, A. M. Howard, M. West, and T. Collins, "Acoustic sonar and video sensor fusion for landmark detection in an under-ice environment," in *Proc. Oceans St. John's*, Sep. 2014, pp. 1–8.

[16] Y. Raaj, A. John, and T. Jin, "3D object localization using forward looking sonar (FLS) and optical camera via particle filter based calibration and fusion," in *Proc. OCEANS MTS/IEEE Monterey*, Sep. 2016, pp. 1–10.

[17] M. Roznere and A. Q. Li, "Underwater monocular image depth estimation using single-beam echosounder," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 1785–1790.

[18] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.

[19] A. Sujiwo, N. University, T. Ando, E. Takeuchi, Y. Ninomiya, and M. Edahiro, "Monocular vision-based localization using ORB-SLAM with LiDAR-aided mapping in real-world robot challenge," *J. Robot. Mechatronics*, vol. 28, no. 4, pp. 479–490, Aug. 2016.

[20] S. J. Haddadi and E. B. Castelan, "Visual-inertial fusion for indoor autonomous navigation of a quadrotor using ORB-SLAM," in *Proc. Latin Amer. Robotic Symp., Brazilian Symp. Robot. (SBR) Workshop Robot. Educ. (WRE)*, Brazil, Nov. 2018, pp. 106–111.

[21] F. Hidalgo, C. Kahlefendt, and T. Bräunl, "Monocular ORB-SLAM application in underwater scenarios," in *Proc. OCEANS-MTS/IEEE Kobe Techno-Oceans (OTO)*, May 2018, pp. 1–4.

[22] Y. Zhang, L. Zhou, H. Li, J. Zhu, and W. Du, "Marine application evaluation of monocular SLAM for underwater robots," *Sensors*, vol. 22, no. 13, p. 4657, Jun. 2022.

[23] A. Cardaillac and M. Ludvigsen, "Marine snow detection for real time feature detection," in *Proc. IEEE/OES Auto. Underwater Vehicles Symp. (AUV)*, Sep. 2022, pp. 1–6.

[24] M. Waszak, A. Cardaillac, B. Elvesæter, F. Rødølen, and M. Ludvigsen, "Semantic segmentation in underwater ship inspections: Benchmark and data set," *IEEE J. Ocean. Eng.*, vol. 48, no. 2, pp. 462–473, Apr. 2023.

ALEXANDRE CARDAILLAC received the Bachelor of Information Technology degree from the School of Digital Innovation, Nantes, France, in 2019, and the M.Sc. degree in artificial intelligence with speech and multimodal interaction from Heriot-Watt University, Edinburgh, U.K., in 2020. He is currently pursuing the Ph.D. degree in engineering with the Department of Marine Technology, Norwegian University of Science and Technology, Trondheim, Norway.

MARTIN LUDVIGSEN (Member, IEEE) was born in 1977. He received the Ph.D. degree in underwater technology from Norges Teknisk-Naturviten Skapelige Universitet (NTNU), Trondheim, Norway, in 2010.

Since 2014, he has been a Professor with the Department of Marine Technology, NTNU, where he is currently a Co-Founder and the Manager of the Applied Underwater Laboratory (AUR-Laboratory). The AUR-Laboratory (https://www.ntnu.edu/web/aur-lab/aur-lab) facilitates research within both engineering disciplines and marine science by providing ROV, AUV, and USV operations. He has long experience at-sea both in arctic waters and in benthic environments associated with Norwegian Midocean Ridge. He has been involved in research projects both in the deep sea, the upper water column, and arctic deploying robotic underwater vehicles. His research interests include the field of underwater vehicles, including perception and interpretation of cameras and sonar data together with autonomy. Adaptive mission planning for one or more vehicles for ocean column mapping has also been a focus point for his research group.