Data Paper

The International Journal of
Robotics Research
2019, Vol. 38(14) 1549–1559
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0278364919883346
journals.sagepub.com/home/ijr

AQUALOC: An underwater dataset for visual–inertial–pressure localization

Maxime Ferrera[1,2], Vincent Creuze[2], Julien Moras[1] and Pauline Trouvé-Peloux[1]

Abstract

We present a new dataset, dedicated to the development of simultaneous localization and mapping methods for underwater vehicles navigating close to the seabed. The data sequences composing this dataset are recorded in three different environments: a harbor at a depth of a few meters, a first archeological site at a depth of 270 meters, and a second site at a depth of 380 meters. The data acquisition is performed using remotely operated vehicles equipped with a monocular monochromatic camera, a low-cost inertial measurement unit, a pressure sensor, and a computing unit, all embedded in a single enclosure. The sensors' measurements are recorded synchronously on the computing unit and 17 sequences have been created from all the acquired data. These sequences are made available in the form of ROS bags and as raw data. For each sequence, a trajectory has also been computed offline using a structure-from-motion library in order to allow the comparison with real-time localization methods. With the release of this dataset, we wish to provide data difficult to acquire and to encourage the development of vision-based localization methods dedicated to the underwater environment. The dataset can be downloaded from: http://www.lirmm.fr/aqualoc/

Keywords

Dataset, underwater robotics, monocular vision, IMU, pressure, SLAM

1. Introduction

Accurate localization is critical for mobile robotics. In open outdoor areas, it can be obtained from the global positioning system (GPS). However, in GPS-denied environments, such as indoor or beneath the sea surface, robots' position must be estimated from other sensors.

In underwater robotics, the localization problem is often solved by coupling high-grade inertial measurement units (IMUs) with compass, Doppler velocity logs (DVLs), and pressure sensors (Paull et al., 2014). Such solutions, classified as dead-reckoning (DR) localization, are highly dependent of the sensors quality and suffer from unbounded drift. While these methods can be employed quite safely for vehicles navigating in the middle of the water column (i.e., in obstacle-free areas), they are not accurate enough for navigation in cluttered areas. In such places, simultaneous localization and mapping (SLAM) methods are preferred. SLAM requires exteroceptive sensors, such as Lidar, sonar, or camera, to measure the 3D structure of the environment. From these data, the localization is estimated while a 3D map is progressively built.

Visual SLAM (VSLAM) and visual–inertial odometry (VIO) have been a hot research topic during the past decades (Cadena et al., 2016). VSLAM consists of estimating localization from visual data, possibly enhanced by complementary sensors, through the mapping of the observed scenes. In ground and aerial robotics, the availability of many public datasets, such as KITTI (Geiger et al., 2012), Malaga (Blanco et al., 2014), or EuRoC (Burri et al., 2016), to cite a few, has greatly affected the development of VSLAM methods. Recent algorithms, relying on monocular cameras (Engel et al., 2018; Forster et al., 2017; Mur-Artal et al., 2015) or on visual–inertial sensors (Leutenegger et al., 2015; Mur-Artal and Tardos, 2017; Qin et al., 2018), have shown impressive results, with centimetric localization accuracy. In underwater robotics, many operations occur near the seabed (biology, oil and gas industry, mine warfare, archeology, etc.), making visual information available. Nonetheless, under such conditions,

[1]DTIS, ONERA, Université Paris Saclay, Palaiseau, France
[2]LIRMM, Université Montpellier, CNRS, Montpellier, France

Corresponding author:
Maxime Ferrera, ONERA, Chemin de la Hunière, Chatillon, 92322, France.
Email: maxime.ferrera@gmail.com

the acquired images suffer from degradation such as turbidity, backscattering, and illumination issues, owing to the properties of the medium. These poor imaging conditions must be accounted for in the development of underwater VSLAM or VIO systems, thus preventing use of the previously cited algorithms (Quattrini Li et al., 2017; Weidner et al., 2017; Zhang et al., 2018). Some previous works have investigated the use of monocular cameras for underwater localization Burguera et al. (2015); Ferrera et al. (2019), sometimes coupled to low-cost IMUs and pressure sensors (Creuze, 2017; Shkurti et al., 2011), sonar (Rahman et al., 2018), or even as a means of detecting loop-closures in DR systems (Kim and Eustice, 2013). However, the limited number of public datasets dedicated to this localization challenge prevent a fair comparison of these methods on common data. Moreover, the fact that these data are difficult to acquire, because of the required equipment and logistics, limits the development of new methods. Bender et al. (2013) proposed a dataset containing the measurements of navigational sensors, stereo cameras, and a multibeam sonar. Mallios et al. (2017) released another dataset dedicated to localization and mapping in an underwater cave from sonar measurements. Images acquired by a monocular camera are also given for the detection of cones placed precisely in order to have a means of estimating drift. However, in both datasets, the acquisition rate of the cameras is too low (<10 Hz) for most VSLAM and VIO methods. Duarte et al. (2016) created a synthetic dataset simulating the navigation of a vehicle in an underwater environment and containing monocular cameras measurements at a framerate of 10 Hz. Many public datasets have also been made available by the oceanography community through national websites (see https://www.data.gov/ or http://www.marine-geo.org). However, these datasets have not been gathered with the aim of providing data suitable for VSLAM or VIO and often lack essential information such as the calibration of their sensors' setup.

In this article, we present AQUALOC, a new dataset aimed at the development of VSLAM and VIO methods dedicated to the underwater environment. The dataset sequences have been recorded using acquisition systems composed of a monochromatic camera, a microelectromechanical system (MEMS)-based IMU, a pressure sensor, and a computing unit for synchronous recordings. These acquisition systems have been embedded on ROVs equipped with lighting systems and navigating close to the seabed. The recorded video sequences exhibit the typical visual degradation induced by the underwater environment such as turbidity, backscattering, shadows, and strong illumination shifts caused by the artificial lighting systems. Three different sites have been explored to create the dataset: a harbor and two archeological sites. The recording of the sequences occurred at different depths, going from a few meters, for the harbor, to several hundred meters, for the archeological sites. The provided video sequences are hence highly diversified in terms of scenes (low-textured areas, very texture repetitive areas, etc.) and of scenarios (exploration, photogrammetric surveys, manipulations, etc.). As the acquisition of ground truth is very difficult in natural underwater environments, we have used the state-of-the-art structure-from-motion (SfM) library Colmap (Schönberger and Frahm, 2016) to compute comparative baseline trajectories for each sequence. Colmap processes the sequences offline and performs a 3D reconstruction to estimate the positions of the camera. This 3D reconstruction is done by matching exhaustively all the images composing a sequence, which allows the detection of many loop closures and, hence, the computation of accurate trajectories, assessed by low average reprojection errors. Along with the computed trajectories, we also provide the list of matched images for each sequence, which could be used to evaluate relocalization or loop-closure detection methods. We further include statistics on the 3D reconstruction to assess their accuracy.

With the release of this dataset, we provide to the community the opportunity to work on data that are difficult to acquire. Indeed, the logistics (ship availability) and the required equipment (deep-sea compliant underwater vehicles and sensors), as well as regulations (official authorizations), can be a barrier preventing possible works on this topic. We are convinced that the availability of this dataset will increase the development of algorithms dedicated to the underwater environment. Both raw and ROS bag formatted field data are provided along with the full calibration of the sensors (camera and IMU). Moreover, the provided comparative baseline makes this dataset suitable for benchmarking VSLAM and VIO algorithms.

The rest of this article is organized as follows. First, we present the design of the acquisition systems used and the calibration procedures employed. Then, an overview of the dataset is given and the acquisition conditions on each site are detailed, highlighting the associated challenges for visual localization. Next, the processing of the data sequences to create a baseline are described. Finally, we detail how the dataset is organized and in which way the data are formatted.

2. The acquisition systems

In order to acquire the sequences of the dataset, we have designed two similar underwater systems. These acquisition systems have been designed to allow the localization of underwater vehicles from a minimal set of sensors in order to be as cheap and as versatile as possible. Both systems are equipped with a monochromatic camera, a pressure sensor, a low-cost MEMS–IMU, and an embedded computer. The camera is placed behind an acrylic dome to minimize the distortion effects induced by the difference between water and air refractive indices. The image acquisition rate is 20 Hz. The IMU delivers measurements from a three-axis accelerometer, three-axis gyroscope, and three-axis magnetometer at 200 Hz. The embedded computer is a Jetson TX2 running Ubuntu 16.04 and is used to record

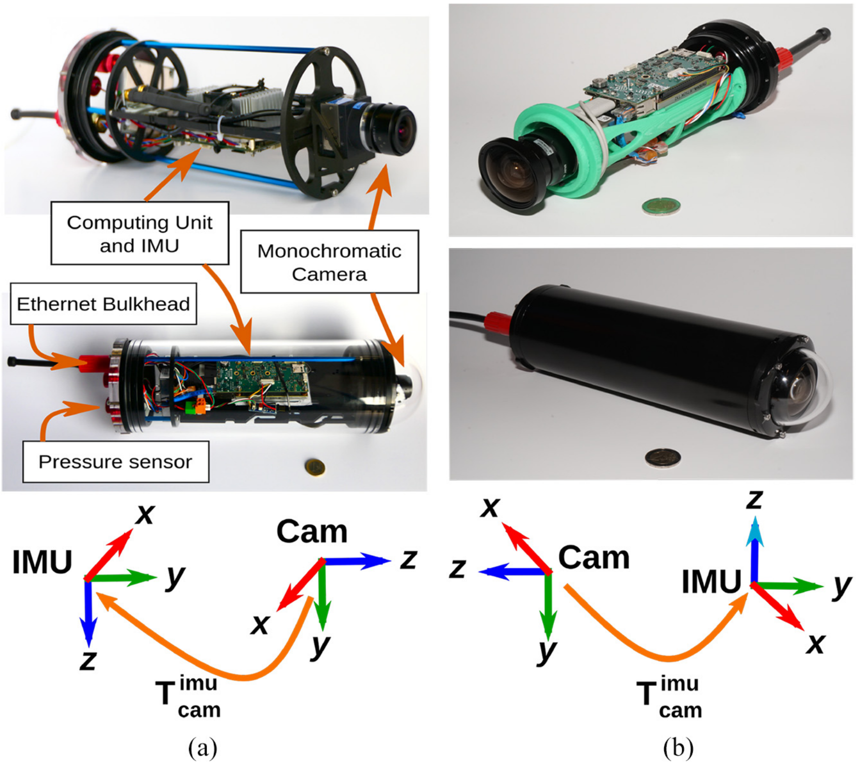


Fig. 1. The acquisition systems equipped with a monocular monochromatic camera, a pressure sensor, an IMU, and a computer along with the sensors' reference frames: (a) System A; (b) System B.

synchronously the sensors' measurements thanks to the ROS middleware. The Jetson TX2 is equipped with a carrier board embedding the mentioned MEMS-IMU and a 1 To NVME SSD to directly store the sensors measurements, thus avoiding any bandwidth or package loss issue. An advantage of the self-contained systems that we have developed, is that they are independent of any robotic architecture and can, thus, be embedded on any kind of remotely operated vehicle (ROV) or autonomous underwater vehicle (AUV). The interface can either be Ethernet or a serial link, depending on the host vehicle's features.

To record data at different depths, we have designed two systems that we will refer to as "System A" and "System B". These systems have the same overall architecture, but they differ on the camera model, the pressure sensor type, and the diameter and material of the enclosure. System A (Figure 1a) is designed for shallow waters and was used to acquire the sequences in the harbor. Its camera has been equipped with a wide-angle lens, which can be modeled by the fisheye distortion model. The pressure sensor is rated for 30 bars and delivers depth measurements at a maximum rate of 10 Hz. System A is protected by an acrylic enclosure, rated for a depth of 100 meters. System B (Figure 1b) was used to record the sequences on the archeological sites at larger depths. Its camera has a slightly lower field of view and the lens can be modeled by the radial-tangential distortion model. It embeds a pressure sensor rated for 100 bars delivering depth measurements at 60 Hz. Its enclosure is made of aluminum and is 400 meters depth rated. The technical details about both systems and their embedded sensors are given in Table 1.

Each camera–IMU setup has been cautiously calibrated to provide the intrinsic and extrinsic parameters required to use it for localization purpose. We have used the toolbox Kalibr (Furgale et al., 2012, 2013) along with an apriltag target to compute all the calibration parameters.

The cameras calibration step allows an estimate of the focal lengths, principal points, and distortion coefficients to be obtained. These parameters can then be used to undistort the captured images and to model the image formation pipeline, with the following notation:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \Pi_{\mathbf{K}}\left(\mathbf{R}_{\mathrm{w}}^{\mathrm{cam}}\mathbf{X}_{\mathrm{w}} + \mathbf{t}_{\mathrm{w}}^{\mathrm{cam}}\right) \tag{1}$$

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} f_x \cdot \frac{x_{\mathrm{cam}}}{z_{\mathrm{cam}}} + c_x \\ f_y \cdot \frac{y_{\mathrm{cam}}}{z_{\mathrm{cam}}} + c_y \end{bmatrix} = \Pi_{\mathbf{K}}(\mathbf{X}_{\mathrm{cam}}) \tag{2}$$

$$\text{with } \mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{X}_{\mathrm{cam}} = \begin{bmatrix} x_{\mathrm{cam}} \\ y_{\mathrm{cam}} \\ z_{\mathrm{cam}} \end{bmatrix}$$

where $\Pi_{\mathbf{K}}(\cdot)$ denotes the projection: $\mathbb{R}^3 \mapsto \mathbb{R}^2$, $\mathbf{K}$ is the calibration matrix, $\mathbf{X}_{\mathrm{w}} \in \mathbb{R}^3$ is the position of a 3D landmark in the world frame, $\mathbf{R}_{\mathrm{w}}^{\mathrm{cam}} \in SO(3)$ and $\mathbf{t}_{\mathrm{w}}^{\mathrm{cam}} \in \mathbb{R}^3$ denote the rotational and translational components of the transformation from the world frame to the camera frame, $\mathbf{X}_{\mathrm{cam}} \in \mathbb{R}^3$ is the position of a 3D landmark in the camera frame, f_x and f_y denotes the focal lengths, and (c_x, c_y) is the

Table 1. Technical details about the acquisition systems.

	Camera sensor	**UEye - UI-1240SE**
System A (Harbor sequences)	Resolution	640 × 512 px
	Sensor	Monochromatic
	Frames per second	20 fps
	Lens	**Kowa LM4NCL C-Mount**
	Focal length	3.5mm
	Pressure Sensor	**MS5837 - 30BA**
	Depth range	0–290 m
	Resolution	0.2 mbar
	Output frequency	5–10 Hz
	Inertial Measurement Unit	**MEMS - MPU-9250**
	Gyroscope frequency	200 Hz
	Accelerometer frequency	200 Hz
	Magnetometer frequency	200 Hz
	Embedded Computer	**Nvidia - Tegra Jetson TX2**
	Carrier board	Auvidea J120 - IMU
	Storage	NVME SSD 1 To
	Housing	**4” Blue Robotics Enclosure**
	Enclosure	33.4 cm × 11.4 cm
	Enclosure Material	Acrylic
	Dome	4” Blue Robotics Dome End Cap
System B (Archeological sequences)	**Camera sensor**	**UEye - UI-3260CP**
	Resolution	968 × 608 px
	Sensor	Monochromatic
	Frames per second	20 fps
	Lens	**Kowa LM6NCH C-Mount**
	Focal length	6 mm
	Pressure sensor	**Keller 7LD - 100BA**
	Depth range	0–990 m
	Resolution	3 mbar
	Output frequency	60 Hz
	Inertial measurement unit	***Same as System A***
	Embedded computer	***Same as System A***
	Housing	**3” Blue Robotics Enclosure**
	Enclosure	25.8 cm × 8.9 cm
	Enclosure Material	Aluminum
	Dome	3” Blue Robotics Dome End Cap

principal point of the camera. The distorted pixel coordinates (u, v) are the projection of $\mathbf{X}_{\text{cam}}$ into the image frame that can be further undistorted using the distortion model and coefficients.

As these parameters are medium dependant, the calibration has been performed in water to account for the additional distortion effects at the dome’s level. The results of the calibration of the fisheye camera can be seen in Figure 2.

The camera–IMU setup calibration allows the extrinsic parameters of the setup to be estimated, that is the relative position of the camera with respect to the IMU, and the time delay between camera’s and IMU’s measurements. This relative position is represented by a rotation matrix $\mathbf{R}_{\text{cam}}^{\text{imu}}$ and a translation vector $\mathbf{t}_{\text{cam}}^{\text{imu}}$. Camera and IMU poses relate to each other through:

$$\mathbf{T}_{\text{cam}}^{\text{w}} = \mathbf{T}_{\text{imu}}^{\text{w}} \mathbf{T}_{\text{cam}}^{\text{imu}} \tag{3}$$

$$\text{with } \mathbf{T}_{\text{cam}}^{\text{imu}} \doteq \begin{bmatrix} \mathbf{R}_{\text{cam}}^{\text{imu}} & \mathbf{t}_{\text{cam}}^{\text{imu}} \\ 0_{1\times 3} & 1 \end{bmatrix} \in \mathbb{R}^{4\times 4}$$

$$\left(\mathbf{T}_{\text{cam}}^{\text{w}}\right)^{-1} = \mathbf{T}_{\text{w}}^{\text{cam}} \doteq \begin{bmatrix} \mathbf{R}_{\text{w}}^{\text{cam}} & \mathbf{t}_{\text{w}}^{\text{cam}} \\ 0_{1\times 3} & 1 \end{bmatrix} \in \mathbb{R}^{4\times 4}$$

where $\mathbf{R}_{\text{cam}}^{\text{imu}} \in SO(3)$, $\mathbf{t}_{\text{cam}}^{\text{imu}} \in \mathbb{R}^3$, $\mathbf{T}_{\text{cam}}^{\text{w}} \in SE(3)$, $\mathbf{T}_{\text{w}}^{\text{cam}} \in SE(3)$, $\mathbf{T}_{\text{cam}}^{\text{imu}} \in SE(3)$, and $\mathbf{T}_{\text{imu}}^{\text{w}} \in SE(3)$. Here $\mathbf{T}_{\text{cam}}^{\text{w}}$ and $\mathbf{T}_{\text{imu}}^{\text{w}}$ respectively represent the poses of the camera and of the body, with respect to the world frame. $\mathbf{T}_{\text{w}}^{\text{cam}}$ is the inverse transformation of $\mathbf{T}_{\text{cam}}^{\text{w}}$ and $\mathbf{T}_{\text{cam}}^{\text{imu}}$ is the transformation from the camera frame to the IMU frame.

Before estimating these extrinsic parameters, the IMU noise model parameters have been derived from an Allan standard deviation plot, obtained by recording the gyroscope and accelerometer measurements for several hours, while keeping the IMU still. These noise parameters are then fed into the calibration algorithms to model the IMU measurements. As these parameters (IMU noises, camera–IMU relative transformation, and measurements time delay) are independent of the medium (air or water), they have been estimated in air. Performing this calibration step in air allowed the fast motions required to correlate the IMU to the camera measurements to be easily performed.

All the calibration results are included in the dataset, that is the cameras’ models (including the intrinsic parameters and the distortion coefficients), the IMUs’ noise parameters, the relative transformation from the camera to the IMU, and the time delay between the camera and the IMU measurements.

3. Dataset overview

As explained in Section 2, System A was used to record the shallow harbor sequences, whereas System B was used on the two deep archeological sites. We propose a total of 17 sequences: 7 recorded in the harbor, 4 on the first archeological site, and 6 on the second site. As each of these environments is in some ways different from the others, we describe the sequences recorded in each environment separately. Table 2 summarizes the specificities of each data sequence. Note that, for each sequence, the starting and ending points are approximately the same. In most of the sequences, there are closed loops along the performed trajectories. Some sequences also slightly overlap, which can be useful for the development of relocalization features.

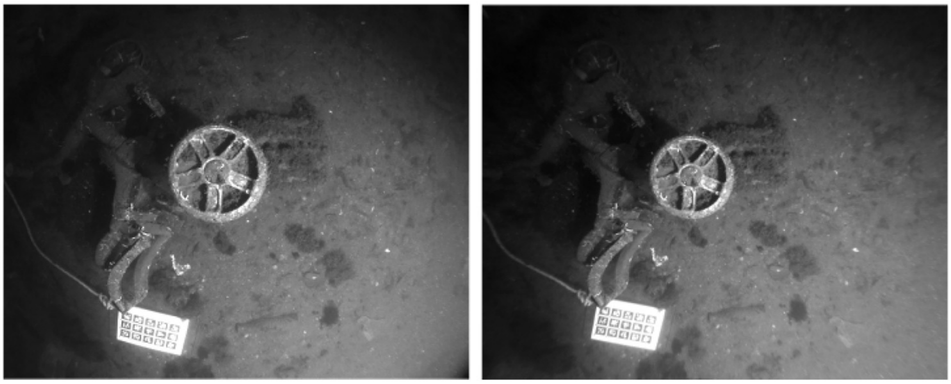


Fig. 2. Distortion effects removal from Kalibr calibration on one of the harbor sequences. Left: raw image. Right: undistorted image.

Table 2. Details on all the AQUALOC sequences and their associated visual disturbances.

Site	Sequence	Duration	Length	Visual disturbances					
				Turbidity	Collisions	Backscattering	Sandy clouds	Dynamics	Robotic arm
Harbor (depth ≈ 4 m) Acquired by System A, embedded on a lightweight ROV	#01	3’49”	39.3 m	X	—	X	—	—	—
	#02	6’47”	75.6 m	X	—	X	—	—	—
	#03	4’17”	23.6 m	X	—	X	—	—	—
	#04	3’26”	55.8 m	X	X	X	—	—	—
	#05	2’52”	28.5 m	X	—	X	—	—	—
	#06	2’06”	19.5 m	X	—	X	—	—	—
	#07	1’53”	32.9 m	X	X	X	—	—	—
First archeological site (depth ≈ 270 m) Acquired by System B, embedded on a medium workclass ROV	#01	14’39”	32.4 m	X	—	X	X	X	X
	#02	7’29”	64.3 m	X	—	X	X	X	—
	#03	5’16”	10.7 m	X	—	X	X	—	—
Second archeological site (depth ≈380 m) Acquired by System B embedded on a medium workclass ROV	#04	11’09”	18.1 m	X	—	X	X	X	X
	#05	3’19”	42.0 m	X	—	X	—	X	—
	#06	2’49”	31.8 m	X	—	X	—	X	—
	#07	9’29”	122.1 m	X	—	X	—	X	—
	#08	7’49”	41.2 m	X	—	X	—	X	—
	#09	5’49”	65.4 m	X	—	X	—	X	—
	#10	11’54”	83.5 m	X	—	X	—	X	—

3.1. Harbor sequences

The harbor sequences were recorded in April 2018. System A was embedded on the lightweight ROV *Dumbo* (*DRASSM-LIRMM*) with the camera facing downward, as shown in Figure 3. The ROV was navigating at a depth of 3–4 meters over an area of around 100 m^2. Although the Sun illuminates this shallow environment, a lighting system was used in order to increase the signal-to-noise ratio of the images acquired by the camera. The explored area was mostly planar but the presence of several big objects made it a real 3D environment, with significant relief.

For each sequence, loops are performed and an apriltag calibration target is used as a marker for starting and ending points. On these sequences, vision is mostly degraded by light absorption, strong illumination variations, and backscattering. In two sequences, visual information even becomes unavailable for a few seconds because of collisions with surrounding objects. Another challenge is the presence of areas with seagrass moving because of the swell. Moreover, the ROV is sensitive to waves and tether disturbances, which results in roll and pitch variations.

3.2. Archeological sites sequences

The archeological sites sequences were recorded in the Mediterranean sea, off Corsica’s shore. System B, designed

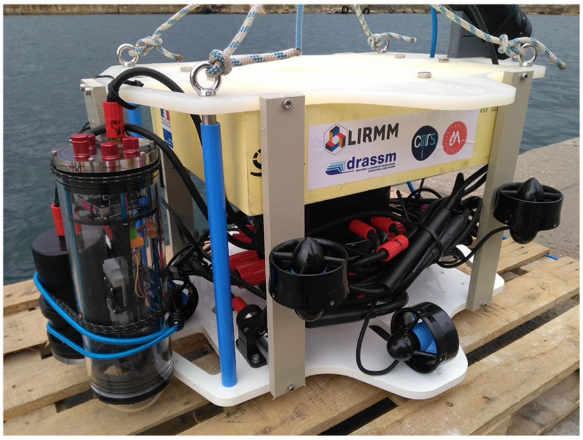


Fig. 3. The ROV *Dumbo* and the acquisition System A, used to record the harbor sequences.

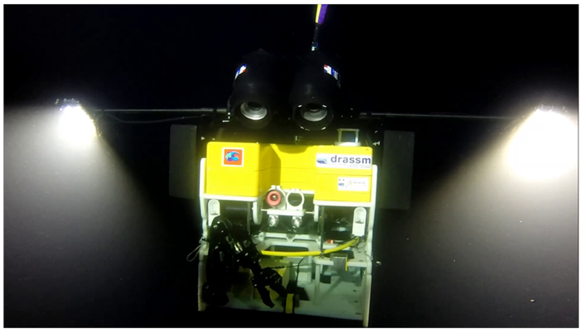
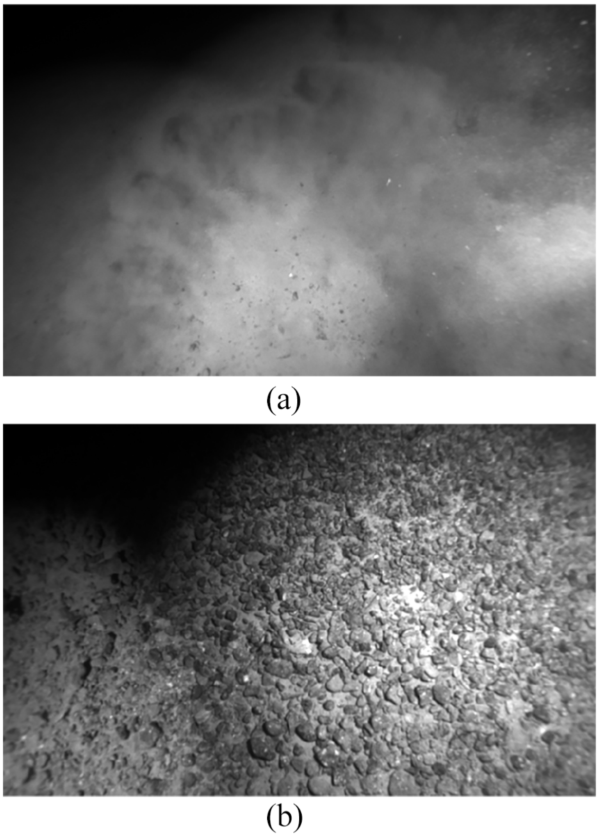


Fig. 4. The ROV Perseo, used on the archeological sites. *(Credit: F. Osada - DRASSM / Images Explorations.)*



(a)

(b)

Fig. 5. Images acquired on the first archeological site (depth: 270 m): (a) sandy cloud; (b) texture repetitive area.

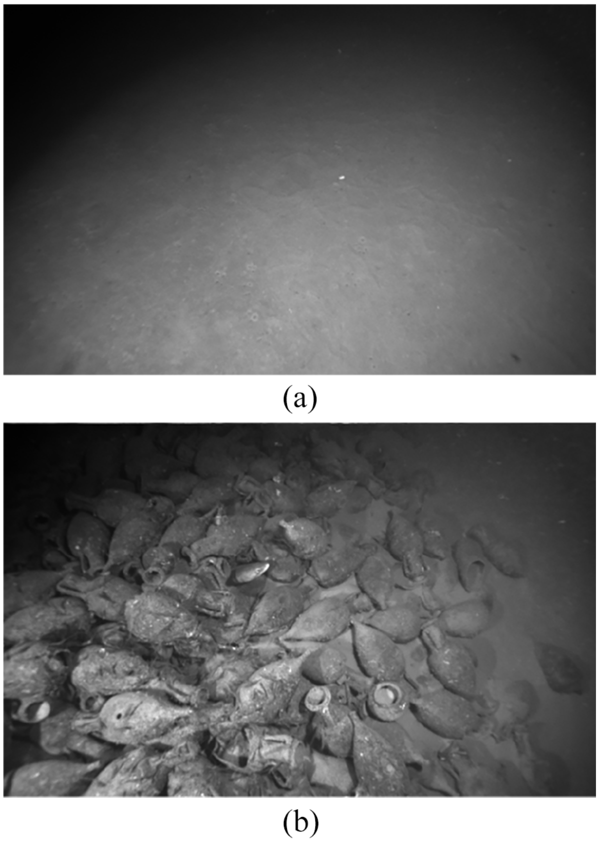
for deep water, was embedded on the *Perseo* ROV (*Copetech SM Company*) displayed in Figure 4. In the way it was attached to the ROV, the camera viewing direction made a small angle with the vertical line (≈20–30°). *Perseo* is equipped with two powerful led lights (250,000 lumens each) and with two robotics arms for manipulation purposes. As localization while manipulating objects is valuable information, to grab an artifact for instance, in some sequences the robotic arms are in the camera's field of view. A total of 10 sequences have been recorded on these sites, with 3 sequences taken on the first site and 7 on the second site.

The first archeological site explored was located at a depth of approximately 270 meters and hosted the remains of an antique shipwreck. Hence, this site is mostly planar and presents mainly repetitive textures, owing to numerous small rocks that were used as ballast in this antique ship (Figure 5a). These sequences are affected by turbidity and moving sand particles, increasing backscattering and creating sandy clouds (Figure 5b). These floating particles are stirred up from the seabed by the water flows of the ROV's thrusters and lead to challenging visual conditions. A shadow is also omnipresent in these sequences in the left corner of the recorded images, owing to the limits of the lighting system.

The second visited archeological site was located at a depth of approximately 380 meters. On this site, a hill of amphorae is present (Figure 6b), the top of which is culminating a few meters above the surrounding seabed level. During these sequences, the ROV was mainly operated for manipulation and photogrammetry purposes. While the amphorae present a highly textured surface, the ROV was also hovering low-textured sandy areas around the hill of amphorae (Figure 6a). Owing to the presence of these amphorae, marine wildlife has been growing on this site. Hence, the environment is quite dynamic, with many fish entering the field of view of the camera and many shrimp moving in the vicinity of the amphorae. In one of the sequences, both arms move in front of the camera. Otherwise, the visual degradation is the same as on the first site.

4. Comparative baseline

As the acquisition of a ground truth is very difficult in natural underwater environments, we have used the state-of-the-art SfM software Colmap (Schönberger and Frahm, 2016) to offline compute a 3D reconstruction for



(a)

(b)

Fig. 6. Images acquired on the second archeological site (depth: 380 m): (a) low texture area; (b) hill of amphorae.

each sequence and extract a reliable trajectory from it. By setting the features extraction parameters very low, we were able to extract enough scale-invariant feature transform (SIFT) features (Lowe, 2004) to robustly match the images of each sequence. Performing a matching of the images in an exhaustive way, that is trying to match each image to all the others, allows a reliable trajectory reconstruction to be obtained as many closed loops can be found (Figure 7). In Table 3, we provide statistics for each sequence about Colmap's 3D reconstructions to highlight the reliability of the reconstructed models. These statistics include the number of images used, the number of estimated 3D points, the average track length of each 3D points (i.e., the number of images observing a given 3D point) and the average reprojection error. The high average track lengths for each sequence (going from 6.7 to more than 20) assess the accuracy of the 3D points' estimation as it leads to high redundancy in the bundle adjustment steps of the reconstruction. Moreover, given these high track lengths, the average reprojection error is a good indicator of the overall quality of a SfM 3D model and for each of the sequences this error is below 0.9 pixels.

The extracted trajectories have been scaled using the pressure sensor measurements and hence provide metric positions. Although these trajectories cannot be considered as being perfect ground truths, we believe that it provides a fair baseline to evaluate and compare online localization methods. Evaluation of such methods can be done using the standard relative pose error (RPE) and absolute trajectory error (ATE) metrics (Sturm et al., 2012).

Furthermore, we have made available the list of overlapping images (i.e., matching) according to Colmap for each sequence. These files could hence be used to evaluate the efficiency of loop-closure or image-retrieval methods.

5. Data sequences format

As explained in the introduction, the sequences are all available as ROS bags and as raw data. The dataset is split into two folders, one for the harbor sequences and the other for the archeological ones.

The dataset repository architecture is as follows.

```
Harbor_site_sequences/
├─ Calibration files/
│  ├─ camera_calib.txt
│  ├─ imu_camera_calib.txt
│  └─ imu_noises.txt
├─ ground truth files/
│  ├─ colmap_traj_sequence_X.txt
│  ├─ ...
│  ├─ colmap_detected_loop_sequence_X.txt
│  └─ ...
├─ sequence_X_bag.tar.gz/
│  ├─ sequence_X.bag
│  └─ ...
└─ sequence_X_raw_data.tar.gz/
   ├─ imu.csv
   ├─ mag.csv
   ├─ images.csv
   └─ images/
      ├─ frameXXXXX.png
      └─ ...
```

The archeological sites sequences do not appear here but are organized in exactly the same manner.

The calibration files are given in the output format of Kalibr (Furgale et al., 2012, 2013).

The trajectories computed by Colmap for each sequence are available as text files and contain the pose in a

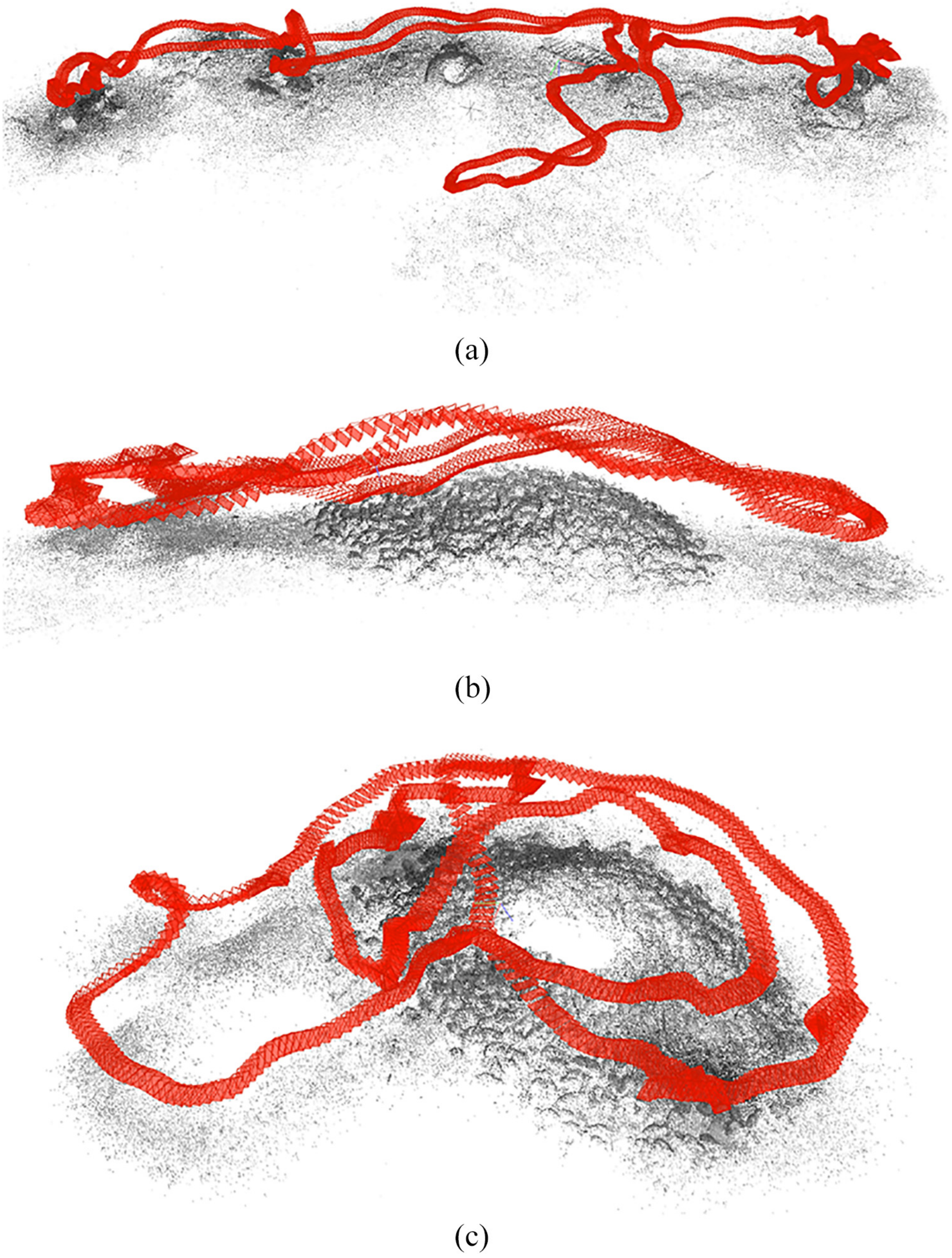


Fig. 7. Examples of trajectories reconstructed with Colmap: (a) Harbor #02; (b) archeological site #07; (c) archeological site #10.

translation-quarternion form. The format of these files is as follows.

#Frame	*tx*	*ty*	*tz*	*qx*	*qy*	*qz*	*qw*
0.	–1.88	2.41	–0.47	0.01	0.06	0.14	0.91
20.	–1.83	2.35	–0.46	0.05	0.64	0.14	0.99
40.	–1.80	2.10	-0.34	0.04	0.58	0.12	0.98
…							

The files containing the loop closures detected by Colmap provide information in the following format:

```
1,1,0,0,1
1,1,1,0,0
0,1,1,0,0
0,0,0,1,1
1,0,0,1,1
…
```

where a 1 indicates an overlap between row *i* and column *j*, with *i* and *j* standing for the frame numbers. Note that

Table 3. Colmap trajectories reconstruction statistics. The number of provided images, the number of reconstructed 3D points, the mean tracking length for the 3D points, and the mean reprojection error for the 3D reconstruction are given for each sequence.

	Harbor sequences							Archeological sites sequences									
	#01	#02	#03	#04	#05	#06	#07	#01	#02	#03	#04	#05	#06	#07	#08	#09	#10
Number of used images	918	1,590	1,031	770	692	508	447	880	445	311	637	200	170	569	470	350	715
Number of 3D points	112,659	305,783	355,130	194,407	236,845	188,807	181,964	196,857	174,514	160,531	249,048	42,877	45,799	251,620	237,882	114,814	329,686
Mean tracking length	14.9	13.2	17.2	9.7	10.7	12.1	9.5	23.5	12.6	8.4	8.5	7.6	6.7	7.4	9.1	7.9	9.2
Mean reprojection error (px)	0.896	0.816	0.713	0.715	0.688	0.733	0.846	0.746	0.621	0.474	0.673	0.601	0.569	0.645	0.616	0.660	0.661

only a subset of the images has been used to compute the offline reconstruction with Colmap (1 image out of 5 for the harbor sequences and 1 out 20 for the archeological sequences). Therefore, the frame number given in these ground-truth files is the number of the corresponding frame in the full sequence.

Concerning the bag files, each sequence is stored in a separate bag containing the following topics.

- **/camera/image_raw**: Images recorded from the camera.
- **/camera/camera_info**: Image width and height info.
- **/rtimulib_node/imu**: Accelerometer and gyroscope measurements.
- **/rtimulib_node/mag**: Magnetometer measurements.
- **/barometer_node/pressure**: Pressure measurements in millibars.
- **/barometer_node/depth**: Depth measurements in meters.
- **/barometer_node/temperature**: Pressure sensor temperature measurements.

In their raw format, each sequence contains the following data.

- **images/**: The directory containing the sequence images.
- **frameXXXXX.png**: The images recorded from the camera.
- **images.csv**: The timestamps related to each image of the sequence.
- **imu.csv**: The accelerometer and gyroscope measurements and their timestamps.
- **mag.csv**: The magnetometer measurements and their timestamps.
- **depth.csv**: The pressure measurements converted into meters and their timestamps.

For each *csv* file, the first row starts with a # and then gives the name of the different fields along with their related measurement units in squared brackets. The following rows contain the values of the measurements. In all these files, the first field is the acquisition timestamp of the measurements. For instance, the depth.csv files appear as follows:

```
#timestamp [ns], depth [m]
1542828791719540119,271.988866935
1542828791735507011,272.01910918
…
```

6. Conclusion

In this article, we have presented a new dataset of subsea monocular video sequences synchronized with inertial and pressure measurements. This dataset is intended for encouraging the development of localization methods for underwater robots navigating close to the seabed. The sequences have been recorded from ROVs in three different environments at different depths: a harbor at a depth of 4 meters, a first archeological site at a depth of 270 meters, and a second site at a depth of 380 meters. The diversity of the recorded environments allowed video sequences to be captured with different visual perturbations typical in underwater scenarios. For each sequence, trajectories have been computed offline using a SfM library and are provided as a baseline for performance comparisons of localization methods. The datasets are available both as ROS bags and as raw data. In future work, we plan to perform new acquisition missions in different underwater environments in order to augment this dataset and increase its diversity.

Acknowledgements

The experiments conducted on the archeological sites have been done in the French waters, in the framework of a research campaign of the DRASSM (French Department of Underwater Archeology - Ministry of Culture) in accordance with the international regulation (UNESCO Convention on the Protection of the Underwater Cultural Heritage, 2001). The authors are grateful to the DRASSM for its logistical support. The authors would also like to thank Anthelme Bernard-Brunel for his help in the design of the acquisition systems and Abderrahmane Kheddar for his helpful remarks.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the CNRS (Mission pour l'interdisciplinarité - Instrumentation aux limites 2018 - Aqualoc project) and Région Occitanie (ARPE Pilotplus project).

ORCID iDs

Maxime Ferrera https://orcid.org/0000-0002-1024-4151
Vincent Creuze https://orcid.org/0000-0002-6813-8562

References

Bender A, Williams SB and Pizarro O (2013) Autonomous exploration of large-scale benthic environments. In: *2013 IEEE International Conference on Robotics and Automation (ICRA)*, Karlsruhe, Germany, pp. 390–396.

Blanco JL, Moreno FA and Gonzalez-Jimenez J (2014) The Málaga Urban Dataset: High-rate stereo and lidars in a realistic urban scenario. *The International Journal of Robotics Research* 33(2): 207–214.

Burguera A, Bonin-Font F and Oliver G (2015) Trajectory-based visual localization in underwater surveying missions. *Sensors* 15(1): 1708–1735.

Burri M, Nikolic J, Gohl P, et al. (2016) The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research* 35(10): 1157–1163.

Cadena C, Carlone L, Carrillo H, et al. (2016) Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics* 32(6): 1309–1332.

Creuze V (2017) Monocular odometry for underwater vehicles with online estimation of the scale factor. In: *IFAC 2017 World Congress*, Toulouse, France.

Duarte AC, Zaffari GB, da Rosa RTS, Longaray LM, Drews P and Botelho SSC (2016) Towards comparison of underwater SLAM methods: An open dataset collection. In: *OCEANS 2016 MTS/IEEE*, Monterey, CA, USA, pp. 1–5.

Engel J, Koltun V and Cremers D (2018) Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(3): 611–625.

Ferrera M, Moras J, Trouvé-Peloux P and Creuze V (2019) Real-time monocular visual odometry for turbid and dynamic underwater environments. *Sensors* 19(3): E687.

Forster C, Zhang Z, Gassner M, Werlberger M and Scaramuzza D (2017) SVO: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics* 33(2): 249–265.

Furgale P, Barfoot T and Sibley G (2012) Continuous-time batch estimation using temporal basis functions. In: *2012 IEEE International Conference on Robotics and Automation (ICRA)*, St. Paul, MN, USA, pp. 2088–2095.

Furgale P, Rehder J and Siegwart R (2013) Unified temporal and spatial calibration for multi-sensor systems. In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, pp. 1280–1286.

Geiger A, Lenz P and Urtasun R (2012) Are we ready for autonomous driving? The KITTI Vision Benchmark Suite. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, pp. 3354–3361.

Kim A and Eustice RM (2013) Real-time visual SLAM for autonomous underwater hull inspection using visual saliency. *IEEE Transactions on Robotics* 29(3): 719–733.

Leutenegger S, Lynen S, Bosse M, Siegwart R and Furgale P (2015) Keyframe-based visual–inertial odometry using nonlinear optimization. *The International Journal of Robotics Research* 34(3): 314–334.

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2): 91–110.

Mallios A, Vidal E, Campos R and Carreras M (2017) Underwater caves sonar data set. *The International Journal of Robotics Research* 36(12): 1247–1251.

Mur-Artal R, Montiel JMM and Tardós JD (2015) ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics* 31(5): 1147–1163.

Mur-Artal R and Tardos JD (2017) Visual–Inertial monocular SLAM with map reuse. *IEEE Robotics and Automation Letters* 2(2): 796–803.

Paull L, Saeedi S, Seto M and Li H (2014) AUV navigation and localization: A review. *IEEE Journal of Oceanic Engineering* 39(1): 131–149.

Qin T, Li P and Shen S (2018) VINS-Mono: A robust and versatile monocular visual–inertial state estimator. *IEEE Transactions on Robotics* 34(4): 1004–1020.

Quattrini Li A, Coskun A, Doherty SM, et al. (2017) Experimental comparison of open source vision-based state estimation algorithms. In: *2016 International Symposium on Experimental Robotics*, pp. 775–786.

Rahman S, Li AQ and Rekleitis I (2018) Sonar visual inertial SLAM of underwater structures. In: *2018 IEEE International*

Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, pp. 1–7.

Schönberger JL and Frahm JM (2016) Structure-from-motion revisited. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA.

Shkurti F, Rekleitis I, Scaccia M and Dudek G (2011) State estimation of an underwater robot using visual and inertial information. In: *2011 IEEE/RSJ Intelligent Robots and Systems (IROS)*, San Francisco, CA, USA.

Sturm J, Engelhard N, Endres F, Burgard W and Cremers D (2012) A benchmark for the evaluation of RGB-D SLAM systems. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vilamoura, Portugal, pp. 573–580.

Weidner N, Rahman S, Li AQ and Rekleitis I (2017) Underwater cave mapping using stereo vision. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, pp. 5709–5715.

Zhang J, Ila V and Kneip L (2018) Robust visual odometry in underwater environment. In: *2018 OCEANS MTS/IEEE Kobe Techno-Oceans (OTO)*, Kobe, Japan, pp. 1–9.