Article

Not peer-reviewed version

Feature-based Place Recognition Using Forward Looking Sonar

Ana Rita Gaspar * and Aníbal Matos

Posted Date: 27 October 2023

doi: 10.20944/preprints202310.1759.v1

Keywords: appearance-based navigation; autonomous underwater vehicles; binary features; BoW; FLS; inspection; loop closure; stonefish

Article

Feature-Based Place Recognition Using Forward Looking Sonar

Ana Rita Gaspar [1,*] and Aníbal Matos [2]

[1] FEUP, INESC TEC, Porto, Portugal; argaspar@inesctec.pt
[2] FEUP, INESC TEC, Porto, Portugal; anibal.matos@inesctec.pt
* Correspondence: argaspar@inesctec.pt

Abstract: Some structures in the port environment still need to be controlled regularly. However, these scenarios present a significant challenge for accurate estimation of the vehicle's position and then detecting similar images. In these scenarios, visibility can be poor which makes the place recognition a hard task as the visual appearance of a local can be compromised. In these operational conditions, imaging sonars are a promising solution. The quality of the acquired images is affected by some factors, but they do not suffer from haze, which is an advantage. In this way, this work proposes a purely acoustic approach to similar image detection based on forward-looking sonar (FLS) data to solve the perception problems in port facilities. In order to simplify the variation of environment parameters and sensor configurations, and given the need for online data for these applications, a port environment was replicated with an ocean simulator: Stonefish. Therefore, experiments were conducted with preconfigured user trajectories to simulate mission inspections, i.e., near structures. The place recognition approach performs better than the results obtained from optical images. The proposed method provides a good compromise in terms of distinctiveness and achieves 87.5% of performance when appropriate constraints and assumptions are made.

Keywords: appearance-based navigation; autonomous underwater vehicles; binary features; BoW; FLS; inspection; loop closure; stonefish

1. Introduction

Port facilities include various structures such as quay walls and adjacent piles that must be inspected for corrosion and damage. Inspection of these structures is usually done by divers and remotely operated vehicles (ROVs). However, this work is dangerous, and ROVs use a cable, which can limit and complicate work in these semi-structured environments. Therefore, autonomous underwater vehicles (AUVs) have been used for these tasks. To do so, they must navigate accurately and to recognize revisited places - loop closure detection - compensating cumulative pose deviations. This decision is based on similarity measures between maps-stored images in an image-only retrieval model to check whether the sensor achieved a revisited scene during sensor motion. However, due to perceptual limitations, navigation near structures is still a challenge. Vision systems are an attractive environmental sensing solution for robust close-range operations because they operate at distances of less than 3 meters, provide rich information, and they are easy to use [1]. However, the underwater environment is dynamic and has no structural features. In addition, this environment is often affected by turbidity or illumination (shallow waters), which often complicates the behavior of navigation and mapping tasks performed by cameras because the perceptual range of optical devices is severely limited in very poor visibility. Such conditions make loop closure detection difficult, and the vehicle may not detect some loops correctly or may detect some erroneous loop closures, causing the trajectory not to be adjusted or to be adjusted incorrectly. A previous paper analyzed the efficiency of a purely visual system for similar image recognition showing that cameras are susceptible to severe haze and brightness conditions, achieving at best a detection rate of 71%, even with enhancement techniques that provide more consistent keypoints [2]. Today, a new type of sonars - active sonars - can emit an acoustic wave and receive the backscatter, providing acoustic images that allow them to perceive the

 doi:10.20944/preprints202310.1759.v1

environment - imaging sonars [3]. Although these sensors suffer from distortion and occlusion effects due to their physical properties, they do not suffer from haze effects, so this category is considered a promising solution for these challenging environmental conditions. Forward Looking Sonar (FLS) and Side Scan Sonar (SSS) are the most used sonars for perception of the environment. FLS highlights because provide a representation of the environment in front of the robot and allows overlapping images during motion. Image matching is the first issue to solve, since it is the key step for pose estimation or place recognition. Due to the characteristics of FLS data, namely low signal-to-noise ratio, low/inhomogeneous resolution, and weak feature textures, traditional feature-based registration methods are not yet designed for acoustic imagery. [4] proposes a pairwise registration of FLS images for the mosaic pipeline based on a Fourier methodology that can provide robustness to all image content against some artifacts commonly associated with acoustic imaging and noise. In 2018, a machine learning method that uses saliency to detect loop closures was proposed for inspecting ship hulls with imaging sonar. To deal with the sparse distribution of sonar images, it is based on the evaluation of the potential information gain and the estimated saliency of the sonar image [5]. Later, a loop closure detector was proposed for a semi-structured underwater environment using only acoustic images acquired by an FLS [6]. A topological relationship between objects in the scene is studied based on a probabilistic Gaussian function. However, the performance of these sonars has greatly improved, and the resolution of their images continues to increase, allowing the FLS to provide comprehensive underwater acoustic images. Therefore, developing efficient approaches to extract visual data from sonar images and understand their performance is critical. Matching algorithms can be based on feature point and region approaches, but considering the FLS properties, and real-time constraints of underwater operations, the feature point matching is more suitable. Considering the need for viewpoint invariant feature descriptors, binary methods are increasingly used for similarity detection. These features require less memory and computation time. Evidence for this was provided in underwater scenes categorized by seafloor features, turbidity, and illumination, where the Oriented FAST and Rotated BRIEF (ORB) descriptor was found to be more effective for detection and matching with the least computation time [7]. Recently, its behavior was also demonstrated for acoustic images using a performance comparison of different feature detectors, with ORB achieving the best overall performance [8]. For quick and effective loop closures based on visual appearance, the bag-of-words (BoW) algorithm is often used for data representation. This approach typically clusters local descriptors using the K-means clustering technique and requires a codebook of visual words. Each local descriptor is assigned to the nearest centroid, and the representation is in the form of a histogram. Its efficiency through inverted index file and hierarchical structures is advantageous [9,10].

In this way, a feature-based place recognition using only sonar images captured by FLS is proposed in this work. The idea, then, is to apply knowledge of visual to acoustic data to evaluate whether they are effective in detecting loops at close range, and then to exploit their potential under conditions where cameras can no longer provide such distinguishable information. To facilitate variation in environment parameters and sensor configurations, and given the lack of online data to meet requirements in this context, a port scenario was replicated based on Stonefish simulator [11].

The paper is structured as follows: Section 2 describes FLS fundamentals and the Stonefish simulator used to acquire underwater images and replicate real-world conditions. Section 3 describes the proposed place recognition algorithm based on forward-looking sonar data. Section 4 describes in detail the performance metrics used. Moreover, both the evaluation of the image description and matching techniques and the behavior of the already seen places for various experiments performed are detailed. Lastly, section 5 describes the main conclusions and planned next steps of this work.

2. Background

In most cases, sensors usually used for outdoor navigation and mapping present several challenges when deployed in an underwater environment. Underwater, there is heavy turbidity, poor lighting conditions and particles in the water make both visual and laser-based range sensors

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1759.v1

useless. Nevertheless, it is essential that vehicles are able for detecting their surroundings, and so attention turns to sonars (SOund Navigation And Ranging), as sound travels well in the water and can travel thousands of meters without loss of energy, making these sensors capable of covering long-term distances and turbid water conditions. Active sonars, namely FLS allow monitoring of the environment by generating an image of the scene in front of the robot with each scan (Imaging Sonars). Therefore, it is important to understand the basics of FLS before we move on to the different steps of the place recognition pipeline.

Therefore, 2.1 describes the operating principles of FLS devices, the generation of acoustic images and the main challenges in handling FLS images, which may affect the following processes such as image matching and consequently loop closure decisions. In addition, 2.2 introduces the Stonefish simulator used to replicate the intended scenario, which includes some environmental parameters and the selected FLS geometry model.

2.1. Forward Looking Sonar

Two-dimensional (2D)-FLS are a new category of sonar capable of providing acoustic images at a high frame rate, and are therefore also called acoustic cameras. The various operating specifications, such as acoustic beam width, operating frequency, acquisition rate, and beam shaping are always associated with the sensor models. However, the operation of sonars keeps: the sonar emits acoustic waves that cover its field of view (FOV) in azimuth (θ) and elevation (ϕ), and the intensity of the returning beam is then determined based on a range (r) and bearing (θ), as shown in Figure 1.

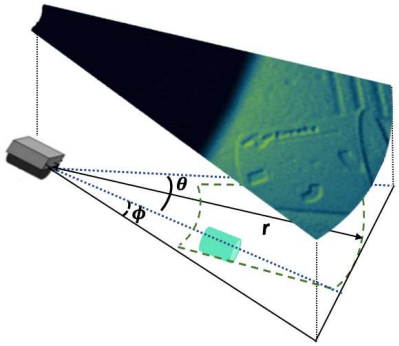

Figure 1. FLS operation: the sound energy returned on the basis of r and θ, and is considered as a map from three-dimensional (3D) points to the null plane.

FLS imaging projects a 3D scene into a 2D image, just like an optical camera: the depth of objects is not lost, but on a sonar image it is not possible to uniquely determine the elevation angle at a given r and θ, i.e., the reflected echo may originate at any point in the reference elevation arc. The images are arranged and mapped in polar coordinates. Thus, the measurements of a raw image correspond to the beams in the angular direction and the range samples in the distance axis. For easier interpretation, the obtained representation is then mapped onto a two-dimensional image based on Cartesian coordinates, resulting in images with uneven resolution. Acoustic images are able to see through turbid environments, but at the cost of a much more difficult type of data. Therefore, there are some issues associated with this type of sonar imaging that can be challenging for inspection tasks, such as:

- **Low resolution:** although FLS are often classified as high-resolution sonars, their image resolution falls far short of that of modern cameras, which typically have millions of pixels. Of course, the resolutions in the cross and down range are crucial for image quality and for distinguishing between closely spaced objects. However, measurement sparsity increases with distance when displayed in Cartesian space, resulting in uneven resolution which imply to degrade the visual appearance of images with weak feature textures;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1759.v1

- **Low signal-to-noise ratio:** even with a large FOV, sonar images exhibit high noise levels - mutual interference causes speckle noise - caused by sampled acoustic echoes, underwater motors near the surface, or other acoustic sensors;
- **Inhomogeneous insonification:** FLSs typically present a Time Varying Gaing (TVG) mechanism with the aim of compensating for transmission losses so that similar targets located at different distances can be perceived with similar intensity. However, changing the angle of incidence or tilt can cause variations in image illumination and other effects that depend on the varying sensitivity of the transducers or lens, which in turn depend on their position in the sonar's FOV. These inhomogeneous intensities can affect the image matching step and, of course, the pose estimation and loop closure decision phases. It is recommended to configure the forward-looking sonar in such a way that there is a small angle between the imaged plane and the bore line (grazing angle), as this allows the largest possible volume to be [12]. Of course, a small angle naturally leads to a larger area without reflected echoes in the image (black area), reducing the effective imaged area, but this configuration will allow vehicles to perform inspection tasks for structures that require close range navigation. This is due to the fact that it also avoids shadows in the images caused by occlusions and significant changes in the visual appearance;
- **Other artifacts:** interfering content may appear in the sonar images, leading to ambiguities during matching: acoustic reflections from the surface, artifacts due to reverberation, or ghost artifacts. However, these interferences can usually be reduced by appropriate configuration and image composition.

2.2. Stonefish Simulator

For a preliminary assessment of the feasibility of using FLS imagery to detect visited places, the Stonefish simulator [1] was used. Its main goal is to create realistic simulations of mobile robots in the ocean, considering the effects of scattering and light absorption. It is an open source C++ library that allows to change the position of the sun in the sky, to simulate optical effects in the water and also to consider the effects of suspended particles. Moreover, it is possible to create specific scenarios, including so-called *"static bodies"* that remain fixed to the origin of the world for the entire duration of the simulation: they are typically used for collision and sensor simulations. Static bodies include a simple plane, simple solids (obstacles), meshes and terrain.

Therefore, a port scenario was replicated to simulate inspection operations in harbour infrastructures that replicate various structures: Quay walls, piers and pillars, i.e. a berth area. The ground is also simulated and consists of some objects commonly found in port facilities, such as garbage, amphorae, anchors and metal grids. Figure 2 shows a wide view of the simulated port facilities scenario. To create the real look of the structures and seabed, a graphical material called "*looks*" must be created that defines how the objects are rendered. All looks are parameterized by reflectance (colour), roughness, metallicity, and reflection factor (0 for no reflections to 1 for mirror) to provide a real simulation of the echoes returned by the sonar. All structures as well as the bottom are considered as rock (material) with some roughness and thus no metallicity factor and no reflections. To add texture to a material, both albedo and normal (or bump) maps are created based on original images to represent the appearance and texture of the scenes, respectively. Figure 3 illustrates the entire rendering process described. The correct setting of the individual maps is crucial for successful rendering, specifically the strength of the bump map. By default, the visibility conditions caused by turbidity (called "*waterType*") and the sun orientation (called "*SunPosition*") have been set so that the simulated scenario looks sufficiently realistic without strict visibility conditions.

---

[1] https://github.com/patrykcieslak/stonefish

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1759.v1

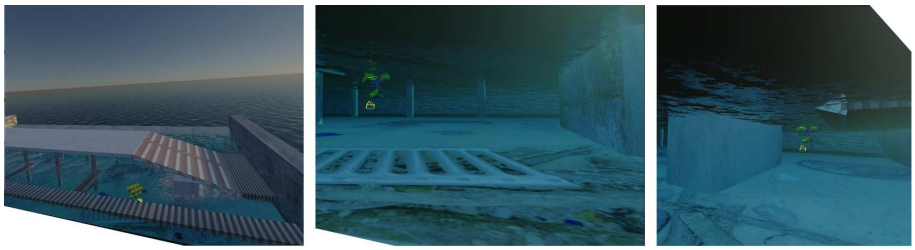

Figure 2. Illustrative images of the simulated port scenario.

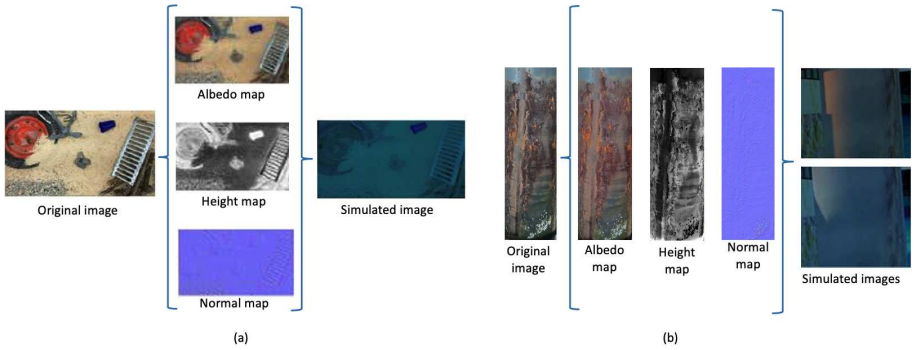

Figure 3. Example of the rendering process to create the terrain (a) and texture of a pile (b).

The simulated AUV - Girona 500 - autonomously executes predefined trajectories between different waypoints. Thus, considering the propulsion system of Girona 500 - five thrusters - a state machine was designed for the AUV to perform an appropriate motion according to the required control motion - e.g. straight ahead or change of direction, with a certain force for smooth navigation to collect reliable data, but also sufficient to respond to the difficulties of the underwater environment (waves, currents, wind, etc.) and even the payload of the vehicle. In this context, an FLS and an odometry sensor (ground truth data) were installed on the vehicle. Both sensors were set with a rate of acquisition of about 7 Hz. To simulate a mission operation, a trajectory near structures (the vehicle moves about 2-3 meters (m) away from the structures) was performed by the AUV at a fixed altitude, with z being the default. The trajectory has a closed loop as the robot travels around the concrete wall and maintains the viewing angle as shown in Figure 4. It moves at 0.65 meters/second (m/s) of speed and decreases its movement speed by 0.25 m/s when it approaches a certain waypoint. If the AUV needs to change its direction of motion to reach an intended waypoint, it turns at 0.4 m/s.

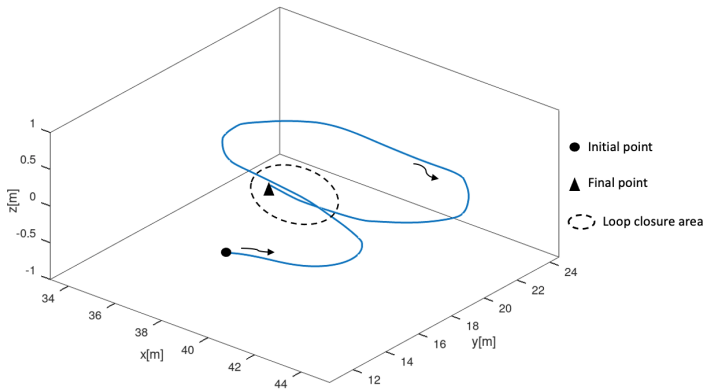

Figure 4. Predefined inspection route of the Girona 500.

The FLS is a top-down sensor, and since it is intended to provide data on the ground near the structures under inspection, its design and configuration have been adapted accordingly. It was configured for a range of up to 3 m (maximum measured range) at a constant standard height of 2 m

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1759.v1

above the ground. To provide good imaging conditions, the sonar was tilted 35°with a horizontal field of view of 40°. In addition, the sonar measurements have 512 beams and 750 bins (range resolution of the sonar image) to mimic the Gemini 720ik sonar that will be used later for real tests. Thus, each sonar image consists of 514x720 pixels. A total of 1067 images were acquired during the mission trajectory. After the simulation is completed, the images and the pose for each one are saved in a folder and a text file, respectively. Figure 5 shows an example of sonar image output.

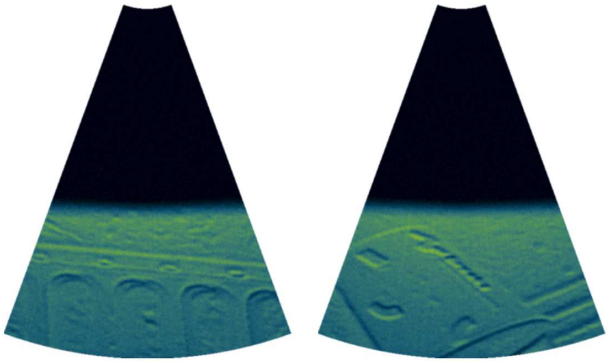

Figure 5. Illustrative example of the images generated by the FLS.

Based on the geometry model shown schematically in Figure 6, each FLS frame covers a width of about 2m (A) and a height of 3m (B), i.e., it represents an area of about 6m². However, there is an area without reflected echoes, which is shown as a black area in Figure 5. Therefore, for each FLS image, a region of interest (ROI) is selected to represent the effectively imaged area of the FLS image. So, a bounding box of 250x350 pixels is used, which means that one FLS image maps an area of about 1.5 m², as illustrated in Figure 7.

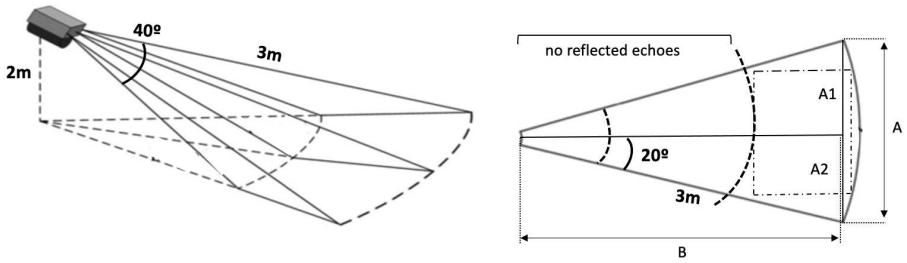

Figure 6. Geometry model used for FLS operation.

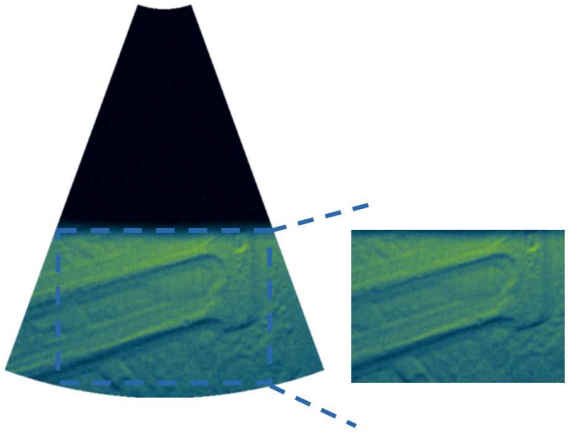

Figure 7. Illustrative example of the effectively mapped area of each FLS image.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1759.v1

3. Acoustic-based Place Recognition

In a practical context, the place recognition is used for searching similar images to a queried image in an image database. It is considered a key aspect for the successful localization of robots, namely to build a map of their surroundings in order to localize themselves (SLAM) [13]. Therefore, this task is also called loop closure and occurs when the robot returns to a point in its trajectory. In this context, correct data association is required so that the robot can uniquely identify landmarks that match those previously seen, from which loop closure can be identified. Thus, a place recognition system must have an internal representation - a map, i.e., a set of distinguishable landmarks in the environment that can be compared to the incoming data. Next, the system must report whether or not the current information is from a known place and if so, which one. Finally, the map is updated accordingly. As a pure image retrieval model, the map consists of the stored images, so the appearance-based methods are the most commonly used. These methods are considered as a potential solution that enables quick and efficient loop closure detections - content-based image retrieval (CBIR), since this scene information does not depend on the pose and consequently on the error estimation. So, *"how can robots decide on the basis of an image of a place whether it is a place they have already seen or not?"*. In order to decide whether an image is a new or not (known) place, a matching between queried and database images is performed using a measure of similarity. Feature extraction is therefore the first step in CBIR to obtain a numerical description. These features must then be aggregated and stored in a data structure in an abstract and compressed form (data representation) to facilitate the similarity search. These measures indicate which places (image contents) are most similar to the current place. This is an important step that affects CBIR performance, as an inappropriate measure will detect fewer similar images and reduce the accuracy of the CBIR system. Therefore, based on the DBoW2 and DLoopDetector [14] libraries, a tree-based similarity detection approach is proposed for the intended context. This method was presented in [2] for optical images and is now adapted for acoustic images. Figure 8 shows the complete process of place recognition based on FLS images.

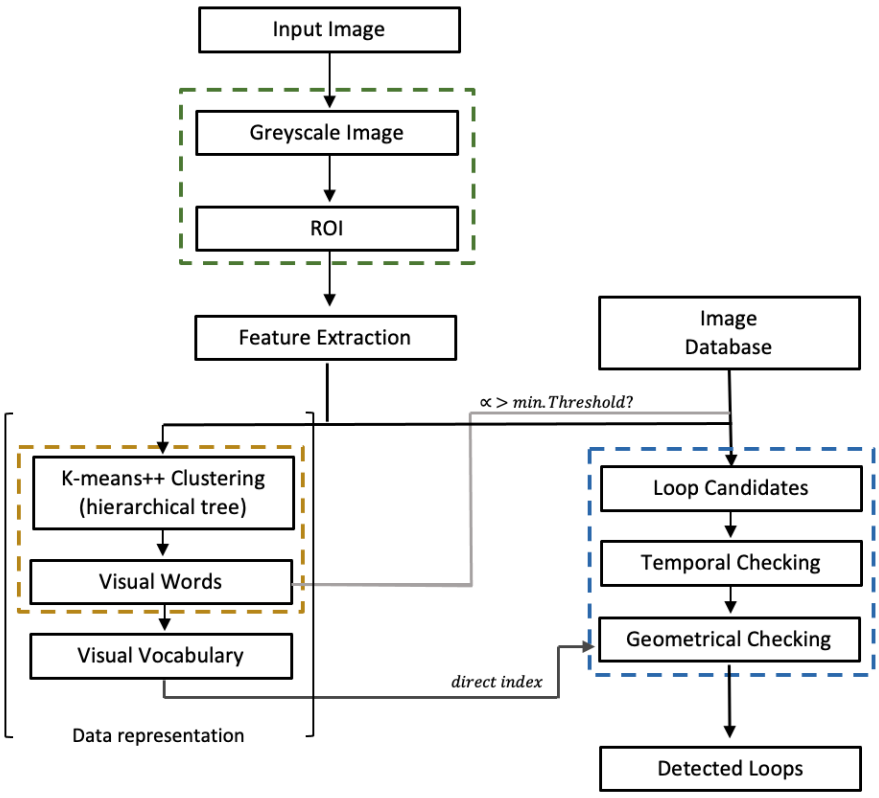

Figure 8. Schematic representation of the developed place recognition method using FLS images.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1759.v1

For each current image, a ROI is used to discard the area without reflected echoes. Thus, each input image is "clipped" in a rectangular area to account for the effective visual information captured by FLS, as described in 2.2. First, an extraction of features and descriptors is performed for each image based on ORB. Then, drawing on these features, an agglomerative hierarchical vocabulary is created using the K-Means++ algorithm: Based on Manhattan distance, clustering steps are performed for each level. In this way, finally, one obtains a tree with W leaves (the vocabulary words), in which each word is assigned a weight related to its relevancy. Inverted and direct indexes are maintained along the BoW to ensure fast comparisons and queries, and then the vocabulary is stored (yellow block in Figure 8). Next, marked by the blue dashed line, to detect similar places, the ROI is also used for each input image, and features based on ORB are extracted and converted into a BoW vector v_t (based on Hamming distance). The database, i.e., images of previously visited places is searched for v_t and, according to weights of each word and their scores (L1 score, i.e., Manhattan distance), a list of matching candidates is generated, represented by the light gray link.Only matches that have a score higher than a similarity threshold α are taking into account. In addition, images that have a similar acquisition time are grouped together (islands), and each group is scored against a time constraint. In addition, each loop candidate must satisfy a geometric test, with RANSAC supported by at least 12 correspondences, *minFPnts*. This value is a common default value for comparing two visual images with different time stamps and therefore may have different perspectives, resulting in a small area of overlap compared to consecutive images. These correspondences are calculated with the direct index using the vocabulary, as represented by the dark gray link.

To measure the robustness of the feature extraction/matching and similarity detection approaches between FLS images, appropriate performance metrics are computed. Their descriptions can be found in 4.1 and 4.2, respectively.

4. Experimental Results

The present section describes the operation of the acoustic place recognition algorithm in an underwater harbor environment. For this purpose, a dataset containing visibility constraints was created using the Stonefish simulator to perform the proposed experiments and evaluations. The performance measures used for evaluating ORB on FLS images and their effectiveness are described and illustrated in 4.1. In 4.2, the performance metrics used to evaluate binary loop closure detection are described. Their behavior in detecting revisited places with different configurations is also shown. An Intel i7 7700K @ 4.5 GHz with 16GB RAM and a NVIDIA GTX 1080 computer were used for the experiments.

4.1. Acoustic Features Effectiveness

Imaging sonars provide increasingly rich acoustic images that allow us to perceive the environment. However, inherent features such as weak textures, interfering content, and low signal-to-noise ratio can impact following imaging phases, particularly feature extraction and matching. Therefore, a functional analysis is considered to analyze the behavior of ORB on acoustic images, based on the amount of ORB keypoints (Kp), number of matches (Nm), and Inliers. The detected Kp on input images is related to visual content, as a higher value usually indicates a better description of the scene. Nevertheless, in some situations there may be pixels that are noisy, leading to misrecognized keypoints. Nm stands for the amount of Kp of a query image corresponding to those in the current image, based on the Hamming distance. Finally, the inliers are the correct correspondences between two images. For this, the RANdom SAmple Consenus (RANSAC) is used from the previously obtained matches [15]: The higher the value, the more similar the compared images are.

Considering standard environmental conditions, different domains - textures - were tested to analyze the behavior of feature extraction and matching algorithms on FLS images. All experiments rely on up to 1500 features per image for the extraction step. Table 1 shows the performance evaluation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1759.v1

for each scenario. A set of 30 consecutive images is used, with each value is obtained by a simple arithmetic mean.

Table 1. Functional evaluation for grade and ground areas surveyed with FLS considering standard visual conditions.

		Grade area	Bottom area
Functional Evaluation	**Kp**	326	
	Nm	326	
	Inliers	90	41

As can be seen, the descriptor finds the same number of keypoints for grade and bottom areas. However, the effective matches are higher for grade region because the bottom area has few stones: an area of low texture. An example of this behavior for both scenarios can be observed in the Figure 9, illustrated by a high quantity of linear lines (correct matches).

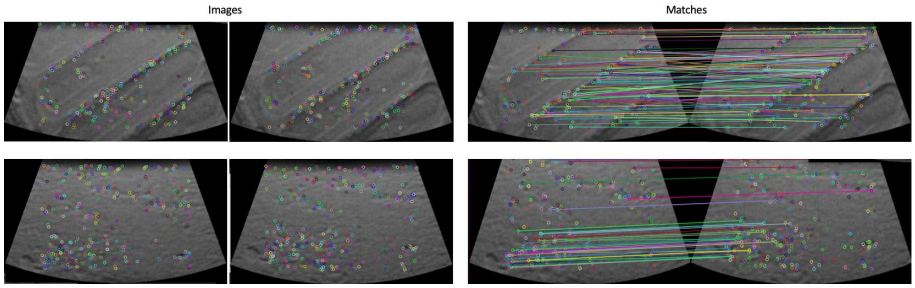

Figure 9. Comparison of keypoint matches in scenarios with (top) and without texture (bottom).

Table 2 shows the obtained functional evaluation considering optical images for both scenarios, considering the same sequence of 30 images. It is obvious that optical images have more keypoints and correct matches between consecutive images when visibility is adequate, since these sensors provide images with more details and have stronger texture perception, since FLS images are based on emitted and returned sounds.

Table 2. Functional evaluation for grade and ground areas captured by a camera considering standard viewing conditions.

		Grade area	Bottom area
Functional Evaluation	**Kp**	935	332
	Nm	914	313
	Inliers	643	151

However, cameras are sensitive to poor texture scenarios, as evidenced by the decline in all functional metrics (bottom area). In fact, performance in such areas was not that far from that obtained with the acoustic images: only 3 times more instead of 7. Moreover, the FLS do not suffer from visibility problems, while the cameras lose performance in such scenarios, as shown by Table 3, in which the camera detects about 80 fewer inliers under turbidity conditions. This suggests that this behavior is not sufficient to achieve inliers between images taken at different times, and thus to detect loops under these conditions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1759.v1

Table 3. Functional evaluation - inliers - for bottom areas detected with FLS and camera considering turbidity conditions.

		Bottom area
Inliers	**FLS**	41
	Camera	70

Table 4 considers poor lighting conditions, and as can be seen, the cameras are also unable to provide many distinguishable features and find as many matches between successive images. As expected, the FLS can account for these conditions without losing performance.

Table 4. Functional evaluation - inliers - for ground areas detected with FLS and camera, considering poor illumination conditions.

		Bottom area
Inliers	**FLS**	41
	Camera	18

4.2. Loop Closure Detection

Given this drop in performance of the visual data compared to FLS images in poor visibility conditions based on successive images, it will be investigated whether this acoustic sensor can help to find similar images in such situations to detect that the robot is already visiting this location and thus provide a correct vehicle position estimation.

To evaluate the behavior of the proposed acoustic place recognition algorithm, a series of ground truth loop closures were generated for the port data described in section 2.2. The pose file created as the ground truth of the trajectory uses a boundary of approximately 1 m to consider two images as the same place based on the effective mapped FLS area. Furthermore, considering the frequency, a mechanism was added to prevent that multiple loops are detected in 1 second (s). Thus, 16 true loop closure situations were considered. The both precision and recall metrics are used for the evaluation the place recognition performance. Thus, the situations where the algorithm is able to recognize a query image as a revisited location with success (TP) or incorrectly (FP) are considered. Likewise, the cases where the method incorrectly does not recognize the requested image as a revisited location (FN) are counted. Therefore, Precision-recall metric is determined: Precision is the ratio of TP to TP and FP, which represents the robustness of correctly detecting a location. Recall is the ratio of TP to TP and FN, i.e., it determines the strength of detecting a place without error. To encounter a suitable combination of the two performance measures, the F1 score (harmonic mean) is also determined.

For all experiments, a fully indexed vocabulary [2] are created. So the scenes are continuous modeled and then the behavior in recognizing revisited areas is analyzed as an incremental learning approach. For all experiments up to 1500 features per image are extracted. For visual images, a similarity threshold (α) of 0.3 was used to assign images as loop candidates because this value was found to best fit to avoid considering widely distant features as the same. Thus, taking into account also this threshold for similarity to find loop candidates in FLS images, one can see the behavior of the location detection method in Figure 10. As can be seen, the algorithm does not incorrectly detect situations where loops close (FP). However, it tends to lose loops along the trajectory, since only 6 loop closures are correctly detected (TP). This fact may mean that the similarity threshold used for the appearance provided by acoustic images may be very high. Table 5 shows the performance values obtained by the algorithm in place recognition considering different similarity thresholds, reaching 100% precision in every cases, i.e., there are no falsely detected (FP) loops, which is crucial in the

[2] The vocabulary is created a priori but using only the images of the trajectory performed.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1759.v1

navigation context. Lowering the similarity requirements will detect more situations where loops close. Nevertheless, the performance improvements in terms of iteration are not very significant. Moreover, the best results are obtained with $\alpha = 0.15$, which seems to be too low a threshold and leads to mistakenly considering any image as a loop candidate, since widely separated points are treated as the same features. Nevertheless, the algorithm fails to close 5 loops (FN) and achieves a maximum recall of 68.75%. This behavior can probably indicate that the features are not robust enough to benefit from the leniency of the requirements.

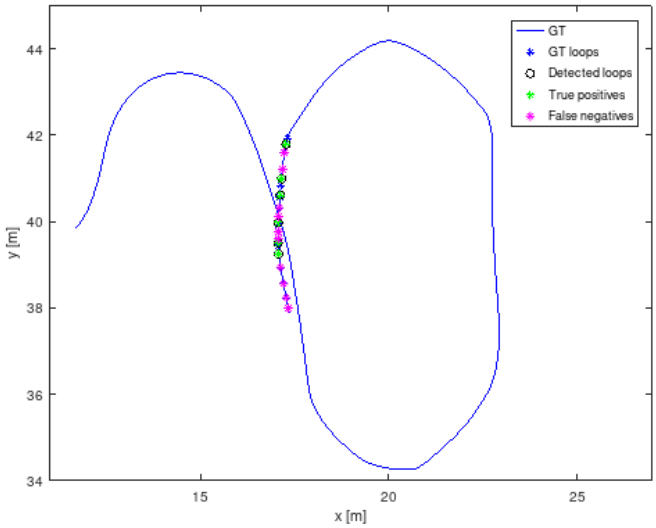

Figure 10. Appearance-based loop closure under standard visibility conditions, taking $\alpha = 0.3$ into account.

Table 5. Performance of loop closure under standard visibility conditions when the similarity threshold α is varied.

	$\alpha = 0.3$	$\alpha = 0.25$	$\alpha = 0.2$	$\alpha = 0.15$
Precision (%)		100		
Recall (%)	37.50	43.75	50.00	68.75
F1-Score (%)	54.55	60.87	66.67	81.48

To decide whether an image (loop candidate) actually represents the same local of a database image, a geometric check is performed between pairs of images, as described in section 3. For optical images, this check must be supported by at least 12 correspondences. Since the inherent properties of FLS sonar can affect the visual appearance of the images (weak textures), Table 6 shows the effects of reducing the minimum number of correspondences, *minFPnts*, to consider two images as the same local. The method allows a minimum of 5 correspondences. For all cases, a similarity threshold α of 0.3 is considered.

Table 6. Performance of loop closure under standard visibility conditions when varying the threshold for minimum correspondences, *minFPnts*.

	minFPnts = 10	***minFPnts = 8***	***minFPnts = 6***
Precision (%)		100	
Recall (%)	62.50	81.25	87.50
F1-Score (%)	76.92	89.66	93.33

By reducing the number of inliers between query (loop candidate) and database images, the algorithm detects more situations where a loop closes. The best results are obtained considering only 6

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1759.v1

correspondences between images, which is an inappropriate value for comparing images taken with different timestamps, and considering that the method allows at least 5 correspondences as a minimum limit. Figure 11 illustrates the behavior of the algorithm for two initial cases (considering 10 and 8 matches for two images that effectively represent the same location). However, above a certain value, the increase is no longer significant: 10, 13, and 14 loop closures are correctly detected, respectively (TP), and there is still not much difference in the variation of this parameter. Thus, there are no more matches that satisfy the requirements of RANSAC.

Comparing the results for $\alpha = 0.3$, using 12 or 10 as the minimum number of correspondences to consider the images as equally local, the performance of the algorithm increases: the number of unrecognized loop closure situations (FN = 10) becomes the number of correctly recognized loops (TP = 10). Therefore, it is crucial that the similarity factor requirements are higher so that any image is no identified with a similarity to the actual image. The geometric check is a final check (like filtering), and therefore it is better to be more benevolent at this stage. Nevertheless, it is obvious that the features and matches are indistinguishable to benefit from the leniency of the requirements. The simplification of both conditions - similarity and matching - is not practical, since distant features are considered as the same point and many images are mistakenly considered as candidates for looping, leading to false detection of loop closure (FP), which is dangerous for the navigation context, since the trajectory is adjusted inappropriately - false pose estimation. Taking into account the experiments and some assumptions, the behavior based on this parameterization - $\alpha = 0.3$ + *minFPnts* = 10 - seems to be the most appropriate for FLS imagery, achieving a precision of 100% and a recall of 62.50%.

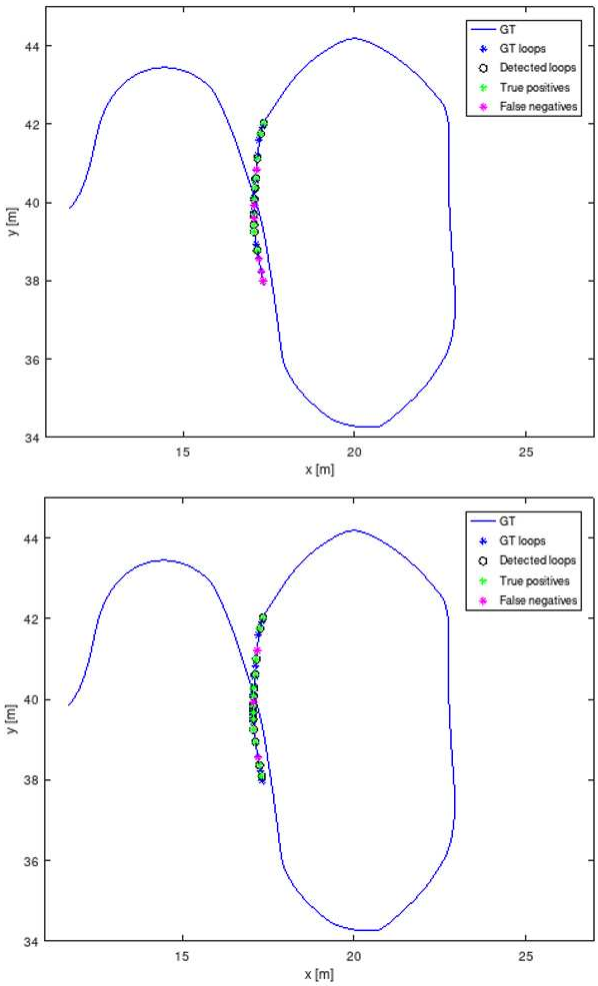

Figure 11. Appearance-based loop closure under standard visibility conditions, considering *minFPnts* = 10 (top) and *minFPnts* = 8 (bottom).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1759.v1

4.2.1. Image Enhancement Procedure

To deal with image degradation - poor visibility - that causes missing image details on optical images, enhancement techniques are used to correct the lack of matching features. There are a few common methods, but CLAHE has proven to be the better choice for underwater scenes for extracting extra information from images with low-contrast, as explored in [2]. This technique emphasizes edges in specific regions and enhances local contrast by dividing each image into distinct, non-overlapping subimages or tiles. For each block, its histogram is clipped and reassigned to prevent excessive emphasis. To avoid block artefacts, bilinear interpolation is then executed of between neighbouring tiles [16]. Thus, there are two key parameters that mainly assure the quality of the image enhancement: the number of tiles (NT), which determines the number of blocks that divide the images into squares of equal size, and the clip limit (CL), which controls the noise gain and is also known as the contrast factor. If you increase the value CL, the image obtained generally becomes brighter, since a larger CL flattens the image histogram. A higher NT increases the dynamic range and contrast of the image. However, selecting a suitable configuration of these parameters is challenging, as they are related to the resolution of the image and the content of the histogram.

Described experiments show that FLS has inherent properties that can degrade data and provide images with weak texture, i.e. features that are not very robust. Therefore, based on the background of optical images and relying on the OpenCV (C++) library, CLAHE is applied to FLS images to understand if these changes can improve the image quality and consequently the feature extraction and matching processes. For this purpose, each image is splitted into blocks of 62 500 pixels, i.e. NT = [2,3]. After applying an enhancer, it is expected that the brightness and details of the images will increase while the naturalness of the image is maintained, but this depends on the visual content. Thus, CL = 1 and CL = 2 were tested. Figure 12 illustrates the modifications in the appearance of the images caused by the individual CL, without and with texture, and shows that the CL increases image contrast and the edges and details are improved. Consequently, the keypoints detected on the enhanced images are higher than on the original images in both cases.

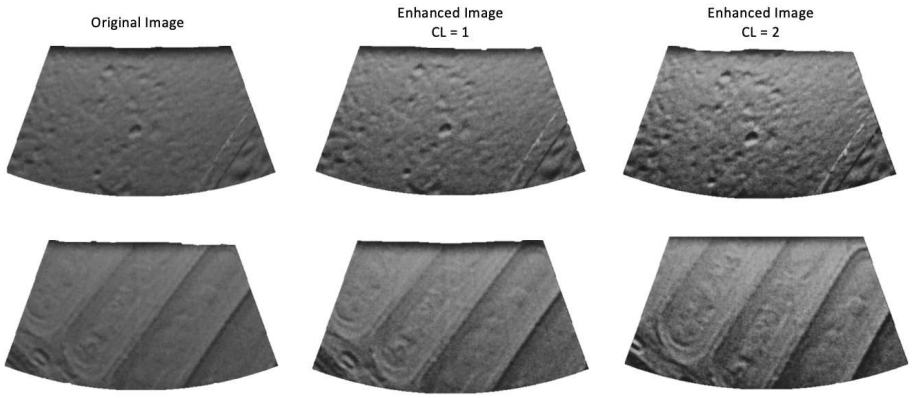

Figure 12. Illustrative example of improvements in acoustic images based on the CLAHE method, using NT = [2,3] and varying CL in images with (bottom) and no texture (top).

But "*are these features strong enough to improve the accuracy of the matching?*". Thus, considering the effect of CLAHE, the behavior of place recognition method was tested for both CL to understand if a better description of the images is given by the enhancement. Table 7 shows the performance achieved in recognising similar places, based on the same test conditions and assumptions from the above experiments. In this context, a similarity threshold α of 0.3 and only 12 and 10 correspondences (*minFPnts)* are considered as the minimum value to consider two images similar. As can be seen, the algorithm achieves 100% of precision in every cases, i.e. there are no misidentified loops (FP) - the features are distinguishable. It would be expected that the recognition rate (recall) would increase as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1759.v1

the contrast (CL), but this occurs only at the cost of decreasing the geometric testing requirements, i.e., the minimum number of inliers to consider a candidate an effective loop.

Table 7. Performance of loop closure under standard visibility conditions resorting CLAHE using NT = [2,3], varying CL and threshold for minimum correspondences, *minFPnts*.

	CL = 1		**CL = 2**	
	minFPnts = 12	***minFPnts = 10***	***minFPnts = 12***	***minFPnts = 10***
Precision (%)		100		
Recall (%)	87.50	81.25	75.00	87.50
F1-Score (%)	93.33	89.66	85.71	93.33

Moreover, the algorithm achieves the same recall in two cases, but in CL = 2 again with a lower value of inliers. So it can be seen that more key points do not necessarily mean distinguishable points: The enhancer can create points that are wrongly recognized as key points because they can be noise, for example. So it seems that the algorithm shows a more stable behavior when CL = 1 and *minFPnts* = 12 is used. It achieves a recall of 87.5%, recognizing 14 loop, and fails only 2 loops, as shown in Figure 13.

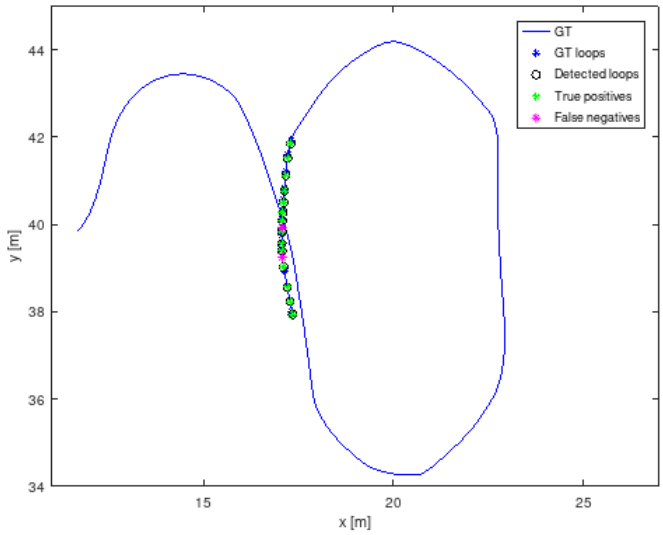

Figure 13. Appearance-based loop closure under standard visibility conditions using CLAHE (CL = 1), taking α = 0.3 and *minFPnts* = 12 into account.

Next, NT = [4,6] is used to find out whether more subdivision of images, i.e., more detailed histograms, has an impact on image description and improves the similarity detection behavior of the algorithm. Table 8 shows the algorithm behavior in recognizing previously visited places, also based on a similarity threshold α of 0.3 and only 12 and 10 matches (*minFPnts*), respectively, as the minimum value to consider two images similar.

Table 8. Performance of loop closure under standard visibility conditions resorting CLAHE using NT = [4,6], varying CL and threshold for minimum correspondences, *minFPnts*.

	CL = 1		**CL = 2**	
	minFPnts = 12	***minFPnts = 10***	***minFPnts = 12***	***minFPnts = 10***
Precision (%)		100		
Recall (%)	68.75	87.50	62.50	93.75
F1-Score (%)	81.48	93.33	76.92	96.75

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1759.v1

In all cases, a precision of 100% is achieved, and the recall increases only with the increase of CL at the expense of decreasing the minimum correspondences for two images to be similar in content, as in the previous case. Looking at CL, the behavior with CL = 2 is better with *minFPnts* = 10, where another loop closure situation is detected. In this case, the same performance as in the previous case (NT = [2,3]) - recall of 87.50% - is obtained with CL = 1, but considering 10 inliers for two images representing the same location, leading us to believe that in this case increasing the number of subdivisions does not describe the image better. With CL = 2 and 10 inliers, the algorithm can achieve 93.75% recall, with only one loop closure detection failing. Thus, the generated image with this value of the contrast threshold does not provide robust key points.

For both experiments - Table 7 and Table 8 - and considering CL = 1 and CL = 2, it is evident that a better description of the image is obtained with NT = [2,3], considering a minimum of 12 inliers. On the other hand, the performance of the algorithm at NT = [4,6] is higher when allowing 10 inliers to consider images similar, and in both cases the difference is due to the detection of another loop closure.

5. Conclusions

In this paper a purely acoustic method for place recognition to overcome the inherent limitations of perception in port scenarios is presented. Poor visibility can complicate the behavior of navigation and mapping tasks performed by cameras. Therefore, forward-looking sonars are a promising solution to extract information about the environment in such conditions, as they do not suffer from these haze effects. These sensors suffer from distortion and occlusion effects, and their inherent characteristics include low signal-to-noise ratio and resolution (which is also inhomogeneous) and weak feature textures, so that conventional feature-based techniques are not yet suitable for acoustic imagery. However, sonar performance has greatly improved and the resolution of these images is steadily increasing, allowing the FLS to provide comprehensive underwater acoustic imagery. The proposed method aims to apply knowledge of visual data to acoustic data to evaluate whether they are effective in detecting loops at close range, and then to exploit their potential under conditions where cameras can no longer provide such distinguishable information. Given the paucity of online data for these applications and to allow variation of environment parameters and sensor configurations, the Stonefish was exploited. Port facilities were simulated in this way to mimic the inspection work commonly performed on structures in these areas. The autonomous vehicle performed a simple trajectory with a loop while the robot moved around the concrete wall. The vehicle navigated between waypoints near structures (approximately 2-3 meters) with the FLS pointed at the ground. Therefore, the sensor was adjusted accordingly, and an odometry sensor was also used to obtain ground truth information. Considering the features that may affect the image quality and thus the subsequent imaging processes such as feature extraction and matching, the effectiveness of the acoustic features was evaluated. For this purpose, a functional analysis was performed measuring the number of detected key points, the corresponding key points, and the effective matches between successive images. The behavior for visual images is also measured to understand the performance of the acoustic images. In general, sonar data provide fewer features than camera images under normal viewing conditions. However, when the appearance of the images is made more difficult, the performance of the FLS is maintained while that of the optical sensor decreases, reaching 18 points, of which the FLS can detect 41. Considering this degradation of visual data performance in poor visibility conditions, which is dangerous in a navigation context, it was investigated whether FLS can help to find similar images in such conditions to detect that the vehicle is already in a certain area, thus allowing a correct estimation of the vehicle position. So, precision and recall metrics were calculated to measure the behavior of the presented acoustic approach. Based on the standard thresholds commonly used for optical images - α = 0.3 and *minFPnts* = 12 - the algorithm does not misdetect situations where a loop closes, i.e. FP = 0 and 100% of precision. However, the strength of the algorithm is low because it tends to lose loops along the trajectory. It detects only 6 situations where loops close, achieving 37.5% of recall. This performance is unsuitable for navigation purposes because there are no trajectory adjustments. Therefore, both the

 doi:10.20944/preprints202310.1759.v1

similarity threshold and minimum of inliers requirements were analyzed. The lower the similarity threshold, the more loop-closing situations are detected by the method. Nevertheless, the performance improvements are not very significant. When the minimum value of matches between images is lowered so that they are considered the same location, it is obvious that the algorithm is not able to recognize some loop closures. However, a balance must be found to maintain quality compromise. Thus, considering 10 inliers as the minimum threshold (with $\alpha = 0.3$), the algorithm detects 10 loop situations and achieves 62.5% of recall and 100% of precision. However, the experiments show that the features are not robust and therefore the matches are not distinguishable to benefit from the mildness of the requirements. Therefore, the effect of an enrichment technique was evaluated. The CLAHE method was used because it proved to be the better choice for extracting additional information from low-contrast images. To improve contrast and highlight image edges, it depends on the parameterization of two parameters: NT, which divides the images into equal square areas, and CL, that controls the noise gain. Looking at the resulting images, we can see that the contrast effectively increases and the details are enhanced, detecting more key points compared to the original images. To understand whether these features are robust enough to increase the number of correct matches, we tested the place recognition behavior for CL = 1 and CL = 2, and NT = [2,3] and NT = [4,6]. The experiments show that increasing these parameters does not lead to a better description of the scene. Under the same assumptions, the most balanced result was obtained with CL = 1 and NT = [2,3]. Under these conditions, the proposed FLS-based approach fails on only 2 loop closures and achieves a recall of 87.5% which is an increase of 50% compared to the original result with the same constraints. Since the FLS images are based on emitted and returned sounds, the FLS behavior is the same regardless of the environmental conditions. This suggests that the FLS supports loop detection in low visibility conditions when the camera no longer provides detailed information.

Future plans are to perform new and more complex inspection trajectories to test the behavior of FLS even when the scene is viewed from different angles. Furthermore, it is planned to identify the conditions and limitations of the interaction of both sensors - camera and FLS - in order to develop a hybrid solution for place recognition in underwater scenes. In addition, an evaluation of this approach in real port facilities is also planned.

Author Contributions: Conceptualization, A.G. and A.M.; Methodology, A.G.; Validation, A.G.; Data—creating and setup, A.G.; Writing—original draft preparation, A.G.; Writing—review and editing, A.G. and A.M; Supervision, A.M.. All authors have read and agreed to the published version of the manuscript.

Funding: This research is financed by FCT - Fundação para a Ciência e a Tecnologia - and by FSE - Fundo Social Europeu through of the Norte 2020 – Programa Operacional Regional do Norte - through of the doctoral scholarship SFRH/BD/146460/2019.

Data Availability Statement: Data available on request.

Acknowledgments: The authors would like to thank the Research Center in Underwater Robotics of the University of Girona for providing the Stonefish simulator [11], which allowed the creation of different scenarios and the executation of the experiments by setting some environmental parameters and sensor configurations.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUV	Autonomous Underwater Vehicle
BoW	Bag-of-Words
CBIR	Content-Based Image Retrieval
CL	Clip Limit
FLS	Forward Looking Sonar
FOV	Field of View
FP	False Positives
FN	False Negatives

Kp	Keypoints
Nm	Number of matches
NT	Number of Tiles
ORB	Oriented FAST and rotated BRIEF
RANSAC	RANdom SAmple Consenus
ROI	Region of Interest
ROV	Remotely Operated Vehicle
SLAM	Simultaneous Localization and Mapping
SSS	Side Scan Sonar
TP	True Positives
TVG	Time Varying Gaing

References

1. Lu, H.; Li, Y.; Zhang, Y.; Chen, M.; Serikawa, S.; Kim, H. Underwater Optical Image Processing: a Comprehensive Review. *Mobile Networks and Applications* **2017**, *22*, 1204–1211, [1702.03600]. doi:10.1007/s11036-017-0863-4.
2. Gaspar, A.R.; Nunes, A.; Matos, A. Visual Place Recognition for Harbour Infrastructures Inspection. OCEANS 2023 - Limerick, 2023, pp. 1–9. doi:10.1109/OCEANSLimerick52467.2023.10244576.
3. Teran Espinoza, A. Acoustic-Inertial Forward-Scan Sonar Simultaneous Localization and Mapping. Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS), 2020.
4. Vilarnau, N.H. Forward-looking sonar mosaicing for underwater environments. 2014.
5. Li, J.; Kaess, M.; Eustice, R.M.; Johnson-Roberson, M. Pose-Graph SLAM Using Forward-Looking Sonar. *IEEE Robotics and Automation Letters* **2018**, *3*, 2330–2337. doi:10.1109/LRA.2018.2809510.
6. Santos, M.M.; Zaffari, G.B.; Ribeiro, P.O.; Drews-Jr, P.L.; Botelho, S.S. Underwater place recognition using forward-looking sonar images: A topological approach. *Journal of Field Robotics* **2019**, *36*, 355–369.
7. Hidalgo, F.; Bräunl, T. Evaluation of Several Feature Detectors/Extractors on Underwater Images towards vSLAM. *Sensors* **2020**, *20*. doi:10.3390/s20154343.
8. Zhou, X.; Yuan, S.; Yu, C.; Li, H.; Yuan, X. Performance Comparison of Feature Detectors on Various Layers of Underwater Acoustic Imagery. *Journal of Marine Science and Engineering* **2022**, *10*. doi:10.3390/jmse10111601.
9. Fuentes-Pacheco, J.; Ruiz-Ascencio, J.; Rendón-Mancha, J.M. Visual simultaneous localization and mapping: a survey. *Artificial Intelligence Review* **2015**, *43*, 55–81. doi:10.1007/s10462-012-9365-8.
10. Sharma, S.; Gupta, V.; Juneja, M. A Survey of Image Data Indexing Techniques. *Artif. Intell. Rev.* **2019**, *52*, 1189–1266. doi:10.1007/s10462-018-9673-8.
11. Cieślak, P. Stonefish: An Advanced Open-Source Simulation Tool Designed for Marine Robotics, With a ROS Interface. OCEANS 2019 - Marseille, 2019. doi:10.1109/OCEANSE.2019.8867434.
12. Su, J.; Tu, X.; Qu, F.; Wei, Y. Information-Preserved Blending Method for Forward-Looking Sonar Mosaicing in Non-Ideal System Configuration. *2023 IEEE Underwater Technology (UT)* **2022**, pp. 1–5.
13. Melo, J.; Matos, A. Survey on advances on terrain based navigation for autonomous underwater vehicles. *Ocean Engineering* **2017**, *139*, 250–264. doi:10.1016/j.oceaneng.2017.04.047.
14. Gálvez-López, D.; Tardós, J.D. Real-Time Loop Detection with Bags of Binary Words. International Conference on Intelligent Robots and Systems, 2011, pp. 51–58. doi:10.1109/IROS.2011.6094885.
15. Kulkarni, S.; Kulkarni, S.; Bormane, D.; Nalbalwar, S. RANSAC Algorithm for Matching Inlier Correspondences in Video Stabilization. *European Journal of Applied Sciences* **2017**, *5*, 20. doi:10.14738/aivp.51.2692.
16. Zuiderveld, K., VIII.5. Contrast Limited Adaptive Histogram Equalization. In *Graphics Gems*; 1994; pp. 474–485. doi:10.1016/b978-0-12-336156-1.50061-6.