# Feature-Based Place Recognition Using Forward Looking Sonar

Some structures in the harbour environment need to be inspected regularly. However, these scenarios present a major challenge for the accurate estimation of a vehicle's position and subsequent recognition of similar images. In these scenarios, visibility can be poor, making place recognition a difficult task as the visual appearance of a local feature can be compromised. Under these operating conditions, imaging sonars are a promising solution. The quality of the captured images is affected by some factors but they do not suffer from haze, which is an advantage. Therefore, a purely acoustic approach for unsupervised recognition of similar images based on forward-looking sonar (FLS) data is proposed to solve the perception problems in harbour facilities. To simplify the variation of environment parameters and sensor configurations, and given the need for online data for these applications, a harbour scenario was recreated using the Stonefish simulator. Therefore, experiments were conducted with preconfigured user trajectories to simulate inspections in the vicinity of 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, achieving 87.5% recall considering appropriate constraints and assumptions for this task given its impact on navigation success. That is, it is based on a similarity threshold of 0.3 and 12 consistent features to consider only effective loops. The behaviour of FLS is the same regardless of the environment conditions and thus it opens new horizons for the use of these sensors as a great aid for underwater perception, namely, to avoid degradation of navigation performance in muddy conditions.

## 1. Introduction

This research presents a purely acoustic method for place recognition that can overcome the inherent limitations of perception in harbour scenarios. Poor visibility can complicate the behaviour 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. Their inherent characteristics include low signal-to-noise ratio and resolution, which is also inhomogeneous, and weak feature textures, so conventional feature-based techniques are not yet used for acoustic imaging. However, sonar performance has greatly improved and the resolution of these images is steadily increasing, allowing FLS to provide comprehensive underwater acoustic images. The proposed method aims to apply what is known about visual data to acoustic data to assess whether it is effective in detecting loops at close range, and then utilise its potential in conditions where cameras can no longer provide such distinguishable information.

## 2. Acoustic-Based Place Recognition

Harbour facilities include various structures such as quay walls and adjacent piles that need to be inspected for corrosion and damage. The inspection of these structures is normally carried out by divers and remotely operated vehicles (ROVs). However, this research is dangerous and ROVs use a cable that can restrict and complicate work in these semistructured environments. Therefore, autonomous underwater vehicles (AUVs) have been used for these tasks. To perform an inspection, they must accurately navigate and recognise revisited places, also known as loop closure detection, to compensate for cumulative pose deviations.

In practise, place recognition is used to search an image database for images similar to a queried image in an image database. It is considered a key aspect for the successful localisation of robots, namely, to create a map of their surroundings in order to localise themselves (SLAM). This task is therefore also referred to as loop closure and takes place when the robot returns to a point on its trajectory. In this context, correct data association is required so that the

robot can uniquely identify landmarks that match those previously seen and by which the loop closure can be identified. A place recognition system must therefore have an internal representation, i.e., a map with a set of distinguishable landmarks in the environment that can be compared with the incoming data. Next, the system must report whether the current information comes from a known place and if so, from which one. Finally, the map is updated accordingly. Since this is an image-only retrieval model, the map consists of the stored images, so appearance-based methods are most commonly used. These methods are seen as a potential solution that enables fast and efficient loop closure detection—content-based image retrieval (CBIR), as this scene information does not depend on pose and consequently on error estimation. So, "how can robots use an image of a place to decide whether it is a place they have already seen or not?". In order to decide whether an image is a new or not (known) location, a matching is made between the queried and database images using a similarity measure. The extraction of features is therefore the first step in CBIR in order 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 content) are most similar to the current place. This is an important step that affects CBIR performance, as an inappropriate measure will recognise fewer similar images and reduce the accuracy of the CBIR system. Therefore, a tree-based approach to similarity detection for the intended context is proposed, as shown in **Figure 1**.
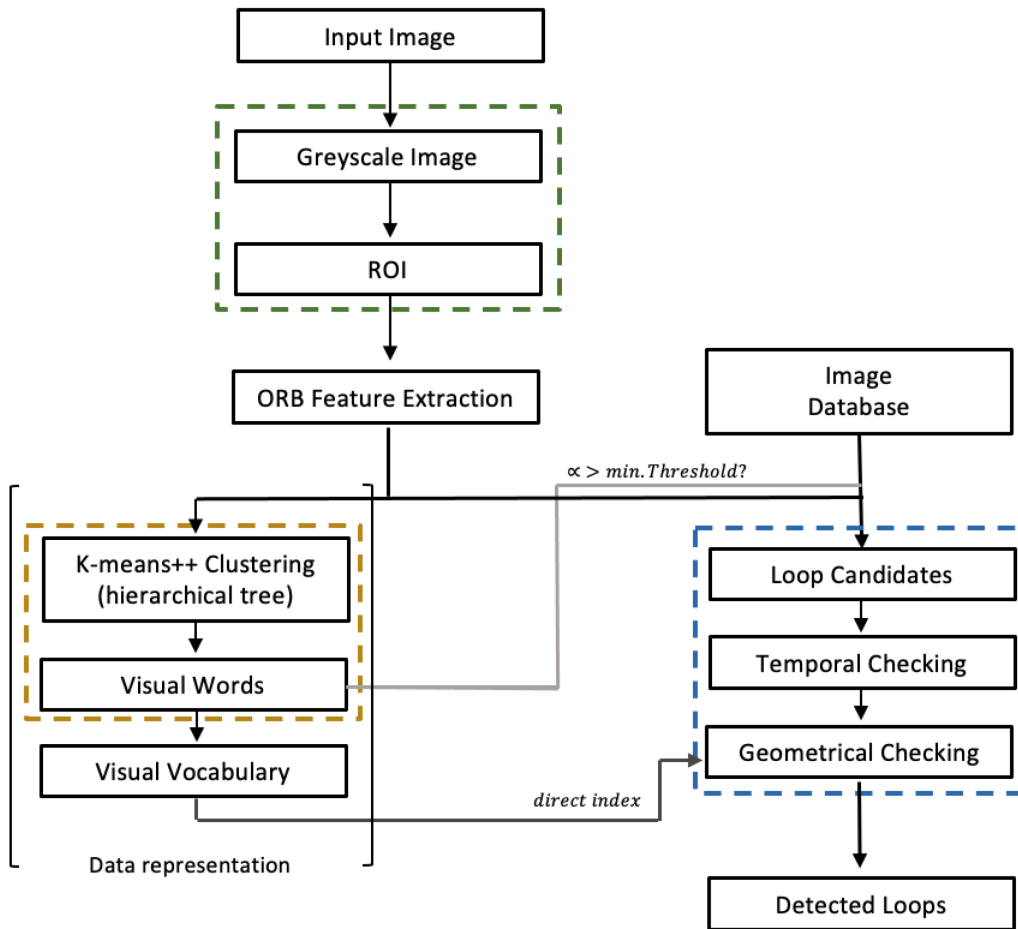


**Figure 1**. Schematic representation of the developed place recognition method using FLS images.

For each current image, an region of interest (ROI) is used to discard the area without reflected echoes. Thus, each input image is "clipped" in a rectangular region to account for the effective visual information captured by FLS. First, features and descriptors are extracted for each image. In this case, binary features are used, namely, ORB features. The binary methods are used because the descriptors for the features must be view invariant. On the other hand, these features require less memory and computation time. Among these methods, ORB stands out as it achieves the best overall performance. More importantly, it proves to be more effective for detection and matching with the least computation time in underwater scenes captured with either cameras covering different environment conditions or acoustic images. It responds better to changes in the scene, especially in terms of orientation and lighting. Based on these features, an agglomerative hierarchical vocabulary is then created using the K-Means++ algorithm; based on the Manhattan distance, clustering steps are performed for each level. In this way, a tree with W leaves (the words of the vocabulary) is finally obtained, in which each word is weighted according to its relevance. Along the BoW, inverted and direct indices are maintained to ensure fast comparisons and queries, and then the vocabulary is stored (yellow block in Figure 1). Next, marked by the blue dashed line, the ROI for each input image is used to recognise similar places, and ORB features are extracted and converted into a BoW vector $v_t$ (based on Hamming distance). The database, i.e., the stored images of

previously visited places, is searched for $v_t$ and based on the weight of each word and its score (L1 score, i.e., Manhattan distance) a list of matching candidates is created, represented by the light grey link. Only matches with a score higher than the similarity threshold an are considered. 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 fulfil a geometric test in which RANSAC is supported by at least 12 correspondences, minFPnts. This value is a common default value for comparing two visual images with different timestamps and therefore possibly different perspectives, resulting in a small overlap area compared to consecutive images. These correspondences are calculated with the direct index using the vocabulary represented by the dark grey link. To measure the robustness of the feature extraction/matching and similarity detection approaches between FLS images, appropriate performance metrics are calculated.

---