International Journal of Advanced Robotic Systems

OPEN ACCESS ARTICLE

Investigation of Vision-based Underwater Object Detection with Multiple Datasets

Regular Paper

Dario Lodi Rizzini[1]*, Fabjan Kallasi[1], Fabio Oleari[1] and Stefano Caselli[1]

1 Universita degli Studi di Parma, Parma, Italy
*Corresponding author(s) E-mail: dlr@ce.unipr.it

Received 02 December 2014; Accepted 17 March 2015

DOI: 10.5772/60526

Abstract

In this paper, we investigate the potential of vision-based object detection algorithms in underwater environments using several datasets to highlight the issues arising in different scenarios. Underwater computer vision has to cope with distortion and attenuation due to light propagation in water, and with challenging operating conditions. Scene segmentation and shape recognition in a single image must be carefully designed to achieve robust object detection and to facilitate object pose estimation. We describe a novel multi-feature object detection algorithm conceived to find human-made artefacts lying on the seabed. The proposed method searches for a target object according to a few general criteria that are robust to the underwater context, such as salient colour uniformity and sharp contours. We assess the performance of the proposed algorithm across different underwater datasets. The datasets have been obtained using stereo cameras of different quality, and diverge for the target object type and colour, acquisition depth and conditions. The effectiveness of the proposed approach has been experimentally demonstrated. Finally, object detection is discussed in connection with the simple colour-based segmentation and with the difficulty of tri-dimensional processing on noisy data.

Keywords Underwater Computer Vision, Object Detection, Image Segmentation

1. Introduction

In recent years, the interest of the scientific community in underwater computer vision has increased, taking advantage of the evolution of sensor technology and image processing algorithms. The main challenges of underwater perception are due to the higher device costs, their complex setup, and the distortion in signals and light propagation introduced by the water medium. In particular, light propagation in underwater environments suffers from phenomena such as absorption and scattering which strongly affect visual perception. While underwater computer vision has been applied to inspection, environment monitoring, mapping and terrain segmentation, underwater object detection has not been thoroughly investigated. In particular, accurate and robust object detection and pose estimation are essential requirements for the execution of manipulation tasks. A robust detection method is able to correctly identify different target objects in different experimental conditions. Once the region of the image corresponding to an object is found, its pose with regard to the observer frame can be estimated in a single image, if the geometry and the size of the object is known, or else by using stereo-processing.

In this paper, we investigate the potential of vision-based object detection algorithms in underwater environments using different datasets, including their contribution to underwater stereo vision processing. Our first contribution is a novel multi-feature object detection algorithm to find human-made artefacts lying on the seabed and which improves and generalizes the work in [1]. The proposed method searches a target object according to a few general criteria, such as salient colour uniformity and sharp contours. The images are initially processed to enhance the contrast in the underwater images. First, the image is partitioned into clusters according to the feature values extracted from each pixel. The current feature vector includes the values of the hue saturation value (HSV) space and its response to the gradient, but it can be easily extended to include pattern-related (e.g., the response to Gabor filters or Eigen transform [2]) and other features. Second, the connected regions of each cluster are classified according to the corresponding angular histogram. The angular histogram of the artefacts is concentrated on one or a few peaks, while the histograms corresponding to the blobs extracted from the natural seabed are usually more distributed. The proposed algorithm relies on these salient properties of common human-made objects, but it is robust to the moderate violation of such hypotheses (e.g., the presence of stripes on general colour uniformity).

The second contribution of this paper is the assessment of the proposed algorithm across different underwater datasets. There are few available datasets focused on objects and acquired using stereo-vision systems in underwater environments. Our experiments were performed on two datasets obtained during the MARIS project (Marine Autonomous Robotics for InterventionS) [3] and a dataset from the Trident project [4]. The differences among the datasets concern the camera quality, the experimental conditions (depth, light conditions, background, sensor guidance, etc.) and the target objects. Despite these differences, the proposed algorithm achieves precise and reliable detection, while elementary colour-based segmentation approaches cannot reliably find a region of interest (ROI). Finally, the positive results achieved in mono-camera detection are discussed in connection with the problems occurring in 3D object recognition. In underwater environments, stereo processing is usually not able to provide a reliable 3D representation of the scene enabling object recognition from shapes. In our evaluation, the target object's pose can be reliably estimated only by exploiting the output of mono-camera processing.

The paper is organized as follows. Section 2 reviews the state of the art in vision-based object detection for underwater environments. Section 3 describes the image processing pipeline and, in particular, the proposed object detection algorithm. Section 4 illustrates the results on object detection and pose estimation in underwater environments. Section 5 discusses problematic issues arising in the underwater context. Section 6 provides some final remarks and observations.

2. Related Work

Computer vision is a major perception modality in robotics and, in particular, for object detection tasks. In underwater environments, however, vision is not as widely used due to the problems arising with light transmission in water. Instead, sonar sensing is largely used as a robust perception modality for localization and scene reconstruction in underwater environments. In [5], Yu et al. describe a 3D sonar imaging system used for object recognition based on sonar array cameras and multi-frequency acoustic signals emissions. An extensive survey of ultrasonic underwater technologies and artificial vision is presented in [6]. However, object detection using sonar imagery is difficult, although some techniques have been proposed. Williams and Groen [7] propose an integral image method to find items differing from the homogeneous background. Sawas et al. [8] use a boosted classifier on Haar features. Underwater laser scanners guarantee accurate acquisition [9]; however, they are very expensive and are also affected by problems with light transmission in water.

Computer vision provides information at lower cost and with a higher acquisition rate compared to acoustic perception. Artificial vision applications in underwater environments include the detection and tracking of submerged artefacts [10], seabed mapping with image mosaicing [11], and underwater SLAM [12]. Garcia et al. [13] compare popular feature descriptors extracted from underwater images with high turbidity, but not for object detection. Aulinas et al. [14] search salient colour regions of interest in order to select stable SURF features such as landmarks in SLAM applications. Without colour segmentation, the data association is unreliable, even for scene description purposes. Stereo vision systems have only been recently introduced in underwater applications due to the difficulty of calibration and the computational performance required by stereo processing. To improve homologous point matching performance, Queiroz-Neto et al. [15] introduce a stereo matching system specific to underwater environments. The disparity of stereo images can be exploited to generate 3D models, as shown in [16, 17]. Leone et al. [18] present a 3D reconstruction method for an asynchronous stereo vision system. Although the 3D reconstruction achieved by underwater stereo vision may be satisfactory to represent a scene, its accuracy is not generally sufficient for the detailed perception required in object detection and recognition.

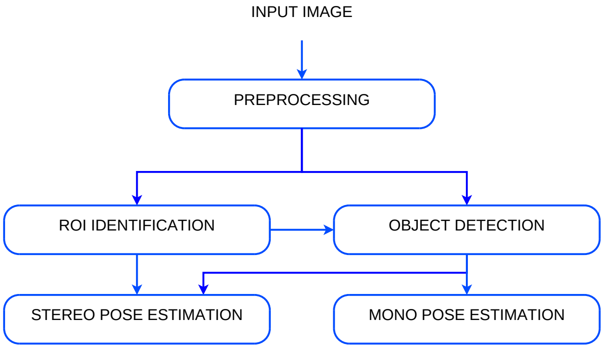


Figure 1. Schema of the object detection and pose estimation algorithm. The algorithm's operations can be combined variously in a pipeline.

Underwater object recognition using computer vision is somewhat difficult due to the lighting conditions of such environments. To cope with the challenging environment, underwater object detection always exploits the peculiar characteristics of the object to be detected, such as shape [19, 20, 10] or colour [21, 22]. Several of these works address pipeline detection problems, since pipelines recur in many underwater applications. Olmos et al. [23] addressed the problem of detecting whether there is a man-made object using contour and texture criteria. However, the proposed method does not return a region in the image containing the object, but simply a binary decision about the object's presence. Several object detection algorithms [22, 21, 24] exploit colour segmentation to find one or more regions of interest and perform a more accurate assessment on the found ROI. Kim et al. [25, 22] present a vision-based object detection method based on template matching and tracking for underwater robots using artificial objects, but all the tests are performed in a swimming pool. Bazeille et al. [21] discuss the colour modification occurring in underwater environments and experimentally assess the performance of object detection based on colour. Since underwater imaging suffers from a short range, low contrast and non-uniform illumination, simple colour segmentation is one of the few viable approaches. In [24], the underwater stereo vision system used in the European project Trident is described. Object detection is performed by constructing a colour histogram in the HSV space of the target object. In the performed experiments, there is an intermediate step between inspection and intervention where the real images of the site to manipulate are available and used for acquiring the target object's appearance [4].

3. Algorithms

Vision-based object detection may be addressed by different approaches according to the input data: through image processing of an image acquired by a single camera, or through more complex shape matching algorithms based on stereo processing. The set of algorithms for underwater object detection proposed in this paper consists of several phases operating at decreasing levels of abstraction and with increasing knowledge about the object to be detected. Figure 1 shows how the different algorithm operations can be combined together into a pipeline. The chosen processing pipeline could depend upon the amount of reliable visual features offered by the object, as well as upon the specific underwater conditions. The initial step aims to detect salient regions with regard to the background representing candidate objects, possibly with no prior knowledge about the object. The output of such a step may be a broad candidate ROI to narrow the searching area in the successive phases, with the classification of the regions as a target object or not (respectively the blocks *ROI Identification* and *Object Detection* in Figure 1). In the first case, no decision about the target's presence is taken from a single frame. Such a decision is delegated either to a more accurate monocular processing or is jointly performed with pose estimation on the 3D point cloud. In the second case, the output of the *Object Detection* module consists of an ROI, the contour of which accurately delimits the object's edges in the image. The point cloud is used only to estimate the pose of an already-detected object (*Stereo Pose Estimation* block in Figure 1). Alternatively, if the geometric model and the dimensions of the object are known, the pose can be estimated from the shape of the object's projection in a single frame. Both the pose estimation methods require a geometric description of the target object.

3.1 Image Pre-processing

Underwater object detection requires the vision system to cope with difficult underwater lighting conditions. In particular, light attenuation in water produces blurred images with limited contrast, and light back-scattering results in artefacts in the acquired images. Object detection becomes even more difficult in the presence of suspended particles or with an irregular and variable background. Hence, for underwater perception, special attention must be paid to algorithmic solutions improving image quality.

The first phase of the algorithmic pipelines in Figure 1 is conceived to compensate the colour distortion incurred by the light propagating in water through image enhancement. No information about the object is used in this phase, since the processing is applied to the whole image. Popular techniques for image enhancement are based on colour restoration [26]. The approach adopted in this paper focuses on strengthening contrast to recover blurry underwater images. A *contrast mask* method is first applied to the component L of the CIELAB colour space of the input image. In particular, the component $L_{in,i}$ of each pixel i is extracted, a median filter is applied to the L -channel of the image to obtain a new blurred value $L_{blur,i}$, and the new value is computed as $L_{out,i} 1.5 L_{in,i} - 0.5 L_{blur,i}$. The effect of the contrast mask is a sharpened image with increased contrast.

Next, in order to re-distribute luminance, contrast-limited adaptive histogram equalization (CLAHE) [27] is performed. The combined application of the contrast mask and CLAHE compensates the light attenuation and removes some of the artefacts in the image. Figure 2 shows an example of the effect of pre-processing for an underwater image. In our experiments, the image enhanced by CLAHE alone is not discernible from one obtained after applying both filters. Hence, the contrast mask computation may be skipped, thereby reducing processing time.

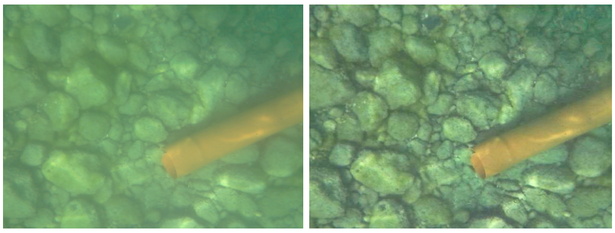


Figure 2. An underwater image before (left) and after (right) the application of a contrast mask and CLAHE

3.2 Mono Camera Processing

The goals of single image processing may include the identification of an ROI containing target objects, the detection of the specific target object and, if the target object's geometry is known, the estimation of its pose. The careful use of monocular processing can significantly improve the effectiveness of the application. Even in an underwater environment, with proper illuminators, shallow and clean waters, and quality cameras, the contours of objects are distinguishable in the image. The target object thus can be detected in a single image or - at least - its detection can be facilitated by restricting the search region to be analysed in later, more expensive steps. Object recognition on a 3D point cloud is computationally expensive and depends upon the quality of the input data obtained by stereo processing. If the object has already been detected based on appearance in mono camera images, the 3D data corresponding to the segmented ROI are used only to estimate the pose of the object. In the following, two rough but fast techniques for the identification of an ROI possibly containing the target object (ROI_{area} and ROI_{colour}) and a more accurate technique for target object detection (ROI_{shape}) are illustrated. Furthermore, a method for the pose estimation of cylindrical objects in a single images is proposed.

ROI Identification

The algorithms described in this section aim to detect an ROI that may represent - or at least contain - an object. The ROI may be searched according to different criteria based on specific features of the object to be found. We have developed two approaches that exploit different assumptions about the properties of the target. The HSV colour space is used to improve colour segmentation [28] since it better represents human colour perception. In particular, to reduce the effect of noise and light distortion, a colour reduction is performed on the H channel of the input image. The algorithms described in this paper use 16 levels of quantized colour. The input image is partitioned into subsets of - possibly not connected - pixels with the same hue level according to the value of the reduced channel H. The rough level quantization is not affected by the patterns generated by light back-scattering.

The first segmentation method, hereafter denoted ROI$_{area}$, is based on the assumption that the unknown object never occupies more than a given portion of the image pixels and that it has a uniform colour. The region corresponding to a given hue level is estimated as the convex hull of the pixels. Only regions the area of which is less than 50% of the image are selected as part of the ROI$_{area}$. This heuristic rule rests on the hypothesis that the object is observed from a distance, such that only the background occupies a large portion of the image.

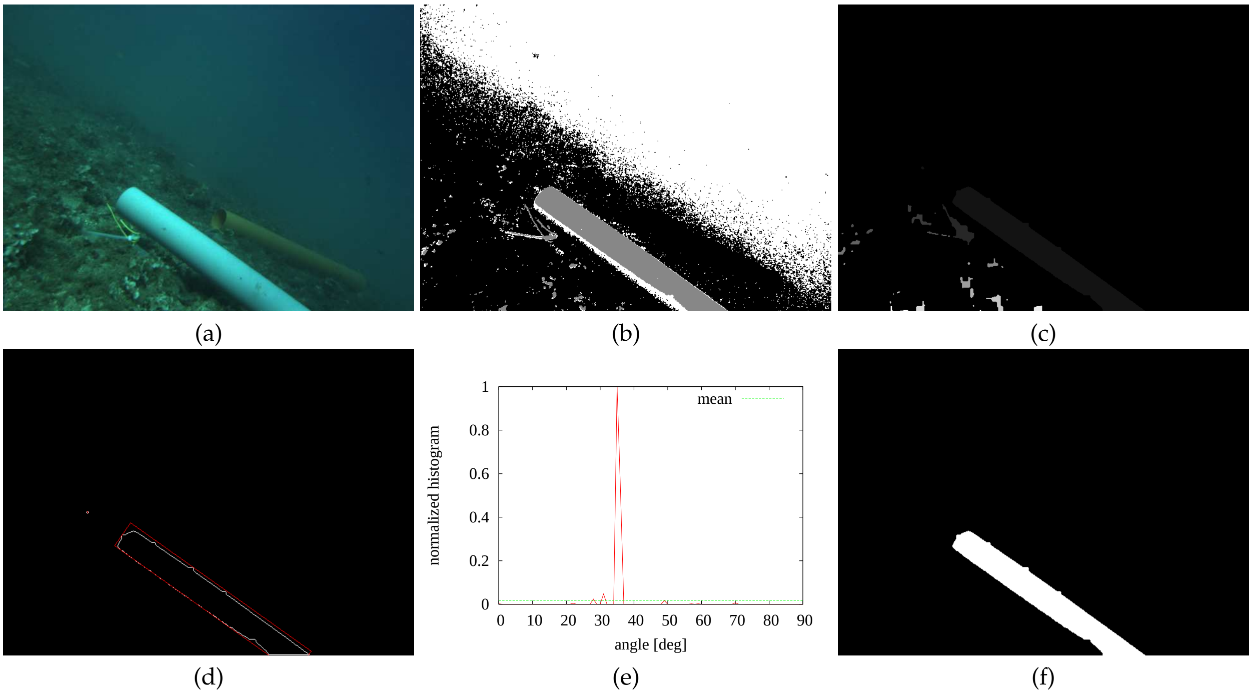
The second approach exploits information about the colour of the target. When the object's colour is known, a more specific *colour mask*(ROI$_{colour}$) can be applied to detect the object with an accurate estimation of the object's contour. Hence, ROI$_{colour}$ is obtained by composing the regions where the colour is close (up to a threshold) to the expected target colour.

The region computed either by ROI$_{area}$ or ROI$_{colour}$ is made available for further processing. Both these ROI estimation techniques exploit only the relative colour uniformity of a texture-less object, but they do not identify a specific object. Furthermore, they tend to overestimate the area that potentially contains the object.

Object Detection

The proposed ROI_{shape} algorithm performs object detection in two steps: *image segmentation* and *contour shape validation*. The goal is the identification of a connected region with straight and sharp contours like typical human-made artificial objects. ROI_{shape} relies on the relative colour uniformity of the object for segmentation and on its regular contour shape for validation. The detection exploits these two salient and relatively general features of artefacts in an underwater environment. In contrast with the previously described coarse segmentation approaches (ROI$_{area}$ and ROI$_{colour}$), ROI_{shape} is able to detect whether the target object belongs to the image before performing pose estimation. The current implementation of the algorithm finds a single object in the image, although it could be adapted to return multiple regions in the image satisfying the above criteria.

The *image segmentation* step classifies each pixel of the image according to its corresponding vector of local features. The input image can be rescaled to a lower size to remove unnecessary details and to reduce the computational complexity of the object detection. The scaling operation acts as a low-pass filter in the image. The initial classification of each pixel p_i is independent of the classification of other pixels. In particular, the feature vector computed for p_i consists of the colour channels of the HSV space, respectively hue h_i, saturation s_i and value v_i, and of the gradient response to a *Sobel* filter g_i. The feature space adopted in this paper is larger than that used in [29] and could be further expanded. Next, the item vectors $\mathbf{f}_i=[h_i,s_i,v_i,g_i]^T$ are clustered according to a k-means algorithm [30]. The number of clusters used in the experiments described in Section 4 is $k=3$ and is independent of the number of objects in the scene. Figure 3(a)-(b) illustrates a typical output of k-means clustering for one of the datasets investigated in this paper: the artificial objects are classified as belonging to the same cluster, whereas the background is split into a light region and a dark region. The image partition achieved by k-means is further refined by computing the connected components of the cluster. Before searching the connected components, a morphological dilation followed by an erosion is performed on the image to avoid over-segmentation. The result is shown in Figure 3(c). The achieved segmentation is correct when there is no


(a) (b) (c)

(d) (e) (f)

Figure 3. Steps of the ROI_{shape} algorithm: (a) the input image; (b) the clusters obtained from k-means; (c) the connected components extracted from one of the clusters; (d) the contour image of one component (within its bounding box); (e) the corresponding angle histogram; (f) the output mask

contact among the objects; otherwise, a more sophisticated and computationally complex technique must be applied. The proposed approach has been successfully applied to different underwater datasets corresponding to typical scenarios of robotic underwater experiments.

The *shape validation* step is applied to each cluster obtained from image segmentation. The algorithm computes the contour of each binary image corresponding to the cluster, as shown in Figure 3(d). Each closed contour represents a cluster-region, and shape matching between the contours and the target shape allows the identification of the target region. The contours of human-made regularly shaped objects - in particular when their projection in the image plane is approximately a rectangle - often consists of parallel edges. Under this assumption, the target region is recognized by detecting parallel line-segments from the contour, e.g., using the *Hough transform*. Hence, a set of segments is extracted from the contour. Each segment j is described by its length l_j and by its supporting line with equation $x\cos\alpha_j + y\sin\alpha_j = r_j$ (coordinates are expressed with regard to the image centre). The detection of line direction is allowed by an angular histogram $\mathcal{H}$ with bin counters $h_s \in \mathbb{N}$ and the intervals $[s\Delta\theta, (s+1)\Delta\theta]$, $s = 0, \ldots, n_h - 1$ and $\Delta\theta = \pi/(2n_h)$. In particular, the segment j increments the corresponding angle bin h_k with a contribution proportional to the square of its length l_j as

$$s = \left\lfloor \frac{(\alpha_j + \pi) \bmod \frac{\pi}{2}}{\Delta\theta} \right\rfloor \qquad (1)$$

$$h_s \leftarrow h_s + \left(\frac{l_j}{\max_i l_i} \right)^2 \qquad (2)$$

The square of the normalized length reduces the influence of the smaller segments resulting from the potential over-segmentation of the contours. Finally, a cluster is classified as an object with a regular shape if the histogram is "peaked", i.e., if it is distributed along a few principal directions. In particular, the validation condition of the clusters is

$$\bar{h} = \frac{1}{n_h} \sum_{s=0}^{n_h - 1} h_s < \sigma_{th} \max_{0 \leq s < n_h} h_s \qquad (3)$$

where σ_{th} is a proper acceptance threshold. An example of a histogram with its mean value is shown in Figure 3(e). The ROI_{shape} algorithm may validate more than a cluster in each image. An additional rule must therefore be provided to select a single cluster according to the features of the specific dataset. For example, in the dataset illustrated by Figure 3(a), the largest segmented regularly-shaped area could be returned. Alternatively, when some part of the robotic exploring agent is included in the image sequence (a condition arising in other datasets), the corresponding area is known and can be excluded from the cluster search area. Of course, ROI_{shape} may also simply return the list of the objects or adopt any task-oriented selection criterion. The execution of ROI_{shape} requires about $150ms$ on an Intel i7-3770 CPU 3.40 GHz and 8 GB RAM.

Mono Camera Pose Estimation

In general, object pose estimation cannot be performed on a single image and requires 3D perception. In this paragraph, we describe pose estimation in mono camera underwater images when the object is known and has a regular shape. We specifically focus on cylinder-like objects, although a similar approach could be developed for box-like objects and other regular 3D shapes. In particular, a cylinder is defined once the cylinder radius c_r and its axis, a line with equation $\mathbf{c}(t)=\mathbf{c}_p+\mathbf{c}_d t$, are given. The contour of a cylinder in the image plane is delimited by two lines with equations $\mathbf{l}_i^T\mathbf{u}=0$ with $i=1{,}2$, where $\mathbf{u}=[u_x, u_y, 1]^T$ is the pixel coordinate vector and $\mathbf{l}_1$, $\mathbf{l}_2$ are the coefficients. Let $\mathbf{l}_0$ be the parameters of the line representing the projection of the cylinder axis in the image. The two lines with parameters $\mathbf{l}_1$ and $\mathbf{l}_2$ are the projections on the image plane of the two planes, which are tangent to the cylinder and contain the camera origin (Figure 4). The line with parameter $\mathbf{l}_0$ is the projection of the plane passing through the cylinder axis and the camera origin. The equations of these three planes in the 3D space are given by

$$\mathbf{l}_i^T(\mathbf{K}\mathbf{p}) = \left(\mathbf{K}^T\mathbf{l}_i\right)^T\mathbf{p} = \mathbf{n}_i^T\mathbf{p} = 0 \qquad (4)$$

where $\mathbf{K}$ is the camera matrix obtained from the intrinsic calibration, $\mathbf{n}_i=\mathbf{K}^T\mathbf{l}_i$ are the normal vectors of the planes corresponding to the lines $\mathbf{l}_i$ with $i=0{,}1{,}2$ (in the following, the normalized normals $\hat{\mathbf{n}}_i=\mathbf{n}_i/\|\mathbf{n}_i\|$ are used), and $\mathbf{p}$ is a generic point in the camera reference frame coordinates. The direction of the cylinder axis is given by direction vector $\mathbf{c}_d=\hat{\mathbf{n}}_1\times\hat{\mathbf{n}}_2$. If the cylinder radius c_r is known, then the distance of the cylinder axis from the camera centre is

$$d = \frac{c_r}{\sin\left(\frac{1}{2}\operatorname{acos}\left(|\hat{\mathbf{n}}_1\cdot\hat{\mathbf{n}}_2|\right)\right)} \qquad (5)$$

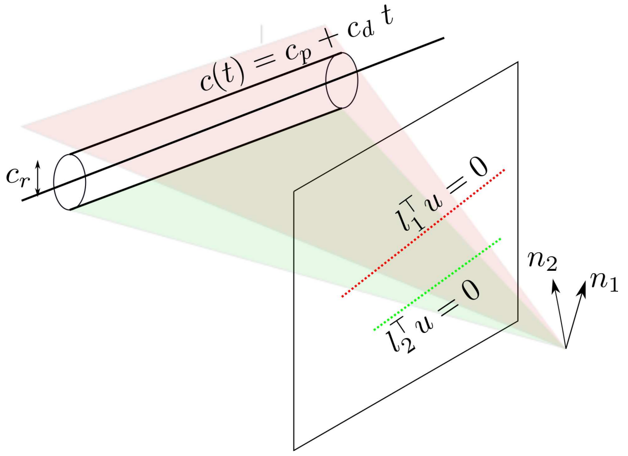


Figure 4. Monocular pose estimation of a cylinder shape using the projection of its tangent planes on the image

The projection of the camera origin on the cylinder axis is equal to $\mathbf{c}_p=d(\mathbf{c}_d\times\hat{\mathbf{n}}_0)$ (if $c_{p,z}<0$, then substitute $\mathbf{c}_p\leftarrow-\mathbf{c}_p$). These geometric constraints allow the estimation of the object pose in space using only a single image. The accuracy of such an estimation depends upon the image resolution and upon the extraction of the two lines. It can be used as an initial estimation or as a validation criterion of the object pose computed on the 3D point cloud generated from stereo vision.

3.3 Stereo Camera Processing

For generic or unknown objects, pose estimation from mono camera images is not possible and 3D data analysis based on stereo processing is required. Moreover, if a coarse segmentation method such as ROI$_{area}$ or ROI$_{colour}$ has been adopted (rather than ROI_{shape}), segmentation must often be refined on 3D data to enable object detection for manipulation or recognition tasks as well as object pose estimation. In these cases, the available ROI is used as a filtering mask to generate a smaller point cloud that represents the 3D scene limited to the object (possibly together with some of the surrounding area). The benefit of restricting the region size where stereo processing is performed is limited when the disparity image is computed using an incremental *block-matching SAD* (sum of absolute differences) algorithm. Since the SAD of a block is computed using the SAD values of adjacent blocks, the advantage of computing the disparity image on the ROI alone is reduced. Indeed, the estimation of point clouds limited to the ROI saves around 15% of the time required for each frame.

3.4 Stereo Camera Pose estimation

In this paragraph, we describe the algorithm for the pose estimation of a regularly-shaped object with stereo processing. Pose is computed with regard to the stereo vision frame. The importance of an ROI is even more apparent in 3D object recognition, since this step requires computationally expensive operations on point clouds. In particular, the ROI can be used to select the point cloud C to search for the target object. If the ROI has not been classified, the distance between the points and the geometric model of the target object provides a criterion for object recognition. Again, our development focuses on the case where the objects to be recognized have a cylindrical shape. Similar developments can be carried out for box-like shapes and other shapes that can be represented by a parametric model. In particular, we represent cylinders in 3D using 7 parameters: the three coordinates of the cylinder axis point $\mathbf{c}_p=[c_{p,x}, c_{p,y}, c_{p,z}]^T$, the axis direction vector $\mathbf{c}_d=[c_{d,x}, c_{d,y}, c_{d,z}]^T$, and the radius c_r. The model matching algorithm simultaneously searches for a subset of the point cloud that better fits a cylindrical shape and computes the value of the cylinder parameters $\mathbf{c}=[\mathbf{c}_p^T, \mathbf{c}_d^T, c_r]^T$. For pose estimation, three algorithms have been implemented: *RANSAC* model-fitting, *particle swarm* *optimization* (PSO) and *differential evolution* (DE) [31]. The latter two are bio-inspired algorithms [29].

The estimated pose is obtained through the geometric alignment of the model of the searched object and the point cloud obtained from stereo processing. The RANSAC, PSO and DE algorithms require a fitness function to measure the consensus of a subset of the point cloud C over a candidate model c. A natural fitness function is the percentage of points $\mathbf{p}_i \in \mathrm{C}$ such that their distance to the cylinder $\mathbf{c}$ is less than a given threshold d_{thr}. The more obvious measure of the displacement between a point $\mathbf{p}_i$ and a cylinder $\mathbf{c}$ is the Euclidean distance

$$\mathrm{d}_E\left(\mathbf{p}_i,\mathbf{c}\right)=\left|\frac{\|\mathbf{c}_p\times\left(\mathbf{c}_p-\mathbf{p}_i\right)\|}{\|\mathbf{l}_d\|}-r\right| \tag{6}$$

However, the Euclidean distance might not take into account some orientation inconsistencies. If the normal vector $\mathbf{n}_i$ on point $\mathbf{p}_i$ can be estimated, the angular displacement between the normal and the projection vector of the point $\mathbf{p}_i$ on the cylinder $\mathbf{c}$ (called $\mathrm{proj}(\mathbf{p}_i,\mathbf{c})$ hereafter) provides

$$\begin{aligned}
\mathrm{d}_N\left(\mathbf{p}_i,\mathbf{n}_i,\mathbf{c}\right)&=\min\left(\alpha_i,\pi-\alpha_i\right)\\
\alpha_i&=\mathrm{arcos}\left(\frac{\mathbf{n}_i\cdot\mathrm{proj}(\mathbf{p}_i,\mathbf{c})}{\|\mathbf{n}_i\|\|\mathrm{proj}(\mathbf{p}_i,\mathbf{c})\|}\right)\\
\mathrm{proj}\left(\mathbf{p}_i,\mathbf{c}\right)&=\mathbf{p}_i-\mathbf{c}_p-\left(\frac{\mathbf{p}_i\cdot\mathbf{c}_d-\mathbf{c}_p\cdot\mathbf{c}_d}{\|\mathbf{c}_d\|^2}\right)\mathbf{c}_d
\end{aligned} \tag{7}$$

The chosen distance function is a weighted sum of two distances

$$\mathrm{d}\left(\mathbf{p}_i,\mathbf{n}_i,\mathbf{c}\right)=w\cdot\mathrm{d}_E\left(\mathbf{p}_i,\mathbf{c}\right)+\left(1-w\right)\cdot\mathrm{d}_N\left(\mathbf{p}_i,\mathbf{n}_i,\mathbf{c}\right) \tag{8}$$

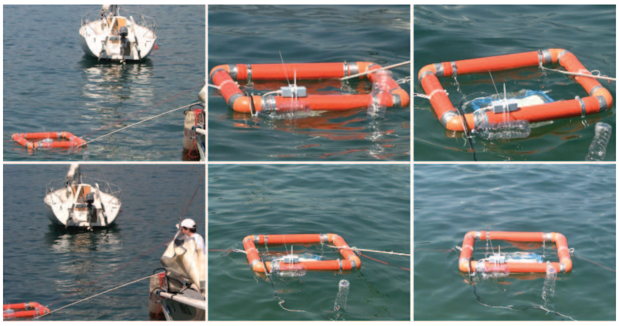


Figure 5. Images of the experimental sessions

4. Experimental Evaluation

4.1 Dataset Acquisition

In this section, we report the evaluation of the vision algorithms for underwater object detection described in section 3 using three datasets. These datasets differ significantly in terms of their acquisition setup and operating conditions. Moreover, one of these datasets has been collected by an independent research group, which is freely available over the Internet. Based on this assessment, we aim to make general remarks on the influence of the acquisition setup and underwater conditions on the performance of vision-based underwater object detection.

The three datasets, *Garda*, *Portofino* and *Soller* are described hereafter. These datasets have been obtained in various scenarios by rather different underwater systems, and have been chosen in order to robustly evaluate the potential of object detection algorithms in the underwater context.

The *Garda* dataset has been acquired using the low-cost stereo vision system described in [32]. This system consists of a low-cost prototype conceived to investigate the performance, power consumption and thermal dissipation trade-offs involved in designing an embedded vision unit for underwater operation. The vision sensors are three Logitech C270 webcams (working resolution: 640x480 @ 7.5 *fps*), chosen due to their low-cost, dimensions, standard USB interface, and the prototypical nature and testing objectives of the system (Figure 6(a)). The cameras can be combined in pairs in order to test three different stereo baselines. The webcams are driven by an off-the-shelf x86 ECU based on a Mini-ITX Intel Desktop Board DN2800MT with an Intel Atom CPU, offering the lowest available TDP (thermal dissipation parameter). Unfortunately, the lack of synchronism between the cameras and the low quality of the sensors have been shown to affect the quality of the disparity images obtained from stereo processing in underwater environments. In contrast, a stereo camera built with the same sensors proved reasonably effective for in-air stereo processing and 3D shape reconstruction [33]. This difference in performance emphasizes the more challenging nature of the underwater environment. Two experimental sessions were conducted at the Lake of Garda to acquire an underwater image dataset in multiple ambient situations and with different objects (Figure 5). The dataset also includes several submerged cylindrical objects (the depth range is from 1.8*m* to 3*m*). In both sessions, the average depth of the camera was about 40*cm* below water level. Figure 7(a) shows an image from this dataset.

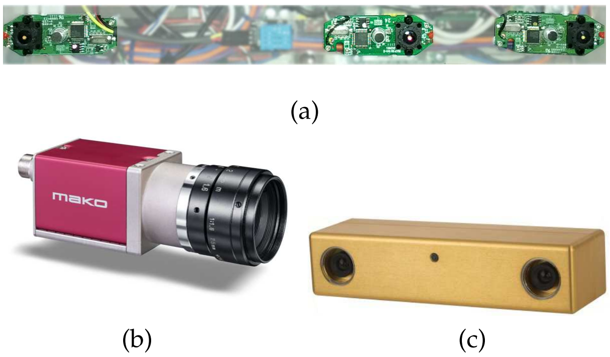


Figure 6. Cameras used for the acquisition of the *Garda* (a), *Portofino* (b) and *Soller* (c) datasets. (a) Multi-stereo head made of three Logitech C270 webcams; (b) Allied Vision Technologies MAKO camera; (c) Point Grey BumbleBee 2.

The *Portofino* dataset represents the first outcome of a new prototype of an underwater stereo vision system currently under development. On the basis of lessons learned from the low-cost prototype, a newly designed system has been developed and tested in shallow water near Portofino, Italy. This system exploits two industrial GigE cameras from Allied Vision Technologies, belonging to the Mako series and providing a resolution of 1292x964 pixels (Figure 6(b)). The embedded ECU, made from a Mini-ITX board and an Intel i7 CPU, represents a good trade-off between computational power and power consumption. The cameras were driven at a rate of 10 fps without communication losses. A sample image of the *Portofino* dataset, at about 10*m* is shown in Figure 7 (b).

Image sequences of the *Soller* dataset were acquired near Soller, Balearic Islands (Spain), by a research team of the European project TRIDENT [24, 4]. Within this project, an autonomous underwater vehicle (AUV) has been developed to perform submarine interventions exploiting a floating manipulator arm and a vision system. The vehicle and issues related to system integration, including the features of the vision system, are described in [24]. The vision unit consists of two Point Grey BumbleBee2 stereo cameras and an ECU for image processing. The two stereo heads have different fields of view (FOVs) and the camera with a wider FOV is used for sea-floor mapping and object detection. In particular, the images of the *Soller* dataset were obtained with the BumbleBee2 BB2-08S2C-25, a very short focal length 8 Mpx camera capable of 97 deg HFOV (Figure 6(c)). With these configurations and a resolution of 1024x768, each pixel corresponds to an area of 1.5 mm^2 at a rough distance of two metres. Figure 7(c) reproduces a sample image from the *Soller* dataset. At the bottom of this image, a part of the arm gripper is clearly visible. This dataset includes many images with the gripper or some part of it in the camera field of view: clearly, self-occlusions must be dealt with. Images were acquired and processed by the on-board ECU, consisting of a mini-ITX board with a CPU Intel i5 @ 2.33 GHz. This configuration was able to acquire and process image pairs at approximately 2 fps [34].

4.2 Mono Camera Object Detection

The aim of the ROI identification algorithms is the identification of the image region containing a significant part of the target object. Such techniques are designed in order to overestimate the region including the target, and must contain a portion of the target to enable the detection operation which follows. As discussed in section 2, alternative approaches like those based on the extraction of local features do not achieve reliable detection.

The performance of ROI_{colour} has been assessed on the Garda, Portofino and Soller datasets and the results of this assessment are shown in Table 1. The frame resolution of each dataset depends upon the camera used in the image acquisition. Furthermore, the datasets are divided into subsets with homogeneous features in the images, i.e., the target object's colour for the Garda and Portofino datasets, and the presence of other objects like the robotic arm for the Soller dataset. A target reference colour is chosen for each group. The output pixels of ROI_{colour} are classified as true positive (TP), true negative (TN), false positive (FP) and false negative (FN) through the comparison with the ground-truth. Table 1 reports the average values of the overlapping results, and the recall and precision for each image subset. In general, the proposed approach performs the ROI identification for the target with high *precision*, except for the Portofino dataset with grey targets and for the Soller dataset with images including both the target object and the arm. The limited precision for the grey target object in the Portofino images depends upon the similarity between the target's colour and the seabed in the background. These results, in terms of colour-based object detection with pixel-level evaluation, appear to be in reasonable agreement with the results from Bazeille et al. in [21]. In essence, contrasting or brilliant colours are reliably detected in underwater images up to a few metres, with very few FPs. There can indeed be a number of FNs, i.e., missed perceptions, which nonetheless can be coped with through repeated observations. In the Soller dataset, the ROI_{colour} computation algorithm tended to treat the target object and the part of the robotic arm in view in the same way, unless the arm zone was suitably discarded from the images (as in the column "Target" in Table 1). The *recall* values are generally low, but the detected portion of the object is sufficient for overall object detection, as pointed out in [35].

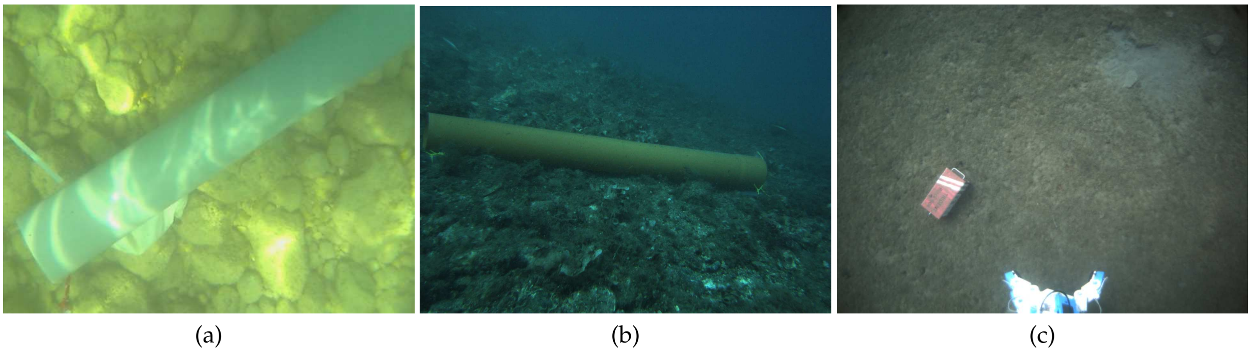


Figure 7. Sample images from the (a) *Garda*, (b) *Portofino* and (c) *Soller* datasets

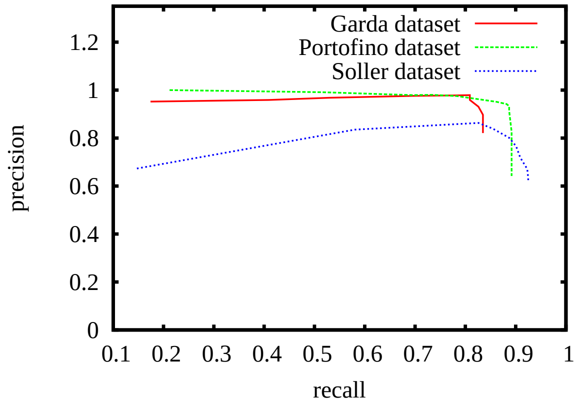


Figure 8. Precision-recall curves of the ROI_{shape} algorithm applied to the Soller and Garda datasets

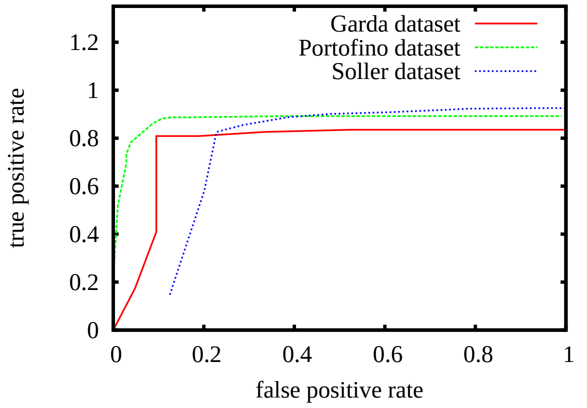


Figure 9. ROC curves of the ROI_{shape} algorithm applied to the Soller and Garda datasets

Dataset	Garda		Portofino		Soller		
	Grey	Orange	Grey	Orange	Target & Arm	Target	Arm
Frames	89	47	107	111	151	354	290
Width	640		1292		1024		
Height	480		964		768		
TP [px]	21117	11411	26407	43775	1688	2548	0
TN [px]	256941	285458	1070788	1163113	772154	776977	760148
FP [px]	334	305	62502	569	6076	360	26284
FN [px]	28807	10026	85791	38030	6513	6547	0
Recall	40.3%	55.9%	29.0%	53.3%	20.6%	27.6%	-
Precision	98.2%	98.2%	49.0%	98.8%	44.1%	93.5%	-

Table 1. The average overlap of the ROI_{colour} area and the target object measured in pixels and the corresponding values of recall and precision. The overlap is measured by true positive (TP), true negative (TN), false positive (FP) and false negative (FN) pixels. The assessment distinguishes the experimental setup according to the target object's colour (for the Garda and Portofino datasets) or to the objects appearing in the scene (in particular, the robotic arm of the Soller dataset).

	Garda	Portofino	Soller
Frame number	136	305	505
sigma _th	0.04	0.04	0.027
TP	93	179	417
TN	19	91	223
FP	2	11	66
FN	22	24	88
Precision	97.9%	94.2%	86.3%
Recall	80.9%	88.2%	82.6%
Accuracy	82.4%	88.5%	80.6%
-FPRate	90.5%	89.2%	77.2%
F-Measure	88.6%	91.1%	84.4%

Table 2. Detection results for the Garda, Portofino and Soller datasets for a specific value of acceptance threshold σ_{th}

Num. Frames	Avg. Distance	Std.Dev. Distance
302	1441 mm	169 mm

Table 3. Mono camera estimated distances

A second set of experiments has been performed for the ROI_{shape} algorithm. This algorithm segments the image into different clusters and classifies each cluster according to its shape. The segmentation into clusters requires a certain similarity among the features extracted for each pixel. Since the ROI_{shape} operates with colour features, the approximate colour uniformity of the target objects is a requirement, although the algorithm is tolerant to the violation of such an hypothesis. The algorithm detects objects with straight and sharp contours, as with many human-made artefacts, especially when they are compared with the less regular borders. The objects used in the considered datasets meet such requirements: grey and orange pipes for the Garda and Portofino datasets, and a red-black box for the Soller dataset. In this assessment, the target is considered to be detected if its ground-truth area overlaps at least 50% with the ROI_{shape} area. Figures 8 and 9, respectively, show the precision-recall (PR) and receiving operating curve (ROC) for the different datasets. Table 2 illustrates the classification values obtained for a specific value of σ_{th} . Note that the PR curve never reaches zero due to the filtering in the segmentation phase: even after increasing the threshold σ_{th} , clearly negative images are not classified as positive and the precision is not compromised.

The PR and ROC corresponding to the Garda and Portofino datasets tend to dominate those obtained for the Soller dataset, at least for a restrictive choice of σ_{th} . This result can be explained by observing that, although rather general, ROI_{shape} has been designed to find and classify pipes with uniform colour. The red-black box from Soller is less

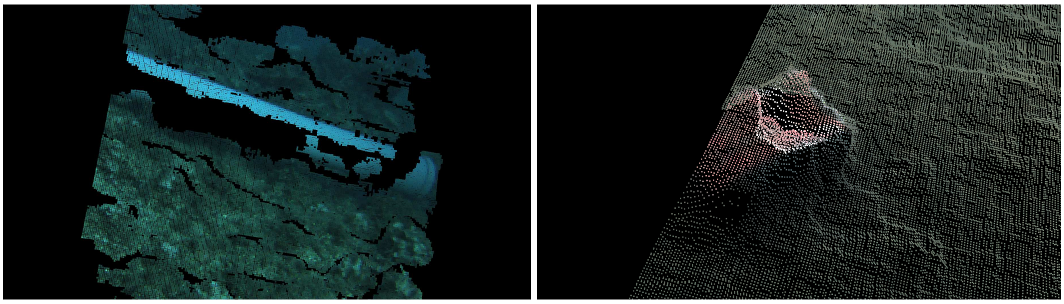


Figure 10. Detailed view of the target object point clouds obtained from the Portofino (left) and Soller (right) datasets. The point cloud can display void regions or shape distortions where the stereo processing fails.

elongated and is marked by white stripes. These features stress, respectively, the shape validation step and the segmentation (e.g., the object could be split into different connected components). Furthermore, the robotic arm belonging to the underwater vehicle appears in the scene and a more advanced method could shadow its projection in the image. However, ROI_{shape} has proven robust to such differences even though they were not predicted during the algorithm's design. The precision in Table 2 is above 85% and its recall is even higher than that obtained for the Garda dataset.

	Grey	Orange
Num. Frames	107	111
True Radius [mm]	45.0	50.0
Avg. [mm]	55.6	85.7
Std. Dev. [mm]	36.9	32.2
Max [mm]	249.5	191.4

Table 4. Cylinder radius estimated in the Portofino dataset for grey and orange pipes

Mono camera images have been used to estimate the pose of a cylindrical pipe - as discussed in section 3.2 - in the Garda dataset images. The algorithm computes all the parameters of the cylinder axis that allow the localization of the target object. However, during experiments at Lake Garda, the embedded system swung rather quickly when attached to the floating support, due to continuous waves (see Figure 5). In such experiments, no ground-truth is usually available and therefore a parameter that is invariant to camera motion is required to assess the precision of the proposed method. The object lies on the lake floor and the camera depth remains approximately constant. Thus, the distance between the camera centre and the cylinder axis in equation (5) approximately meets this prerequisite. Table 3 illustrates the average distance and the standard deviation of the axis computed in a sequence of 302 frames. The standard deviation of 17 *cm* is due to both the estimation error and the slight variation of distances caused by waves.

4.3 Stereo Vision Processing

All three datasets considered in this work provide pairs of camera frames acquired by stereo cameras. The Garda dataset differs from the other two since it uses three cameras and the acquisition is not triggered by a signal. The resulting disparity images and the corresponding point clouds are rather sparse and inaccurate. In a few cases, when the target object is relatively still with respect to the cameras, it is possible to compute the point cloud and to apply the 3D object recognition algorithms illustrated in section 3.4 in the Garda dataset. The method operates only on those points corresponding to an ROI obtained, for example, with ROI_{colour}. Figure 11 shows an example of recognition based on alignment. The success rate of this approach is generally low, with a precision value of around 60% and recall around 40%. The classification is mainly due to ROI identification.

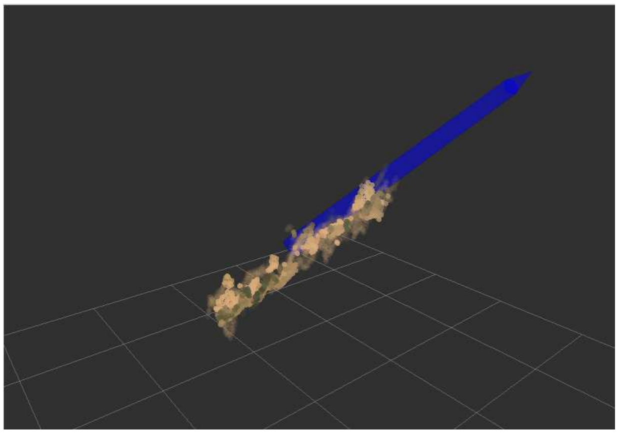


Figure 11. An example of pose estimation by matching the raw point cloud from the Garda dataset (orange) and a cylinder model (blue)

The Portofino and Soller datasets provide high-quality and synchronized camera frames. Figure 10 illustrates two examples of point clouds obtained by each of the two collections. The images are globally dense, although there are empty regions where the disparity image fails or is incorrect. For example, the holes are located in correspondence with the uniform texture of the pipe in Figure 10(left). Moreover, the resulting 3D reconstruction is sometimes distorted (Figure 10(right)). Figure 12 shows two examples

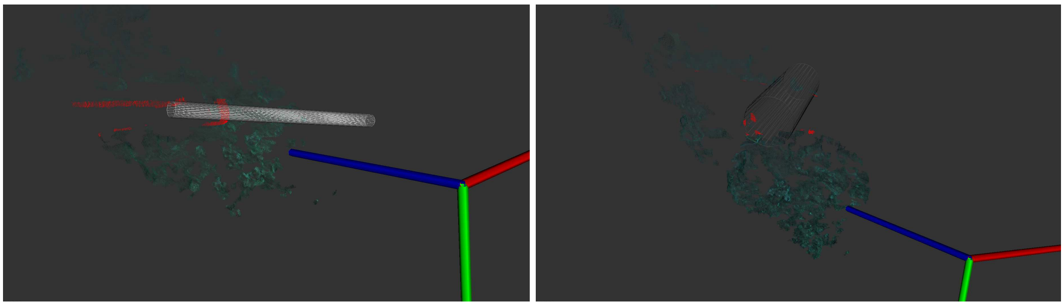


Figure 12. Pose estimation results of a cylinder fitting in the 3D point cloud for the Portofino dataset. The points in the ROI are in red, the cylinder is computed by model-fitting with RANSAC and the reference frame corresponds to the left camera of the stereo system.

of RANSAC cylinder-fitting. Thanks to the previous object detection, the estimated pose is close to the correct object location. When there are enough 3D points lying in the ROI (Figure 12(left)), the cylinder axis is computed with acceptable accuracy. Otherwise, the resulting pose is rather inaccurate (Figure 12(right)). The assessment of the object's dimensions and, in particular, of its radius in Table 4 leads to the same conclusion. The estimated radius is close to the correct value, with a standard deviation of around 3 *cm* for both the grey and the orange pipes. This result depends upon both the reliable monocular object detection and the density of the stereo disparity image. The complete object recognition based on the point cloud remains difficult due to the inaccurate underwater perception of objects. However, in underwater environments stereo cameras enable the evaluation of a 3D target object's position once the object has been already detected in the corresponding single image.

5. Discussion

The experimental assessment was performed across several datasets acquired at different depths (from 2 *m* to 10 *m* depth), light conditions and objects. Not all the techniques presented in section 3 can be applied to all the datasets. The ROI_{shape} method has been successfully used to detect target objects by exploiting colour uniformity and shape regularity in all three datasets. These assumptions are standard characteristics in an underwater context, as illustrated by the discussion of related work in section 2.

Popular approaches to object detection, like feature constellation methods, are less reliable for underwater computer vision. The standard SIFT key-point feature [36] was tested on sample images from the Portofino dataset. Figure 13 shows a few examples of the resulting feature association between the model to be found and an image containing the same object. The results are clearly unreliable. Although features are in general less stable with texture-less objects, like the orange pipe, the associations are strongly affected by the different luminance of the target (e.g., in the leftmost example of Figure 13 the model features are matched with another pipe). Thus, the feature constellation method, which depends upon the association between the object model's features and the extracted ones, is not reliable.

The 3D data obtained from stereo processing have been used to estimate the target object pose for the Garda and Portofino datasets. The point clouds resulting from the stereo camera are not dense or detailed enough to enable more sophisticated approaches than geometric model-fitting. The precision of the pose estimation depends upon the density of the disparity map, while the density of the disparity map in turn relies upon the correspondence of homologous points in the two images. Only the accurate assessment of the ROI containing the target object avoids erroneous matches between the geometric model and the candidate 3D points. When an accurate ROI is available and there are enough points belonging to the target, the estimated values of the object's dimensions (i.e., the radius) and pose are close to the correct ones (Figure 12(left)). Unfortunately, and quite often, the object point cloud is not sufficiently populated to achieve an acceptable estimation (see Figure 12(right)).

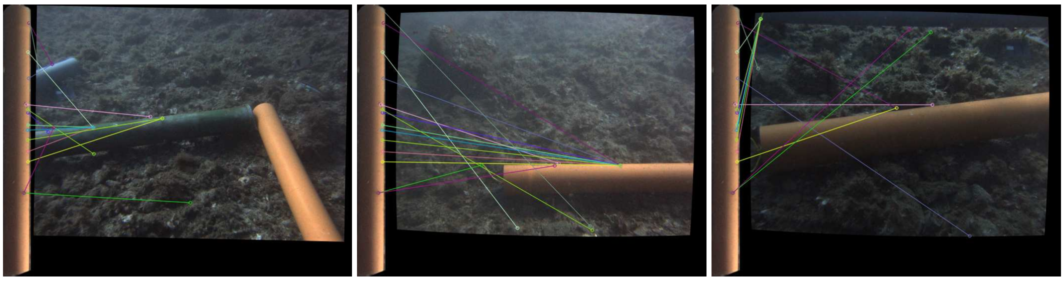


Figure 13. Results of the constellation method with SIFT feature extraction and matching applied to the Portofino dataset at different light conditions. The matches are represented by coloured segments between the features extracted from the object model image (on the left) and those extracted from each image. The resulting associations are rather unreliable.

6. Conclusion

This paper has investigated vision-based object detection algorithms in underwater environments using multiple datasets. We have presented a complete algorithmic pipeline for underwater object detection and pose estimation and, in particular, a novel multi-feature object detection algorithm to find human-made artefacts. The proposed method operates according to rather general hypotheses on the salient colour uniformity and sharp shape of the target object, and has been effectively used to search different items in several underwater scenarios. Once the target is detected, its pose with regard to the camera can be estimated using the region of disparity image corresponding to such a target. The performance of the proposed algorithm has been assessed using three underwater datasets. The camera quality, the experimental conditions (depth, light conditions, background, sensor guidance, etc.) and the types of targets are different in the used datasets. However, the object detection algorithm has proven robust to such differences, and achieves satisfactory values of precision and recall in underwater environments. In the design of a stereo vision system for underwater objects' manipulation, the importance of monocular detection is apparent in connection with the unreliability of simple segmentation techniques and with the shape distortion occurring in stereo camera reconstruction.

7. Acknowledgements

This article is a revised and expanded version of a paper entitled *Bio-inspired Object Detection and Pose Estimation Algorithms for Underwater Environments* presented at the Workshop-Conference on Bio-inspired Robotics - International Advanced Robotics Program (IARP), ENEA-Frascati, Roma (IT), May 14-15, 2014 [1].

This research has been supported by MIUR in the MARIS project. We would like to thank the Center for Underwater Robotics Research of the University of Girona (Spain), the Robotic Intelligence Lab of the University of Jaume-I and the ISME of the University of Genova for making the Trident Soller dataset freely available.

8. References

[1] F Kallasi, F. Oleari, M. Bottioni, D. Lodi Rizzini, and S. Caselli. Bio-inspired object detection and pose estimation algorithms for underwater environments. In *Workshop-Conference on Bio-inspired Robotics of International Advanced Robotics Program (IARP)*, pages 1–6, May 2014.

[2] A. T. Targhi, E. Hayman, J.-O. Eklundh, and M. Shahshahani. The eigen-transform and applications. In *Proceedings of the 7th Asian conference on Computer Vision - Volume Part I*, ACCV'06, pages 70–79, Berlin, Heidelberg, 2006. Springer-Verlag.

[3] G. Casalino, M. Caccia, A. Caiti, G. Antonelli, G. Indiveri, C. Melchiorri, and S. Caselli. Maris: a national project on marine robotics for interventions. In *Mediterranean Conference on Control & Automation*, 2014.

[4] J. J. Javier Fernandez, M. Prats, P.J. Sanz, J. C. Garcia, R. Marin, M. Robinson, D. Ribas, and P. Ridao. Grasping for the seabed: Developing a new underwater robot arm for shallow-water intervention. *IEEE Robot. Automat. Mag.*, 20(4):121–130, 2013.

[5] S.-C. Yu, T.-W. Kim, A. Asada, S. Weatherwax, B. Collins, and Junku Yuh. Development of high-resolution acoustic camera based real-time object recognition system by using autonomous underwater vehicles. In *OCEANS 2006*, pages 1–6, 2006.

[6] P. Jonsson, I. Sillitoe, B. Dushaw, J. Nystuen, and J. Heltne. Observing using sound and light: a short review of underwater acoustic and video-based methods. *Ocean Science Discussions*, 6(1):819–870, 2009.

[7] D.P. Williams and J. Groen. A Fast Physics-Based, Environmentally Adaptive Underwater Object Detection Algorithm. In *Proc. of OCEANS*, pages 1–7, 2011.

[8] J. Sawas, Y.R. Petillot, and Y. Pailhas. Cascade of Boosted Classifiers for Rapid Detection of Underwater Objects. In *ECUA 2010 Istanbul Conference*, pages 1–8, 2010.

[9] Alan Gordon. Use of laser scanning system on mobile underwater platforms. In *Autonomous Underwater Vehicle Technology, 1992. AUV'92., Proceedings of the 1992 Symposium on*, pages 202–205. IEEE, 1992.

[10] M. Narimani, S. Nazem, and M. Loueipour. Robotics vision-based system for an underwater pipeline and cable tracker. In *OCEANS 2009 - EUROPE*, pages 1–6, 2009.

[11] T. Nicosevici, N. Gracias, S. Negahdaripour, and R. Garcia. Efficient three-dimensional scene modeling and mosaicing. *Journal of Field Robotics*, 26(10), 2009.

[12] R. Eustice, H. Singh, J. Leonard, M. Walter, and R. Ballard. Visually navigating the rms titanic with slam information filters. In *Proceedings of Robotics: Science and Systems*, Cambridge, USA, June 2005.

[13] R. Garcia and N. Gracias. Detection of interest points in turbid underwater images. In *IEEE OCEANS*, pages 1–9, 2011.

[14] J. Aulinas, M. Carreras, X. Llado, J. Salvi, R. Garcia, R. Prados, and Y.R. Petillot. Feature extraction for underwater visual SLAM. In *TODO: FIX*, pages 1–7, 2011.

[15] J.P. Queiroz-Neto, R. Carceroni, W. Barros, and M. Campos. Underwater stereo. In *Computer Graphics and Image Processing, 2004. Proceedings. 17th Brazilian Symposium on*, pages 170–177, 2004.

[16] V. Brandou, A.-G. Allais, M. Perrier, E. Malis, P. Rives, J. Sarrazin, and P.-M. Sarradin. 3D reconstruction of natural underwater scenes using the stereovision system iris. In *OCEANS 2007 - Europe*, pages 1–6, 2007.

[17] R. Campos, R. Garcia, and T. Nicosevici. Surface reconstruction methods for the recovery of 3D models from underwater interest areas. In *OCEANS, 2011 IEEE - Spain*, pages 1–10, 2011.

[18] A. Leone, G. Diraco, and C. Distante. Stereoscopic system for 3-d seabed mosaic reconstruction. volume 2, pages 541–544, Sep. 2007.

[19] P. Primo Zingaretti and S.M. Zanoli. Robust real-time detection of an underwater pipeline. *Engineering Applications of Artificial Intelligence*, 11(2):257 – 268, 1998.

[20] G.L. Foresti and S. Gentili. A hierarchical classification system for object recognition in underwater environments. *Oceanic Engineering, IEEE Journal of*, 27(1):66–78, Jan 2002.

[21] S. Bazeille, I. Quidou, and L. Jaulin. Color-based underwater object recognition using water light attenuation. *Intel Serv Robotics*, 5:109–118, 2012.

[22] D. Lee, G. Kim, D. Kim, H. Myung, and H.-T. Choi. Vision-based object detection and tracking for autonomous navigation of underwater robots. *Ocean Engineering*, 48:59â€“68, 2012.

[23] A. Olmos, E. Trucco, and D. Lane. Automatic man-made object detection with intensity cameras. In *OCEANS '02 MTS/IEEE*, volume 3, pages 1555–1561, Oct 2002.

[24] M. Prats, J.C. Garcia, S. Wirth, D. Ribas, P.J. Sanz, P. Ridao, N. Gracias, and G. Oliver. Multipurpose autonomous underwater intervention: A systems integration perspective. In *Mediterranean Conference on Control & Automation*, pages 1379–1384, July 2012.

[25] D. Kim, D. Lee, H. Myung, and H.-T. Choi. Object detection and tracking for autonomous underwater robots using weighted template matching. In *OCEANS, 2012 - Yeosu*, pages 1–5, 2012.

[26] C. Ancuti, C.O. Ancuti, T. Haber, and P. Bekaert. Enhancing underwater images and videos by fusion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 81–88, 2012.

[27] S.M. Pizer, E.P. Amburn, J.D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B.T.H. Romeny, and J.B. Zimmerman. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, Image Processing*, 39(3):355–368, September 1987.

[28] S. Sural, G. Qian, and S. Pramanik. Segmentation and histogram generation using the HSV color space for image retrieval. In *International Conference on Image Processing.*, volume 2, pages II–589–II–592 vol.2, 2002.

[29] F. Kallasi, D. Lodi Rizzini, J. Aleotti, and S. Caselli. Bio-Inspired Object Detection and Pose Estimation Algorithms for Underwater Environments. In *Workshop-Conference on Bio-inspired Robotics, International Advanced Robotics Program (IARP)*, pages 1–7, 2014.

[30] R.O. Duda, P. E Hart, and D.G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[31] R. Ugolotti, Y. S.G. Nashed, P. Mesejo, S. Ivekovi, L. Mussi, and S. Cagnoni. Particle swarm optimization and differential evolution for model-based object detection. *Applied Soft Computing*, 13(6):3092–3105, 2013.

[32] F. Oleari, F. Kallasi, D. Lodi Rizzini, J. Aleotti, and S. Caselli. Performance Evaluation of a Low-Cost Stereo Vision System for Underwater Object Detection. In *World Congress of the International Federation of Automatic Control (IFAC)*, pages 3388–3394, 2014.

[33] F. Oleari, D. Lodi Rizzini, and S. Caselli. A low-cost stereo system for 3d object recognition. In *Intelligent Computer Communication and Processing (ICCP), 2013 IEEE International Conference on*, pages 127–132, 2013.

[34] E. Simetti, G. Casalino, S. Torelli, A. Sperinde, and A. Turetta. Floating underwater manipulation: Developed control methodology and experimental validation within the trident project. *Journal of Field Robotics*, 31(3):364–385, 2014.

[35] C. Barngrover, S. Belongie, and R. Kastner. Jboost optimization of color detectors for autonomous underwater vehicle navigation. In P. Real, D. Diaz-Pernil, H. Molina-Abril, A. Berciano, and W. Kropatsch, editors, *Computer Analysis of Images and Patterns*, volume 6855 of *Lecture Notes in Computer Science*, pages 155–162. Springer Berlin Heidelberg, 2011.

[36] D.G. Lowe. Distinctive image features from scale-invariant keypoints. 60(2):91–110, 2004.