Hindawi Publishing Corporation
Journal of Sensors
Volume 2016, Article ID 4265042, 16 pages
http://dx.doi.org/10.1155/2016/4265042

Research Article

Robotic Visual Tracking of Relevant Cues in Underwater Environments with Poor Visibility Conditions

Alejandro Maldonado-Ramírez and L. Abril Torres-Méndez

Robotics and Advanced Manufacturing Group, CINVESTAV Campus Saltillo, 25900 Ramos Arizpe, COAH, Mexico

Correspondence should be addressed to L. Abril Torres-Méndez; abriltorresm15@gmail.com

Received 26 March 2016; Revised 18 June 2016; Accepted 28 June 2016

Academic Editor: Youcef Mezouar

Using visual sensors for detecting regions of interest in underwater environments is fundamental for many robotic applications. Particularly, for an autonomous exploration task, an underwater vehicle must be guided towards features that are of interest. If the relevant features can be seen from the distance, then smooth control movements of the vehicle are feasible in order to position itself close enough with the final goal of gathering visual quality images. However, it is a challenging task for a robotic system to achieve stable tracking of the same regions since marine environments are unstructured and highly dynamic and usually have poor visibility. In this paper, a framework that robustly detects and tracks regions of interest in real time is presented. We use the chromatic channels of a perceptual uniform color space to detect relevant regions and adapt a visual attention scheme to underwater scenes. For the tracking, we associate with each relevant point superpixel descriptors which are invariant to changes in illumination and shape. The field experiment results have demonstrated that our approach is robust when tested on different visibility conditions and depths in underwater explorations.

1. Introduction

Visual tracking of relevant regions in scenes with poor visibility is an important problem in robotic vision research. In particular, for the autonomous robotic exploration of natural underwater structures (e.g., coral reefs), it is fundamental to perform a closer, cautious, and a noninvasive analysis of the changes that occur in the structure of interest to assist in the research of marine biologists. Usually, human intervention is required to indicate which regions are of interest for monitoring by remotely operating the underwater vehicle. As this can be quite demanding, the need of using an Autonomous Underwater Vehicle (AUV) is very appealing. Moreover, the visual and control algorithms need to be quite robust and run in real time in order to be effective. In recent years, several systems capable of collecting information, dynamically or statically, in underwater environments have been developed. In the case of AUVs, great efforts have been made to provide them with sufficient autonomy to perform specific tasks. Thus, the main challenge is to transfer to the robotic agent the ability of recognizing what regions are of interest for monitoring and to keep those regions on view for a certain period of time to be able to obtain useful visual data for its posterior analysis. However, as these targets or regions of interest may be located far from the vehicle, they need to be detected from the distance. The rapid attenuation of electromagnetic radiation in water limits the range of optical sensors. Also, the existence of variable lighting and the presence of suspended particles (also known as marine snow) cause geometrical and color distortions that result in poor visibility. Furthermore, the structure (in terms of geometric shape) of coral reefs is practically null. Since underwater environments are highly unstructured and constantly changing environments, one of the main problems that still remains open is the accurate estimation of the robot's position and orientation. This makes the detection and tracking of visual cues difficult. Considering the mentioned problems, if the goal is to cautiously explore the fragile marine life that exists in coral reefs, it is necessary to first detect visual targets that are relevant for the exploration and then

robustly track them so that the robot movements are not erratic or abrupt. In other words, the tracking must be stable enough to allow for smooth control movements in the robotic system.

We are interested in allowing an AUV to conduct an exploration of coral reefs according to how a human diver would do it: that is, the route to follow is guided by the features in the environment that catch her attention. It turns out that for underwater environments using this type of exploration there exists limited research work in the literature. For example, in [1], a method is presented to classify the captured images by the robot according to the degree of novelty contained in the features. The novelty parameter is an indicator used to control the speed of the robot along a predefined path. An extension of this work is presented in [2], where the movement of the robot is controlled to be directed to areas in the image with more visual content, causing the robot to move to areas containing coral reef and ignore the areas where only sand is present. One important thing to note is that, in an exploration mode, it is crucial not to limit the movements of the robot to a previous specified path; instead, the approach used should allow for a more natural scanning. In this sense, a diver (sufficiently curious and fearless) exploring a coral reef for the first time will be guided by what catches her attention, despite not having prior information about what she could find.

In this research work, we present a real-time visual-based framework to robustly detect and track relevant features from the distance with the aim of exploring coral reefs. The real-time performance in robotics applications is fundamental since the tracked features will help to direct the exploration trajectories in subsequent captured images while estimating the relative pose of the robot. We build upon our previous work [3, 4]. In [4], a visual attention model, adapted to underwater scenes, was presented for the first time. The inputs were a set of videos taken underwater. Although the visually relevant cues were likely to be detected on subsequent frames, it was not enough to keep track of a particular relevant cue for long. Moreover, it only worked when water conditions were optimal, thus failing when poor visibility conditions were present. In [3], we characterized the colors of relevant features by using a perceptually uniform color space. We compared the CIE *Lab* and the $L\alpha\beta$, which were able to define a super-color-pixel descriptor to describe a relevant region by using its chromatic channels only. The color opponent processing (*blue-yellow* and *green-red*) makes the recovering of color underwater easy, in particular red and yellow tones, by enhancing their contrast wrt the blue/green tonalities of sea waters.

In this paper, we have extended our previous work in many aspects. First, we give a detailed description of each of the stages involved in our Aquatic Visual Attention (AVA) model as well as improvements to have a better saliency map in terms of the compactness of the relevant regions. Second, we have compared the performance of the proposed framework. On one hand, we compare the quality in the detection of regions of interest of our AVA model in underwater scenes at different depths with the classic Neuromorphic Vision Toolkit method. On the other hand, we compare the robustness of superpixels descriptors for tracking the most relevant region of interest with other methods of object tracking.

The contribution of this paper is a novel computational visual attention model built to work on underwater environments, namely, coral reefs. The proposed visual attention model focuses on detecting as well as tracking relevant regions. The purpose of having a tracker is to lead the motion of an Autonomous Underwater Vehicle (AUV) in an exploration task. This way the AUV should be able to detect, without human intervention or any kind of precise information of a particular region, which part of the coral reef could draw the attention for a human and move towards it.

The outline of the paper is as follows. Section 2 presents background on the perception of color in underwater scenes and also on visual attention models. Section 3 describes our model and its implementation. The experimental results, comparison of the performance of the proposed framework, and discussion are presented in Section 4. Finally, in Section 5, the conclusions and future work are given.

2. Background

2.1. Underwater Perception of Color. Poor visibility conditions underwater affect the perception of color. This is due to the attenuation of light, water conditions, distance to objects, depth, and other factors [5]. Visibility in foggy days is very similar to that of underwater images. The effect is that near objects are clearer while distant objects gradually disappear. This effect is illustrated in Figure 1 by comparing images of the same natural scene under foggy and normal day conditions. The mountains in the back of Figure 1(a) cannot be seen in Figure 1(b).

Color perception in common sea water diminishes according to the distance or depth where the object of interest is located. In most cases, the color in objects that are more than 10 meters of distance are almost indistinguishable (see Figure 2). As for depth, the first color to disappear is red; beginning as soon as 3 m of depth there is almost not red light left from the sun. From 5 m to 10 m, the range from orange to yellow lights is lost. By 25 m, only blue light remains [5]. Figure 3 shows an example of an image of our AUV at different depth and water conditions. We know that the color of our robot is red by the sides. By verifying the color of the intensity pixels in a small window (zoomed in), we see that the color is very different from red, ranging from dark red to dark blue. However, the processes carries out in our brain adjust the colors up to certain grade.

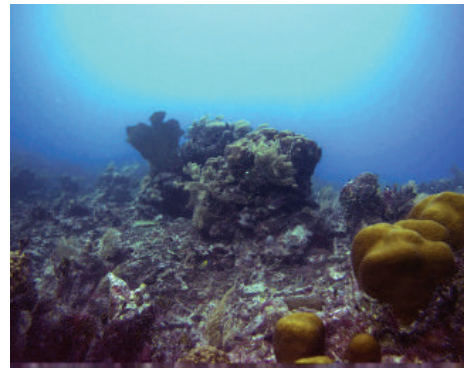
2.2. Perceptually Uniform Color Spaces for Color Discrimination. Natural structures underwater, such as the formations of coral reef, are rich in color and texture. They may have certain shapes, but they do not always follow a specific pattern or geometry. Thus, if we want to have a descriptor for a given feature, the only cue to detect and recognize would be color. In this trend, the discrimination of color is the problem we want to solve. This is different to the color

(a) Image in a sunny day

(b) Image in a foggy day

Figure 1: The same outdoor natural scene under different weather conditions. In (b), the effects of fog clearly show how objects in the distance gradually disappear (e.g., mountains in the back, clearly seen in (a), are gone). This effect is similar in underwater scenes (see Figure 2). Taken from [3].

(a)



(b)

Figure 2: Examples showing the effect of distance perception in underwater. Same as in foggy days, distant objects gradually disappear. However, an additional effect is that near objects appear bigger than they actually are.

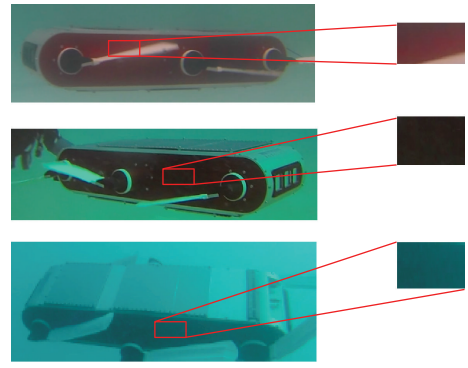


Figure 3: Example of color perception at different depth and water conditions.

restoration problem, in which a good result is basically that with a *natural look* of color appearance, but there is not guarantee that the true original color has been recovered whatsoever.

To discriminate color, one needs to measure the differences among the entire range of visible colors in a way that matches perceptual similarity as good as possible. This task can be simplified by the use of perceptually uniform color spaces, in which a small change of a color will produce the same change in perception anywhere in the color space. This is due to the fact the chromatic channels are spaced further apart. Examples of perceptual uniform color spaces are the CIE *Lab* and the *Lαβ*. On one hand, the CIE *Lab* model was specifically developed to describe all the color that the human eye can perceive [6] and it was designed to preserve the perceptual color distance. Thus, the Euclidean distance is an accurate representation of the perceptual color difference. The *a* channel values represent the relative light purplish red (magenta) or greenness of each pixel. Shifting the curve upwards builds up magentas and weakens greens. The *b* channel does the same for yellow versus blue. Altering the slope of these curves changes color contrast, while adjusting parts of the curve selectively changes different ranges of colors. On the other hand, the *Lαβ* is a decorrelated principal component color space. This color space was derived from a large ensemble of hyperspectral images of natural scenes using the first-order statistics of the images. Because of its decorrelation property of three channels, the *Lαβ* space has been used for color mapping in terrestrial applications [7, 8] and just recently it was used for underwater applications for color correction [9] with good results.

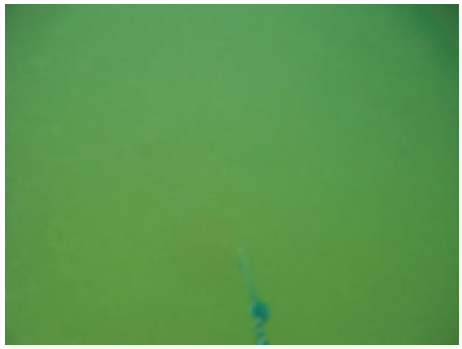
2.3. Experiments: Underwater Color Discrimination. We carried out experiments to visually compare how color can be discriminated when using the RGB, HSV, CIE *Lab*, and

(A) Input image

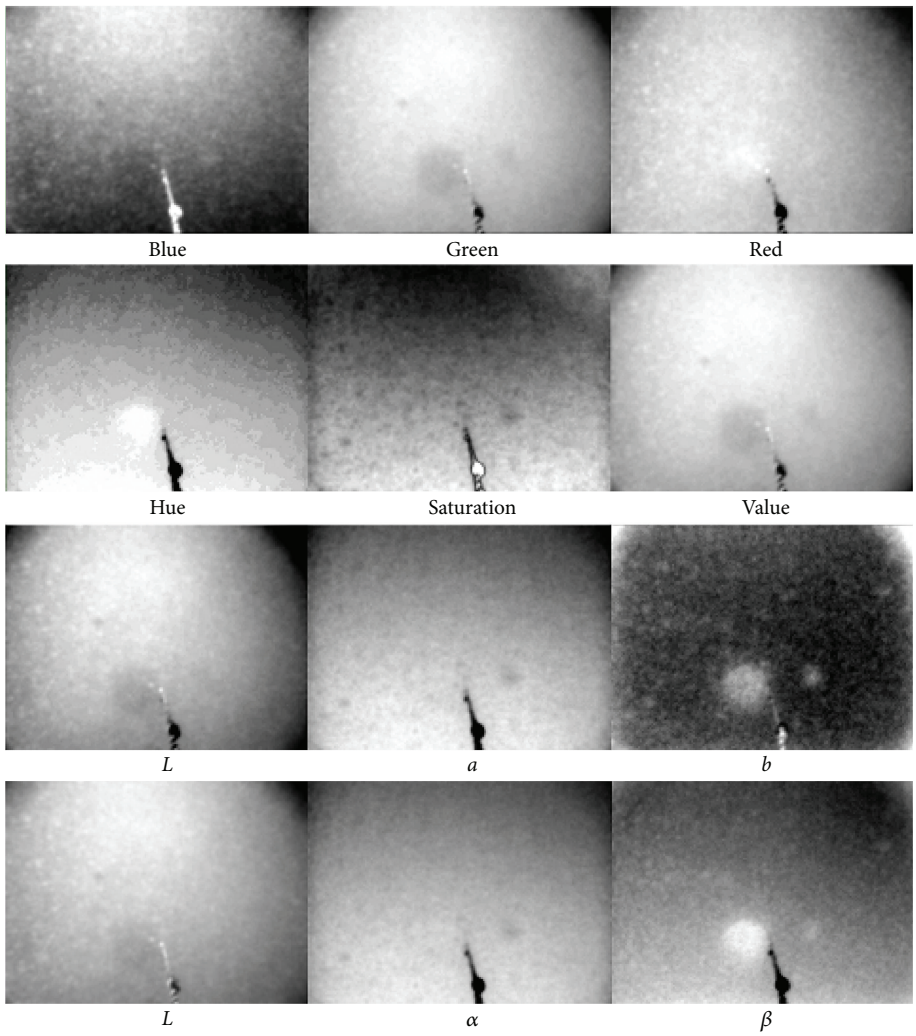
(B) The channels of the RGB, HSV, CIE *Lab* and *Lαβ* color spaces

Figure 4: Chromatic channels of different color spaces applied to an outdoor scene. Note that the red channel of RGB and the *a* channel of CIE *Lab* are in the last column in order to visually facilitate the comparison. Taken from [3].

the *Lαβ* color spaces. It is important to remind that our goal is to see how red and yellow tonalities are detected. We are neither doing a restoration of the color nor enhancing the color in images. The underwater images were taken on three different sea waters, from the Caribbean and the Yucatan peninsula. As it was previously mentioned, the advantage of using opponent color spaces is because for this type of images; one of the opponent colors is basically the color of water, that is, a bluish or greenish tone. Since colors are usually defined in terms of human observation, the evaluation of the performance of an algorithm that involves color information is a more qualitative aspect than a quantitative one. Figures 4 and 5 show examples of using different color spaces in an outdoor and underwater images under poor visibility



(A) Input image



(B) The channels of the RGB, HSV, CIE *Lab* and *Lαβ* color spaces

Figure 5: Chromatic channels of different color spaces applied to an underwater scene. Note that the red channel of RGB and the *a* channel of CIE *Lab* are in the last column in order to visually facilitate the comparison. Taken from [3].

conditions which are illustrated. Figure 4 depicts in the first row the input image taken outside water; then in the next rows, the three channels of the RGB, HSV, CIE *Lab*, and *Lαβ* color spaces, respectively, are shown. In similar arrangement of images, Figure 5 shows the three channels of each color space when applying to underwater images in poor visibility conditions.

It can be observed that all color spaces discriminate red and yellow colors in the images. However, in underwater, only the CIE *Lab* and *Lαβ* color spaces were able to discriminate the red color of the ball. This conclusion arises from a visually qualitative comparison.

2.4. Visual Attention Models. Visual attention is a selective process that allows us to determine what draws our attention according to the visual stimuli we receive from our environment. Several works have been done in the area of neuropsychology to understand how humans pay attention to what we see. Even today, there are several theories about how the human visual attention system works. Based on those theories, various computer models have been proposed. Studies about visual attention originally emerged in the area of psychology and neurophysiology over a century ago [10], when scientists began to develop theories and models to explain it. But it was not until 1987, in the work presented by Koch and Ullman [11], when the first model of a biologically inspired computational attention was published. After this work many more were proposed, being the work by Itti et al. [12] the most relevant to date. A comprehensive survey of visual attention and its implementation in computer systems can be found in [13].

One of the motivations for incorporating attention capabilities in systems that process huge amount of information is to reduce the amount of the data to be processed. This can be achieved by taking only the information. In the area of computer vision it is particularly noticeable, as images contain thousands, even millions of pixels. The problem of reducing image information has been addressed in various ways. To mention a few, there exist methods that are based on the detection of points of interest, such as the Harris' corner detector [14], SURF [15], or the well-known SIFT [16]. Also, there are detectors of lines, ellipses, and circles [17, 18]. Another approach that has also been applied involves the predictive methods, which use information regarding the task to be performed to limit the amount of information to be processed.

Two of the more popular attention models, due to their easy implementation, flexibility, and fast computation, are the Neuromorphic Vision Toolkit (NVT) proposed by Itti et al. [12] and the attention system called Visual Object detection with a computational attention system (VOCUS) by Frintrop et al. [19]. The Focus of Attention (FoA) is the place in the image that draws the attention of the system. Itti et al. [12] searched for the FoA by using a Winner-Take-All neural network. Frintrop et al. [19] find the point with the highest saliency value by scanning every point, and the most salient region is determined by seed region growing.

Recently, visual attention models have been used in robotic applications [20], and in underwater applications to primarily assist marine biologists in their review of underwater videos. For example, Walther et al. [21] and Edgington et al. [22] detect objects and potentially interesting visual events for humans in order to label the frames of a video stream as interesting or boring. In both research works, the NVT [12] model is used. The videos used in those works were recorded by a Remotely Operated Vehicle (ROV).

Barat and Rendas [23] present a visual attention system for detection of manufactured objects. Their model is based on the minimum description length test for detecting the motion of contrasting neighboring regions. After that, a statistical technique is adapted to determine the boundary of the object. Correia et al. [24] use intensity, motion, and edge maps as features for their visual attention model to detect the Norway lobsters and help scientist to quantify them.

In all these works, the visual attention models are used for aiding humans in the task of analyzing video streams. In our case, we want the visual attention model to direct the robot motion through the automatic detection and tracking of features that could be of interest for a human during an exploration. Particularly, we are interested in transferring abilities to an AUV in order to detect regions of interest without human supervision while successfully navigating the environment. For the case of autonomous underwater exploration the visual attention algorithm requires real-time performance. Moreover, as hardware limitations in underwater robots are still an issue, the algorithms should have a low computational cost.

3. The Proposed Method

In this section, the method we propose for detecting and tracking relevant features in underwater scenes is described. Our approach for detection of relevant features uses some key ideas of Itti's and Frintop's visual attention models [12, 19]. A computational visual attention algorithm detects relevant regions in an image emulating the human visual attention.

Traditionally, the detection of relevant features relies on a saliency map—a gray-scale image in which the brightest part is the most relevant in terms of features such as intensity, color, and orientation. Given that the existing natural objects in underwater scenes lack specific orientation and shape, our attention model strongly relies on color information. However, the inherent poor visibility and color degradation of sea water are critical at distances and depths greater than 10 meters. For that reason, it is important to select an appropriate color space to achieve an effortlessly underwater image enhancement. We use the CIE *Lab* color space.

The most relevant regions can be found by selecting the location with the highest value in the saliency map. In a sequence of underwater images of the same scene, it is common that the location associated with the highest value of saliency changes drastically from one frame to the next. This is due to the variations in the illumination and/or local water conditions. Thus, if the location of the region of interest in the image domain is going to lead the motion of the vehicle

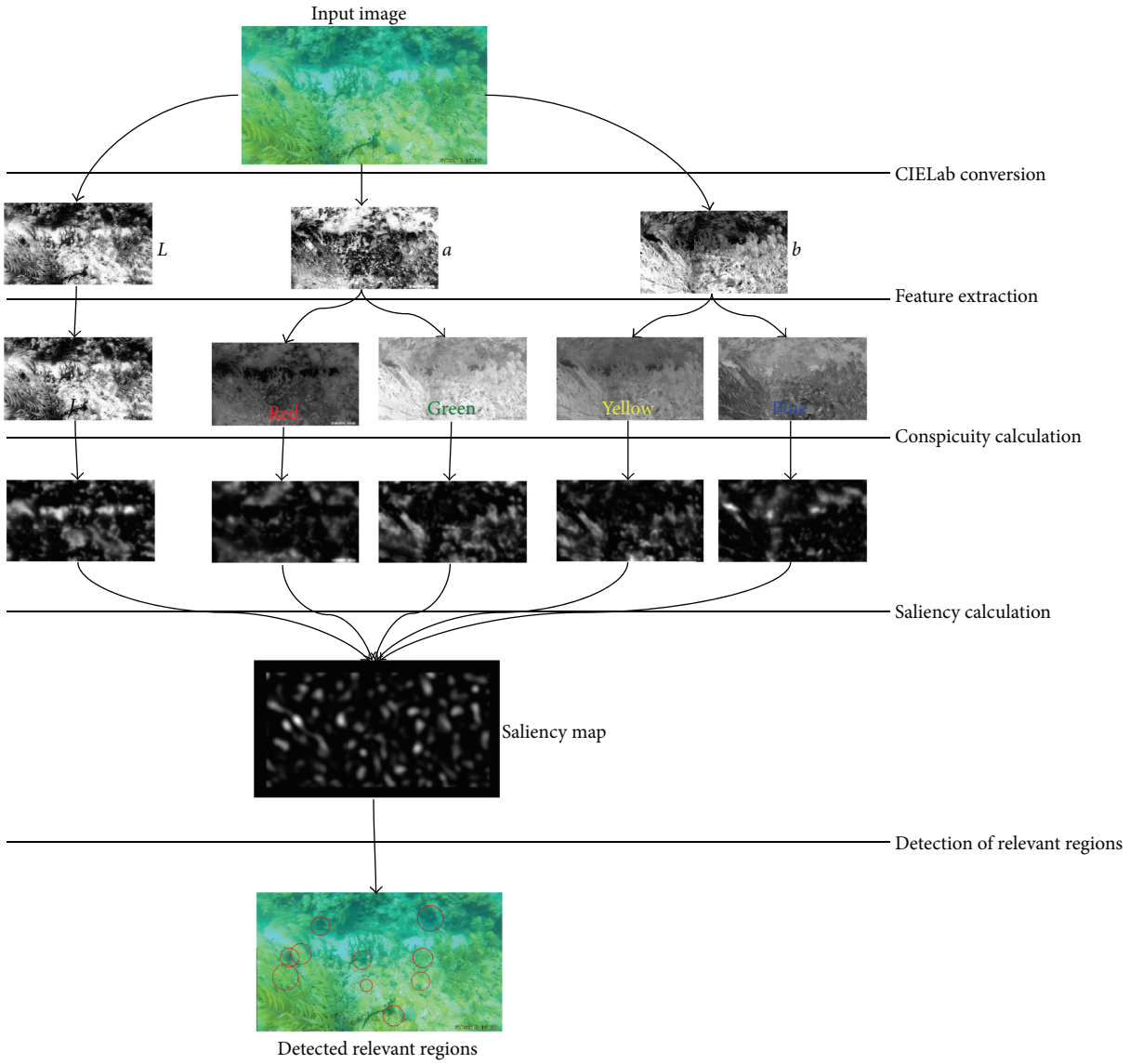


Figure 6: A general overview of the proposed method for detecting relevant regions.

in the space domain, then a robust tracking of the same or very similar region (in position and appearance) is crucial to minimize the erratic motion of the vehicle.

In the following sections, we describe in more detail each of the steps involved in our visual attention model. In Figure 6, a general overview of the proposed method for detecting relevant regions is depicted.

3.1. Preprocessing of the Image. The input image is scaled to a proper size (typically 0.25 of the original size). Then, the image is converted to the CIELab color space. In Section 2.3 some advantages of this color space as well as some examples can be found.

3.2. Getting the Features Maps. We use intensity and color (red, yellow, green, and blue) as features. The intensity map corresponds to L-channel of the CIELab image. The colors are extracted from a and b channels, as described in [25], as follows:

$$\mathbf{F}_i(x, y) = V_{\max} - \left\| ab(x, y) - p \right\|, \quad (1)$$

where $F_i(x, y)$ is ith feature map, $V_{\max} = 255$ in 8-bit depth images, $p = (a_d, b_d)$ is the desired color to extract in terms of the chromatic channels, and $ab(x, y)$ is the ab-channel of the image. The color feature maps are gray-scale images in which the intensity indicates how near is the desired color to the original color of the pixel. We do not use the orientation feature in our model, as it mainly works well in structured environments (e.g., man-made environments).

3.3. Getting the Conspicuity Maps. The conspicuity map is a gray-scale image where the most relevant regions (in terms of a feature) appear brighter than other regions. The first step to calculate these maps is to build a Gaussian pyramid for each

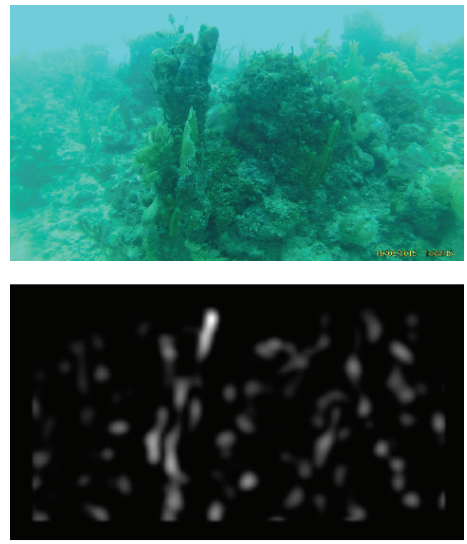


Figure 7: Example of the saliency map obtained from an underwater image.

feature. A Gaussian pyramid is built by applying a Gaussian filter and then downsampling the image in half. If we apply this process again to the resulting image, we can construct the other levels of the pyramid. The number of levels used in the pyramid depends on the size of the input image and the size of the relevant regions to be found. Bigger regions require more level in the pyramid to be effectively detected. We use a 5-level pyramid: that is, five scales $s_m = \{1, 0.5, 0.25, 0.125, 0.0625\}$.

An important aspect to consider in any computational visual attention system is highlighting the relevant part for each feature map. This is usually done by using a center-surround mechanism (also called *center-surround difference*), which is inspired in cells of the human visual receptive field [26]. In our approach, these differences are implemented as convolution. Let $\mathbf{P}(d)$ be the image in the level d of the pyramid for a given feature, then the center-surround differences are applied as follows:

$$\mathbf{P}'(d,\sigma) = \mathbf{P}(d) - \mathbf{K}(\sigma) * \mathbf{P}(d),$$
$$\mathbf{K}(\sigma) = \frac{1}{(2\sigma+1)^2 - 1} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & 0 & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{(2\sigma+1)\times(2\sigma+1)}, \quad (2)$$

where σ defines the size of the mask and $*$ is the convolution operator between an image and the mask. For each level of the pyramid two maps are obtained, $\mathbf{P}'(d,3)$ and $\mathbf{P}'(d,4)$.

The resulting images from the application of the center-surround differences are resized to 0.25 of the size of the original image. After that, all the images from the same feature pyramid are added to a single image $\mathbf{C}$, called conspicuity map.

It is important to note that, contrary to [12, 25], in which the created conspicuity map involves all colors, we calculate a conspicuity map for *each* of the color features. This allows us, in the posterior stages, to indicate which colors have more relevance during the exploration.

3.4. Getting the Saliency Map. The saliency map is a gray-scale image, in which the most relevant parts appear brighter. To obtain this map, a Difference of Gaussians (DoG) is applied to each conspicuity map. After that, a weighted sum of the resulting maps (normalized in the range $[0, 1]$) is computed. Formally, the saliency map is calculated as follows:

$$\mathbf{S} = \sum_i w_i \cdot \mathrm{DoG}(\mathbf{C}_i), \quad (3)$$

where index i represents each of the conspicuity maps obtained from each feature. By assigning different weight values w_i to each map, we can give a preference to a particular color tonality. The weighted sum can be seen as a simple way to incorporate a top-down attention. Unlike VOCUS, in which a training image containing the object to search is used, our model does not need images of a particular object. In any case, we just need to have some information about the possible dominant color of an object of interest. An example of a saliency map can be seen in Figure 7.

3.5. Searching of Relevant Points. Once the saliency map is calculated, a search for q more relevant points or *regions of interest* (RoI) is carried on. As in *VOCUS*, a sequential search of the highest values over all image pixels is done. Also, to avoid repeating the location of points, we apply an inhibition of return approach. This way, the area surrounding each of the relevant points is inhibited and the next relevant point will be far from the previous one, allowing for a sparse distribution of relevant regions. Figure 8 shows an example of the RoIs detected in an image. Unlike our previous work [3, 4] where a fixed area around a given point is inhibited, in this work a Seeded Region Growing method [27] is used over the saliency map to determine a circle that encloses the area to be inhibited.

Figure 8: Example of relevant regions detected (enclosed by red circles) in an underwater scene.

3.6. Superpixel-Based Descriptors for Tracking of Relevant Regions. From the set of regions of interest detected with the AVA algorithm, the *Focus of Attention* (FoA) is the one with the highest value. Thus, the FoA represents the region that caught the attention the most in an underwater scene.

For some applications, once a FoA is selected, it is important to keep track of it in the following images in a sequence. As our purpose is to explore an underwater environment, our AVA model must keep track of the same (or very similar) FoA in subsequent frames as much as possible, if and only if this region is still among the most relevant ones. We are interested in this behavior because it will lead the actions of a robot during an exploration task. Having abrupt changes of the FoA's location from one frame to the next one may cause an erratic motion.

To track a region or a point in an image, a descriptor is needed. We propose to use a superpixel-based descriptor. A particular advantage that superpixels offer is that they adapt their shape to enclose similar characteristics of a region, in terms of color and position. Thus, if we associate with each relevant region to be tracked the superpixel characteristics they belong to, we are assuring a local robust description.

The procedure is as follows. The input image is segmented in M superpixels using the SLIC algorithm [28] with $M \ll N$, where N the number of pixels in the input image. Each superpixel is a set of pixels with similar features and it is characterized by a 5-dimensional vector of the form $[L_s, a_s, b_s, x_s, y_s]$, where L, a, b are the mean color values of the pixels belonging to a given superpixel in the CIELab color space and (x_s, y_s) is the centroid of the superpixel. A relevant region is described by the vector $\mathbf{s}$ composed from the components a_s, b_s, x_s, and y_s from the superpixels it belongs to. It can be noted that L_s component is not taken into account because the illumination in this kind of environments can change from frame to frame.

Once we have the descriptors for each of q most relevant regions, we choose the closest one (the most similar) to the descriptor of the FoA from the previous frame. The chosen region becomes the FoA of the current frame. The distance (similarity) measure between two superpixel-based descriptors, $\mathbf{s}_j$ and $\mathbf{s}_k$, is based on the SSD metric as in [28], without the luminance part:

$$D\left(\mathbf{s}_j, \mathbf{s}_k\right) = \sqrt{\left(\frac{d_c}{N_c}\right)^2 + \left(\frac{d_s}{N_s}\right)^2}, \tag{4}$$

where

$$\begin{aligned} d_c &= \sqrt{\left(a_j - a_k\right)^2 + \left(b_j - b_k\right)^2}, \\ d_s &= \sqrt{\left(x_j - x_k\right)^2 + \left(y_j - y_k\right)^2}, \end{aligned} \tag{5}$$

where N_c and N_s are normalization factors for the distance in the color and image space, respectively. These values were set as described in [29].

Figure 9 illustrates the use of superpixels to achieve a stable tracking of similar FoAs in a region of interest. If the distance from the closest saliency descriptor to the previous FoA descriptor is greater than a defined threshold μ, the distances are ignored and the point with the highest saliency value is chosen as the new FoA.

4. Experimental Results

In this section, we present the experimental results to validated the parts of the proposed approach. First, we show the outcome of the comparison of detected relevant regions by humans and the proposed system. Then we compare the relevant regions detected by our approach (AVA) and the Neuromorphic Visual Toolkit (NVT) [12]. After that, a comparison in terms of tracking is shown. Finally, we present the outcome of using the proposed approach to guide the motion of an underwater robot in an exploration task.

4.1. Relevant Regions Detected by Humans. A comparison between the regions considered as relevant by a group of people and by the proposed approach is presented. The purpose of this experiment is to show that our visual attention algorithm is able to detect regions that have the potential to draw the attention of a human. Thus, the AUV can autonomously explore the underwater environment in terms of what a human could consider relevant.

We asked 32 people (16 men and 16 women between 20 and 30 years of age with no experience in coral reefs) to select (by clicking on the screen) the region that attracts their attention the most in a set of underwater images containing various scenes of coral reef. Then, we applied our algorithm on the same set of images. Two regions are considered coincident if their circles of radius r centered at the relevant region present an overlapping greater than 80%. Figure 10

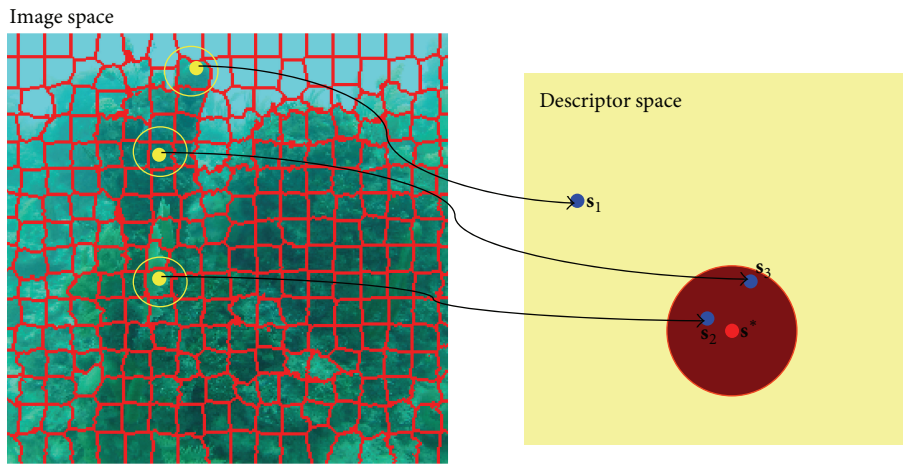


FIGURE 9: Finding the next Focus of Attention. The descriptor for each detected relevant region is obtained. The next Focus of Attention is the closest descriptor to the previous FoA's descriptor $\mathbf{s}^*$. For a descriptor $\mathbf{s}_j$ to be considered as a FoA candidate its distance to $\mathbf{s}^*$ should be less than a given threshold μ (represented as the circle around $\mathbf{s}^*$).

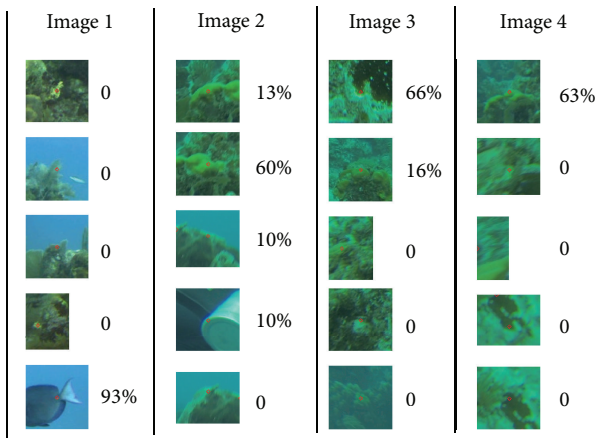


FIGURE 10: Regions of interest (RoI) detected by our method and the percentage of coincidence made by a group of 32 people. Each column shows patches containing the five most relevant regions detected by our system in the corresponding image. More than half of the group choose as relevant at least one of the areas detected with our visual attention system.

depicts the obtained results. Each row of the array of images in the figure contains the five most relevant regions detected by AVA and the percentage of people that considered the same region as relevant.

In the presented results, more than half of the group choose as relevant at least one of the areas detected with our visual attention system. This study shows us that our model approximates the way a person will select regions of interest in coral reefs environments. This is important since we want our robot to explore the coral reef as a diver visiting it for the first time.

4.2. Comparison of Detected Regions. In order to measure the performance of our method in terms of detecting relevant regions on underwater scenes, we carried on an analytical comparison of our results with those obtained using the NVT method [12]. This method was used as implemented in the Saliency Tool Box (STB) (the STB can be found in http://www.saliencytoolbox.net/) [30]. For the STB, the default configuration was used. The features used by our algorithm are the intensity and color (red, green, yellow, and blue). For our method, we set the weights of all the conspicuity maps equal to 1.

For this study, we need to determine if the relevant regions detected by the computational attention methods can be considered of interest for a human. This can be done by using a person's judgement. However, this criterion can be very subjective and time consuming for a large set of images. We decided to simplify the evaluation and assumed that the interesting regions should appear on parts of the coral reef: that is, the areas that visually correspond only to water are not considered of interest. First, to divide the image into water and nonwater regions, we applied an adapted version of the robust superpixel-based classifier proposed in [31].

This classifier is used to segment the floor in indoor environments for mobile robot reactive navigation. We have adapted this classifier so it can segment water instead. One of the advantages is that it can be trained online with the current water conditions, and once it is running, it can automatically adapt to possible changes in tonality. All this makes the classifier quite robust. A classification example is depicted in Figure 11.

To perform the comparison test, both algorithms were set to detect the five most relevant regions on each of the 1550 frames in six video sequences. The videos contain a great variety of water conditions, depths, and scenarios of the coral reef of Costa Maya, Mexico. It is important to mention that many of the images in the sequence present challenging situations, for example, high brightness from the sun, bluish and greenish tonalities in case of images taken at deeper locations, and blurriness due to camera motion. All the detected regions that fell into the nonwater area were counted as relevant. In Table 1, the results obtained are shown.

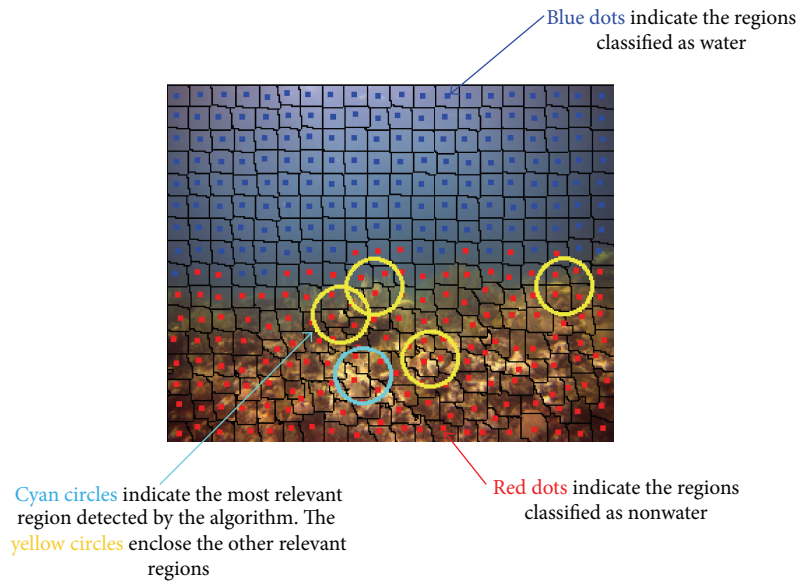


Figure 11: Example of a classification of water (blue dots superpixels) and nonwater (red dots superpixels) regions. Also in this image, five relevant regions detected by our visual attention algorithm are shown. The most relevant region is enclosed in a cyan circle whereas the rest are enclosed in yellow circles.

It can be seen that the percentage of regions detected as interesting by our method is greater than the percentage when using NVT. Let us not forget that these results are over the five most relevant regions detected by both algorithms. We carried on another test, in which only the first most relevant regions were considered. If this relevant region was considered as interesting then it was counted as correct. Table 2 shows the percentages of the interesting regions for the two algorithms for comparison purposes. Although the proposed algorithm percentage is higher than the NVT, the difference is minimal. For this case, however, we have noted (by visually inspecting the detected regions) that many of the relevant points detected by the NVT method were on areas containing only sand or rock formations of brown or black color, which are not considered of interest in an exploration task.

In Figure 12, some images from the video sequences with the relevant regions detected by the algorithms are shown. Qualitatively, in terms of relevance, it can be seen that some of the regions detected by the NVT algorithm are on water or on irrelevant parts like sand or shadows. Also, it can be noted, in the sixth row of both figures, that the regions detected by our algorithm tend to be in the coral reef despite of the abrupt illumination changes due to the sun.

As was shown in Tables 1 and 2, the detected regions by our algorithm tend to be part of the coral reef in more occasions than the detected regions by the NVT algorithm. The difference is notorious when the five most relevant regions were counted. This fact could be useful when we want to lead an autonomous robotic exploration to gather video-observations of this kind of environments (coral reefs), because if more regions are detected in the coral reef then the autonomous agent will go to that place instead of moving to a zone where there is only water.

Table 1: Comparison of the five most relevant regions detected as relevant or interesting (nonwater regions) by using the NVT and the proposed method. In the last row the total number of images and the average of the percentage of detected relevant regions are specified.

Seq.	Frames	Depth [m]	% of *interesting* regions (NVT)	% of *interesting* regions (AVA)
1	168	7.7	75.20	95.85
2	243	7.8	66.91	95.80
3	153	7.8	83.00	99.60
4	181	11.8	57.79	92.15
5	163	7.1	74.47	98.28
6	242	11.3	59.92	90.76
	1150		**69.04**	**94.90**

4.3. Tracking of Relevant Regions. In this section, a comparison between the tracking of a region by using the superpixel descriptors and a keypoint-based descriptor is done. As keypoint detector and descriptor we have used SURF [15], SIFT [16], and ORB [32]. The implementation of the descriptors is the one available on OpenCV. To find the correspondence between keypoints we have use a robust matcher which is available in [33]. The keypoint descriptors and detectors are used with default configuration. For AVA, the yellow and red features have preference through the weights.

For this test, we evaluate the length of tracking, that is, the number of consecutive frames that a given region is tracked in a sequence of images. The region to be tracked is the most relevant as considered by the proposed visual

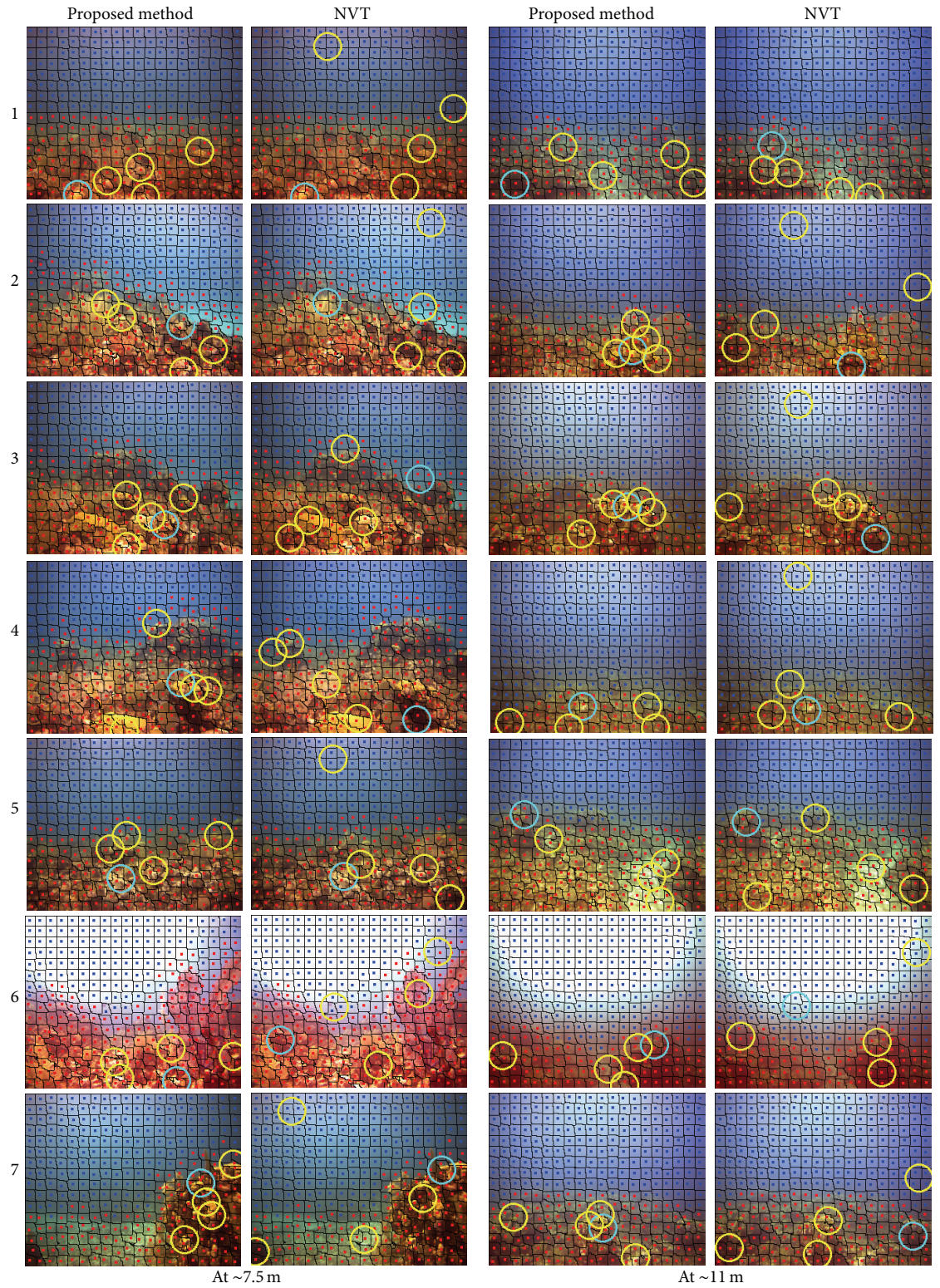


Figure 12: Some of regions detected as interesting by our visual attention algorithm and the NVT. The videos sequences were taken approximately between 7.5 m and 11 m of depth.

Table 2: Comparison of the most relevant region detected as relevant (nonwater regions) by using the NVT and the proposed method. In the last row the total number of images and the average of the percentage of detected relevant regions are specified.

Seq.	Frames	Depth [m]	% of *relevant* region (NVT)	% of *relevant* region (AVA)
1	168	7.7	90.47	91.07
2	243	7.8	94.23	95.06
3	153	7.8	99.34	100
4	181	11.8	88.39	95.5
5	163	7.1	95.02	98.15
6	242	11.3	87.02	91.60
	1150		**92.41**	**95.23**

Table 3: Tracking length comparison of SURF, SIFT, and ORB against the proposed approach.

Tracker	Processing time per frame	Tracking length
SURF	227.8 ms	64.67%
SIFT	258.6 ms	74.35%
ORB	13.6 ms	32.65%

attention algorithm. The image sequences are taken from different videos recorded by a diver while exploring a coral reef.

It is important to remark that the complexity of the AVA algorithm is $O(N)$, where N is the total number of pixels in the image. The average processing time, in a 2.1 GHz dual-core processor, for an image of 480×270 is 122 ms. A total of 8545 images comprises the sequences used in this test.

In Table 3, the average percentages of length of tracking between the proposed method and the keypoint-based trackers and the average processing times are shown. The percentage indicates how a long keypoint-based method's tracking length is in comparison with the AVA tracking length. For example, the tracking length of the SIFT-based tracker is 74.3% of the AVA tracking length. We have normalized all the percentages with respect to the AVA tracking length because it was the method that gets the longer tracking length.

Although the SIFT-based tracker is the one with almost the same tracking length as AVA, it is approximately twice slower. The faster tracker is the one based on ORB; however, it is also the one with the smallest tracking length. From the results of Table 3, it can be noted that the proposed approach outperforms the other methods when tracking regions in underwater environments, particularly coral reefs.

With respect to the processing time, our method can process in average 8 frames of size 480×270 per second. It is important to consider that the current implementation of our method is not yet optimized in terms of software. However, we have found that the current processing frame rate can be good enough to work when exploring an underwater environment because this task tends to be executed with slow motions of the AUV.

From the presented results, it can be remarked that the proposed method can detect and track regions that are likely to draw the attention of humans in coral reefs. This makes our approach suitable for using it to guide an exploration in terms of regions of potential interest for humans.

4.4. Field Trials. For experimental tests we use an amphibious robot named Mexibot of the AQUA family [34]. In water, the robot's propulsion is based on six fins that can provide motion in 5 degrees of freedom up to depths near 35 meters. Mexibot's medium size ($60 \times 45 \times 12$ cm) allows for easy maneuverability, which is important in the time response on the robot's control, when navigating with the purpose of closely monitoring an unstructured environment.

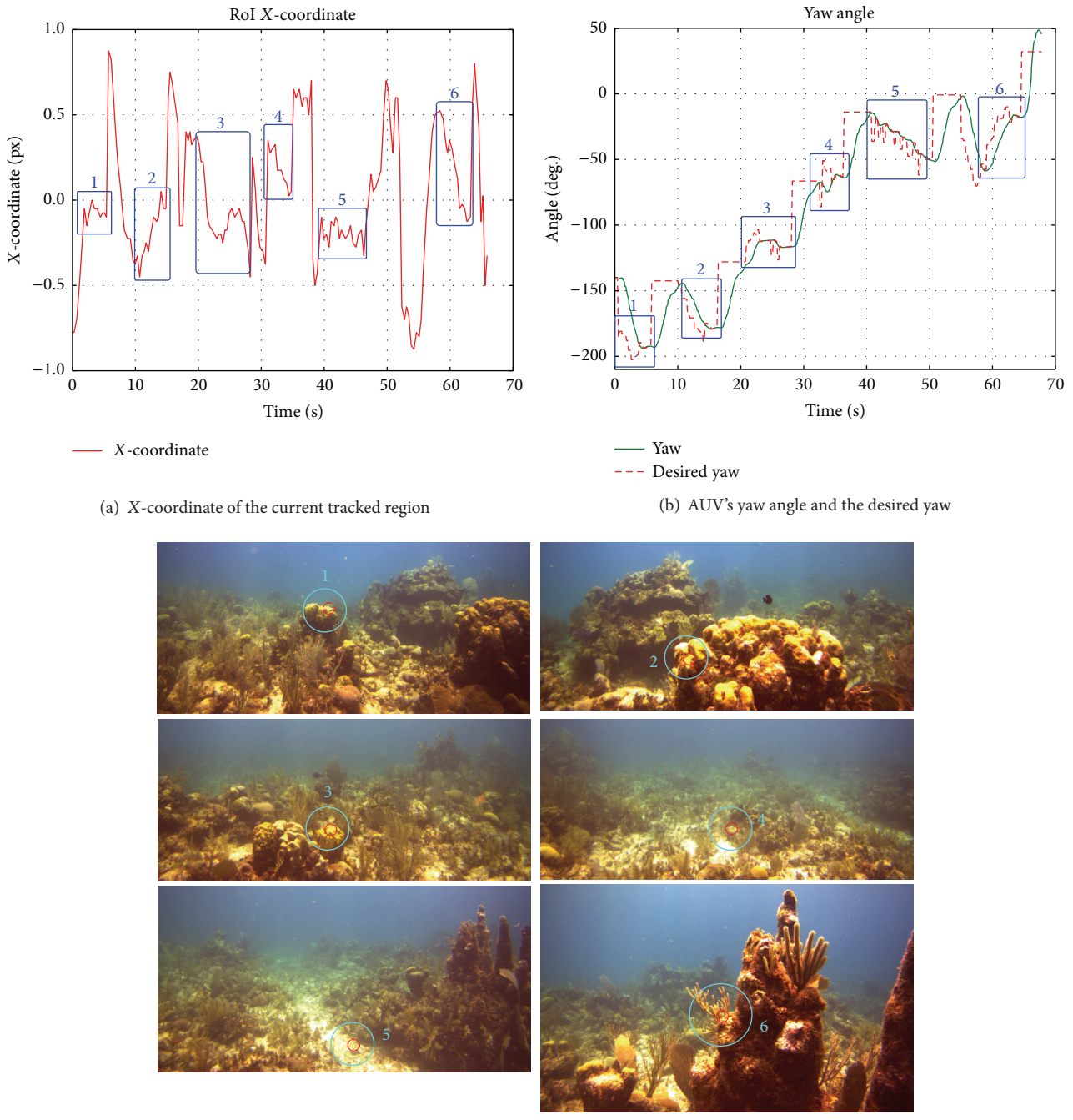
All the trials were performed in an area that belongs to the second largest coral reef system, located in Costa Maya, Mexico. The coral reef ecosystem in this zone has a wide diversity of living organisms (flora and fauna) with a great variety in colors. It also has variable conditions in terms of depth and visibility. We performed the experiments in a depth range from 5 to 18 m.

During the field trials several exploration tests were performed. Most of the tests were set to a two-minute duration as we needed to verify their performance under different conditions. In Figure 13, the results from an exploration are shown. During this test the AUV was programmed to turn 90° around its Z axis every certain time. This had two purposes: the first one is for safety, to avoid a possible collision between the robot and the coral reef. In the moment of the tests the AUV did not have an implemented method for collision avoidance. The second purpose is for testing the capabilities of the proposed approach to detect and track new regions. This way the AUV should detect and track a different region every certain time.

It can be seen in Figure 13 that the AUV effectively changes its yaw angle in order to track the region detected by the visual attention algorithm. The boxes in Figures 13(a) and 13(b) enclose the period during which the same RoI was tracked by the AUV. It can be seen that these regions were followed during several seconds by the AUV until before the 90° turn. These results show that the proposed approach can be used to guide the motion of a AUV for exploring an unknown environment.

5. Conclusions and Future Work

We have presented ongoing research on the detection and tracking of invariant features that are considered relevant during the exploration of a coral reef habitat. The main goal is to perform an autonomous cautious exploration and gather high quality image data with a robotic system that could be directly deployed into the environment, with few or no prior information of it. It is important to highlight that



(a) X-coordinate of the current tracked region

(b) AUV's yaw angle and the desired yaw

(c) Examples of the tracked regions during the test

FIGURE 13: Obtained results during a field trial in a coral reef. In (a) and (b) we can observe the X-coordinate of the region of interest as well as the yaw angle of the AUV, respectively. In (c) images from the tracked RoI can be seen.

the system is trained to adapt itself to the local water and illumination conditions in an online manner. The integrated framework is fast enough to perform the exploration while fitting to the control navigation requirements of the system. Future research will focus on the incorporation of a notion of forward movement to estimate how far the robot is from a certain region as well as adding texture information on the detection of regions of interest in order to reduce errors in the selection of relevant regions (e.g., sand or rock regions are not of interest for exploration).

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The authors would like to thank CONACyT for their support and project funding. The authors also thank Mar Adentro Diving, Mahahual, for their support during their sea trials.

References

[1] Y. Girdhar, P. Giguère, and G. Dudek, "Autonomous adaptive exploration using realtime online spatiotemporal topic modeling," *International Journal of Robotics Research*, vol. 33, no. 4, pp. 645–657, 2014.

[2] Y. Girdhar and G. Dudek, "Exploring underwater environments with curiosity," in *Proceedings of the 11th Conference on Computer and Robot Vision (CRV '14)*, pp. 104–110, IEEE, Montreal, Canada, May 2014.

[3] A. Maldonado-Ramírez and L. A. Torres-Mndez, "Using supercolor pixels descriptors for tracking relevant cues in underwater environments with poor visibility conditions," in *Proceedings of the IEEE Workshop on Visual Place Recognition in Changing Environments (ICRA '15)*, May 2015.

[4] A. Maldonado-Ramírez, L. Torres-Méndez, and E. Martínez-García, "Robust detection and tracking of regions of interest for autonomous underwater robotic exploration," in *Proceedings of the 6th International Conference on Advanced Cognitive Technologies and Applications*, pp. 165–171, Venice, Italy, May 2014.

[5] Y. Y. Schechner and N. Karpel, "Clear underwater vision," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, vol. 1, pp. I536–I543, IEEE, Washington, DC, USA, June 2004.

[6] CIE, *Recommendations on Uniform Color Spaces, Color Difference Equations, Psychometric Color Terms*, vol. 2, no. 15 (E.-1.3.1), CIE Publication, 1971.

[7] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer Graphics and Applications*, vol. 21, no. 5, pp. 34–41, 2001.

[8] L. F. M. Vieira, E. R. D. Nascimento, F. A. Fernandes Jr., R. L. Carceroni, R. D. Vilela, and A. D. A. Araújo, "Fully automatic coloring of grayscale images," *Image and Vision Computing*, vol. 25, no. 1, pp. 50–60, 2007.

[9] G. Bianco, M. Muzzupappa, F. Bruno, R. Garcia, and L. Neumann, "A new color correction method for underwater imaging," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 40, no. 5, pp. 25–32, 2015.

[10] W. James, *The Principles of Psychology*, 1890.

[11] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of Intelligence: Conceptual Structures in Cognitive Neuroscience*, L. Vaina, Ed., vol. 188 of *Synthese Library: Studies in Epistemology, Logic, Methodology, and Philosophy of Science*, pp. 115–141, Springer, Amsterdam, The Netherlands, 1987.

[12] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[13] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: a survey," *ACM Transactions on Applied Perception*, vol. 7, no. 1, article 6, 2010.

[14] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the Alvey Vision Conference*, vol. 15, p. 50, Manchester, UK, 1988.

[15] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: speeded up robust features," in *Computer Vision-ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds., vol. 3951 of *Lecture Notes in Computer Science*, pp. 404–417, Springer, Berlin, Germany, 2006.

[16] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the 7th IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157, IEEE, 1999.

[17] R. O. Duda and P. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Communications of the ACM*, vol. 15, no. 1, pp. 11–15, 1972.

[18] C. Akinlar and C. Tonal, "EDCircles: real-time circle detection by Edge Drawing (ED)," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '12)*, pp. 1309–1312, Kyoto, Japan, March 2012.

[19] S. Frintrop, G. Backer, and E. Rome, "Goal-directed search with a top-down modulated computational attention system," in *Pattern Recognition*, W. Kropatsch, R. Sablatnig, and A. Hanbury, Eds., vol. 3663 of *Lecture Notes in Computer Science*, pp. 117–124, Springer, Berlin, Germany, 2005.

[20] M. Begum and F. Karray, "Visual attention for robotic cognition: a survey," *IEEE Transactions on Autonomous Mental Development*, vol. 3, no. 1, pp. 92–105, 2011.

[21] D. Walther, D. R. Edgington, and C. Koch, "Detection and tracking of objects in underwater video," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, vol. 1, pp. I-544–I-549, July 2004.

[22] D. Edgington, K. Salamy, M. Risi, R. E. Sherlock, D. Walther, and C. Koch, "Automated event detection in underwater video," in *Proceedings of the in OCEANS 2003*, vol. 5, pp. P2749–P2753, San Diego, Calif, USA, September 2003.

[23] C. Barat and M.-J. Rendas, "A robust visual attention system for detecting manufactured objects in underwater video," in *Proceedings of the OCEANS*, pp. 1–6, Singapore, May 2006.

[24] P. L. Correia, P. Y. Lau, P. Fonseca, and A. Campos, "Underwater video analysis for norway lobster stock quantification using multiple visual attention features," in *Proceedings of the 15th European Signal Processing Conference*, pp. 1764–1768, IEEE, Poznan, Poland, 2007.

[25] S. Frintrop, *Vocus: a visual attention system for object detection and goal-directed search [Ph.D. dissertation]*, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany, 2006.

[26] S. Palmer, *Vision Science, Photons to Phenomenology*, The MIT Press, 1999.

[27] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, pp. 641–647, 1994.

[28] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels," Tech. Rep., EPFL, Lausanne, Switzerland, 2010.

[29] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2281, 2012.

[30] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[31] F. G. Rodríguez-Telles, L. A. Torres-Méndez, and E. A. Martínez-García, "A fast floor segmentation algorithm for visual-based robot navigation," in *Proceedings of the 10th International Canadian Conference on Computer and Robot Vision (CRV '13)*, pp. 167–173, May 2013.

[32] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *Proceedings of the International Conference on Computer Vision (ICCV '11)*, pp. 2564–2571, IEEE, Barcelona, Spain, November 2011.

[33] R. Laganière, *OpenCV 2 Computer Vision Application Programming Cookbook: Over 50 Recipes to Master This Library of Programming Functions for Real-Time Computer Vision*, Packt Publishing, 2011.

[34] G. Dudek, P. Giguere, J. Zacher et al., "Aqua: an amphibious autonomous robot," *Computer*, vol. 40, no. 1, pp. 46–53, 2007.