remote sensing

MDPI

Review

The Use of Saliency in Underwater Computer Vision: A Review

Marco Reggiannini * and **Davide Moroni**

Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo", Consiglio Nazionale delle Ricerche, 56124 Pisa, Italy; davide.moroni@isti.cnr.it

* Correspondence: marco.reggiannini@isti.cnr.it

Abstract: Underwater survey and inspection are tasks of paramount relevance for a variety of applications. They are usually performed through the employment of optical and acoustic sensors installed aboard underwater vehicles, in order to capture details of the surrounding environment. The informative properties of the data are systematically affected by a number of disturbing factors, such as the signal energy absorbed by the propagation medium or diverse noise categories contaminating the resulting imagery. Restoring the signal properties in order to exploit the carried information is typically a tough challenge. Visual saliency refers to the computational modeling of the preliminary perceptual stages of human vision, where the presence of conspicuous targets within a surveyed scene activates neurons of the visual cortex, specifically sensitive to meaningful visual variations. In relatively recent years, visual saliency has been exploited in the field of automated underwater exploration. This work provides a comprehensive overview of the computational methods implemented and applied in underwater computer vision tasks, based on the extraction of visual saliency-related features.

Keywords: visual saliency; underwater computer vision; underwater image understanding; multi-sensor survey

check for updates

Citation: Reggiannini, M.; Moroni, D. The use of saliency in underwater computer vision: A review. *Remote Sens.* **2021**, *13*, 22. https://dx.doi.org/10.3390/rs13010022

Received: 10 November 2020
Accepted: 20 December 2020
Published: 23 December 2020

1. Introduction

Saliency is a crucial concept in neuroscience. In particular, visual saliency refers to a preattentive stage of human visual perception that enables an observer to gain awareness about the different relevance of the regions appearing in an observed scenario. Associating an interest score to each point in an observed scene provides the basis for higher-level tasks of the vision system, such as object recognition and classification. Saliency has been conceptually modeled by neuroscientists and employed by automation engineers in order to endow a robotic platform with autonomous perceptual skills.

This type of modeling has been exploited to implement a visual attentive system deployed onboard of robot vehicles. The proven capability to perform land surveys through a computational attentive system also attracted the interest of maritime engineers, motivated by the pursuit of automating the underwater mapping task and making it more safe and efficient. So far, several attempts and models have been developed and thoroughly described in the literature [1], but, as far as the underwater exploration domain is concerned, only a few have been effectively addressed.

Established techniques for in-air object detection usually fail in the underwater scenario. The complex environment makes the survey operations extremely hard to accomplish. The sun radiation only penetrates a few meters in the water medium. It undergoes an attenuation process that reduces the radiation components in an irregular and non-homogeneous way (see Figure 1) depending on the light wavelength (see, e.g., Reference [2]). This phenomenon results in a severe distortion of the spectral content of the image. In addition, the optical image formation is affected by radial and tangential distortions due to the light propagation through the surrounding medium and the camera lens. This further issue results in a distorted reproduction of the target geometry, which should be corrected by preliminary calibration of the optical system. Moreover, optical images captured in the underwater environment are affected by visibility degradation due to partial polarization and a hazing effect resulting from the light scattering inside the water medium. Suitable filtering techniques may be applied to restore the actual color properties of the recorded scene while, at the same time, obtaining a haze-based visual depth estimation [3–6].

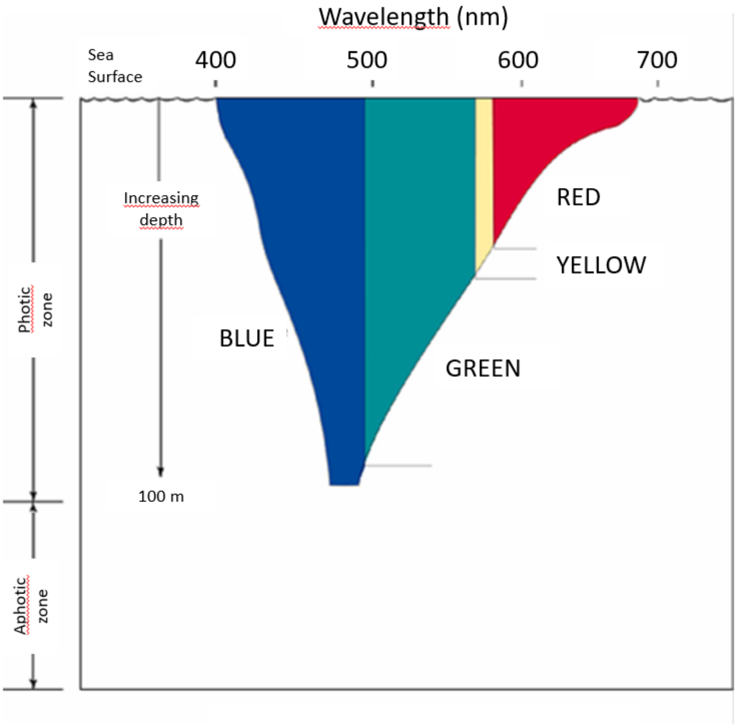

Figure 1. Light absorption in water.

On the other hand, acoustic sensing represents a powerful technique for underwater observations. Compared to the electromagnetic radiation-based sensing, acoustic waves propagate much deeper into the water medium, allowing the survey to cover large areas within reasonable amounts of time. In this case, drawbacks are represented by the propagation losses caused by multiple phenomena, such as energy absorption due to spherical divergence or the geometrical distortions of the signal related to the peculiar transmission and reception of the acoustic signal. As a consequence, preliminary processing is required to restore the geometrical properties of the collected signal and to filter out multiplicative-like disturbances, such as speckle noise (see Reference [7]), which are typical factors causing a low signal to noise ratio in acoustic sensing.

In this framework, classical computer vision approaches are typically prone to false alarms or missed detections. Saliency has become more and more popular in the underwater observation field since it represents a reliable tool to identify meaningful spots in the payload data stream. It can be exploited both as an immediate detector, as well as a preliminary stage for signal enhancement purposes in a processing pipeline.

Saliency modeling shows a weak dependence on the physical peculiarities of the transmission medium. In fact, no primary role is assigned to the physical medium within the formal framework of the visual attention models considered in this work. Nonetheless, saliency has proven interesting performances in several circumstances, involving the exploration of diverse environmental scenarios by means of multiple sensing modalities. Primary image analysis based on visual saliency, such as image segmentation or foreground detection, demonstrated that the extraction of informative content without any a priori knowledge of the surveyed scenario is a viable task.

To consistently cover the literature domain related to the main topic of the review, a dedicated search has been performed, through the engine *ISI Web Of Science* (www.webofknowledge.com), by exploiting the keywords "*underwater*" and "*saliency*". Search criteria neglected any constraint on scientific literature rankings, such as thresholds on impact factor or the like. In order to ensure a fitting set of papers, the output returned by the engine search has been filtered through a preliminary reading of the papers' summary sections (abstract and conclusions). This selection allowed to shortlist a reference bibliography of 49 papers, representing the starting set for the review. Later, the references pool has been extended to include all those articles that were believed to be meaningful for the review purpose, enlarging the set to more than 60 papers. This has been finally extended to 100 papers, in order to include references that were needed to ensure self-consistency.

The paper is arranged as follows: Next, Section 2 concerns an outline of the most relevant saliency models oriented to the underwater exploration; Sections 3–5 concern a detailed survey of saliency-based methods, respectively, dedicated to object detection and segmentation (Section 3), navigation and mapping (Section 4), and image enhancement and restoration (Section 5); Section 6 reports about the currently existing databases that can be exploited as benchmarks for testing saliency-based implementations. Section 7 develops a discussion about the reviewed methods concluding the paper.

2. Saliency Models

Saliency is a general concept which is ubiquitous in computer vision and image analysis, and that has found application in several domains. As previously stated, the concept was born in the neuroscience and has grown in biological vision and bio-inspired methods to artificial vision. Indeed, in primates, intermediate and higher visual processes appear to select a subset of the available sensory information before further processing, presumably to lower the burden of image understanding. This sort of selection is implemented in the form of a spatially circumscribed region of the visual field, the so-called *focus of attention* [8], that appears to be *saliency-driven*. However, during the years, the concept has moved in new and different directions, not necessarily linked with biological vision. In the literature, there are now main general (i.e., not application- nor domain-specific) approaches to saliency estimation. A brief, non-exhaustive description of the leading models is reported here below. For a deeper treatment of the topic of saliency, we refer the reader to the general surveys in Reference [1,9] and to the technical introduction in Reference [10].

2.1. Feature-Based Saliency

The first attempt to develop a biologically-inspired model of human attention can be found in Reference [8]. Here, the authors, exploiting the Feature Integration Theory developed in Reference [11], model saliency by means of a conspicuity map obtained through the combination of a few visual features, derived from the image intensity, color, and orientation. In detail (also see Figure 2), feature maps are obtained by first computing a Gaussian pyramid for each of the mentioned image property, then performing the difference between layers in the pyramid, corresponding to representations of the same map at different scales. Biologically speaking, this mimics the center-surround operation carried out by the visual receptive fields.

In the original model, a Gaussian pyramid with 9 scales $\sigma = \{0, \ldots, 8\}$ is built and differences are computed between the scales k and $k + \delta$ for $k = 2, 3, 4$ and $\delta = 3, 4$. When considering the intensity map, this thus yields 6 features per pixel. For color, the opponent double system is adopted, considering the red/green and green/red double opponency map and the blue/yellow and yellow/blue double opponency map, yielding a total of $2 \times 6 = 12$ features related to color. Four orientation maps at angles 0, $\pi/4$, $\pi/2$, $3/4\pi$ are finally considered, yielding $4 \times 6 = 24$ orientation features. Therefore, the total number of features considered in the original model is 42. Notice that variations in the number of features are feasible and that the model can be adapted to other kinds of images, such as monochrome images and acoustic maps.

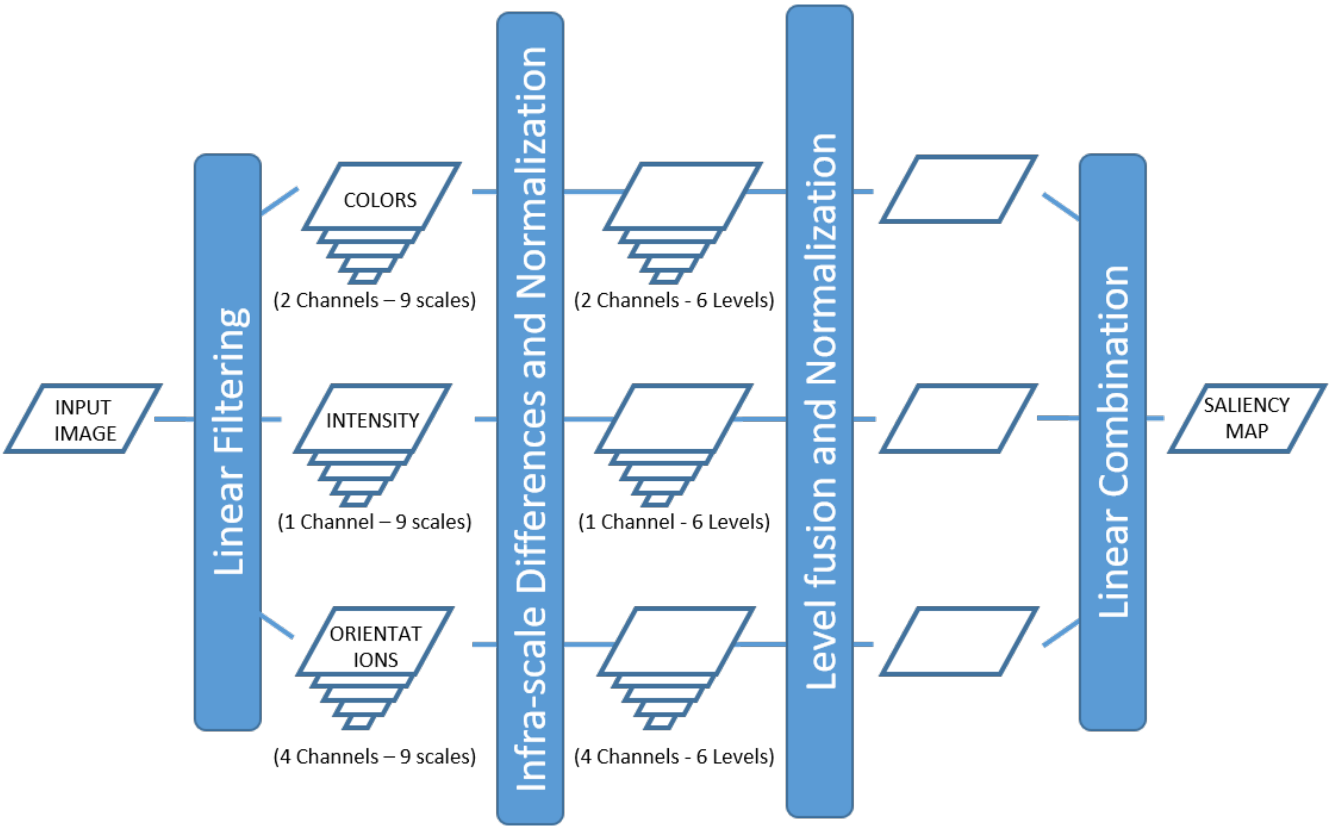

Figure 2. Itti, Koch, and Niebur Architecture.

For each image property, the resulting features are properly summed in the so-called conspicuity maps, and the three resulting maps are eventually integrated into a final saliency map (see Figure 3 for an example). Large values of the saliency map correspond to interesting points in terms of visual perception.

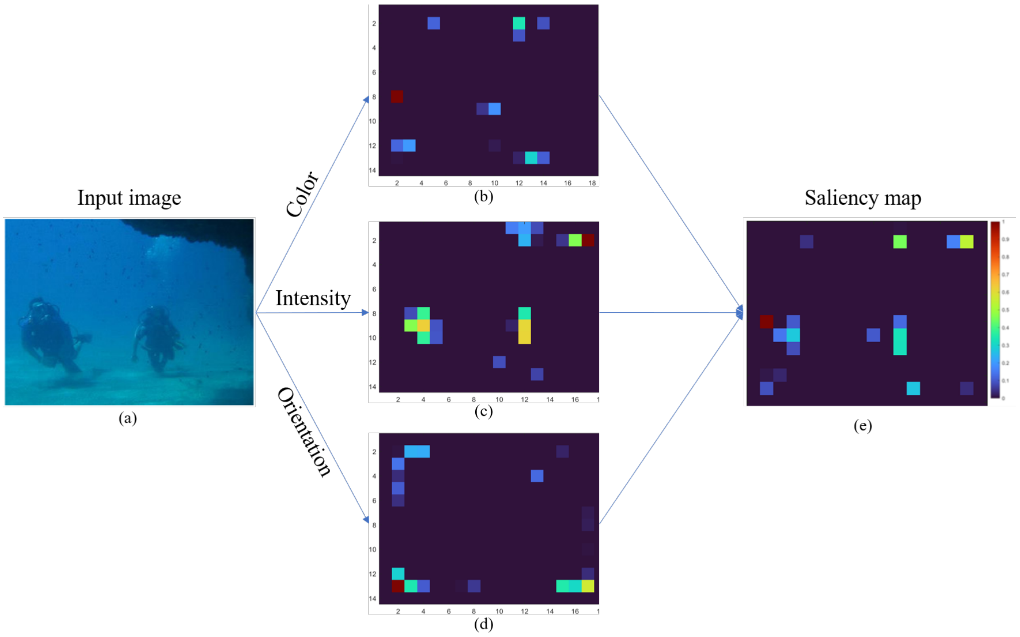

Figure 3. Example of application of Itti's model: (**a**) original image, (**b**) color conspicuity maps, (**c**) intensity conspicuity maps, (**d**) orientations conspicuity maps, (**e**) saliency map. All units are arbitrary units.

Based on the generated map, the attentional process is performed through the proper identification within the saliency map of the most conspicuous regions. This is modeled by means of a 2D winner-take-all neural network, with a number of nodes equaling the number of saliency map pixels. Indeed, each node is fed by a single pixel; hence, pixels with the largest saliency values activate the corresponding neuronal response and accordingly steer the focus of attention orientation. At the same time, the whole network is reset, and a transient inhibition is activated in the saliency map region corresponding to the current focus of attention, in such a way that the attentional model is prevented from selecting again the already identified spots. See Figure 4 for an example of attended locations.

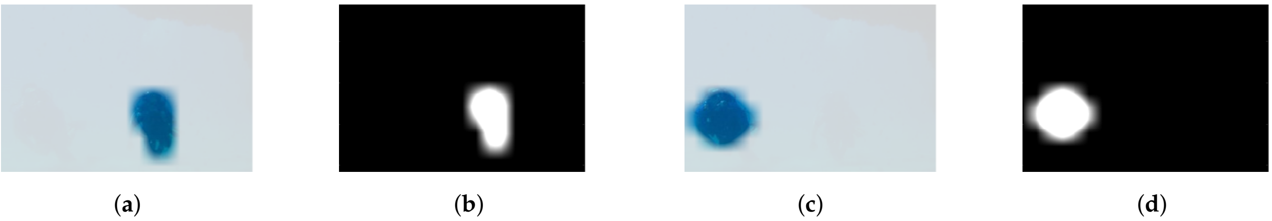

Figure 4. Example of attended locations computed from the original image depicted in Figure 3: (**a**) attended location deriving from center-surround operations on intensity between scales $\sigma = 3, 6$ and (**b**) corresponding attended mask; (**c**) attended location deriving from center-surround operations on intensity between scales $\sigma = 4, 7$ and (**d**) corresponding attended mask.

Since it does not require top-down guidance, the Itti and Koch method allows for a massively parallel and fast selection of interesting locations in an image to be used for different purposes.

More recently, a novel and advanced method inspired by biological vision in aquatic mammals has been proposed in Reference [12] for color images. The authors start their investigation observing that most of the saliency-based methods that achieve good results in-air suffer from poor performance in underwater scenarios. In contrast, specific features of underwater color images have not been fully explored yet. Indeed, besides color and intensity, underwater images present an additional feature, i.e., the already mentioned haze effect that can be used to estimate an unscaled visual depth (see Reference [3–5] and the more recent work of Reference [6] which provides superior results underwater). This piece of 3D information appears to be used by marine animals [13]. Their biological vision system has a significant depth sensitivity which makes the short-range objects visually more salient than the distant ones. Furthermore, the short-range underwater objects are exponentially enhanced in visual saliency, resulting in a nonlinear depth sensitivity. Based on this analysis, they propose the following formulation for the comprehensive saliency S:

$$S = (D_{\text{color}} + D_{\text{intensity}} + D_{\text{depth}}) \exp(r), \tag{1}$$

where $\exp(r)$ is the depth adjustment factor, and the terms D_j for $j \in \{\text{color, intensity, depth}\}$ are point-to-point visual contrast measures computed on color, intensity, and depth maps. This new biologically-inspired method, tested on a rather small dataset made of 200 images from 50 underwater scenes gathered from YouTube videos, has outperformed existing methods. Nevertheless, also considering its recent publication, it has not been yet considered in larger studies or complex applications, nor has it been validated on reference public benchmark.

2.2. Spectral Residual-Based Saliency

Spectral residual-based saliency has been proposed in Reference [14] as an efficient and computationally simple model from an operational point of view. In this paper, it is first remembered that natural images share common similarities for what concerns their spectral properties. In particular, it is observed that the power spectrum of any natural image follows a $1/f$ trend, where f represents spatial frequency. Hence, if plotted in log-log axes, the spectrum behaves linearly (see figure 5). The authors state that the natural image information results from the combination of a common spectral component, coinciding with the average linear behavior of the log-log power spectrum, plus a specifically individual component. According to this approach, saliency is related to this latter component, which is then naturally defined as related to the deviation of the spectrum from the average linearity. Saliency map is thus obtained by (i) computing the image power spectrum, (ii) subtracting the linear component, obtained by averaging a large amount of natural images spectra, and, finally, (iii) applying inverse Fourier transform to obtain the conspicuity map in the spatial domain.

While the approach is effective on natural images taken in the atmosphere, however, it has been observed by Reference [15] that the spectral characteristics of underwater images are different and the general log-log-spectrum has a strong initial decrease and a heavier tail. For this reason, Feng et al. have proposed some corrections to the original method, splitting the spectrum into sectors and applying different heuristic weights to each sector. The obtained results are superior with respect to the Itti and Koch method when used for proto-object detection in an image (validation on a set of 120 images, acquired in real scenarios and selected from YouTube videos).

Among the methods that address saliency as a frequency domain problem, it is worthwhile to mention the approach followed in Reference [16] which proposes a different bottom-up paradigm for detecting visual saliency, characterized by a scale-space analysis of the amplitude spectrum of natural images. Here, the saliency map is obtained by reconstructing the 2D signal using the original phase and the amplitude spectrum, filtered at a scale selected by minimizing saliency map entropy. With respect to the spectral residual approach which uses only a feature map (namely the intensity map), the authors suggest the use of the Hypercomplex Fourier Transform (HFT) [17] in order to include more features and obtain better performance. Different features can be taken into account, for example, regarding color and motion information. Indeed, the HFT is based on quaternions and, as such, up to 4 real features can be included at once:

$$f(x,y) = w_1 f_1 + w_2 f_2 i + w_3 f_3 j + w_4 f_4 k, \tag{2}$$

where $w_1, \ldots, w_4$ are weights, $f_1, \ldots, f_4$ are feature maps, and i, j, k are the fundamental quaternion units. The original paper of Reference [16] uses the following features for RGB images:

$$\begin{aligned} f_2 &= (R+G+B)/3, && (3) \\ f_3 &= \mathcal{R} - \mathcal{G}, && (4) \\ f_4 &= \mathcal{B} - \mathcal{Y}, && (5) \end{aligned}$$

where R, G, B are the red, green, blue channels, and $\mathcal{R} = R - (G+B)/2$, $\mathcal{G} = G - (R + B)/2$, $\mathcal{B} = B - (R+G)/2$, $\mathcal{Y} = (R+G)/2 - R - G/2 - B$. Thus, the features are based on the opponent color space representation of the input image. The features f_1 can be used to effectively integrate a motion feature $\mathcal{M}$ in case of video analysis (see, e.g., Reference [18]). On the basis of such hypercomplex representation of multiple feature maps, several spectra are computed using the HFT, and a saliency map is built in a way similar to the spectral residual approach, selecting the proper scales automatically for detecting salient areas.

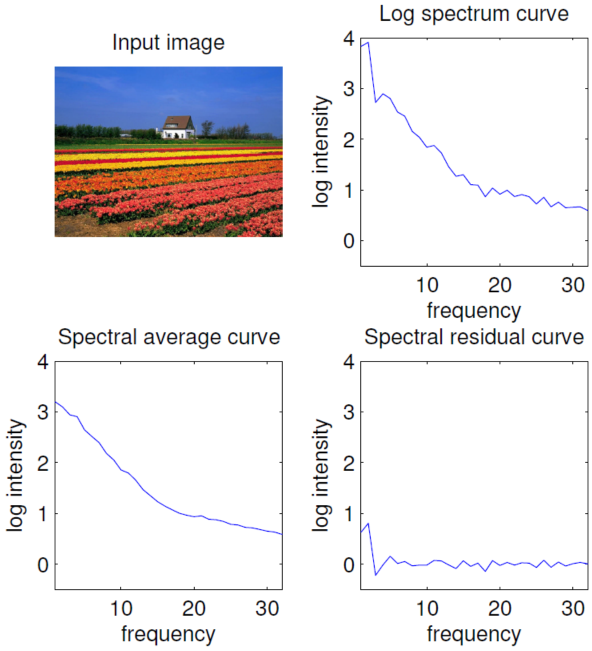

Figure 5. Average power spectrum trend in natural images (taken from Reference [14]).

2.3. Information Content-Based Saliency

Another popular means to extract salient pixels in images is based on information theory concepts, such as the information entropy of a signal. To the purpose of identifying interesting spots, it has been observed that visual features extracted from the neighborhood of salient spots, e.g., intensity histogram, exhibit flatter distribution with respect to the distribution of features extracted from non-informative or cluttered scenarios. According to this observation, it is natural to consider the entropy of an image patch as a quantitative measurement of its information content, in this case corresponding to its saliency. This approach has been largely developed in the computer vision domain, as discussed in Reference [19]. Given a point $\mathbf{x}$ and a scale s, the Shannon entropy is defined through the probability of having a given value for feature d, in which possible values pertain to a set D. Namely:

$$H_D(s,\mathbf{x}) = -\sum_{d\in D} p(ds,\mathbf{x}) \log p(ds,\mathbf{x}), \tag{6}$$

where $p(ds,\mathbf{x})$ is the probability density function for feature d conditioned on s and $\mathbf{x}$. Salient points are detected by finding the scale value for which a maximum entropy value is observed. On the other hand, it is true that entropy peaks can be observed at different scales; hence, the most relevant one must be selected correctly. This is obtained by taking into account the statistics of neighboring pixels, in which properties vary as a function of scale, as well. Thus, the entropy expression is modified introducing a weight that accounts for the variation in the magnitude value of p, as a function of scale only in the nearby of the considered peak.

$$W_D(s,\mathbf{x}) \propto \sum_{d\in D} \big|p(ds_p,\mathbf{x}) - p(ds_p - 1,\mathbf{x})\big|. \tag{7}$$

The saliency map is finally defined as:

$$Y_D(s,\mathbf{x}) = H_D(s_p,\mathbf{x}) \times W_D(s_p,\mathbf{x}). \tag{8}$$

As a conclusion to Section 2, notice that the described models are the most primarily employed for saliency estimation oriented to underwater exploration and analysis. A focused search performed on dedicated scientific web platforms returns a plethora of works that can be roughly categorized based on the task to be performed. The spectrum of applications of visual saliency concepts to underwater environment exploration is almost entirely covered by Object Detection and Segmentation, Navigation and Mapping, and Image Enhancement and Restoration. A survey concerning these usages is reported, respectively, in Sections 3–5.

3. Object Detection and Segmentation

Object detection is probably the primary application for which saliency estimation is conceived. The main goal is to automatically discriminate between background and foreground regions, hence identifying the portion of the image that potentially represents a target of interest for the survey purposes. This section presents a synthesis of the most relevant methods available in the literature, starting with early works concerning foreground detection and identification of proto-objects and subsequent developments (Section 3.1) and going through methods incorporating or taking advantage of frame differencing for video analysis (Section 3.2). Section 3.3 is dedicated to advanced and refined active contours formulations which make a pivotal use of saliency, while the somewhat higher level vision tasks connected to recognition, classification, and analysis are discussed in Section 3.4.

3.1. Foreground Detection and Proto-Objects

Foreground detection refers to the identification of the area in an image occupied by the objects of interest that *stand out* on the background. Therefore, its determination is deeply linked with the attentional process. The resulting foreground area can be sometimes arranged and grouped to form simplified models of the object to be detected (*proto-objects*). The seminal saliency concepts described by Itti, Koch, and Niebur in Reference [8] found a first interesting application to the underwater context in the paper by Reference [20], where the authors propose a method to annotate automatically video frames. The goal is to rough cut-out the (typically extensive) video stream sections that capture non-relevant scenarios (mainly empty scenes or non-interesting objects, such as marine debris) and to identify the segments where relevant objects are observed (e.g., animals or interesting artefacts). The proposed method, based on the computation of a saliency map through the classical approach [8], is exploited to identify meaningful locations in the image and to extract distinctive features, i.e., geometrical and morphological properties of the detected objects, that may be employed for subsequent classification purposes. Compared to human annotation performances, the discriminating power of the presented method turns out to be effective (see Receiver Operating Characteristics curves presented in Reference [20]).

A similar approach is adopted in Reference [21], where the authors propose a sea creatures classification method based on the processing of optical data captured during Autonomous Underwater Vehicle (AUV) missions. After preliminary processing for signal enhancement purposes, the saliency map of the surveyed scene is computed in order to assess the presence of meaningful objects. Then, the candidate targets are further processed and assigned with a classification label returned by a Convolutional Neural Network (CNN), in which architecture is based on AlexNet [22]. Therefore, in this paper, a saliency-based approach is used as a robust method for the proposal of candidate regions to be further processed by specifically trained models. It can be observed that the performance of saliency in proposing candidate regions is general and does not depend on the specific recognition task, while the performance of the subsequent steps achieving classification is task-specific and depends on the actual training of the CNN. The approach has been validated using 37,394 candidate area images, extracted from 3866 seafloor images captured by an AUV.

Atallah et al. developed one of the first works exploiting saliency related concepts for object detection in the underwater domain [23]. In their approach, given a point in the image, saliency is defined as a function of a feature descriptor (for example, the entropy value computed within the neighborhood of the point) and of the corresponding scale value, i.e., the size of the pixel neighborhood employed for the descriptor computation. Accordingly, saliency can be defined as the maximum entropy value at a given point and at a given scale value (see Equations (6)–(8)). This modeling accounts for the fact that an image region can be interpreted as salient or not depending on the scale value at which its descriptive features are computed. This means that, at a given spatial point, saliency varies with scale, so the maximization must also be performed considering this further variability. Besides, entropy can exhibit more than one peak at different scales. The maximum is selected by observing the variability of the detected peaks in their neighborhood and selecting the one that stands out most against the neighboring values. The method is suited for the processing of ultrasound data acquired by side scan sonars. Validation has been performed in 10 m of water in a sheltered cove with a uniform, fine-grained planer sand substrate on which a test site of material types was set out.

Wang et al. [24] address the issue of simultaneous object detection and segmentation in underwater optical images. They propose to compute the saliency map of a corresponding underwater image starting from the approach outlined by Itti et al. [8]. This entails the computation of 3 feature maps which highlight the information carried by the starting map in terms of color, intensity, and orientation, as explained in Section 2.1 and depicted in Figure 2. Adopting the original Itti approach would result in a saliency map given by the summation $S = I_{\text{color}} + I_{\text{intensity}} + I_{\text{orientation}}$. Wang et al. observe that, as a matter of fact, the objects' saliency may result from a non-homogeneous combination of the three mentioned factors. Hence, the authors propose a modified definition of saliency, through a weighted combination:

$$S = \sum_j \alpha_j I_j, \quad \text{where} \quad \sum_j \alpha_j = 1, \tag{9}$$

where $j \in \{\text{color, intensity, orientation}\}$ represents each of the three considered image features. The coefficient α_j takes into account the difference between the foreground and the background areas, for each feature typology. Each α_j is quantitatively estimated comparing, by means of the Bhattacharyya distance [25], the corresponding feature histograms obtained from the foreground and background areas. In particular, the Bhattacharyya distance evaluation returns a scalar value which statistically represents the difference between the foreground and background distributions of each feature, hence describing the discriminating power of that feature. Then, the computed value is exploited to define the combination coefficients in Equation (9) and, accordingly, the final saliency map.

In Reference [26], Huo et al. propose a system to perform object detection and 3D reconstruction of targets observed in optical videos. They employ saliency first to identify the salient regions in the image, and to exploit the result to perform foreground object segmentation later. This approach allows us to reduce the usually large computational cost of a segmentation procedure and to enhance the robustness of the following 3D reconstruction process. To this aim, brightness and texture features are computed. Then, pixels are clustered in super-pixels adopting a similarity criterion based on the computed features. Later, the obtained super-pixels undergo a further clustering process to return the final segmentation output.

3.2. *Temporal Information and Object Tracking*

Temporal information can be used or even incorporated in saliency computation for detecting changing and, thus, potentially relevant objects. Vice versa, high-salient spots can be used to detect and track objects in video streams. Chen et al. [27] adopted a saliency-based approach to perform underwater object detection through optical image processing in videos. They propose to estimate the saliency map related to an image through the spectral residual computation method (see Section 2.2). The proposed method is substantially inspired by Reference [14], apart for what concerns a pre-processing stage, which has been included in the pipeline. In particular, the authors suggest applying a first frame-difference algorithm in order to enhance the discrimination power of the following foreground detector. A similar approach is presented in the already mentioned paper of Reference [15]: there, after computing saliency, object detection is achieved by a standard fuzzy c-means clustering [28], which produces proto-object instances in which precision, however, might be insufficient for applications demanding refined object contours.

In Reference [29], Kumar et al. propose a method to perform event detection in optical video streams captured in the underwater environment. The proposed method exploits saliency concepts starting from the classical model [8]. In particular, the saliency map is computed, and then a thresholding operation is locally performed in order to obtain Local Patch Saliency (LPS) regions. These disjoint regions cover the entire image domain and individually represent local conspicuity maps. Then, the authors apply the morphological closing operation (see, e.g., Reference [30]) to restore the compactness property of patterns in which spatial domain extends over multiple patches. This operation is applied to individual frames. To recognize an event that occurs throughout a sequence of consecutive frames, a process to model the background of the scene, called *Adaptive Saliency Subtraction*, is applied. Following this approach, the background model is identified with the non-salient regions, but it is periodically updated to take into account the varying environment. In Reference [31], the same research group proposes an underwater moving object detection technique by visual saliency estimation based on multiple frames difference. The basic idea is to use temporal information to generate the motion saliency map in order to detect moving objects while suppressing the noise present in the background. More in detail, a continuous symmetric difference of adjacent frames is computed and used to generate full resolution saliency map of the current frame to highlight moving objects with higher saliency values. Range filters are used to get edges of an object, while morphological operators are used to suppress the noise present in the foreground. The proposed algorithm is tested for performance evaluation by performing various experiments under different conditions on videos acquired by Central Scientific Instruments Organization (CSIO) in Chandigarh, an India-based national laboratory dedicated to research, design, and development of scientific and industrial instruments. Video frames have 704 × 480 pixels and generally contain about 10^4 frames. The method does not require user interaction. Nevertheless, some basic parameters (e.g., the size of the used morphological operators) are hard-coded and seem to work in general scenarios, as demonstrated by visual and statistical parameters evaluated by simulation of different videos.

Recently, representation learning has been used for saliency modeling and estimation in the context of object detection and tracking. In Reference [32], the authors propose novel solutions to issues arising in the framework of deep-sea event detection, tracking, and data summarization, implementing a saliency-based system for object detection purposes. The saliency descriptor is provided as the output of a CNN, returning the probability for every pixel of being salient or not. The model is based on a variation of the Holistically-Nested Edge Detector (HED) proposed in Reference [33], where several short connections to the skip-layer structures are introduced. Namely, a top-down view with 5 convolutional layers is built, with each layer having a short connection to the saliency map in output. In this way, the network captures rich multi-scale feature maps at each layer, which are suitably integrated into the global saliency map thanks to the short connections. The methods are tested on original videos collected by the Chinese sea exploration vehicle Jiaolong, during several real tests at sea, where each test consists of more than ten hours, including diving, sailing, and floating. The saliency detector is used as a tracker to follow objects of interest in the video stream, providing quite effective results. Unfortunately, comparison is only presented with respect to other two, non-underwater specific, trackers.

3.3. *Saliency in Active Contour Segmentation*

Active contours, snakes, and level sets are known as powerful methods to obtain accurate object segmentation, which turns out useful whenever fine object recognition or shape analysis must be performed. In this context, saliency often plays an ancillary role, as in Reference [34], where the authors propose an object detection method based on the analysis of a set of sequential images captured by underwater cameras. In this paper, rough object detection is first performed by looking for the most conspicuous points in a co-saliency map, an extended version of the saliency map taking into account a whole set of M images. The pixels are first clustered through a K-means algorithm on a single-image level, and, later, the pixels in the remaining part of the image dataset are labeled based on the computed cluster centers. The saliency map estimation starts from the extraction of three primary features, computed taking into account the entire set of images, for each of the K identified clusters: (i) the *contrast* feature of k-th cluster ($k = 1 : K$), which represents the mean deviation of cluster center μ_k from the rest of the cluster centers, (ii) the *spatial deviation* feature, a scalar value indicating the average distance from the image center, of pixels belonging to cluster k, and (iii) the *correspondence* feature, which describes the clusters distribution over the images sequence; in this latter case, an M-bin histogram is computed, with each entry relating to the number of pixels associated to cluster k in image j, and the *correspondence* feature is defined as proportional to the inverse value of the histogram variance. Then, the computed features are used to define a prior probability of the k-th cluster occurrence in the image dataset: $p(C^k) \sim \prod_i w_i(k)$, where $w_i(k)$ represents the i-th feature descriptor computed for the k-th cluster. Assuming that the pixel saliency $p(\mathbf{x})$, with $\mathbf{x}$ conditioned to be part of cluster k, follows a Gaussian distribution $p(xC^k)$, the pixel saliency is derived as follows:

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}, C^k) = \sum_{k=1}^{K} p(\mathbf{x}C^k)p(C^k). \tag{10}$$

After this quite elaborate pre-processing stage centered on the identification of salient areas, a refined segmentation technique based on active contours is applied to eventually segment the object.

A similar role of saliency within active contours approaches is presented in Reference [35], where the goal is the segmentation of human-made artefacts, especially for archaeological surveys. In this case, saliency—computed thanks to the Itti's model—is used to create a rough object segmentation to initialize the active contours.

In Reference [36], a more in-depth use of saliency is made as a step towards automatic shape extraction and segmentation of objects in an effort to automatize the analysis of underwater footage produced by AUVs. Mainly, the saliency map is used to produce an edge indicator function to be employed as the data-driven term in a geodesic active contour model [37]. Indeed, traditional active contour models exploit the gradient of the image in the stopping criterion for extracting the shape of the objects. However, it turns out that, in the unconstrained underwater environment, gradient information is destroyed due to the frequent temporal and spatial changes, becoming unusable. Based on this consideration, the image gradient alone cannot be used as a stopping condition in these scenarios. Saliency, in the Itti and Koch model [8], is instead a characteristic which has nice invariance properties with respect to the surrounding environment; therefore, it can be used to discriminate both objects and locations in a scene that stand aside the other locations in the background. In the traditional active contour formulation, the edge indicator function in term of the gradient of the original image I is defined as:

$$g(I) = \frac{1}{\sqrt{1 + \alpha \nabla G_\sigma * I}}, \tag{11}$$

where G_σ is a Gaussian kernel with standard deviation σ, and α is a positive parameter. Notice that, in smooth areas of the image I, we have $g(I) \approx 1$, while $g(I)$ assumes values close to zero in the presence of image discontinuities. The definition of edge indicator in terms of saliency is similar:

$$Sg(I) = \frac{1}{\sqrt{1 + \alpha \nabla G_\sigma * S_{\mathrm{map}}(I)}}, \tag{12}$$

where $S_{\text{map}}(I)$ denotes the saliency map computed by Reference [8] (also see Section 2.1). Notice that $Sg(I) \approx 1$ where saliency is constant, while $Sg(I)$ assumes values close to zero when saliency has strong variations. The proposed method is tested on RGB images, performing an extensive quantitative and qualitative analysis with respect to several image segmentation metrics on a dataset of videos obtained from Ocean Networks Canada (ONC) [38,39]. Initialization seeds for the active contours are obtained using the method reported in Reference [29], which is also discussed in Section 3.2 of the present paper.

Another approach to object segmentation based on active contours is presented in Reference [40], where saliency is used as a data-driven term to favor convergence of the contours to the boundaries of the desired object of interest. The authors employ a formulation of active contours, based on the so-called level-set approach, in which the contours to be sought are represented as the zero set of a real function ϕ having the image as its domain. Instead of deforming the contour, in the level set approach, the entire function ϕ is modified, usually minimizing a criterion expressed in an energy form:

$$E(\phi) = E_{\text{reg}}(\phi) + E_{\text{ext}}(\phi),$$

where $E_{\text{reg}}(\phi)$ is a regularization term favoring smoothness and avoiding pathologies in the function ϕ, while $E_{\text{ext}}(\phi)$ is a data-driven term. In the proposed paper, $E_{\text{ext}}(\phi)$ takes into account two features of underwater images, i.e., *transmission* and *saliency*. The first is based on the haze effect, that, while degrading the information of underwater images and making it challenging to distinguish object boundaries, provides, as already mentioned, an unscaled sight distance as a byproduct. Being a sort of depth information, the transmission feature might be used to distinguish the foreground from the background and, in the present paper, its computation is based on the so-called dark channel prior and its variations for underwater imaging (see, e.g., Reference [41,42]). Saliency is instead computed using the HFT (see Section 2.2). Such a pair of features is then inserted into a modification of the Chan-Vese model [43] for level sets, which favors segmentation into regions homogeneous with respect to saliency and transmission. The approach is tested on real underwater images, available on YouTube (www.youtube.com), that were collected to establish a benchmark for experimental evaluations. About 200 images and 50 scenes are included, which were manually labeled by 10 volunteers. The authors compare the object identification obtained by level set segmentation with general methods based on saliency, such as Reference [8,16,44–46]. The quantitative performance of the object detection is evaluated with respect to six criteria, namely Precision (Pr), Similarity (Sim), True Positive Rate (TPR), F-score (FS), False Positive Rate (FPR), and Percentage of Wrong Classifications (PWC). On the benchmark dataset, the level set approach outperforms all the other cited approaches. This makes the approach the best studied and documented among those related to object detection and segmentation. Besides, the dataset, featuring good variability concerning context and the ambient light, is of interest since, in principle, it can ensure a comprehensive evaluation of the methods. Unfortunately, it seems to be not freely available.

3.4. Object Recognition, Classification, and Analysis

Finally, in this last section, works centered on object detection but going further to object recognition and analysis are reported. In Reference [47], the authors present an architecture, based on saliency, which can be exploited to implement a classification algorithm. In this framework, the authors adopt a saliency definition relating to information theory, i.e., based on the entropy of the underlying probability distribution of pixel intensities (or derived quantities, such as texture). High entropy values are correlated to more unpredictable values, hence more salient pixels. Entropy is used here to compute the gain in information associated with the intensity value of every pixel, as well as a global measure to describe the information content of the entire image. The latter operation allows us to perform the comparison with the rest of the images in the considered dataset and to describe the relative unpredictability of the considered image. The saliency map is used to identify interesting regions in underwater optical videos. While traditional methods (e.g., based on Support Vector Machines) outperform the saliency-based method on data exhibiting similarities with that exploited for the training task, the opposite is true when the saliency-based classifier applies to entirely different data (for example, from different surveyed sites).

Chuang et al. [48] propose a system aimed at fish recognition purposes. It is actually based on a non-rigid part model which learns fish properties within an unsupervised learning framework. Here, saliency is exploited as a tool to perform a smart initialization of the learning algorithm, in order to prevent the minimization algorithm from getting stuck in local optima during the learning task. Saliency usage is motivated by the observation that parts perceived as conspicuous by the human eye mostly coincide with the actually salient ones. In this work, saliency is defined through the Phase Fourier Transform of the image, as described in Section 2.2. The main reason for this choice lies in the need to have a fast and straightforward computational tool, in order to enable the processing of large amounts of data. The proposed method is tested with respect to the object recognition task on the Fish4Knowledge dataset [49].

Template matching for object detection and recognition using saliency is presented in Reference [50]. In their work, the authors propose a methodology based on transformable template matching for sonar data. As in other cases, it is first found and discussed that usual schemes for template matching fail in the underwater scenarios due to the specific aspects of sonar data. For instance, cross-correlation approaches give poor results since the scattering of objects depends on incidence angles. In the authors' approach, first, an object is reconstructed from sonar video sequences based on the analysis of acoustic shadows. Then, at processing time, the target regions, i.e., where to look for the objects, are identified by employing the fast saliency detection techniques based on spectral residual (see Section 2.2), significantly improving efficiency by avoiding an exhaustive global search. After detection, the salient region is expanded to reach the same size as the template. The method has been evaluated on a real dataset acquired by a dual-frequency side scan sonar. The dataset is available upon request.

A somewhat different line of research concerning saliency for object detection and analysis is proposed by Kin-Man Lam et al. in [51]. Considering the quaternionic representation of three-channel color images given by:

$$q = Ri + Gj + Bk, \tag{13}$$

where R, G, B are the red, green, and blue channel, and they used a combination of quaternionic metrics [52], taking into account chrominance and intensity changes, to construct differential features that reflect directional information in an image. The normalized directional features are then fused to form an integrated directional map referred to as Quaternionic Distance Based Weber Descriptor (QDWB). In Reference [53], QDWB is also combined with other features, such as Pattern Distinctness (PD) [54] and Local Contrast (LC) [55], which makes it possible to perform a more robust estimation of saliency. Tests are conducted on a small dataset of RGB images which has been made publicly available (also see Section 6), providing a set of validated image descriptors for object detection purposes.

Saliency has also found a role in texture detection and classification. For instance, in Reference [56], the authors propose a framework for an ensemble of probabilistic distance measures based on the analysis of saliency which has good discriminative capabilities in detecting the seabed type and identifying the presence of Sabellaria colonies, mussels, rocks and sand. The approach is tested on a limited set of synthetic acoustic maps generated from a small set of real images.

4. Navigation and Mapping

Visual saliency in the navigation and mapping framework is employed as a bottom-up attention model enabling the identification of trajectories that ensure the largest possible information gain. This turns out to be of paramount relevance in localization and mapping tasks, for example, aiming at the estimation of the vehicle's pose based on a set of pairwise camera shots. Authors adopting this approach refer qualitatively to saliency as the ability of two images to be registered. For example, this ability depends on the amount of texture richness observed and measured within an image, as well as on the rarity of the observed patterns, considering their occurrence throughout the entire set of collected images.

A somewhat inverse approach takes into account the problem of detecting salient points by performing a prediction based on the statistical behavior of a large pool of human subjects. For example, the participants are asked to freely interact with a virtual reconstruction of a given environment and their actions are recorded. Then, the goal is to estimate the underlying probabilistic relationships between the measured users' actions and the environment conspicuity features, encoded in the corresponding saliency map. Examples of this manifold attitudes towards underwater navigation and mapping are discussed in the following.

4.1. Entropy-Based Visual Attention for Localization & Mapping

In the underwater domain, localization, mapping and navigation represent topics of paramount relevance, playing a pivotal role in the planning and fulfilment of a robotic survey mission. Unlike the in-air counterpart, underwater Simultaneous Localization And Mapping (SLAM, see Reference [57]) cannot rely on effective localization tools, such as the Global Positioning System. The usual approach to determine the vehicle position is to collect and properly process measurements of motion-related quantities. This can be fulfilled through dedicated sensors that continuously provide navigation data, such as Inertial Measurement Unit or Doppler Velocity Log. The captured data is processed through data fusion algorithms to return estimates of the vehicle position eventually. Such a dead reckoning approach is prone to severe uncertainties and usually affected by relevant drifts in the trajectory estimation. A limitation of the error drift is possible by forcing the vehicle's trajectory to intersect specific points, called loop closure points, in which informative properties are rich enough that they are identified as crucial key points in the trajectory planning. By forcing the vehicle trajectory to cross these points repeatedly, it is possible to bind the position estimation error. Thus, the positioning system accuracy largely depends on the robustness of the criterion adopted to select loop closure spots. To these purposes, saliency is a widely exploited concept. It is usually employed as a tool to identify relevant key points for vehicles' trajectories, as well as a method to identify potential candidates for object detection.

Concerning underwater vehicle navigation based on SLAM, Kim et al. propose, in Reference [58,59], to include a saliency estimation stage to identify frames, in an optical video stream, that may be employed as loop closure candidates. To this purpose, saliency is introduced with a twofold definition, a local one which describes the richness in terms of texture content of a given frame, and a global one which describes the conspicuity of a frame compared to the entire captured set. The second definition supports and completes the first in the sense that texture-rich frames can be negligible in case they exhibit attributes that frequently appear throughout the entire video stream. The authors borrow saliency conceptual definitions from the framework of automatic document analysis. Visual features (SIFT [60], SURF [61]) are extracted from the input image and associated to *words*. Hence, an image, analogous to a document, can be described in terms of the set of words (*bag-of-words*), i.e., the related complex of descriptive visual features, occurring in it. Local saliency, i.e., saliency map computed for an individual frame, is accordingly defined as the entropy of the histogram of words extracted from that image. On the other hand, the saliency map for a given frame compared to the entire set of frames contained in the video stream is assessed through an informative descriptor, called the *Term Frequency—Inverse Document Frequency*, also borrowed from the document analysis research field. In this case, saliency denotes the rarity of the considered frame compared to the rest of the video. The corresponding quantitative descriptor is computed by taking a census of the occurrence of the words. Adopting this approach, frames exhibiting rare features, e.g., highly rich in terms of texture content, are ranked with high scores.

The same authors propose, in Reference [62] and later extended in Reference [63], a method to implement exploration and survey of an underwater scene. Saliency is again exploited to identify candidate points that can be efficiently employed for loop closure revisiting purposes. The navigation instructions involve the alternation of revisiting and exploration actions, performing one or the other according to the balance between uncertainty in the vehicle's pose and the ratio between the surveyed area and the full area to be covered.

The concepts introduced by Kim et al. have been further exploited by the research community. For example, in Reference [64], the authors propose a system dedicated to navigation purposes based on Reference [59]. The main novelty concerns the construction of the vocabulary. In Reference [59], every candidate word, actually identified by a vector of features extracted from the image, is quantitatively compared with the already existing ones by means of an Euclidean inner product such that, in case the vectors diverge sufficiently (e.g., when the cosine of the angle between the compared vectors is larger than a given threshold), the candidate word is included in the vocabulary. This is, therefore, progressively populated, in contrast to the approach employed in Reference [64], where the vocabulary is built in a unique offline operation, adopting the method of Density-Based Spatial Clustering of Applications with Noise (DBSCAN, [65]).

In Reference [66], Li et al. propose a system to perform underwater SLAM by exploiting a forward-looking sonar as the only perceptual device. The acquired set of sonar frames undergoes a preliminary selection based on saliency content, in order to identify those frames that include a large amount of information and discard meaningless frames which would only result in increased computational burden. In this work, saliency is defined as a global image feature, which can be learned through a CNN. The CNN is trained to provide a feature extractor that is robust for sonar-based localization. In this case, a value in the saliency map corresponds to the image texture diversity. It is, hence, exploited to capture the local variation of the extracted image features and to identify candidate frames for loop closure proposals.

In Reference [67], Kaeli proposes a system to perform anomaly detection in underwater sonar imagery by adopting a saliency-based estimation criterion. The image is processed by means of a variety of operators. First, a Laplacian of Gaussian is applied to identify stable image features, such as corners. Then, the image is fed to a filter bank in order to associate every pixel with a histogram of features. Every filtering action is repeatedly performed varying the image scale. Once each pixel has an associated multi-scale histogram, this is compared with the surroundings' histogram through an absolute norm difference. The maximum difference value, by definition resulting from a multi-scale maximization process, determines the saliency value of the related pixel. Regions that appear locally heterogeneous will return low saliency values, while the opposite will occur in case of spots that exhibit outstanding details.

4.2. Bottom-Up Visual Attention in Underwater Mapping

As previously mentioned, it is possible to qualitatively state that a given spatial region is salient in case an imagery signal captured within that spatial domain exhibits some peculiar anomalies that enable an observer to perceive that spot as interesting with respect to the surroundings. This naive definition has the nice quality of being independent of the nature of the employed sensor technology. This implies that saliency can be considered as a powerful descriptor for a large variety of sensing approaches and technologies. In particular, underwater exploration often entails the exploitation of acoustic imaging which is typically a favorite choice, compared to optical sensing, given the potentially large exploration range and a lower energy release in the water medium, compared on equal values of range from a source. On the other hand, it is suitable to point out that saliency is sometimes employed as a tool to address the perception of relevant spots in the data also in case of low signal to noise ratio or poor visibility conditions, which represent typical circumstances in the underwater scenario. In this unfavorable framework, saliency is exploited to focus on the most relevant spots, in which perception is hindered by environmental factors, but still carries sufficiently valuable information to be exploited for automated exploration. In this regard saliency has also been used to achieve a more robust and automatic registration of underwater data, especially in the case of acoustic images, so as to obtain large-scale seafloor reconstruction and mosaicing.

For instance, in Reference [68], the author addresses the problem of underwater image registration using an approach based on landmarks matching. It is argued that classical detectors, such as the Harris corner detector [69], can be used to identify a number of feature points to be used as a set of landmarks for assessing the displacement vector field between a pair of images. Nevertheless, being based on curvature analysis, Harris detector and similar ones produce a number of points which might be insufficient in some regions of the image. For this reason, the author proposes to integrate curvature-based feature points with feature points extracted as the maximum of the saliency map. In particular, he employs a variation of Itti's model, described in Reference [70], suitably adjusted to be used on one-channel acoustic images. It is found that the saliency map is a complementary way to fill the displacement vector field, firstly only determined by curvature-based points, yielding a more precise and dense estimation and, in turn, superior quality of image registration. Extending and refining this work, Chailloux et al. present, in Reference [71], a method to perform image registration and mosaicing for side-scan sonar underwater imagery. In particular, they propose to identify corresponding pixels between two images by introducing a novel similarity measure based on the fusion between two information descriptors, Correlation Ratio and Mutual Information [72]. In this framework, saliency is employed to detect key points in the image that can be used to estimate the geometrical transformation mapping one reference image onto a test image. Thus, saliency, actually computed adopting the Itti's model [8], is exploited as an initialization tool, enabling focus on the spots in the image that contain the largest amount of information. Similar approaches are described in Reference [73], where a saliency map is extracted by processing acoustic images after shadow removal and used to identify feature points based on SURF [61], and, in Reference [74], where the HFT (Section 2.2) is applied to compute the saliency map of RGB images after haze removal via the dark channel prior [41].

4.3. Visual Saliency through Data-Mining

The collection and analysis of large amounts of information related to a given scenario can shed some light on its saliency properties. In this regard, an interesting assumption is that human visual saliency can be modeled by proxy, observing the exploratory behavior of a conspicuous number of human subjects. A quantitative measurement of this behavior is feasible using eye tracking systems or, in case a computing machine is employed to interface between the human subject and the observed scene, by recording the input exploration commands given by the monitored subjects. Based on the assumption that, unless differently instructed, humans focus their attention at what they find interesting, this approach allows to infer the hidden relationship between the unintentional behavior of the users and the conspicuity of the tested environment.

An interesting approach to estimate the visual saliency of a tri-dimensional scenario is proposed by Johnson-Roberson et al. In Reference [75,76]. Here, the authors discuss a data-driven method to saliency detection and estimation in underwater optical imagery. In particular, the authors conceived a crowd-sourcing experiment based on the involvement of a large set of participants, recruited through the internet and supplied with a mobile application platform. Through this mobile software, every participant is provided with a 3D reconstructed model of a certain underwater scene. The users are then asked to freely interact with the model, and their virtual exploration actions are recorded. Later, the camera motion parameters (pan, tilt, zoom), stored during the explorations, are processed to perform statistical analysis (e.g., histograms of the exploration actions). This allows direct monitoring of the users' focus of attention. Intuitively, the most frequently observed locations and zoom actions refer to potentially interesting spots that, hence, may be labeled as salient ones. The authors propose a method based on Hidden Markov Models (HMMs) [77] to label the collected information as salient or non-salient. Through this approach, two time-series are modeled, the *Observable States* O, relating to the measured exploratory parameters, and the *Hidden States* Y, corresponding to the non observable variables, in this case associated to the saliency features of the observed scene. By measuring the O states, the HMM framework allows us to estimate the probability of a salient state occurrence given that particular observation:

$$P(Y|O) \sim P(O|Y) \cdot P(Y). \tag{14}$$

The classification of the observed state in salient or non-salient is performed by maximizing the Equation (14) with respect to Y. The value for which $P(Y|O)$ is maximum determines the corresponding value in the saliency map.

The proposed method represents an interesting approach which paves the way for further exploitation of crowd-sourcing in visual attention issues. Indeed, ground-truth saliency is usually a hard reference to have available. Its generation as the output of a statistical-based process provides an interesting solution, which is worth being further explored.

5. Image Enhancement and Restoration

In the underwater domain, visual perception is hindered by hostile environmental circumstances. As mentioned previously, the electromagnetic wave intensity drops down after a few meters due to absorption effects, and the radiation captured by the camera is attenuated, in terms of its frequency content, as a consequence of the path covered throughout the water medium. This results in hazy, low-contrast, and color-distorted images, which can hardly be exploited for image analysis purposes. The following section reports about methods based on the use of saliency concepts for signal enhancement purposes.

Achanta et al. [78] propose a method to restore color and contrast image properties through the exploitation of a saliency-based technique. In particular, they start by proving that the most popular saliency estimation methods do not fully exploit the entire frequency content of an input image. Indeed, they observe that almost every state-of-the-art method includes, at some point of the proposed processing pipeline, a downsampling module which necessarily implies a loss of information. This loss affects the details and resolution properties of the output result. Hence, the authors propose a novel method based on the preliminary band-pass filtering of the image through a Difference of Gaussians (DoG) [30] operator, with properly chosen parameters, such that the image spectral content is largely preserved. The saliency map is defined as the absolute difference between the image mean value I_μ and a blurred version of the image, obtained by applying the DoG filter:

$$S(x,y) = I_\mu - I_{w_{hc}}(x,y), \tag{15}$$

where w_{hc} represents the high-frequency cutoff value of the blurring filter. The resulting saliency map may be later employed for specific purposes, such as segmentation and detection of the foreground area.

The described method has often been adopted by researchers interested in the restoration of underwater images. Restoration of a single input image usually starts from the definition of a second layer of input images, derived from the original one through specific filtering operations. White balance, noise reduction, or contrast stretching are typically employed in this stage, which represents a preliminary attempt to restore or partially enhance the information content of the image. The following stage concerns the processing of the derived input images to obtain the so-called *weight maps*. These are the output result of specific filters applied to the input images in order to highlight meaningful features in which observation is considered as an inherent evidence of the image quality. Weight maps may represent local or global contrast, exposure, and also the pixel saliency, where the latter is often computed through the Achanta's approach. In particular, this stage usually consists in the blending of the derived input images by means of the computed weight maps, eventually returning the restored image. For example, in Reference [79,80], the authors propose two different methods to enhance the visibility of underwater images, both based on the Achanta's approach. In these papers, the saliency map is exploited, together with luminance and chromatic maps, as a weight map, and it is employed in the processing pipeline to highlight salient regions and make them more prominent in the final output.

Likewise, in Reference [81], the authors propose a method to compute the amount of backscattering light that reaches the observing sensor in order to later perform image restoration. In particular, they claim to have developed a novel method based on the local estimation of the backscattering component as opposed to the typical approach found in the corresponding literature, where backscattered light is usually estimated as a global value, assumed to be uniform over the entire image. In this framework, saliency is employed in the final multi-scale fusion procedure, which returns the actual output. Three input maps are fed into this blending stage: two represent estimates of the backscattering light performed at two different scales, while the third one is obtained by applying a Laplacian operator, in order to incorporate fine-detail information in the resulting image. The restored image eventually results from the blending of the three estimated inputs. In this case, the weight maps exploited for the fusion process are represented by a saliency map (computed in the Achanta's way), a saturation map and a contrast map. Similarly, in Reference [82], the authors propose a method to perform underwater image restoration through multi-scale multiple images fusion. As in the previous cases, the original image is first processed to correct color and to enhance the contrast properties. This leads to multiple input images that are eventually blended through local and global contrast maps, and a saliency map, computed through the Achanta's model. The main difference, with respect to the previously described methods, lies in the implementation of the final blending stage, which is based on a l_p-norm decomposition instead of a more classical multi-scale pyramid decomposition.

The presence of *caustics* represents another possible artefact that affects the quality of underwater images. In shallow water, the light rays refracted by the water surface may concentrate in different areas casting a characteristic time-varying texture on underwater objects and seabed, which are known as caustics. Suppression or reduction of caustics is seen as a method for image restoration for recovering the actual appearance of objects of interest. Saliency has a role also in this context. Indeed, in Reference [83], saliency is used as a mean to remove caustics from underwater images, thus enhancing their quality. With respect to other papers cited in this section, an ad hoc method to highlight regions with caustics is proposed based on the use of a small and easy trainable CNN, named *SalienceNet*. The network is trained, with an input consisting of synthetic images (produced using Maya software [84]) containing caustics and the corresponding masks as ground-truth, to produce saliency maps of the likelihood of caustics occurring at each pixel. A second CNN, taking in input the original image and the saliency map, is introduced to achieve photorealistic removal of caustics. To prove the results of their work, the authors have processed several real-world underwater videos, where, lacking a proper ground-truth, the efficiency of the SalienceNet has been evaluated a posteriori in terms of reconstruction error after caustics removal.

Finally, saliency can also be exploited as a tool to provide a quantitative descriptor of the quality of an image. In Reference [85], the authors explore different metrics to assess the quality of sonar images captured in the underwater environment. They introduce, in the processing pipeline, a saliency estimation module, computed according to several models available in the literature, which enables to increase the performance of the metrics exploited for image quality evaluation.

6. Resources and Benchmarking

As it might be apparent from the sections above, very often, evaluation and comparison of saliency detection methods for the underwater scenarios has been conducted on small datasets, collected ad hoc by the authors of the specific studies. In order to cope with this issue, a relevant platform, dedicated to the validation of a segmentation algorithm based on general saliency extraction, has been proposed by Achanta et al. [78]. The authors start from a previous database proposed in Reference [86], containing more than 20,000 records, where objects' locations in the images are coarsely identified by means of bounding boxes. Then, they select a subset of 1000 images and generated ground-truth binary images where foreground and background are finely segmented, i.e., identifying the objects' accurate contour and area. Together with the new database, they also provide a suite of popular algorithms for object segmentation that can be exploited to compare the performances of a proposed algorithm with a representative sample of the available state-of-the-art methods.

More recently, in Reference [87] and in its extension in Reference [88], a database specifically designed for the benchmarking of saliency estimation methods for underwater object detection is proposed. The database, named Marine Underwater Environment Database (MUED), is collected in an artificial pool mimicking the variabilities in illumination, background, and pose that are normally encountered in the real environment. Besides, water turbidity is artificially changed by adding soil and milk to water. A total of 8600 underwater images (resolution 648 × 486) of 430 distinct objects is contained in the database. Reference [88] presents also a baseline evaluation with a wide range of known general methods, including Graph Regularization [55], Patch-Distinctness [54], Dense and Sparse Reconstruction [89], Nonlinearity covariance [46], Multiscale Super-Pixel [90], Cellular Automata [91], QDWB [51], and PD and LC [53] (the last methods being proposed by the same team who published the database). Methods are compared with respect to the task of object detection computing the precision, recall, and F-score of detected salient object rectangle with respect to the ground-truth rectangle. Similarly, mean absolute error, overlapping ratio, and area under ROC curve (AUC) are computed. Most of the methods seem to have good results in specific scenarios, while PD has stable and good performance in all the contexts. The dataset is freely available in Zenodo, split into two parts (see Reference [92,93]), and contains the ground-truth as a text file with the left, top, right, and bottom coordinates of each rectangle. An additional freely available resource is the Underwater Image Enhancement Benchmark (UIEB) dataset [94]. The dataset, initially conceived for benchmarking image enhancement approaches, is used in Reference [95] for evaluating a mild variation of the general saliency detection method based on PD [54]. UIEB contains 950 real-world underwater images, 890 of which have the corresponding reference images. The remaining 60 underwater images are to be considered as challenging data for image enhancement and restoration. Although it is a structured dataset, notice, however, that no ground-truth is provided regarding saliency estimation and object detection. A good reference dataset for large-scale comparison is represented by the Fish4Knowledge dataset [96], a huge collection of underwater videos from multiple cameras, used, for instance, in Reference [48]. Although the dataset was not conceived bearing saliency estimation in mind, it provides about 200 TB of videos from 3 sites, from a total number of 9 cameras. The video footage spans over 3 years and permits evaluation of algorithms in disparate sea conditions. Ground-truth is provided for different vision tasks and, especially, for target detection against complex underwater background. Finally, it is appropriate to admit that resources to perform benchmarking on sonar data processing are currently limited. To the best of our knowledge, significant datasets are unavailable, thus implying tough issues in the pursuit of fair methods comparison.

7. Discussion and Conclusions

Essentially, saliency is a filter that enables an observer to identify meaningful spots in the scene. Emerging as a biological model of preattentive vision, it naturally fits in the object detection task, which is typically a main goal in every computer vision application of saliency. One of the appealing attributes of saliency is that of being an unsupervised technique, hence enabling to extract informative content from a completely unknown context. This makes saliency an adaptive cross-cutting feature and an attractive tool for diverse research branches. Moreover, the conceptual structure of saliency modeling is weakly affected by different application scenarios, such as its implementation within underwater or in-air survey tasks. At most, the modeling may require small amendments on specific reference parameters, such as the tuning of the curve describing the average power spectrum trend for natural underwater images (see Section 2.2). Necessary changes made, the main model core remains intact, proving to be a tool that keeps being usable, regardless of the experimental circumstance.

As shown in this paper, the concept of saliency has found relevant and disparate applications for dealing with underwater vision, a domain in which usual tasks become more challenging and demanding of specific algorithms and approaches. A summary of the methods that were surveyed in this paper is presented in Table 1, by identifying their primary purpose, the approach used, or proposed for saliency computation and the type of data and validation reported by their authors.

Object detection is one of the predominant applications. The main reason is that saliency can cope with unstructured and unpredictable scenarios in order to detect interesting areas of the image in which relevant (but possibly unknown) objects are to be detected. Further, saliency formulations have the flexibility to cope with underwater peculiarities (e.g., limited light penetration, low contrast, scattering, and spectral distortion) and to adapt to different modalities (e.g., acoustic imaging). Saliency has also fitted into several segmentation schemes, including refined schemes based on advanced level set formulations. It can act as a quick way to have robust initialization of active contours, as well as a theoretical measure to steer data-driven contour evolution. Recent approaches have shown the emergence of 3D features of the underwater images into the visual task. Indeed, one of the artefacts of underwater images, namely the haze effect, can be turned into usable information by deriving a sort of unscaled depth map. RGB images can therefore be turned into RGB-D maps. From the introduction of this estimated channel, several new applications of saliency have spurred, since depth is a clue to distinguish foreground from background, also under the principles of vision for mammal underwater creatures.

Saliency acquired consolidated relevance also in the underwater computer vision domain, which differs with respect to the analogous on the in-air circumstances for amplified challenges and difficulties. Underwater navigation took considerable advantage of saliency estimation, primarily because of the capability to identify the most conspicuous spots in a surveyed area, therefore allowing an appropriate selection of loop closure points and eventually make environmental mapping more robust. Understanding an underwater scenario is challenging because of the previously mentioned issues related to severe signal distortions and degradation. Often, the saliency models typically adopted for vehicle maneuvering or adaptive path planning do not rely on straight deterministic features, as opposed to the quantities typically employed in the terrestrial framework (color, intensity). Indeed, for navigation purposes, it is quite a consolidated approach to adopt entropy-based models that allow us to pick up the degree of complexity and unpredictability of the signal. Nevertheless, Itti's model is usually employed in the navigation field to focus on outstanding spots in which appearance keeps being quite constant throughout the data collection campaign, such as key spots in side-scan sonar swaths that may be used for map stitching and mosaicing.

Concerning image restoration, saliency is usually employed as a weight map that allows highlighting of relevant spots in the represented scene and eventually emphasizing the amount of information carried by the image.

Despite the number of papers and the growing interest in saliency witnessed by the very recent scientific production, there is still a lack of public resources for benchmarking and comparison that can be considered a de facto standard for algorithm evaluation. As described, most of the papers use their own datasets, often gathered from publicly available videos, but in which actual content is not explicated in a way sufficient for reconstructing them. As surveyed, some more general datasets are appearing, though not specific for saliency evaluation but as general-purpose ones, and it might be desirable that they acquire the role of reference benchmarks in the near future.

Original saliency modeling [8] results from the combination of several feature maps, which are appropriately chosen and precisely defined. The re-emergence of automatic learning in computer vision through deep neural network architectures indicates novel opportunities. Indeed, most of the classical methods adopted local and global contrast as a mean to derive saliency maps, where the definition of contrast is based on various types of handcrafted image features (e.g., color, intensity and histogram) at the pixel or superpixel level. In the underwater scenario, different ad hoc handcrafted features were envisaged, also for coping with different sensing modalities, such as acoustic imaging. The advances in deep learning are today favoring a shift from handcrafted features to computed features discovered through the so-called representation learning. During this survey, some attempts based on this paradigm shift have been already included, namely approaches based on CNNs to detect salient areas in acoustic data [66] and remove caustics from optical images [83]. However, it appears that the full potential that can be obtained by deep learning has not yet been uncovered. For instance, works for in-air images, such as Reference [97,98], have not yet a counterpart or specific application in the underwater field. In these last approaches, multi-level and multi-scale deep features can be extracted from the images, thanks to CNNs, and used either as deep contrast features or directly as a saliency map. Notice that, in these approaches, the original seminal ideas contained in Itti's model are retained if we allow standard features to be substituted by deep features and linear combination by fusion strategies based on more complex networks, such as fully-connected layers. On the one hand, this means that massive deep feature application to classical underwater vision problems that were tackled by saliency should be investigated. On the other hand, it suggests that saliency, although being a somewhat subjective notion, is a powerful concept that has still much to say in underwater vision.

Table 1. Summary of surveyed methods.

Paper	Year	Purpose	Saliency Model	Data			Evaluation	
				RGB Images	*Video*	*Acoustic Data*	*Real*	*Synthetic*
			Object Detection and Segmentation					
Edgington et al. [20]	2003	Object detection	Itti		✓		✓	
Ahn et al. [21]	2018	Object detection and CNN-based classification	Itti	✓			✓	
Atallah et al. [23]	2005	Object detection	Entropy-based			✓	✓	
Wang et al. [24]	2013	Detection & Segmentation	Itti	✓			✓	
Chen et al. [27]	2014	Object detection	Spectral residual	✓			✓	
Chuang et al. [48]	2016	Initialization of object recognition	Phase Fourier Transform	✓			✓	
Zhu et al. [34]	2017	Detection & Segmentation	Saliency map based on contrast, position, and correspondence		✓		✓	
Sanchrez-Torres et al. [99]	2018	Segmentation	Ad hoc based on morphological operators	✓			✓	
Huo et al. [26]	2018	Detection & 3D Reconstruction	Aggregation of salient superpixels	✓			✓	
Kumar et al. [36]	2019	Shape reconstruction using edge-based active contours	Itti	✓	✓		✓	
Chen et al. [40]	2019	Segmentation using region-based active contours	HFT	✓			✓	
Barat et al. [35]	2010	Segmentation using active contours featuring saliency in initialization	Itti	✓			✓	
Kumar et al. [31]	2019	Moving object detection	Multiple frames difference		✓		✓	
Zhu et al. [50]	2019	Template Matching	Spectral residual			✓	✓	
Jian et al. [51]	2018	Object detection	QDWB	✓			✓	
Jian et al. [53]	2018	Object detection	QDWB + PD + LC	✓			✓	
Johnson-Roberson et al. [47]	2010	Classification	Entropy-based	✓			✓	
Cong et al. [32]	2019	Saliency-based Object Detection	Saliency map obtained by Deep Convolutional Neural Network	✓	✓		✓	
Harrison et al. [56]	2011	Texture segmentation	Co-occurence matrices and ensemble of distance			✓		✓
			Navigation and Mapping					
Kim et al. [58]	2011	Navigation & Mapping through Local/Global Saliency estimation	Entropy-based	✓	✓		✓	
Kim et al. [59]	2013	Navigation & Mapping through Local/Global Saliency estimation	Entropy-based	✓	✓		✓	
Kim et al. [62]	2015	Navigation & Mapping through Local/Global Saliency estimation	Entropy-based	✓	✓		✓	
Ozog et al. [63]	2015	Navigation & Mapping through Local/Global Saliency estimation	Entropy-based	✓	✓		✓	
Geng et al. [64]	2016	Navigation & Mapping	Entropy-based			✓	✓	
Li et al. [66]	2018	Simultaneous Localization and Mapping	Entropy-based			✓	✓	
Johnson-Roberson et al. [75]	2014	Saliency Estimation through Crowdsourcing	Gaze-tracking & Hidden Markov Model estimation	✓			✓	
Johnson-Roberson et al. [76]	2015	Saliency Estimation through Crowdsourcing	Gaze-tracking & Hidden Markov Model estimation	✓			✓	
Kaeli et al. [67]	2016	Anomaly detection	Entropy-based			✓	✓	
Kumar et al. [29]	2019	Saliency estimation for object detection	Itti		✓		✓	
Chailloux [68]	2005	Image registration based on landmarks	Ittis' model variation			✓	✓	
Chailloux et al. [71]	2011	Saliency estimation for large scale mapping	Itti			✓	✓	
Fu et al. [73]	2015	Saliency estimation for feature point detection	Local contrast			✓	✓	
Zhang et al. [74]	2016	Feature point detection and matching	HFT	✓			✓	
			Image Enhancement and Restoration					
Achanta et al. [78]	2009	Salient region detection	Difference of Gaussian-based band pass filtering	✓			✓	
Fang et al. [79]	2013	Underwater Image restoration	Difference of Gaussian-based band pass filtering	✓			✓	
Singh et al. [80]	2016	Underwater Image restoration	Difference of Gaussian-based band pass filtering	✓			✓	
Ancuti et al. [81]	2016	Underwater Image restoration	Salient region detection	✓			✓	
Jianhua et al. [82]	2019	Underwater Image restoration	Salient region detection	✓			✓	
Forbes et al. [83]	2019	Image restoration	Convolutional neural network-based saliency estimation	✓			✓	✓
Zhang et al. [85]	2019	Image Quality Evaluation	Several models are employed			✓		✓

Funding: This paper is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101000825 (NAUTILOS).

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

References

1. Borji, A.; Itti, L. State-of-the-Art in Visual Attention Modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 185–207.
2. Duntley, S.Q. Light in the sea. *JOSA* **1963**, *53*, 214–233.
3. Chiang, J.Y.; Chen, Y.C. Underwater image enhancement by wavelength compensation and dehazing. *IEEE Trans. Image Process.* **2011**, *21*, 1756–1769.
4. Galdran, A.; Pardo, D.; Picón, A.; Alvarez-Gila, A. Automatic red-channel underwater image restoration. *J. Vis. Commun. Image Represent.* **2015**, *26*, 132–145.
5. Li, C.; Quo, J.; Pang, Y.; Chen, S.; Wang, J. Single underwater image restoration by blue-green channels dehazing and red channel correction. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 1731–1735.
6. Łuczyński, T.; Birk, A. Underwater image haze removal with an underwater-ready dark channel prior. In Proceedings of the IEEE OCEANS 2017-Anchorage, Anchorage, AK, USA, 18–21 September 2017; pp. 1–6.
7. Richards, M.A.; Scheer, J.A.; Holm, W.A.; Beckley, B.; Mark, P.; Richards, A. (Eds.) *Principles of Modern Radar: Basic Principles*; Institution of Engineering and Technology: London, UK, 2010.
8. Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259.
9. Frintrop, S.; Rome, E.; Christensen, H.I. Computational visual attention systems and their cognitive foundations: A survey. *ACM Trans. Appl. Percept. (TAP)* **2010**, *7*, 1–39.
10. Frintrop, S. Computational visual attention. In *Computer Analysis of Human Behavior*; Springer: Berlin, Germany, 2011; pp. 69–101.
11. Treisman, A.M.; Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **1980**, *12*, 97–136, doi:10.1016/0010-0285(80)90005-5.
12. Chen, Z.; Gao, H.; Zhang, Z.; Zhou, H.; Wang, X.; Tian, Y. Underwater salient object detection by combining 2d and 3d visual features. *Neurocomputing* **2020**, *391*, 249–259.
13. Marshall, J.; Carleton, K.L.; Cronin, T. Colour vision in marine organisms. *Curr. Opin. Neurobiol.* **2015**, *34*, 86–94.
14. Hou, X.; Zhang, L. Saliency detection: A spectral residual approach. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 18–23 June 2007; pp. 1–8.
15. Feng, H.; Yin, X.; Xu, L.; Lv, G.; Li, Q.; Wang, L. Underwater salient object detection jointly using improved spectral residual and Fuzzy c-Means. *J. Intell. Fuzzy Syst.* **2019**, *37*, 329–339.
16. Li, J.; Levine, M.D.; An, X.; Xu, X.; He, H. Visual saliency based on scale-space analysis in the frequency domain. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 996–1010.
17. Ell, T.A. Quaternion-Fourier transforms for analysis of two-dimensional linear time-invariant partial differential systems. In Proceedings of the 32nd IEEE Conference on Decision and Control, San Antonio, TX, USA, 15–17 December 1993; pp. 1830–1841.
18. Guo, C.; Zhang, L. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Trans. Image Process.* **2009**, *19*, 185–198.
19. Kadir, T.; Brady, M. Saliency, Scale and Image Description. *Int. J. Comput. Vis.* **2001**, *45*, 83–105, doi:10.1023/A:1012460413855.
20. Edgington, D.R.; Salamy, K.A.; Risi, M.; Sherlock, R.; Walther, D.; Koch, C. Automated event detection in underwater video. In *Oceans 2003. Celebrating the Past... Teaming Toward the Future (IEEE Cat. No. 03CH37492)*; IEEE: New York, NY, USA, 2003; Volume 5, pp. P2749–P2753.
21. Ahn, J.; Nishida, Y.; Ishii, K.; Ura, T. A Sea Creatures Classification Method using Convolutional Neural Networks. In Proceedings of the 2018 18th International Conference on Control, Automation and Systems (ICCAS), PyeongChang, Korea, 17–20 October 2018; pp. 420–423.
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90, doi:10.1145/3065386.
23. Atallah, L.; Shang, C.; Bates, R. Object detection at different resolution in archaeological side-scan sonar images. *Eur. Ocean.* **2005**, *1*, 287–292.
24. Wang, H.; Dong, X.; Jie, S.; Wu, X.; Chen, Z. Saliency-Based Adaptive Object Extraction for Color Underwater Images. *Appl. Mech. Mater.* **2013**, *347–350*, doi:10.2991/iccsee.2013.661.
25. Bhattacharyya Distance. Encyclopedia of Mathematics. 2020. Available online: http://encyclopediaofmath.org/index.php?title=Bhattacharyya_distance&oldid=46047 (accessed on 21 December 2020).
26. Huo, G.; Wu, Z.; Li, J.; Li, S. Underwater Target Detection and 3D Reconstruction System Based on Binocular Vision. *Sensors* **2018**, *18*, 3570, doi:10.3390/s18103570.
27. Chen, Z.; Wang, H.; Shen, J.; Dong, X. Underwater Object Detection by Combining the Spectral Residual and Three-Frame Algorithm. In *Advances in Computer Science and its Applications*; Jeong, H.Y., Obaidat, M.S., Yen, N.Y., Park, J.J.J.H., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 1109–1114.

28. Bezdek, J.C. *Pattern Recognition with Fuzzy Objective Function Algorithms*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
29. Kumar, N.; Sardana, H.; Shome, S.; Mittal, N. Saliency Subtraction Inspired Automated Event Detection in Underwater Environments. *Cogn. Comput.* **2019**, *12*, doi:10.1007/s12559-019-09671-x.
30. Gonzalez, R.C.; Woods, R.E. *Digital Image Processing*, 4th ed.; Pearson: London, UK; Prentice Hall: Upper Saddle River, NY, USA, 2018.
31. Underwater moving object detection by temporal information. *Int. J. Recent Technol. Eng. (IJRTE)* **2019**, *8*.
32. Cong, Y.; Fan, B.; Hou, D.; Fan, H.; Liu, K.; Luo, J. Novel Event Analysis for Human-Machine Collaborative Underwater Exploration. *Pattern Recognit.* **2019**, *96*, 106967, doi:10.1016/j.patcog.2019.106967.
33. Hou, Q.; Cheng, M.M.; Hu, X.; Borji, A.; Tu, Z.; Torr, P.H. Deeply supervised salient object detection with short connections. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3203–3212.
34. Zhu, Y.; Hao, B.; Jiang, B.; Nian, R.; He, B.; Ren, X.; Lendasse, A. Underwater image segmentation with co-saliency detection and local statistical active contour model. In Proceedings of the OCEANS 2017-Aberdeen, Aberdeen, UK, 19–22 June 2017; pp. 1–5.
35. Barat, C.; Phlypo, R. A fully automated method to detect and segment a manufactured object in an underwater color image. *EURASIP J. Adv. Signal Process.* **2010**, *2010*, 1–10.
36. Kumar, N.; Sardana, H.; Shome, S. Saliency based shape extraction of objects in unconstrained underwater environment. *Multimed. Tools Appl.* **2018**, *78*, doi:10.1007/s11042-018-6849-9.
37. Paragios, N.; Deriche, R. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 266–280.
38. Barnes, C.; Best, M.; Bornhold, B.; Juniper, S.; Pirenne, B.; Phibbs, P. The NEPTUNE Project-a cabled ocean observatory in the NE Pacific: Overview, challenges and scientific objectives for the installation and operation of Stage I in Canadian waters. In Proceedings of the IEEE 2007 Symposium on Underwater Technology and Workshop on Scientific Use of Submarine Cables and Related Technologies, Tokyo, Japan, 17–20 April 2007; pp. 308–313.
39. Gebali, A.; Albu, A.B.; Hoeberechts, M. Detection of salient events in large datasets of underwater video. In Proceedings of the 2012 Oceans, Hampton Roads, VA, USA, 14–19 October 2012; pp. 1–10, doi:10.1109/OCEANS.2012.6404996.
40. Chen, Z.; Sun, Y.; Gu, Y.; Wang, H.; Qian, H.; Zheng, H. Underwater Object Segmentation Integrating Transmission and Saliency Features. *IEEE Access* **2019**, *7*, 72420–72430.
41. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 2341–2353.
42. Sathya, R.; Bharathi, M.; Dhivyasri, G. Underwater image enhancement by dark channel prior. In Proceedings of the IEEE 2015 2nd International Conference on Electronics and Communication Systems (ICECS), Coimbatore, India, 26–27 February 2015; pp. 1119–1123.
43. Vese, L.A.; Chan, T.F. A multiphase level set framework for image segmentation using the Mumford and Shah model. *Int. J. Comput. Vis.* **2002**, *50*, 271–293.
44. Harel, J.; Koch, C.; Perona, P. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, UK, 2007; pp. 545–552.
45. Gu, K.; Zhai, G.; Lin, W.; Yang, X.; Zhang, W. Visual saliency detection with free energy theory. *IEEE Signal Process. Lett.* **2015**, *22*, 1552–1555.
46. Erdem, E.; Erdem, A. Visual saliency estimation by nonlinearly integrating features using region covariances. *J. Vis.* **2013**, *13*, 11–11.
47. Johnson-Roberson, M.; Pizarro, O.; Williams, S. Saliency ranking for benthic survey using underwater images. In Proceedings of the 2010 11th International Conference on Control Automation Robotics Vision, Singapore, 7–10 December 2010; pp. 459–466.
48. Chuang, M.; Hwang, J.; Williams, K. A Feature Learning and Object Recognition Framework for Underwater Fish Images. *IEEE Trans. Image Process.* **2016**, *25*, 1862–1872.
49. Boom, B.J.; Huang, P.X.; He, J.; Fisher, R.B. Supporting ground-truth annotation of image datasets using clustering. In Proceedings of the IEEE 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 1542–1545.
50. Zhu, J.; Siquan, Y.; Han, Z.; Tang, Y.; Wu, C. Underwater Object Recognition Using Transformable Template Matching Based on Prior Knowledge. *Math. Probl. Eng.* **2019**, *2019*, 1–11, doi:10.1155/2019/2892975.
51. Jian, M.; Qi, Q.; Dong, J.; Sun, X.; Sun, Y.; Lam, K.M. Saliency detection using quaternionic distance based weber local descriptor and level priors. *Multimed. Tools Appl.* **2018**, *77*, 14343–14360.
52. Lan, R.; Zhou, Y.; Tang, Y.Y. Quaternionic weber local descriptor of color images. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *27*, 261–274.
53. Jian, M.; Qi, Q.; Dong, J.; Yin, Y.; Lam, K.M. Integrating QDWD with pattern distinctness and local contrast for underwater saliency detection. *J. Vis. Commun. Image Represent.* **2018**, *53*, 31–41.
54. Margolin, R.; Tal, A.; Zelnik-Manor, L. What makes a patch distinct? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 1139–1146.
55. Yang, C.; Zhang, L.; Lu, H. Graph-regularized saliency detection with convex-hull-based center prior. *IEEE Signal Process. Lett.* **2013**, *20*, 637–640.

56. Harrison, R.; Birchall, R.; Mann, D.; Wang, W. A novel ensemble of distance measures for feature evaluation: Application to sonar imagery. In *International Conference on Intelligent Data Engineering and Automated Learning*; Springer: Berlin, Germany, 2011; pp. 327–336.
57. Durrant-Whyte, H.; Bailey, T. Simultaneous localization and mapping: Part I. *IEEE Robot. Autom. Mag.* **2006**, *13*, 99–110.
58. Kim, A.; Eustice, R. Combined visually and geometrically informative link hypothesis for pose-graph visual SLAM using bag-of-words. In Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, USA, 25–30 September 2011; pp. 1647–1654, doi:10.1109/IROS.2011.6048439.
59. Kim, A.; Eustice, R.M. Real-Time Visual SLAM for Autonomous Underwater Hull Inspection Using Visual Saliency. *IEEE Trans. Robot.* **2013**, *29*, 719–733.
60. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110.
61. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359.
62. Kim, A.; Eustice, R.M. Active visual SLAM for robotic area coverage: Theory and experiment. *Int. J. Robot. Res.* **2015**, *34*, 457–475.
63. Ozog, P.; Carlevaris-Bianco, N.; Kim, A.; Eustice, R. Long-term Mapping Techniques for Ship Hull Inspection and Surveillance using an Autonomous Underwater Vehicle. *J. Field Robot.* **2015**, *24*, doi:10.1002/rob.21582.
64. Geng, Y.; Wang, Z.; Shi, C.; Nian, R.; Zhang, C.; He, B.; Shen, Y.; Lendasse, A. Seafloor visual saliency evaluation for navigation with BoW and DBSCAN. In Proceedings of the OCEANS 2016-Shanghai, Shanghai, China, 10–13 April 2016; pp. 1–5.
65. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD'96, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996*; AAAI Press: Palo Alto, CA, USA, 1996; pp. 226–231.
66. Li, J.; Kaess, M.; Eustice, R.M.; Johnson-Roberson, M. Pose-Graph SLAM Using Forward-Looking Sonar. *IEEE Robot. Autom. Lett.* **2018**, *3*, 2330–2337.
67. Kaeli, J. *Real-Time Anomaly Detection in Side-Scan Sonar Imagery for Adaptive AUV Missions*; In Proceedings of the 2016 IEEE/OES Autonomous Underwater Vehicles (AUV), Tokyo, Japan, 6–9 November 2016; pp. 85–89, doi: 10.1109/AUV.2016.7778653.
68. Chailloux, C. Region of interest on sonar image for non symbolic registration. In Proceedings of OCEANS 2005 MTS/IEEE, Washington, DC, USA, 17–23 September 2005; pp. 810–814.
69. Harris, C.G.; Stephens, M.; others. A combined corner and edge detector. In Proceedings of the Fourth Alvey Vision Conference, Manchester, UK, 31 August–2 September 1988; pp. 147–151, doi: 10.5244/c.2.23.
70. Chauvin, A.; Hérault, J.; Marendaz, C.; Peyrin, C. Natural scene perception: Visual attractors and image neural computation and psychology. In *Connectionist Models of Cognition and Perception, Proceedings of the Seventh Neural Computation and Psychology Workshop, Brighton, UK, 17–19 September 2001*; Bullinaria, J.A., Lowe, W., Eds.; World Scientific Publishing Co Pte Ltd.: Singapore, 2002.
71. Chailloux, C.; Le Caillec, J.; Gueriot, D.; Zerr, B. Intensity-Based Block Matching Algorithm for Mosaicing Sonar Images. *IEEE J. Ocean. Eng.* **2011**, *36*, 627–645.
72. Mitchell, H.B. *Image Fusion*; Springer: Berlin/Heidelberg, Germany, 2010.
73. Fu, L.; Wang, Y.; Zhang, Z.; Nian, R.; Yan, T.; Lendasse, A. A shadow-removal based saliency map for point feature detection of underwater objects. In Proceedings of the OCEANS 2015-MTS/IEEE Washington, Washington, DC, USA, 19–22 October 2015; pp. 1–5.
74. Zhang, L.; He, B.; Song, Y.; Yan, T. Underwater image feature extraction and matching based on visual saliency detection. In Proceedings of the OCEANS 2016-Shanghai, Shanghai, China, 10–13 April 2016; pp. 1–4.
75. Johnson-Roberson, M.; Bryson, M.; Douillard, B.; Pizarro, O.; Williams, S. Crowdsourced Saliency for Mining Robotically Gathered 3D Maps Using Multitouch Interaction on Smartphones and Tablets; In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, 31 May–5 June 2014; IEEE: New York, NY, USA, 2014; pp. 6032–6039, doi:10.1109/ICRA.2014.6907748.
76. Johnson-Roberson, M.; Bryson, M.; Douillard, B.; Pizarro, O.; Williams, S. Discovering salient regions on 3D photo-textured maps: Crowdsourcing interaction data from multitouch smartphones and tablets. *Comput. Vis. Image Underst.* **2015**, *131*, 28–41, doi:10.1016/j.cviu.2014.07.006.
77. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **1989**, *77*, 257–286.
78. Achanta, R.; Hemami, S.; Estrada, F.; Susstrunk, S. Frequency-tuned salient region detection. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1597–1604.
79. Fang, S.; Deng, R.; Cao, Y.; Fang, C. Effective Single Underwater Image Enhancement by Fusion. *J. Comput.* **2013**, *8*, doi:10.4304/jcp.8.4.904-911.
80. Singh, R.; Biswas, M. Adaptive histogram equalization based fusion technique for hazy underwater image enhancement. In Proceedings of the 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Tamil Nadu, India, 15–17 December 2016; pp. 1–5.
81. Ancuti, C.; Ancuti, C.O.; De Vleeschouwer, C.; Garcia, R.; Bovik, A.C. Multi-scale underwater descattering. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 4202–4207.
82. Wang, J.; Wang, H.; Gao, G.; Lu, H.; Zhang, Z. Single Underwater Image Enhancement Based on Lp-norm Decomposition. *IEEE Access* **2019**, 1, doi:10.1109/ACCESS.2019.2945576.

83. Forbes, T.; Goldsmith, M.; Mudur, S.; Poullis, C. DeepCaustics: Classification and Removal of Caustics From Underwater Imagery. *IEEE J. Ocean. Eng.* **2019**, *44*, 728–738.
84. Autodesk Maya. 2020. Available online: https://www.autodesk.com/products/maya/ (accessed on 21 December 2020).
85. Zhang, H.; Li, S.; Chen, W.; Liu, Y. The Influence of Different Saliency on Full-Reference Sonar Image Quality Evaluation. *IOP Conf. Ser. Mater. Sci. Eng.* **2019**, *569*, 052093, doi:10.1088/1757-899x/569/5/052093.
86. Liu, T.; Sun, J.; Zheng, N.; Tang, X.; Shum, H. Learning to Detect A Salient Object. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 18–23 June 2007; pp. 1–8.
87. Jian, M.; Qi, Q.; Dong, J.; Yin, Y.; Zhang, W.; Lam, K.M. The OUC-vision large-scale underwater image database. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 1297–1302.
88. Jian, M.; Qi, Q.; Yu, H.; Dong, J.; Cui, C.; Nie, X.; Zhang, H.; Yin, Y.; Lam, K.M. The extended marine underwater environment database and baseline evaluations. *Appl. Soft Comput.* **2019**, *80*, 425–437.
89. Li, X.; Lu, H.; Zhang, L.; Ruan, X.; Yang, M.H. Saliency detection via dense and sparse reconstruction. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 2–8 December 2013; pp. 2976–2983.
90. Tong, N.; Lu, H.; Zhang, L.; Ruan, X. Saliency detection with multi-scale superpixels. *IEEE Signal Process. Lett.* **2014**, *21*, 1035–1039.
91. Qin, Y.; Lu, H.; Xu, Y.; Wang, H. Saliency detection via cellular automata. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 110–119.
92. Jian, M.; Qi, Q. Underwater Images Part A. 2019. Available online: https://zenodo.org/record/2542305#.X-INxNhKiUk (accessed on 21 December 2020).
93. Jian, M.; Qi, Q. Underwater Images Part B. 2019. Available online: https://zenodo.org/record/2542307#.X-INxdhKiUk (accessed on 21 December 2020).
94. Li, C.; Guo, C.; Ren, W.; Cong, R.; Hou, J.; Kwong, S.; Tao, D. An underwater image enhancement benchmark dataset and beyond. *IEEE Trans. Image Process.* **2019**, *29*, 4376–4389.
95. Cui, Z.; Wu, J.; Yu, H.; Zhou, Y.; Liang, L. Underwater Image Saliency Detection Based on Improved Histogram Equalization. In *International Conference of Pioneering Computer Scientists, Engineers and Educators*; Springer: Berlin, Germany, 2019; pp. 157–165.
96. Boom, B.J.; He, J.; Palazzo, S.; Huang, P.X.; Beyan, C.; Chou, H.M.; Lin, F.P.; Spampinato, C.; Fisher, R.B. A research tool for long-term and continuous analysis of fish assemblage in coral-reefs using underwater camera footage. *Ecol. Inform.* **2014**, *23*, 83–97.
97. Li, G.; Yu, Y. Visual saliency detection based on multiscale deep CNN features. *IEEE Trans. Image Process.* **2016**, *25*, 5012–5024.
98. Huang, K.; Gao, S. Image saliency detection via multi-scale iterative CNN. *Vis. Comput.* **2020**, *36*, 1355–1367.
99. Sanchez-Torres, G.; Ceballos-Arroyo, A.; Robles-Serrano, S. Automatic measurement of fish weight and size by processing underwater hatchery images. *Eng. Lett.* **2018**, *26*, 461–472.