Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2023.028262
Article

Tech Science Press

An Efficient Method for Underwater Video Summarization and Object Detection Using YoLoV3

Mubashir Javaid[1], Muazzam Maqsood[2], Farhan Aadil[2], Jibran Safdar[1] and Yongsung Kim[3,*]

[1]Department of Computer Science, UET, Taxila, Pakistan
[2]Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock, Pakistan
[3]Department of Technology Education, Chungnam National University, Daejeon, 34134, Korea
*Corresponding Author: Yongsung Kim. Email: kys1001@cnu.ac.kr
Received: 06 February 2022; Accepted: 24 March 2022

Abstract: Currently, worldwide industries and communities are concerned with building, expanding, and exploring the assets and resources found in the oceans and seas. More precisely, to analyze a stock, archaeology, and surveillance, several cameras are installed underseas to collect videos. However, on the other hand, these large size videos require a lot of time and memory for their processing to extract relevant information. Hence, to automate this manual procedure of video assessment, an accurate and efficient automated system is a greater necessity. From this perspective, we intend to present a complete framework solution for the task of video summarization and object detection in underwater videos. We employed a perceived motion energy (PME) method to first extract the keyframes followed by an object detection model approach namely YoloV3 to perform object detection in underwater videos. The issues of blurriness and low contrast in underwater images are also taken into account in the presented approach by applying the image enhancement method. Furthermore, the suggested framework of underwater video summarization and object detection has been evaluated on a publicly available brackish dataset. It is observed that the proposed framework shows good performance and hence ultimately assists several marine researchers or scientists related to the field of underwater archaeology, stock assessment, and surveillance.

Keywords: Computer vision; deep learning; digital image processing; underwater video analysis; video summarization; object detection; YOLOV3

1 Introduction

About two-thirds of the earth's surface is covered with oceans and these are homelands for many organisms in the oceans. To regulate the climate conditions and contribute to the oxygen cycle, these marine organisms are very helpful [1]. Scientists and researchers started the ocean's studies to determine and understand how these resources present in the oceans will affect the climate conditions and other critical matters. Data collection and its analysis play a very vital role in these studies. Exploration surveys are one of the productive and functional methods to collect data on oceanic habitats. There are various

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

applications in which the data of these surveys are used in an extensive manner such as conducting censuses of population [2], geological mapping [3,4], and archeology [5]. There are many kinds of these surveys conducted where digital imagery-related surveys are the most important ones. Some important equipment and cameras for example sleds or trawl nets are used during recording and data collection of large flat areas such as in Nephrops surveys [2]. Remotely controlled vehicles (ROV) are another robust option with a camera for recording. At the end of these video collection surveys, a manual analysis is performed on all recorded data of videos by scientists [2]. The main reason behind the limitations of manual analysis includes that videos are lengthy in duration starting from several minutes in archaeological surveys [5] up to several hours in shark surveys [6]. This causes fatigue and tediousness in process of reviewing for the scientists. Further, the poor visibility of underwater videos is also a major limitation of manual analysis. Instead of only considering the problem of quantity and quality of video data, there exist some other problems such as different scientists producing different results according to their analysis. Due to this, an entire procedure of analysis needs to repeat painstakingly. To solve problems connected with the manual analysis of these videos, current research has been pursued to automate this manual analysis using computer vision and digital signal processing techniques. Some of these techniques include image enhancement, content analysis, and content summarization [5]. Some particular sets of items of interest can also be detected automatically by utilizing the content analysis techniques from the survey video e-g crabs [7] and lobsters [8]. Hence, one of the main objectives of this research is to automate the manual analysis of underwater videos using deep learning-assisted computer vision-based techniques. Hence, video summarization through keyframes comes to be a very accurate solution to this problem. One of the approaches to detect keyframes is based on the matching procedure of pixels in two subsequent frames in a video, referred to as detection of shot boundaries [9]. Some other approaches to keyframe extraction include perceptual features e-g color, motion, and objects. In an object-dependent approach, a threshold value is used to predict keyframes in videos by calculating the difference among the number of regions presents in the succeeding and preceding frames [10].

Hence, it is concluded that different video summarization-based techniques are adopted to select the keyframes from a video to summarize it in different domains. However, there exists very little amount of research on underwater video summarization [11]. On the other hand, a lot of object detection studies to detect different kinds of objects from underwater videos are proposed [12–14]. These object detection techniques include traditional thresholding-based methods [15,16] and deep learning methods including single-shot [12], Faster-RCNN [17], and YoLo-based [14] object detectors. However, these thresholding-based methods are less accurate and their performance drops with diverse varying backgrounds as well as with low contrast and blurry images. On the other hand, deep learning methods are more accurate than these methods. Therefore, the proposed framework is based on a deep learning strategy to perform object detection [18–21]. Nevertheless, utilizing and improving these deep learning-based methods is a continual area of research. Therefore, in comparison with existing work, we suggest more advanced version of the object detector model namely YoLoV3 to detect underwater creatures. YoLoV3 employs DrakNet-53 as backbone architecture which consists of 53 layers of convolution instead of DarkNet-19 which is used in YoLoV2 and ResNet-based architectures. These DarkNet architectures are faster than ResNet architectures and hence ultimately increase the speed of the model. Furthermore, during training, it employs the logistic classifiers to predict the classes and hence provides more accurate results in comparison with other YoLo versions and object detectors. On the same line, this study provides a more complete framework to perform both video summarization and object detection from underwater videos. Hence, this study more thoroughly facilitates marine research by giving a complete end-to-end solution. More precisely, in the first stage, we have extracted the keyframes using the PME method [22] to summarize the large size underwater videos by discarding all irrelevant frames, following on, image enhancement operations are performed over the resulting keyframes to remove blurry and shady effects.

The reason for choosing the PME model is that the patterns of motions in a video are also modeled in this algorithm since motion is the most noticeable element in events of video. Through this part of the proposed framework, a summarized form of video is generated. Subsequently, an object detector-based on YoLoV3 which is fine-tuned using pre-trained DarkNet-53 weights is employed to perform the object detection of under-water creatures such as crabs, fishes, including small, large size, and Jellyfishes, as well as assist in analyzing the habitat of different species in the sea or oceans. Moreover, this study will also help to extract information for stock assessment of different species. The suggested approach is validated on a publicly available underwater video dataset namely the brackish dataset and encouraging results have been obtained. This research has the following contributions:

- A complete end-to-end automated framework for underwater video summarization and object detection is proposed to assist marine researchers in conducting different tasks
- An objector model namely YoLoV3 with pre-trained weights of DarkNet53 is fine-tuned over the extracted keyframes to analyze the habitat of underwater species
- The proposed method performs well in carrying out video summarization and object detection

The remaining sections of the paper are categorized as follows, Section 2 provides a complete literature review on analysis of underwater vides, Section 3 describes the proposed method, Section 4 reports and explains various results and experiments while the last section provides the conclusion. Some sample images (video frames) of the brackish dataset are shown in Fig. 1.

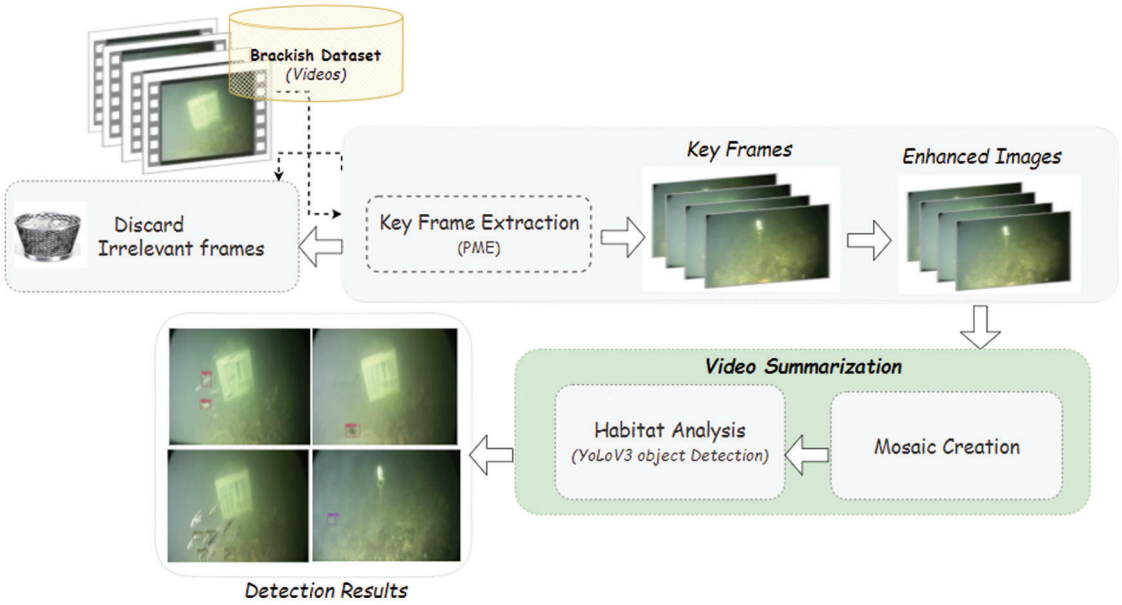

Figure 1: A pictorial overview of the proposed methodology

2 Related Work

In this section, we review some existing literature on video summarization approaches and later on we discuss some literature on underwater video analysis in terms of object detection of underwater creatures.

Video abstraction is the process of generating a representation of long videos which is concise and effective. It has numerous applications such as large volume video browsing and retrieval [23]. The video storage efficiency and effectiveness are also improved by this process [24]. Video summarization and video skimming are two major groups of video abstraction. Video summarization is also called still or

static image abstraction, static video abstract, or static storyboard [25]. On the other hand, video skimming is called moving image abstraction or moving/dynamic storyboard [26,27]. The most important content from videos sequences is preserved in both of the approaches so that for end-users a more comprehensible and understandable description is presented. In the community of computer vision, the summarization of video is an active research area. There are various categories and applications in which it is used such as Wildlife videos [28], TV documentaries [26], and sports videos [29]. There are six techniques of video summarization that include feature selection, event detection methods, trajectory analysis, clustering algorithms, shot selection, and the use of mosaics. In most cases, two or more techniques are used in combined form for example clustering with feature selection [25,26,30]. All these approaches differ from each other based on how they extract the feature vector for the representation of each frame of the video sequence [25,26,31]. Moreover, saliency maps and features based on motion are also utilized [27,31]. Every proposed method of feature vectors has some limitations such as for each particular frame only coarse information is maintained in color histograms-based approaches. For instance, when the motion in video sequences is very large then there is a failure of motion-based features. Moreover, for textured and cluttered backgrounds the performance of saliency maps-based methods is very poor. In comparison with these, in this study, we have performed the task of video summarization for underwater large-size videos to highlight the important events to reduce the manual process of analyzing these videos. Further, we have employed the PME method to first extract the keyframes from underwater videos.

Furthermore, if we look into the literature of underwater video analysis then, a monocular camera is used by Zhou et al. for tracking fish underwater [32]. An Autonomous underwater vehicle (AUV) is employed by Forney et al. [33] and Clark et al. [34] to target the leopard sharks. A stereo video with low frame and low contrast quality is studied by Chuang et al. [35] for tracking of fishes. Through these methods or devices, a large set of underwater videos are analyzed to target underwater creatures such as fishes and sharks. However, manual analysis of these large-scale videos is time consuming and hectic task since they contain a lot of redundant information. A shape-based level set technique is also used to detect underwater fishes by Ravanbakhsh et al. [36]. Copepods underwater are detected by Leow et al. [37] by using neural networks. A hierarchical classifier is designed by Huang et al. [38] for the recognition purpose of live fishes. An extreme learning machine (ELM) is used in the dynamic model hypothesis to track fish trajectory in the work of Nian et al. [39]. Gabor filter is used in the work of Zhou et al. [32] for tracking fish. All of these methods show good performances in the detection and classification of underwater creatures. However, there is still a gap that exists in improving these methods using a more advanced set of algorithms. Further, it is very challenging to overcome all these limitations such as ubiquitous noise and uneven illumination. Further, in the present era, deep learning-based object detection techniques are very commonly used in almost different domains [40,41]. A hybrid approach for object detection and classification is also proposed in [42] to improve the performance of object detection models. To detect salient objects from complex backgrounds, a deep learning approach along with graph-based segmentation is proposed in [43] to improve object detection. Hence, in comparison with these methods, this study analyzes the underwater videos using the deep learning method to track different underwater objects under a single framework such as fishes including very small and large size fishes, crabs, and jellyfishes. This suggested method is based on the computer-vision-assisted deep learning technique of YoLoV3 and is a more advanced approach than Gabor filters [32], ELM [39], neural networks [37], and hierarchical classifiers [38]. Further, in the scenario of video summarization, there exists some research works for event summarization of underwater videos [11,44]. Currently, image processing techniques based on different transforms such as dual watermarking are also proposed to maintain the security of data in the smart grid [45]. Audio watermarking for the security of audio data is also designed for the telemedicine domain [46].

3 Proposed Methodology

The pictorial representation of the proposed methodology is shown in Fig. 1. We first extract the keyframes from the underwater videos collected from the Brackish dataset using the Perceived motion energy (PME) model [22]. After extracting the keyframes the remaining frames are discarded. These keyframes provide the summarized form of underwater videos. Subsequently, we remove the shady and blurry effects from the video. For this, adaptive histogram equalization is applied to improve the visibility of the images. We have also created the mosaic from the underwater videos. Following on, the enhanced keyframes are used as an input to the Yolov3 objection detection model which is fine-tuned using DarkNet-53 architecture to detect underwater species. The detail of each step is given below:

3.1 Keyframe Extraction Using PME

In the first step, we have employed the PME method to extract the keyframes from underwater videos collected from the brackish dataset. Generally, the salient content of the video sequences is represented by keyframes. For many video-related tasks such as browsing, indexing, and retrieval, an appropriate abstraction is provided by these frames [47]. With the help of keyframes, the amount of data is reduced that is needed for video indexing and browsing. Generally, motion is a very common salient feature, and also it is very useful to determine the keyframes. Hence, in this work, the keyframes in the video are extracted by the triangle model of the Perceived motion energy (PME) model. This model observes the motion patterns to extract the keyframes. The frames are selected as keyframes that are present at turning points. There are two main turning points for frames, one is at motion acceleration and the other is motion declaration. This method extracts more representative keyframes from a given video sequence without any threshold criteria and hence it is a fast method. In this method, the motion activities in video shots are represented by the triangle model of PME. The model sub-segments the given video shots. These sub-segments differ from each other depending upon the movement patterns in terms of two parameters i-e acceleration and deceleration. The salience of visual action which is relative is reflected in accumulated PME and thus it can be employed as a major criterion to sort motion patterns according to their importance. Motion data is first extracted to build the model directly from the streams of the videos. In each macroblock of the B frame, there exist two vectors of motion for the compensation. This is also called the Motion vector field (MVF). For the entire frame, the vectors of motion's average magnitude $Mag(t)$ are computed by Eq. (1):

$$Mag(t) = \frac{\frac{\sum MixFE_{n(i,j)}(t)}{N} + \frac{\sum MixBE_{n(i,j)}(t)}{N}}{2} \tag{1}$$

In the above Eq. (1), the vectors of forwarding motion are represented by $MixFE_{n(i,j)}(t)$ while the vectors of backward motion are represented by $MixBE_{n(i,j)}(t)$. The total number of macroblocks is denoted by N. The values (i, j) denote the macroblock positions while the E denotes the energy. The computation of $MixFE_{n(i,j)}(t)$ and $MixBE_{n(i,j)}(t)$ is similar to $MixE_{n(i,j)}(t)$ [48]. The term $MixE_{n(i,j)}(t)$ involves the information about camera and object motions. The $\alpha(t)$ represents the dominant motion direction and its percentage can be defined as:

$$\alpha(t) = \frac{max(AH(t,\ k),\ k \in [1,\ n])}{\sum_{k=1}^{n} AH(t,\ k)} \tag{2}$$

The quantization of the 2π angle is done up to n angle ranges. After that, with n bins, the angle of the histogram is created overall vectors of forward motion represented by $AH(t, k)$ in the above Eq. (2) where the k belongs to $[1, n]$. In all directions of motions, the dominant direction bin is represented by the term

$max(AH(t, k)$. The term "*max*" indicates to select of the maximum values between ($AH(t, k)$ and k. The value of n is set to be 16. The PME of a given frame B is computed as:

$$PME(t) = Mag(t) \times \alpha(t) \tag{3}$$

In the above Eq. (3), the percentage of dominant motion direction is denoted by $\alpha(t)$ while the value of $Mag(t)$ is computed using Eq. (1) which denotes the vectors of motion's average magnitude. Moreover, to filter out the noise, a temporal filter is also applied with temporal window size W_t. In this method, sorting is performed according to magnitudes in the window. The values present at the last of the lists are cropped while for the remaining values in the list the average is taken. The resulting values form the mixture energy $MinEn_{(i,j)}$. Both object and camera motion energy is included as:

$$MinEn_{(i,j)} = \frac{1}{(M - 2 \times |\alpha M| \times W_t^2)} \sum_{m=\alpha M+1}^{M-\alpha M} Mag_{(i,j)}(m) \tag{4}$$

In the above Eq. (4), the total magnitudes which are present in the window are represented by M. The largest integer which is not greater than αM is denoted by $\lfloor \alpha M \rfloor$. The term W_t denotes the size of the temporal window. The values which are sorted in the list are called magnitude values denoted by $Mag_{(i,j)}(m)$. The value of parameter α whose values lie in the range $0 \leq \alpha \leq 0.5$ is known as the trimming parameter. All those samples which are not included in the accumulating computation are controlled by this parameter. More detail of this process can be found in [48]. For every B frame, the PME value is computed. This process is completely enough for selecting keyframes from given video shots.

3.2 Image Enhancement and Mosaic Creation

After extracting the keyframes from a video using the PME model, we have applied the image enhancement method to enhance the visibility of the resulting frames. It is due to the reason that detecting the objects of interest accurately requires the quality of images to be enhanced since analyzing blurry and shady images may result in poor performances. More precisely, in this research study, the visibility in these resulting keyframes will be improved by applying the adaptive histogram equalization. For analyzing underwater habitat, we will create an image mosaic from videos after removing blurry/ Shady effects and image enhancement. Mosaic will be created by mapping each frame to a common reference frame in the sequence.

3.3 Habitat Analysis

In the last stage, with respect to burrows, a habitat is analyzed with the help of object detection technique. The object detector YOLOV3 [49] is applied here for purpose of analyzing. The input of the YoLoV3 is the enhanced keyframes of the underwater videos. Moreover, the YoLoV3 does not require the pre-, and post-processing steps for final outputs and works as a single network model to detect multiple objects from an image using one single pass. Hence the process is called unified detection. The separate chunks of the object detection algorithm are integrated into a single network model. Yolo model looks only once at the input to predict the output irrespective of location and type of object in the image. An end-to-end training is used in the Yolo hence increasing the speed of the network. The architecture of the Yolo model consists of 24 convolution layers. After that, two fully connected layers are deployed. Instead of using max-pooling layers for down-sampling an image, the strided convolutions are used in the network for downscaling. The input image is first divided into 7×7 grids. For detecting each object, the grid cell is responsible in the case, when the object falls into the center of the grid cell. The bounding box and probabilities of classes are predicted for each grid by the model.

3.3.1 Bounding Box Prediction:

For each prediction from the model, the four coordinates are predicted as shown in Fig. 2. More precisely, the bounding box with four different locations is predicted by the model. Further, with the help of the sigmoid function denoted by sigma, the center coordinates of the bounding box are predicted in relation to the position of application of the filter. For every bounding box, the four coordinates are t_x, t_y, t_w and t_h. On the left top corner of the image by (c_x, c_y), if the cell is offset and the height and width of the bounding box are represented by p_w and p_h then the prediction of the model is given by:

$$b_x = \sigma(t_x) + c_x \tag{5}$$

$$b_y = \sigma(t_y) + c_y \tag{6}$$

$$b_w = p_w e^{t}_{w} \tag{7}$$

$$b_h = p_h e^{t}_{h} \tag{8}$$

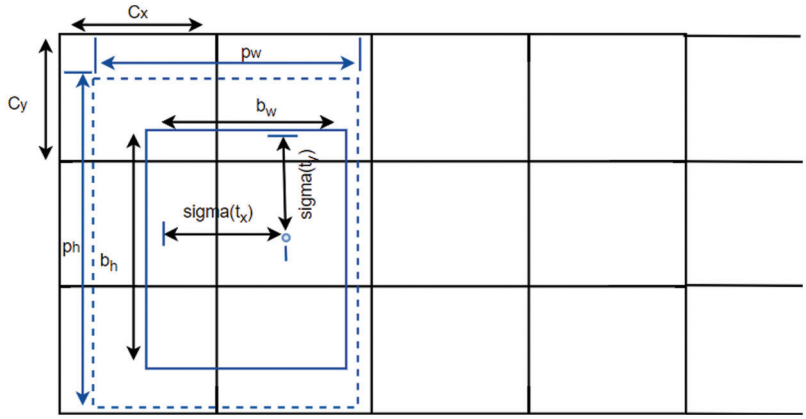

Figure 2: Predictions in form of bounding boxes

In the above Eqs. (5)–(8), b_w and b_h denote the width and height of the bounding box predicted by the YoLoV3 model. Moreover, the sum of squared error is used for training purposes.

If for coordinate prediction the ground-truth value is, $\hat{t}_*$ then the ground-truth value is the gradient which is calculated with the help of the ground-truth box is subtracted from the prediction $\hat{t}_* - t_*$. All the equations given above are inverted for computing the value of ground truth. The logistic regression is used to predict the objective score of every bounding box in the Yolo model. In case the bounding box prior is overlapped by more than one ground-truth object then this value becomes 1.

3.3.2 Predictions of Classes

By using multi-label classification, the classes present in the bounding boxes are predicted by each box. The logistic classifiers are used in the network instead of using softmax. For predictions of classes, the loss function named binary-cross entropy is used. A sample example of prediction is shown in Fig. 3.

3.3.3 Scale Based Predictions

More precisely, in Fig. 3 the prediction of the YOLO model is depicted in which the input image of underwater is divided into the $S \times S$ grid. In case, when the object falls within a grid cell then for the detection of the object this grid cell is responsible. On the other hand, the Yolo v3 model uses three different scales for predicting boxes. Similar to the feature pyramid networks [50] model, all these scales

are used for the process of feature extraction. Different convolution layers are inserted in the model from the base feature extractor. A 3D tensor is predicted which contains three values i-e objectness score, predictions, and bounding box. Moreover, in the Yolo model, feature maps resulting from earlier layers of the model are also used and integrated with up-sampled features with help of concatenation operation. From the up-sampling features, a more semantic formation is extracted while the earlier feature maps are used to get fine-grained information. These integrated feature maps are further processed by adding more convolution layers. By this operation, the size becomes twice but has no impact on the tensor. For the prediction of the final scale, the same pattern is repeated one more time. The bounding box priors are further by using k-means clustering.

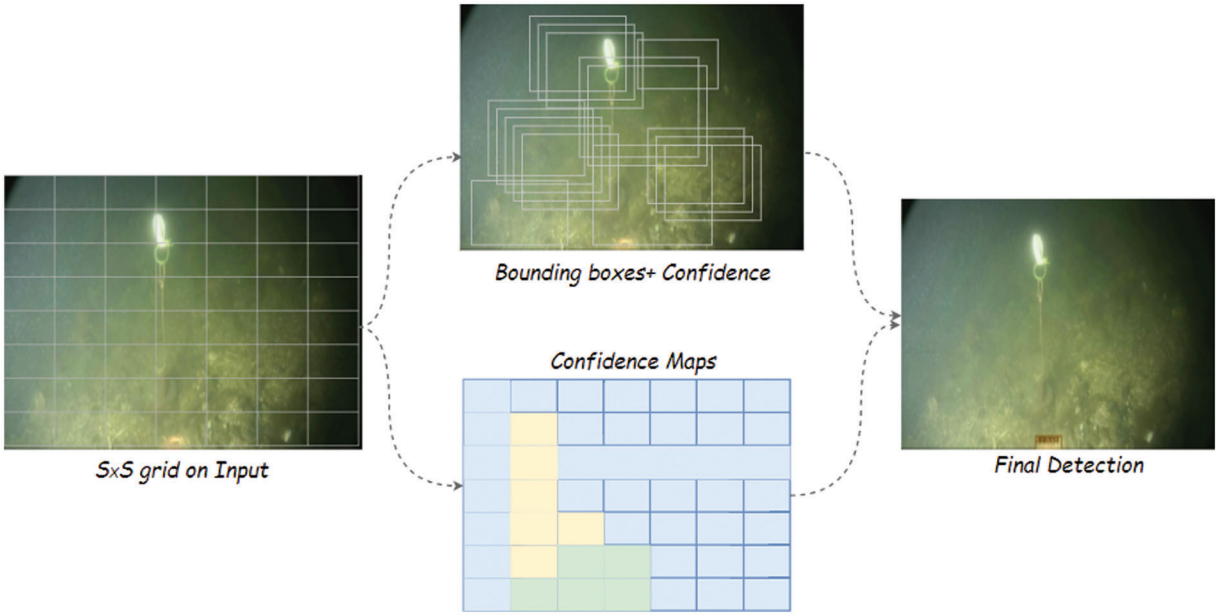

Figure 3: The object detection model: Regression is used to mimic the detection. At first, the input image is divided into $S \times S$ grid and for every cell of the grid, the model predicts B bounding boxes along with the values of confidence and probabilities of classes C. A tensor $S \times S \times (B*5 + C)$ is used to represent the predictions

3.4 Feature Extraction

In yolov3, the Darknet-53 is used for feature extraction while the Darknet-19 is used in yolov2. However, in yolov3, more successive layers of size 3×3 and 1×1 are added to the stuff of the network. The network becomes larger with 53 convolution layers and with more shortcut connections. Here in our work, we used the pre-trained weights of Darknet-53 on our custom dataset to perform object detection on the resulting keyframes of videos. Moreover, the hyperparameters of the model include the learning rate, which is set to 0.0001, the number of epochs is 100, and a batch size of 8 samples.

4 Experiments and Results

In this section, we discussed the results of the suggested framework along with their analysis. In addition, the details related to the dataset are also added.

4.1 Dataset

There exist many underwater datasets [51–53] for research purposes. However, in this research study, we have used the brackish dataset which is the first and freely accessible dataset that contains underwater videos from the European seas. This dataset involves the bounding box ground truth labels of different objects in the water i-e fish, crabs, and other aquatic species. The data has been collected in Limfjorden, a brackish waterway that flows around Aalborg in northern Denmark. The total number of videos in the dataset is 80 along with bounding box labeling in different formats such as YoLo Darknet, MS COCO, and AAU. The fishes in water are labeled in eight kinds of classes including fish, small fish, crab, shrimp, jellyfish, and starfish. All the videos are divided into different folders depending upon the number of times with which they are labeled. The whole dataset is already divided into three non-overlapping sets namely, train, validation, and tests in ratio 80/10/10. Different file names i-e train.txt, test.txt, and validation.txt are provided in which the names of the frames of video of every set are specified.

4.2 Experiments

To assess the performance of the proposed framework, we have validated it over the brackish dataset. At first, the frames of brackish video for each category which includes fish, small fish, crab, shrimp, jellyfish, and starfish are extracted. Later, we employed the PME approach to select the important keyframes from each video. Those extracted keyframes from underwater videos are shown in Fig. 4. These keyframes play an important role in video summarization. If more significant keyframes from a video that contains interesting events are identified, then the summarized video will be more informative. One of the important elements in highlighting these events is that they are in motion and ultimately act as an essential feature to determine the keyframes [22]. Furthermore, the summarized video in the form of keyframes should reflect the salient visual content of the video. Hence, to extract more relevant frames by involving motion patterns, we employed the PME method. Through this method, more relevant keyframes are extracted which ultimately makes the summarization accurate as well as improves its performance. Following on, all these frames are applied with image enhancement operation to enhance their visibility as shown in Fig. 4 (bottom row). As obvious from these images that the resulting enhanced images are bright with high contrast. This step makes it easy for the model to detect different objects as these enhanced underwater images are given as an input to the YoLoV3 object detection model. Since there exist very small objects, such as small size fishes, whose color and texture appear to be submerged with water due to inadequate illumination. Hence, we improve this degradation in illumination as well as contrast by enhancing the underwater images. Moreover, the YoloV3 with pre-trained DarkNet53 weights is fine-tuned over the train set frames. After training, we have validated the trained model over the test. The results of detection for the category of “Jellyfish” are shown in Tab. 1. In the above Tab. 1, it is observed that the suggested model detects the underwater species namely “Jellyfish” in 63 frames out of 116 frames in the first video. More precisely, the corrected labeled frames were only 12 in this video. Similarly, for videos 2, 4, and 5 the total number of correctly labeled frames is 0. However, the total number of non-labeled frames is much higher than labeled frames. Likewise, we access the performance of the detector over the second category called Crab. The results of the trained model in detecting the crabs from underwater videos are given in Tab. 2. It is observed that correctly labeled frames for category “crab” are more than the category “Jellyfish”.

As seen from Tab. 2, for video 1 and video 5, the correctly labeled frames are 19 and 28 respectively. Furthermore, the ratio of non-labeled frames is more than labeled frames in this category. Similarly, the detection results in terms of correctly labeled frames for the category “big fish” are also given in Tab. 3. It is observed that with this category the model performs better than the previous two categories namely ‘Jellyfish” and “Crab”. This is due to the reason that big fishes are large size objects and are easier to be detectable by the model. Moreover, the textural appearance of jellyfish appears to be very light until it is

very close to the observing camera and hence the result of jellyfish detection is less accurate than big fish. Similarly, the same is the case with the crabs. The crabs appear to be idle on the sea's bottom in the majority of the videos hence causing difficulty for the model to be detectable since it is difficult to observe the crabs. The total number of highest correctly labeled frames for this category is 258 for test video 3. However, for the last video, the total number of correctly annotated frames is zero. Later on, we observe the performance of the object detector over the "small fish" as given in Tab. 4. For this category, we have only three test set videos.

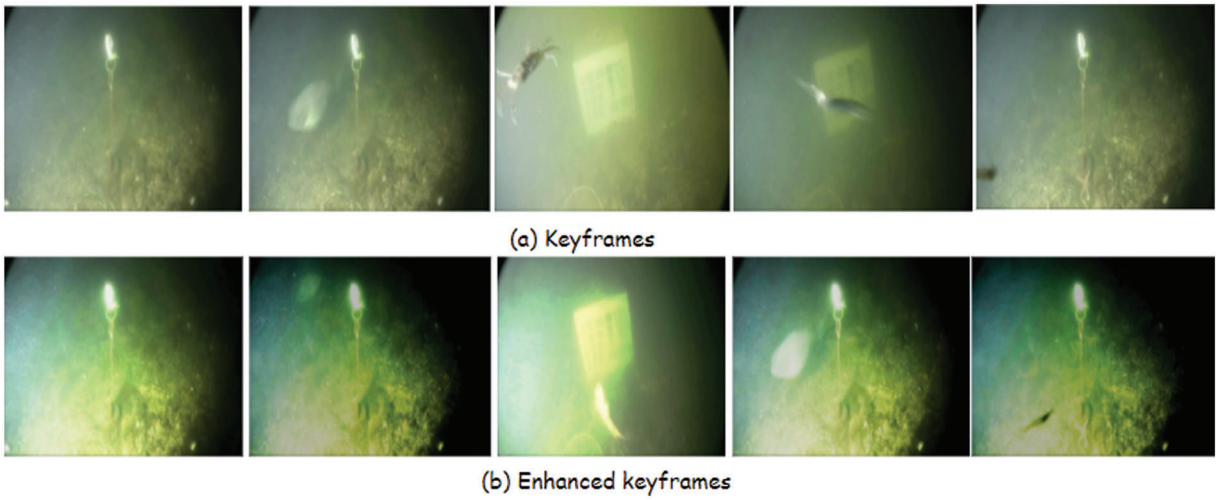

Figure 4: Some sample images of extracted keyframes and enhanced keyframes

Table 1: Detection results of YOLoV3 on the category "Jelly Fish"

Category = "Jelly Fish"				
Video	Total frames	Labeled frames	Non-labeled frames	Correct
1	116	63	41	12
2	108	16	92	0
3	113	6	102	5
4	111	10	101	0
5	146	0	108	0

Table 2: Detection results of YOLoV3 on category "Crab"

Category = "Crab"				
Video	Total frames	Labeled frames	Non-labeled frames	Correct
1	120	24	77	19
2	190	0	108	0
3	183	0	148	0
4	139	0	74	0
5	114	0	47	28
6	139	0	122	0

Table 3: Detection results of YOLoV3 on the category "Big Fish"

Category = "Fish big"				
Video	Total frames	Labeled frames	Non-labeled frames	Correct
1	222	69	24	129
2	182	16	11	155
3	461	139	89	258
4	158	79	17	55
5	190	111	53	25
6	149	11	98	40
7	267	3	203	61
8	200	46	66	88
9	179	18	115	46
10	162	2	3	157
11	185	14	4	167
12	145	0	101	0

In the last, we have accessed the performance of the proposed model over the fish small shrimp category. The results of the detection model over this category in terms of correctly labeled and unlabeled frames are given in Tab. 5. It is observed that the object detector model performs better with categories of fish i-e small, big, and shrimp than the crab and jellyfish. More specifically, the total number of correctly labeled frames in this category for video is 17, similarly, for video 2 it is 64, for video 3 it is 17, and so on. Nevertheless, for video 5 the total number of correctly labeled frames is 0, similarly, for video 5, the total number of correctly labeled frames is only 3. Category by category observation of detection results shows that model shows good results in almost all of the categories except for the accurate detection of "Jelly Fish". In addition to these, some detection results in terms of output bounding boxes are shown in Fig. 5. It is observed that the model performs efficiently in the detection of different underwater species, especially very small size species. As shown in the first image of Fig. 5, a model is detecting a very small object namely small fish.

Table 4: Detection results of YOLoV3 on the category "Small Fish"

Category = "Small Fish"				
Video	Total frames	Labeled frames	Non-labeled frames	Correct
1	120	29	91	0
2	610	91	72	447
3	256	27	54	175

It is obvious from the image that it is very difficult to observe these objects even with the naked eye, but the suggested model shows encouraging results. Similarly, the second image in the first row also depicts the detection results of very small underwater species. However, the third image in row 1, the second and fourth image in row 2 depict the results of detection with large size species underwater.

Table 5: Detection results of YOLoV3 on the category "Small Fish Shrimp"

Category = "Small Fish SHRIMP"				
Video	Total frames	Labeled frames	Non-labeled frames	Correct
1	109	60	31	17
2	205	63	76	64
3	171	69	77	17
4	181	87	48	41
5	202	12	190	0
6	183	19	94	70
7	131	13	32	86
8	121	0	82	3

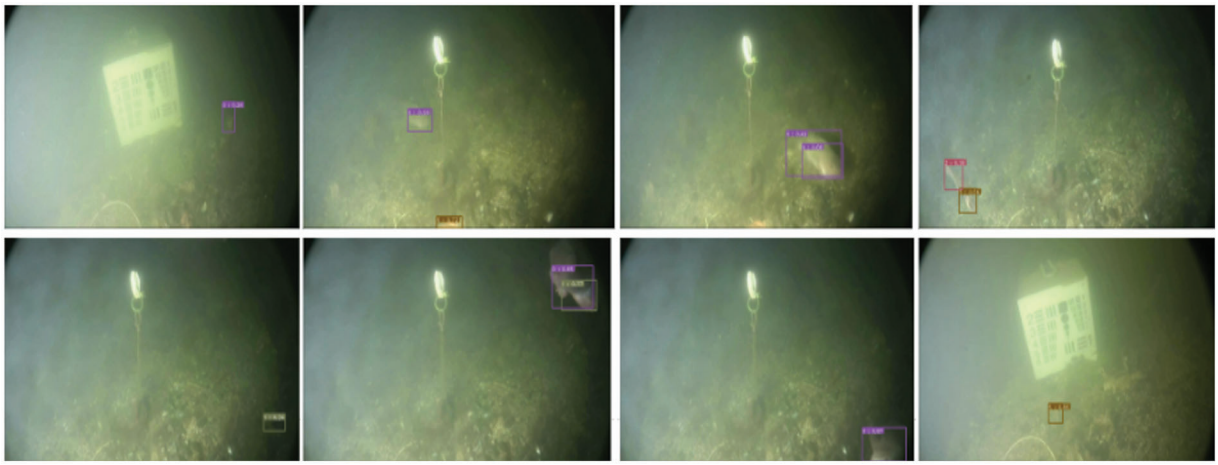

Figure 5: Detection results of the YoloV3 object detection model

4.3 Discussions and Comparisons

Currently, different research scientists related to marine management have adopted different technologies such as remote sensing, digital cameras, remotely operated underwater vehicles (AUV), and Unmanned undersea vehicles (UUV) to track different activities underwater as surveillance, stock assessment, and habitat analysis. During their navigation in the seas and oceans, these gadgets acquire large-scale videos. However, analyzing these large size videos manually to extract the information regarding key events such as stock assessments and habitat analysis of different underwater species is a challenging task for marine researchers. On the other hand, in underwater videos, the accurate detection of different species is also a difficult task due to poor illumination and lighting conditions underwater. Hence, to solve these challenges, this research study intends to provide a solution to automate the manual process of video examination through the video summarization process. Currently, in existing studies, there is very little amount of research on the summarization of videos [11] for under video. On the other hand, the frameworks for the detection of underwater creatures through object detection models are proposed in different studies [12–14]. However, in comparison with them, this study provides a complete end-to-end solution for underwater video analysis by performing both video summarization and object detection to carry out habitat analysis. Video summarization is one of the emerging tools and strategies

which not only provides the condensed version of the video but the most relevant parts of the videos are also preserved as well as also assisting in providing benefits to organizations that are deeply involved in video processing and searching. Through video summarization, important frames of video referred to as keyframes are extracted and their integration leads to a summarized video. In this study, we have employed the PME method to extract the keyframes from underwater videos.

Furthermore, to handle the light and illumination problems we have to apply the image enhancement operations over the resulting keyframes to enhance their visibility. Following that, the keyframes are used to train the object detection model to recognize different types of species underwater to perform habitat analysis. For this purpose, we have employed the YoloV3 object detection model that utilizes less costly backbone architecture namely DarkNet-53 than previous versions of YoLo that utilizes ResNet-based architectures. Instead of training it from scratch, we have performed the fine-tuning of the model on our custom dataset. It is observed that the suggested framework provides complete automation to underwater video analysis by first summarizing the content of the video followed by image enhancements and object detection in videos to carry out habitat analysis and exhibits the encouraging outcomes. Furthermore, when compared to prior research studies, then the proposed framework is a more complete solution to automate the process of underwater video analysis. The comparison with existing methods is shown in Tab. 6.

Table 6: Comparison with Existing methods in underwater video analysis

Authors	Approach	Video summarization	Underwater object detection	Performance
Kavitha et al. [11]	Statistical feature extraction	Yes	No	0.1 (false negative ratio)
Cao et al. [12]	Single Shot Detector (SSD)	No	Yes (only crab class)	0.99 (mAP)
Chen et al. [13]	Monocular vision	No	Yes	0.965 (precision)
Zhang et al. [54]	YoloV4	No	Yes (6 species)	0.9265 (mAP)
Xu et al. [14]	YoLo object detector	No	Yes (Only fishes)	0.5392 (mAP)
Proposed	**PME + YoLoV3**	**Yes**	**Yes (6 species)**	**0.946 (mAP)**

For instance, Kavitha et al. [11] proposed an underwater video summarization approach in which the frames of video are converted into wavelet sub-bands followed by calculating the standard deviation among two successive frames. Similarly, Cao et al. [12] proposed a Single-shot detector (SSD) based object detection model for the detection of underwater species. They modified the backbone architecture of SSD with MobileNetv2 and replaced the traditional layers of convolution with depth-wise convolutions. Likewise, Chen et al. [13] suggest a monocular vision-based sensor-reliant object detection method to detect underwater objects. Instead of only employing visual features such as color and intensity, their work also incorporates the information of light transmission. Moreover, Xu et al. [14] proposed a deep learning-based object detection method for the tracking and monitoring of Jellyfish in seas and oceans. Zhang et al. [54] employs YoLoV4 object detector to perform underwater object detection. Further, it is worth noting that the above-mentioned studies utilize various datasets for underwater objects identification. Furthermore, some research studies have only detected a few underwater species of e-g crabs and only fishes. In comparison to these existing methods, this study demonstrates an extended version in which six different underwater species are targeted. More

specifically, this study involves both summarization and habitat analysis with a PME method and object detection approach to entirely automate the system and provide a complete framework to assist the marine research community and management.

5 Conclusion

Currently, the attention of the industrial and scientific community especially marine researchers towards underwater video analysis through computer vision approaches have been risen owing to advancements in the field of computer vision, artificial intelligence, and digital image processing technologies. Hence, to assist the marine researchers, an automated deep learning-based complete solution is proposed in this study to allow for quick analysis of these under-water videos. This can be accomplished by video summarization strategy using the Perceived motion energy (PME) method which extracts the keyframes from the underwater videos and later on these frames are enhanced to remove the blurriness. Subsequently, an object detection algorithm namely YoLoV3 is employed to perform habitat analysis of underwater species. This involves the detection of different objects underwater such as crabs, jellyfishes, small and big size fishes, etc. The YoLoV3 model is fine-tuned using pre-trained DarkNet53 weights to perform object detection. It is observed that the proposed framework shows the best outcomes and has the potential to assist marine researchers in conducting their studies regarding different tasks. In the future, the proposed framework is extended with other object detection approaches with different backbone networks along with attention mechanisms to further improve the performance.

Funding Statement: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1G1A1099559).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] P. Cloud and A. Gibor, "The oxygen cycle," *Scientific American*, vol. 223, no. 3, pp. 110–123, 1970.

[2] N. Campbell, H. Dobby and N. Bailey, "Investigating and mitigating uncertainties in the assessment of scottish nephrops norvegicus populations using simulated underwater television data," *ICES Journal of Marine Science*, vol. 66, no. 4, pp. 646–655, 2009.

[3] F. M. Caimi, D. M. Kocak and V. L. Asper, "Developments in laser-line scanned undersea surface mapping and image analysis systems for scientific applications," in *OCEANS 96 MTS/IEEE Conf. Proc. the Coastal Ocean-Prospects for the 21st Century*, Fort Lauderdale, FL, USA, pp. 75–81, 1996.

[4] K. Lebart, E. Trucco and D. Lane, "Real-time automatic sea-floor change detection from video," in *OCEANS 2000 MTS/IEEE Conf. and Exhibition. Conf. Proceedings (Cat. No. 00CH37158*), Providence, RI, USA, pp. 1337–1343, 2000.

[5] Y. Y. Schechner and N. Karpel, "Recovery of underwater visibility and structure by polarization analysis," *IEEE Journal of Oceanic Engineering*, vol. 30, no. 3, pp. 570–587, 2005.

[6] M. E. Bond, E. A. Babcock, E. K. Pikitch, D. L. Abercrombie, N. F. Lamb *et al.,* "Reef sharks exhibit site-fidelity and higher relative abundance in marine reserves on the mesoamerican barrier reef," *PloS One*, vol. 7, no. 3, pp. e32983, 2012.

[7] M. Mehrnejad, A. B. Albu, D. Capson and M. Hoeberechts, "Detection of stationary animals in deep-sea video," in *2013 OCEANS-San Diego*, San Diego, CA, USA, pp. 1–5, 2013.

[8] P. Lau, P. Correia, P. Fonseca and A. Campos, "Estimating Norway lobster abundance from deep-water videos: An automatic approach," *IET Image Processing*, vol. 6, no. 1, pp. 22–30, 2012.

[9] E. Ardizzone and M. La Cascia, "Automatic video database indexing and retrieval," in *Representation and Retrieval of Video Data in Multimedia Systems*, Paris, France, Springer, pp. 29–56, 1997.

[10] C. Kim and J. -N. Hwang, "An integrated scheme for object-based video abstraction," in *Proc. of the Eighth ACM Int. Conf. on Multimedia*, New York, United States, pp. 303–311, 2000.

[11] J. Kavitha and P. A. J. Rani, "Design of a video summarization scheme in the wavelet domain using statistical feature extraction," *International Journal of Image, Graphics and Signal Processing*, vol. 4, pp. 60–67, 2015.

[12] S. Cao, D. Zhao, X. Liu and Y. Sun, "Real-time robust detector for underwater live crabs based on deep learning," *Computers and Electronics in Agriculture*, vol. 172, pp. 105339, 2020.

[13] Z. Chen, Z. Zhang, F. Dai, Y. Bu and H. Wang, "Monocular vision-based underwater object detection," *Sensors*, vol. 17, pp. 1784, 2017.

[14] W. Xu and S. Matzner, "Underwater fish detection using deep learning for water power applications," in *2018 Int. Conf. on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA, pp. 313–318, 2018.

[15] A. Saini and M. Biswas, "Object detection in underwater image by detecting edges using adaptive thresholding," in *2019 3rd Int. Conf. on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, pp. 628–632, 2019.

[16] M. Girija, M. Rajasekar and S. Nithiya, "Real time live fish object detection and tracking in under water stereo videos," *Assessment*, vol. 1, pp. 21–24, 2014.

[17] S. Huang, M. Huang, Y. Zhang and M. Li, "Under water object detection based on convolution neural network," in *Int. Conf. on Web Information Systems and Applications*, Hong Kong, pp. 47–58, 2019.

[18] M. Bukhari, K. B. Bajwa, S. Gillani, M. Maqsood, M. Y. Durrani *et al.,* "An efficient gait recognition method for known and unknown covariate conditions," *IEEE Access*, vol. 9, pp. 6465–6477, 2020.

[19] M. Maqsood, M. Bukhari, Z. Ali, S. Gillani, I. Mehmood *et al.,* "A residual-learning-based multi-scale parallel-convolutions-assisted efficient cad system for liver tumor detection," *Mathematics*, vol. 9, no. 10, pp. 1133, 2021.

[20] M. Maqsood, S. Yasmin, I. Mehmood, M. Bukhari and M. Kim, "An efficient da-net architecture for lung nodule segmentation," *Mathematics*, vol. 9, no. 13, pp. 1457, 2021.

[21] R. Ashraf, S. Afzal, A. U. Rehman, S. Gul, J. Baber *et al.,* "Region-of-interest based transfer learning assisted framework for skin cancer detection," *IEEE Access*, vol. 8, pp. 147858–147871, 2020.

[22] T. Liu, H. -J. Zhang and F. Qi, "A novel video key-frame-extraction algorithm based on perceived motion energy model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 10, pp. 1006–1013, 2003.

[23] M. Ajmal, M. H. Ashraf, M. Shakir, Y. Abbas and F. A. Shah, "Video summarization: Techniques and classification," in *Int. Conf. on Computer Vision and Graphics*, Warsaw Poland, pp. 1–13, 2012.

[24] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 3, no. 1, pp. 3–es, 2007.

[25] S. E. F. De Avila, A. P. B. Lopes, A. da Luz Jr. and A. de Albuquerque Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.

[26] J. Almeida, N. J. Leite and R. d. S. Torres, "Vison: Video summarization for online applications," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 397–409, 2012.

[27] Q. -G. Ji, Z. -D. Fang, Z. -H. Xie and Z. -M. Lu, "Video abstraction based on the visual attention model and online clustering," *Signal Processing: Image Communication*, vol. 28, no. 3, pp. 241–253, 2013.

[28] S. -P. Yong, J. D. Deng and M. K. Purvis, "Key-frame extraction of wildlife video based on semantic context modeling," in *The 2012 Int. Joint Conf. on Neural Networks (IJCNN)*, Brisbane, QLD, Australia, pp. 1–8, 2012.

[29] J. -q. Ouyang and R. Liu, "Ontology reasoning scheme for constructing meaningful sports video summarisation," *IET Image Processing*, vol. 7, no. 4, pp. 324–334, 2013.

[30] P. Mundur, Y. Rao and Y. Yesha, "Keyframe-based video summarization using delaunay clustering," *International Journal on Digital Libraries*, vol. 6, no. 2, pp. 219–232, 2006.

[31] A. Divakaran, K. A. Peker and H. Sun, "Video summarization using motion descriptors," in *Storage and Retrieval for Media Databases 2001*, San Jose, CA, United States, pp. 517–522, 2001.

[32] J. Zhou and C. M. Clark, "Autonomous fish tracking by rov using monocular camera," in *The 3rd Canadian Conf. on Computer and Robot Vision (CRV'06)*, Quebec, Canada, pp. 68, 2006.

[33] C. Forney, E. Manii, M. Farris, M. A. Moline, C. G. Lowe *et al.,* "Tracking of a tagged leopard shark with an auv: Sensor calibration and state estimation," in *2012 IEEE Int. Conf. on Robotics and Automation*, Saint Paul, MN, USA, pp. 5315–5321, 2012.

[34] C. M. Clark, C. Forney, E. Manii, D. Shinzaki, C. Gage *et al.,* "Tracking and following a tagged leopard shark with an autonomous underwater vehicle," *Journal of Field Robotics*, vol. 30, no. 3, pp. 309–322, 2013.

[35] M. -C. Chuang, J. -N. Hwang, K. Williams and R. Towler, "Tracking live fish from low-contrast and low-frame-rate stereo videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, pp. 167–179, 2014.

[36] M. Ravanbakhsh, M. Shortis, F. Shaifat, A. S. Mian, E. Harvey *et al.,* "An application of shape-based level sets to fish detection in underwater images," in *GSR*, CEUR Workshop Proceedings, Melbourne, Australia, 2014.

[37] L. K. Leow, L. -L. Chew, V. C. Chong and S. K. Dhillon, "Automated identification of copepods using digital image processing and artificial neural network," *BMC Bioinformatics*, vol. 16, pp. 1–12, 2015.

[38] P. X. Huang, B. J. Boom and R. B. Fisher, "Hierarchical classification with reject option for live fish recognition," *Machine Vision and Applications*, vol. 26, pp. 89–102, 2015.

[39] R. Nian, B. He, B. Zheng, M. Van Heeswijk, Q. Yu *et al.,* "Extreme learning machine towards dynamic model hypothesis in fish ethology research," *Neurocomputing*, vol. 128, pp. 273–284, 2014.

[40] Z. Zou, Z. Shi, Y. Guo and J. Ye, "Object detection in 20 years: A survey," *arXiv preprint arXiv:1905.05055*, 2019.

[41] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen *et al.,* "Deep learning for generic object detection: A survey," *International Journal of Computer Vision*, vol. 128, pp. 261–318, 2020.

[42] M. Aamir, Y. -F. Pu, Z. Rahman, W. A. Abro, H. Naeem *et al.,* "A hybrid proposed framework for object detection and classification," *Journal of Information Processing Systems*, vol. 14, pp. 1176–1194, 2018.

[43] Z. Rahman, Y. -F. Pu, M. Aamir and F. Ullah, "A framework for fast automatic image cropping based on deep saliency map detection and Gaussian filter," *International Journal of Computers and Applications*, vol. 41, pp. 207–217, 2019.

[44] Y. Cong, B. Fan, D. Hou, H. Fan, K. Liu *et al.,* "Novel event analysis for human-machine collaborative underwater exploration," *Pattern Recognition*, vol. 96, pp. 106967, 2019.

[45] Q. Chen and M. Xiong, "Dual watermarking based on wavelet transform for data protection in smart grid," in *2016 3rd Int. Conf. on Information Science and Control Engineering (ICISCE)*, Beijing, China, pp. 1313–1316, 2016.

[46] X. Zhang, X. Sun, X. Sun, W. Sun and S. K. Jha, "Robust reversible audio watermarking scheme for telemedicine and privacy protection," *Computers, Materials & Continua*, vol. 71, pp. 3035–3050, 2022.

[47] P. Aigrain, H. Zhang and D. Petkovic, "Content-based representation and retrieval of visual media: A state-of-the-art review," *Multimedia Tools and Applications*, vol. 3, pp. 179–202, 1996.

[48] Y. -F. Ma and H. -J. Zhang, "A new perceived motion based shot content representation," in *Proc. 2001 Int. Conf. on Image Processing (Cat. No. 01CH37205)*, Thessaloniki, Greece, pp. 426–429, 2001.

[49] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[50] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan *et al.,* "Feature pyramid networks for object detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Hawai'i Convention Center, pp. 2117–2125, 2017.

[51] G. Lic and W. Ren, "An underwater image enhancement benchmark dataset and beyond," *IEEE Transactions on Image Processing*, vol. 29, pp. 4376G4389, 2020.

[52] TURBID: *An underwater Dataset*, Available at: http://amandaduarte.com.br/turbid/. (Accessed on: 26 Nov 2021).

[53] Brackish Dataset: Available at: https://www.kaggle.com/aalborguniversity/brackish-dataset. Accessed on: 26 Nov 2021).

[54] M. Zhang, S. Xu, W. Song, Q. He and Q. Wei, "Lightweight underwater object detection based on YOLO v4 and multi-scale attentional feature fusion," *Remote Sensing*, vol. 13, no. 22, pp. 4706, 2021.