Detection of Marine Animals in a New Underwater Dataset with Varying Visibility

Conference Paper · June 2019

CITATIONS
16

READS
928

5 authors, including:

Malte Pedersen
Aalborg University
10 PUBLICATIONS **46** CITATIONS
SEE PROFILE

Rikke Gade
Aalborg University
30 PUBLICATIONS **760** CITATIONS
SEE PROFILE

Thomas B. Moeslund
Aalborg University
392 PUBLICATIONS **10,330** CITATIONS
SEE PROFILE

Niels Madsen
Aalborg University
114 PUBLICATIONS **1,774** CITATIONS
SEE PROFILE

Detection of Marine Animals in a New Underwater Dataset with Varying Visibility

Malte Pedersen, Joakim Bruslund Haurum, Rikke Gade, Thomas B. Moeslund
Visual Analysis of People (VAP) Laboratory, Aalborg University
mape@create.aau.dk, joha@create.aau.dk, rg@create.aau.dk, tbm@create.aau.dk

Niels Madsen
Section of Biology and Environmental Science, Aalborg University
nm@bio.aau.dk

Abstract

The increasing demand for marine monitoring calls for robust automated systems to support researchers in gathering information from marine ecosystems. This includes computer vision based marine organism detection and species classification systems. Current state-of-the-art marine vision systems are based on CNNs, which in nature require a relatively large amount of varied training data. In this paper we present a new publicly available underwater dataset with annotated image sequences of fish, crabs, and starfish captured in brackish water with varying visibility. The dataset is called the Brackish Dataset and it is the first part of a planned long term monitoring of the marine species visiting the strait where the cameras are permanently mounted. To the best of our knowledge, this is the first annotated underwater image dataset captured in temperate brackish waters. In order to obtain a baseline performance for future reference, the YOLOv2 and YOLOv3 CNNs were fine-tuned and tested on the Brackish Dataset.

1. Introduction

More than 70% of the Earth is covered by water and our oceans plays a vital role for humans all around the globe. In order to reduce declination of biodiversity and uphold sustainable fisheries, it is important to keep our oceans healthy. A necessary step towards a better understanding of marine life and ecosystems is to monitor and analyze the impact human activities have on our waters both on a local, regional, and international scale [28].

As underwater cameras and technology in general become more accessible, automatic computer vision based methods are being developed for efficient detection and classification of marine animals and plants, which can be of great aid for marine researchers in analyzing and monitoring our oceans. While underwater images captured in pure water can be handled much like regular images captured above water, other factors must be taken into account when processing images from natural waters. The optical properties of natural water depend on the absorption and scattering of light. While the light scattering in pure water only depends on the temperature and pressure, natural waters show much larger temporal and spatial variations due to the varying content of dissolved and particulate matter. These particles affect both scattering and absorption of light and are often visible as noise in the images [20, 32].

Within scientific communities, it is common practice to evaluate the performance of methods on the same dataset to allow for fair comparison and benchmarking. However, to be able to develop robust algorithms for, e.g., analysis of marine life and ecosystems, the datasets need to represent the natural variations in optical properties seen in natural waters across the world.

To the best of our knowledge, there exists no large scale labeled dataset of temperate coastal or estuarine environments that allows for the development of methods for detecting and classifying marine species in such waters. In particular, none of the publicly available datasets are captured in European marine environments. This is needed to develop robust methods for all water types and marine species, and furthermore, it is needed in order to reach the goal of Good Environmental Status (GES), stipulated by the European Union's Marine Strategy Framework Directive (MSFD) [1, 11].

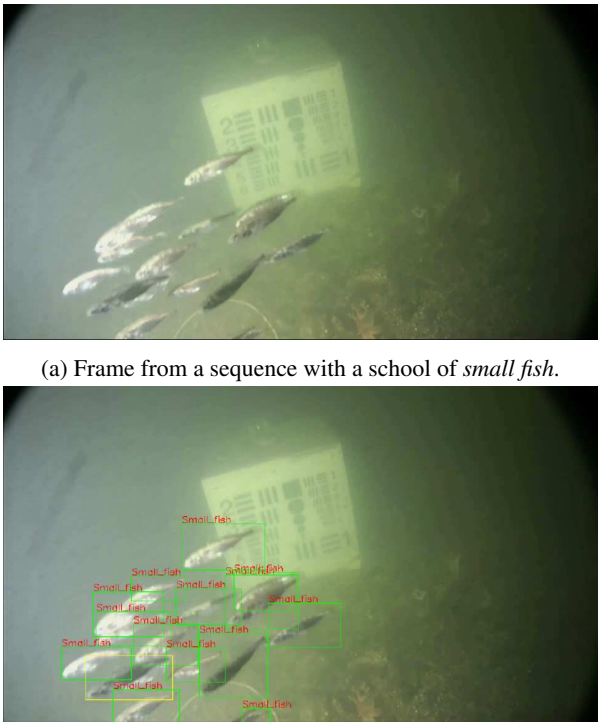
With this paper, we strive to accommodate this need by releasing a publicly available dataset that has been captured over several weeks in temperate brackish water. Hence, it includes natural variation derived from, e.g., time of day, weather conditions, and activities in the water. The dataset is the first stage of a long-term marine monitoring project

where three cameras continuously capture video of the marine life near the bottom of Limfjorden in Denmark, nine meters below the water surface. The aim of this paper is to present this new annotated dataset which, due to its uniqueness, is an important addition to existing annotated marine image datasets. Furthermore, we evaluate two existing state-of-the-art detection methods on the new dataset and present the results as a baseline for future reference.

1.1. Contributions

- A publicly available underwater dataset[1] containing bounding box annotated sequences of images containing *big fish, small fish, starfish, shrimps, jellyfish,* and *crabs* captured in a brackish strait with varying visibility.
- An overview of annotated underwater image datasets.
- A baseline evaluation of state-of-the-art detection methods on the presented dataset.

An example from the proposed dataset can be seen in Figure 1, where a frame from a sequence with a school of *small fish* is presented with and without annotations.



(a) Frame from a sequence with a school of *small fish*.

(b) Same frame as above, but with annotations drawn on the image.

Figure 1: A frame example from the proposed dataset. The same frame is shown with and without annotations.

2. Related Work

As the oceans cover everything from the dark abyssal zone to the sunny coral reefs, the variation between underwater datasets and their associated detection methods can be significant. This section presents an overview of some of the areas where computer vision algorithms have been developed to assist in the huge task of monitoring our oceans and an overview of annotated marine image datasets.

The term *marine vision* will be used in the remainder of this paper as an umbrella term covering the methods and algorithms developed for the purpose of assisting in monitoring marine environments.

2.1. Marine Vision Methods

Detection and classification of fish has been addressed by Villon *et al*. [42] who investigates the performance of a traditional Support Vector Machines (SVM) classifier trained on Histogram of Oriented Gradients (HOG) features for classifying coral reef fish and compares it with the performance of a fine-tuned Convolutional Neural Network (CNN). Their tests show that the CNN outperforms the traditional classification methods. The same conclusion is reached by Salman *et al*. [38] who compares traditional classification methods such as SVM, k-Nearest Neighbours (k-NN), and Sparse Representation Classifier (SRC) with CNN. They achieve an average classification rate of more than 90% on the LifeCLEF14 [22] and LifeCLEF15 [23] fish datasets using CNN and generally a significantly lower rate using the traditional methods. Siddiqui *et al*. [40] reaches state-of-the-art performance on fish species classification using a very deep CNN with a cross-layer pooling approach for enhanced discriminative ability in order to handle the problem of limited labelled training data.

Another interesting marine vision area is scallop detection which has been investigated by Dawkins and Gallager [13]. Using multiple features and a series of cascaded Adaboost classifiers, they developed one of the most prominent scallop detection algorithms. A more recent attempt was presented by Rasmussen *et al*. [34] who tested variations of the YOLOv2 CNN trained for scallop detection. They achieved high accuracy while being able to run in real-time for keeping up with live recordings from an Autonomous Underwater Vehicle (AUV).

Coral reefs are of great interest to marine biologists worldwide but they are difficult and tedious to monitor. In order to assist biologists, Mahmood *et al*. [30] fine-tuned a VGGNet using a subset of the Benthoz-15 dataset [6] and used it for automatically analyzing the coral coverage of three sites in Western Australia. Another approach was investigated by Beijbom *et al*. [5] who achieved state-of-the-

[1] https://www.kaggle.com/aalborguniversity/brackish-dataset

art performance by fusing standard reflectance images with fluorescence images of corals in a 5 channel CNN.

2.2. Annotated Marine Image Datasets

A thing that is common for state-of-the-art detection methods, and deep learning methods in particular, is that they are in need of relatively large amounts of training data. One of the most popular underwater datasets for fish detection and species classification is the F4K dataset [15]. It was recorded from 10 cameras between 2010 and 2013 in Taiwan and it has been used for multiple detection and classification algorithms [39, 38, 18, 25, 9, 42]. The F4K dataset is large and consists of videos and images with complex scenes, various marine species and lots of annotations making it an obvious benchmark dataset. It was also used as part of the LifeCLEF tasks [22, 23, 21].

Another large dataset is the Jamstec E-Library of Deep-sea Images (J-EDI) [17], which consists of videos and images of deep sea organisms captured by Remotely Operated underwater Vehicles (ROV). The images of the J-EDI dataset are annotated on an image level and have been used to train CNNs for detection of deep sea organisms [29, 26]. Two other datasets with focus on fish are the Croatian Fish Dataset [19] which consists of cropped images of 12 different fish species and the QUT Fish Dataset [2] which consists of fish images both in and out of water.

However, as already mentioned, it is not only fish that are of interest within marine vision. Another critical field is monitoring of benthic organisms, such as scallops and corals. The HabCam dataset [41, 10] consists of 2.5 millions annotated images of mainly scallops, but also fish and starfish. The images have been captured along the continental shelf off the east coast of the USA and was used in the work presented by Dawkins and Gallager [13].

The BENTHOZ-2015 dataset [6] is a benthic dataset recorded along the coasts of Australia and used for classifying corals [30]. The Tasmania Coral Point Count [16] was recorded in 2008 during 22 dive missions using an AUV off the South-East coast of Tasmania and has been used for kelp detection [31].

Other annotated coral reef datasets include the Moorea Labeled Corals [4] which is an annotated subset of the Moorea Labeled Corals Long Term Ecological Research project in French Polynesia and the Eilat Fluorescence dataset [5], which experiments with a combination of standard reflectance and fluorescence images in order to improve the coral classification rate of CNNs. Both datasets have been recorded using a custom variation of a photo-quadrat [33].

A collection of datasets has been published for the data challenge of the workshop "Automated Analysis of Marine Video for Environmental Monitoring" in 2018 and 2019 [43]. These datasets include a part of the HabCam dataset [10], as well as four other datasets, MOUSS, AFSC, MBARI, and NWFSC and the images are annotated with either keypoints or bounding boxes.

A table summarizing the underwater datasets can be found in Table 1 along with the proposed dataset, named the Brackish Dataset, which will be described in further details in Section 4.

3. Camera Setup

The setup used to capture the proposed dataset consists of three cameras and three lights. The devices are placed in a grid-wise manner on a stainless steel frame as illustrated in Figure 2. However, the position and orientation of the devices in the figure are not representative for the arrangement used to capture the dataset.

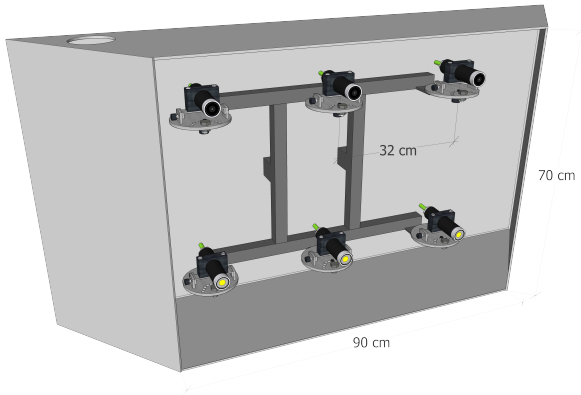


Figure 2: The setup consists of a stainless steel frame with three cameras and three lights.

The cameras use a 1/3" Sony ExView Super HAD Color CCD imaging sensor and a 2.8 mm lens with a resolution up to 1080×1920 pixels and a framerate up to 30 fps with H.264 compression. The lamps are LEDs emitting light with 1900 lumens. Each camera and light is fitted in a cylindrical waterproof casing, which can resist a water pressure of approximately 10 bar. The diameter of the casing is 30 mm and the length is 128 mm.

As both the cameras and lights are placed in the same type of waterproof casing the setup is easily configurable, since all six positions in the steel frame can hold either lights or cameras. The design of the steel frame allows divers to adjust the orientation of lights and cameras under water. This is illustrated in Figure 3, where *knob1* can be pulled in order to adjust the device vertically and *knob2*, hidden beneath the mount, can be pulled to adjust it horizontally. The two bolts on top of the mount can be loosened in order to change or replace a device.

The setup is permanently mounted on a pillar of the Limfjords-bridge connecting Aalborg and Nørresundby in

	Environment	Recording type	Visibility	Sensor	Images	Labeling
F4K - Complex [24]	Reef	Stationary	Varying	RGB	14 videos	Bounding box
F4K - Species [8]	Reef	Stationary	Varying	RGB	27,370	Masks
F4K - Trajectories [7]	Reef	Stationary	Varying	RGB	93 videos	Bounding box
J-EDI [17]	Deep sea	ROV	Clear	RGB	1,500,000	Image level
Croatian Fish Dataset [19]	-	Various	Varying	RGB	794	Bounding box
QUT Fish Dataset [2]	-	Various	Clear	RGB	3,960	Bounding box
HabCam [10]	Shelf sea	Towing	Clear	Stereo	2,500,000	Bounding box
Benthoz-15 [6]	Reef	AUV	Clear	Stereo	9,874	Points
Tasmania Coral Point Count [16]	Reef	AUV	Clear	Stereo	1,258	Points
The Moorea Labeled Corals [4]	Reef	Photoquadrat	Clear	RGB	2,055	Points
Eilat Fluorescence [5]	Reef	Photoquadrat	Clear	RGB	212	Points
MOUSS [43]	Ocean floor	Stationary	Clear	Gray	159	Bounding box
AFSC [43]	Ocean	ROV	Clear	RGB	571	Points
MBARI [43]	Ocean floor	-	Clear	RGB	666	Bounding box
NWFSC [43]	Ocean floor	ROV	Clear	RGB	123	Points
The Brackish Dataset (Proposed)	Brackish strait	Stationary	Varying	RGB	14,518	Bounding box

Table 1: An overview of annotated underwater image datasets.

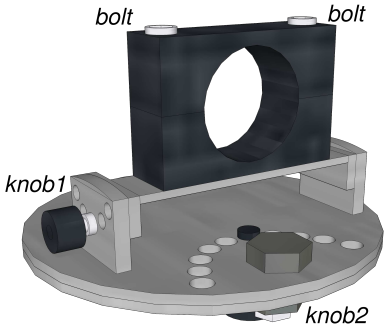


Figure 3: Adjustable mount for cameras and lights.

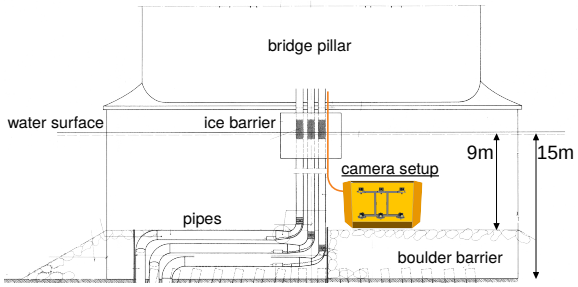


Figure 4: Drawing of the bridge pillar, boulder barrier, and the placement of the camera setup. The drawing is not to scale.

Denmark, at a depth of around nine meters. A barrier of boulders are placed around the pillar for protection, but it also functions as a habitat for various marine species. The setup is therefore placed above this barrier as illustrated in Figure 4, which is not to scale. The barrier is 6 meters high and slopes down towards the fairway between the pillars.

All cameras and lights are connected with cables which provides power and connects the devices to a Digital Video Recorder (DVR) system placed on the bridge. The DVR system is connected to the Internet, allowing for remote real time control and streaming from the cameras.

4. Dataset

Video data from the three cameras have been recorded since February 2019, currently resulting in more than 4,000 hours of video data. As the turbidity of the water and the activity from marine animals vary to a large degree, it is only a fraction of the recordings that is of interest seen from a computer vision perspective. A varied subset of 89 video clips captured between February 20 and March 25 has therefore been handpicked to be the foundation of the proposed dataset. All the chosen clips were captured from a single camera placed in the center position in the bottom of the frame, pointing towards the seafloor. One light, located directly to the left of the camera, seen from the camera's point of view, has been turned on at all times. The light is oriented towards the seafloor, but away from the camera's field of view in order to reduce backscatter.

The videos were categorized based on the main activity of the respective video and subsequently manually annotated with a bounding box annotation tool [3] resulting in a total of 14,518 frames with 25,613 annotations. The distribution of the annotations can be seen in Table 2 where the

number of annotations and amount of videos, where each class occur, are presented. It should be noted that multiple classes can occur in the same video.

Class	Annotations	Video Occurrences
Big fish	3,241	30
Crab	6,538	29
Jellyfish	637	12
Shrimp	548	8
Small fish	9,556	26
Starfish	5,093	30

Table 2: Overview of the share of annotations and amount of video occurrences for the six respective classes.

Professor Niels Madsen, who is a marine biologist with expert knowledge on the local marine environment, has inspected the videos in order to help identify the various types of marine animals. However, due to the turbid recordings, and the relatively similar visual appearance of multiple fish species, it is extremely difficult to determine the exact species. The fish have therefore been coarsely classified as being either *big fish* or *small fish*.

The objects tagged as *big fish* are in most cases lumpsuckers (*Cyclopterus lumpus*) and can be seen in a gray/green or reddish variant depending on whether it is a male or female. However, the sculpin (*Myoxocephalus scorpius*) also visits the site and it can be difficult to tell the difference between the two when the water is turbid.

The *small fish* are in most cases sticklebacks (*Gasterosteus aculeatus*) when schools of fish appear in front of the camera. Other fish that have been observed include gobies (*Pomatoschistus*), European sprat (*Sprattus sprattus*), herrings (*Clupea harengus*), and eelpouts (*Zoarces viviparus*).

Most images contain various sizes of particles in the water, from dissolved matter to floating leaves and seaweed. These objects are not of immediate interest and are considered as noise in the images.

An object has been placed in the scene in front of the camera for other research purposes. For the first couple of weeks a floating device is visible (seen in figure 5a), while the videos from the last weeks contain a concrete block with a visible test pattern of size 40x40cm (seen in figure 5c).

Example frames with all object classes of the Brackish Dataset can be seen in Figure 5. The images also show the variation of turbidity between the videos, e.g., figure 5h shows high turbidity, while 5c has low turbidity.

Limfjorden, where the videos have been recorded, is a 180 km long strait located between The North Sea and Kattegat. At the recording location the strait is approximately 500 m wide, up to 15 m deep, and at a distance of approximately 20 km to Kattegat. The water in the strait is brackish, which is a mixture between saltwater from the seas and freshwater from streams which ends up in the strait. The strength of the currents and winds in the strait can become relatively high and stir up sediments, which increases the turbidity. On other occasions, especially during summer, the water can be calm, resulting in a layered split between the heavy saline sea water and the lighter fresh water on top. The mean water temperature per month measured 1 meter below the surface can vary from 0.5°C to 18 °C.

5. Evaluation

Two state-of-the-art CNN based object detectors (YOLOv2 and YOLOv3 [36, 37]) are tested on the new dataset in order to obtain baseline results for future reference. Both networks have been fine-tuned in the Darknet framework [35]. YOLO is a convolutional neural network based single-shot object detector, which divides each image into regions and predicts bounding boxes and corresponding probabilities for each region. The probability, as a measure of confidence for a detection of a certain class, is used for weighting of each bounding box. As a single-shot object detector, YOLO processes the entire image and predicts all relevant bounding boxes in a single pass through the network, allowing for a high image per second processing rate.

5.1. Training

A brief explanation of the differences between the two pre-trained object detectors are:

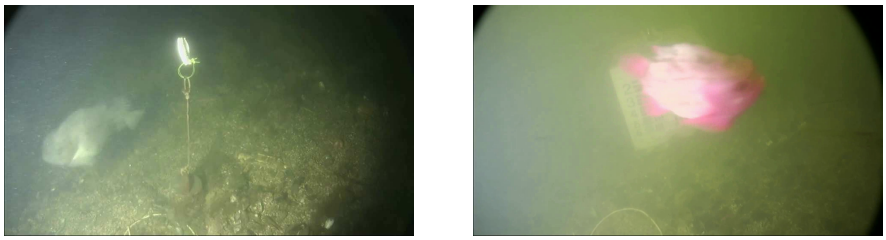
- The YOLOv2 detector is obtained from the VIAME toolkit [12] and is pre-trained on ImageNet and fine-tuned on fish datasets from NOAA Fisheries Strategic Initiative on Automated Image Analysis.
- The YOLOv3 detector is used in its original version pre-trained on the Open Images dataset [37].

The YOLOv2 detector is already fine-tuned on underwater images, but contains only the two classes: *Vertebrates* and *Invertebrates*. Therefore, further fine-tuning is needed in order to be able to evaluate this model on the proposed dataset.

The YOLOv3 detector is trained on the Open Images dataset, which contains 601 classes where five of those are relevant: *fish, starfish, jellyfish, shrimp*, and *crab*.

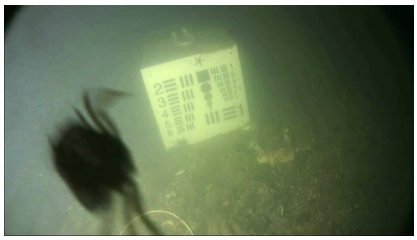
Both models have subsequently been fine-tuned on the proposed dataset, described in section 4, and tested with both Open Images and Brackish dataset categories on the proposed dataset. It should be noted that the only difference between the two categories is that the *fish* class in Open Images contains both the *big fish* and *small fish* from the Brackish categories.

The dataset is split randomly into 80 % training, 10 % validation and 10 % test data. Each network is trained for 30,000 iterations using their original training regime, only adjusting the batch size and setting the input size to 416×416, and with the earliest layer weights frozen.

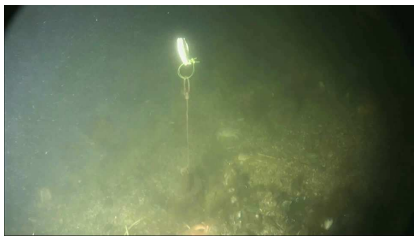


(a) *Big fish*

(b) *Big fish*



(c) *Crab*



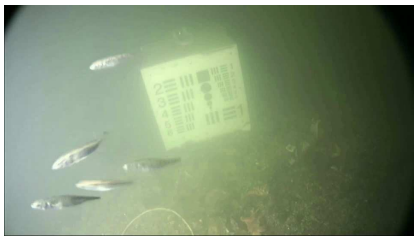
(d) *Crab*

(e) *Jellyfish*

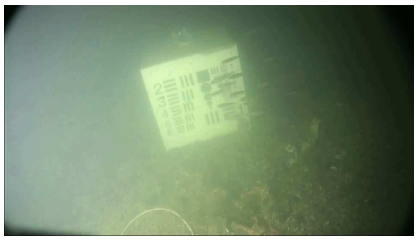
(f) *Jellyfish*

(g) *Shrimp*

(h) *Shrimp*



(i) *Small fish*



(j) *Small fish*

(k) *Starfish*

(l) *Starfish*

Figure 5: Example-frames from the dataset, which illustrate the large in-class variation as well as the variation in turbidity.

5.2. Results

Object detectors are commonly evaluated based on the mean Average Precision (mAP) metric, which is the average precision calculated per category, averaged over all categories. The prediction bounding boxes are filtered by their Intersection over Union (IoU) with the ground truth bounding boxes. The MS COCO dataset [27] is the front running dataset used for object detection, segmentation, and more, which evaluates the mAP under different conditions.

The primary metric is the *AP@[IoU = 0.5:0.95]*, which is referred to as *AP*. *AP* is calculated as the averaged mAP, where the mAP values are calculated with an IoU threshold of [0.5, 0.55, 0.6, ..., 0.90, 0.95]. A metric with the IoU threshold set to 0.5 is also calculated, denoted AP_{50}, which is the primary metric of another large object detection dataset, PASCAL VOC [14]. For all the metrics, only the top 100 most confident predictions per image are included.

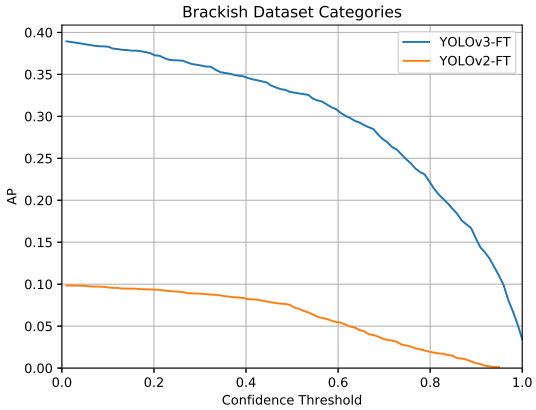


Figure 6: The *AP* metric as a function of the thresholded detection confidence, when using the Brackish dataset categories, for the fine-tuned models.

The *AP* is plotted against a prediction confidence threshold for the proposed dataset in Figure 6. It can be seen that as the threshold is increased, the *AP* decreases. As the decrease is monotonic, it indicates that a large part of the correct predictions are with a low confidence. Therefore, a low confidence threshold of 0.01 is chosen when evaluating the trained networks.

The *AP* and the AP_{50}, which are the primary metrics of MS COCO and PASCAL VOC, are used as performance indicators. The networks have been evaluated on the new Brackish dataset with the Open Images categories, see Table 3, and the Brackish categories, see Table 4.

The results show that fine-tuning on the proposed dataset increases performance dramatically, but also that the achieved performance can still be improved.

Furthermore, a look into the per-class performance of the fine-tuned YOLOv3 network shows that the *starfish* category has a significantly higher score than the others. This is assumed to be due to both the distinctive shape and because the starfish rarely moves in the videos, leading to multiple annotations of starfish which are near identical.

	Categories	*AP*	AP_{50}
YOLOv3	Open Images	0.0022	0.0035
YOLOv2 fine-tuned	Open Images	0.0748	0.2577
YOLOv3 fine-tuned	Open Images	**0.3947**	**0.8458**

Table 3: Results of the models evaluated on the Open Images categories and compared by the *AP* and AP_{50} metrics.

	Categories	*AP*	AP_{50}
YOLOv2 fine-tuned	Brackish	0.0984	0.3110
YOLOv3 fine-tuned	Brackish	**0.3893**	**0.8372**

Table 4: Results of the models evaluated on the Brackish categories and compared by the *AP* and AP_{50} metrics.

The low score of *shrimp* and *small fish* is assumed to be due to their relatively small size and fast movement, which causes motion blur and loss of features.

Class	*AP*	AP_{50}
Big fish	0.4621	0.8999
Crab	0.4205	0.9271
Jellyfish	0.3746	0.8205
Shrimp	0.3238	0.7662
Small fish	0.2449	0.6229
Starfish	0.5102	0.9867

Table 5: Per category results for the fine-tuned YOLOv3 model with the proposed Brackish dataset categories.

6. Future Work

The camera setup used to capture the proposed dataset has been developed in a way that allows for easy maintenance, adjustments, and replacement. It is a permanent setup that will be used to monitor the various species that visit the area during the seasons. At the moment the three cameras are pointing diagonally downwards toward the riverbed, but there will be made ongoing modifications in order to capture different types of dataset, including stereo sequences.

The proposed dataset is part of an ongoing research project where data is logged 24 hours a day from all three cameras. However, the recordings are stored in a compressed format in order to reduce the amount of data. In the future, the plan is to expand the current dataset with uncompressed recordings.

The largest species that has been observed on the recordings is the harbor seal (*Phoca vitulina*), which is a protected animal in national waters and of great interest for marine researchers. The seal is not a part of the proposed dataset due to the few encounters so far, but hopefully will be in the future as more data is gathered and specific species are added to the dataset in close collaboration with local marine biologists.

7. Conclusion

A new bounding box annotated image dataset of marine animals, recorded in brackish waters, is presented in this paper. The dataset consists of 14,518 frames with 25,613 annotations of the six classes: *big fish, small fish, crab, jellyfish, shrimp,* and *starfish*. To the best of knowledge, the proposed dataset is unique, as it is the only annotated image dataset captured in temperate brackish waters.

Two state-of-the-art CNNs (YOLOv2 and YOLOv3) has been fine-tuned on the proposed Brackish Dataset and evaluated in order to create a baseline for future reference. The YOLOv2 object was pre-trained on Imagenet and fine-tuned to fish datasets and it was obtained from the VIAME toolkit [12]. The YOLOv3 detector was the original version pre-trained on the Open Images dataset [37]. The evaluation is based on the primary metrics of the MS COCO and PASCAL VOC, which are both based on the mean Average Precision (mAP). The fine-tuned YOLOv3 network achieved the best performance with $AP \approx 39\%$ and $AP_{50} \approx 84\%$, allowing for improvements to be made.

The proposed Brackish Dataset has been made publicly available at `https://www.kaggle.com/aalborguniversity/brackish-dataset`

References

[1] Andrej Abramic, Daniel Gonzalez, Emanuele Bigagli, Anne Che-Bohnenstengel, and Paul Smits. INSPIRE: Support for and requirement of the marine strategy framework directive. *Marine Policy*, 92:86–100, 2018.

[2] Kaneswaran Anantharajah, ZongYuan Ge, Chris McCool, Simon Denman, Clinton Fookes, Peter Corke, Dian Tjondronegoro, and Sridha Sridharan. Local inter-session variability modelling for object classification. In *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2014.

[3] Chris H. Bahnsen, Andreas Møgelmose, and Thomas B. Moeslund. The aau multimodal annotation toolboxes: Annotating objects in images and videos. *arXiv preprint arXiv:1809.03171*, 2018.

[4] Oscar Beijbom, Peter J. Edmunds, David I. Kline, B. Greg Mitchell, and David Kriegman. Automated annotation of coral reef survey images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012.

[5] Oscar Beijbom, Tali Treibitz, David I. Kline, Gal Eyal, Adi Khen, Benjamin Neal, Yossi Loya, B. Greg Mitchell, and David Kriegman. Improving automated annotation of benthic survey images using wide-band fluorescence. *Scientific Reports*, 6(1), 2016.

[6] Michael Bewley, Ariell Friedman, Renata Ferrari, Nicole Hill, Renae Hovey, Neville Barrett, Oscar Pizarro, Will Figueira, Lisa Meyer, Russ Babcock, Lynda Bellchambers, Maria Byrne, and Stefan B. Williams. Australian sea-floor survey data, with images and expert annotations. *Scientific Data*, 2:150057, 2015.

[7] Cigdem Beyan and Robert B. Fisher. Detecting abnormal fish trajectories using clustered and labeled data. In *2013 IEEE International Conference on Image Processing*. IEEE, 2013.

[8] Bastiaan J. Boom, Phoenix X. Huang, Jiyin He, and Robert B. Fisher. Supporting ground-truth annotation of image datasets using clustering. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012.

[9] Zheng Cao, Jose C. Principe, Bing Ouyang, Fraser Dalgleish, and Anni Vuorenkoski. Marine animal classification using combined CNN and hand-designed image features. In *OCEANS 2015 - MTS/IEEE Washington*. IEEE, 2015.

[10] Northeast Fisheries Science Center. Habitat mapping camera (HABCAM), 2017. `https://inport.nmfs.noaa.gov/inport/item/27598`.

[11] Alessandro Crise, Maurizio Ribera d'Alcalà, Patrizio Mariani, George Petihakis, Julie Robidart, Daniele Iudicone, Ralf Bachmayer, and Francesca Malfatti. A conceptual framework for developing the next generation of marine OBservatories (MOBs) for science and society. *Frontiers in Marine Science*, 5, 2018.

[12] Matthew Dawkins, Linus Sherrill, Keith Fieldhouse, Anthony Hoogs, Benjamin Richards, David Zhang, Lakshman Prasad, Kresimir Williams, Nathan Lauffenburger, and Gaoang Wang. An open-source platform for underwater image and video analytics. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017.

[13] Matthew Dawkins, Charles Stewart, Scott Gallager, and Amber York. Automatic scallop detection in benthic environments. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE, 2013.

[14] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[15] Robert B. Fisher, Yun-Heh Chen-Burger, Daniela Giordano, Lynda Hardman, and Fang-Pang Lin. *Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data*. Springer Publishing Company, Incorporated, 1st edition, 2016.

[16] Australian Centre for Field Robotics. Tasmania coral point count. `http://marine.acfr.usyd.edu.au/datasets/`.

[17] Japan Agency for Marine-Earth Science and Technology. JAMSTEC E-library of deep-sea images, 2016. `http://www.godac.jamstec.go.jp/jedi/e/`.

[18] Snigdhaa Hasija, Manas Jyoti Buragohain, and S. Indu. Fish species classification using graph embedding discriminant

analysis. In *2017 International Conference on Machine Vision and Information Technology (CMVIT)*. IEEE, 2017.

[19] Jonas Jäger, Marcel Simon, Joachim Denzler, Viviane Wolff, Klaus Fricke-Neuderth, and Claudia Kruschel. Croatian fish dataset: Fine-grained classification of fish species in their natural habitat. In *Procedings of the Machine Vision of Animals and their Behaviour Workshop 2015*. British Machine Vision Association, 2015.

[20] Nils G. Jerlov. *Marine Optics*, volume 14. Elsevier, 1976.

[21] Alexis Joly, Hervé Goëau, Hervé Glotin, Concetto Spampinato, Pierre Bonnet, Willem-Pier Vellinga, Julien Champ, Robert Planqué, Simone Palazzo, and Henning Müller. Lifeclef 2016: Multimedia life species identification challenges. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 286–310, Cham, 2016. Springer International Publishing.

[22] Alexis Joly, Hervé Goëau, Hervé Glotin, Concetto Spampinato, Pierre Bonnet, Willem-Pier Vellinga, Robert Planqué, Andreas Rauber, Robert Fisher, and Henning Müller. Lifeclef 2014: Multimedia life species identification challenges. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, pages 229–249, Cham, 2014. Springer International Publishing.

[23] Alexis Joly, Hervé Goëau, Hervé Glotin, Concetto Spampinato, Pierre Bonnet, Willem-Pier Vellinga, Robert Planqué, Andreas Rauber, Simone Palazzo, Bob Fisher, and Henning Müller. Lifeclef 2015: Multimedia life species identification challenges. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 462–483, Cham, 2015. Springer International Publishing.

[24] Isaak Kavasidis, Simone Palazzo, Roberto Di Salvo, Daniela Giordano, and Concetto Spampinato. An innovative web-based collaborative platform for video annotation. *Multimedia Tools and Applications*, 70(1):413–432, 2013.

[25] Xiu Li, Min Shang, Jing Hao, and Zhixiong Yang. Accelerating fish detection and recognition by sharing CNNs with objectness learning. In *OCEANS 2016 - Shanghai*. IEEE, 2016.

[26] Yujie Li, Huimin Lu, Jianru Li, Xin Li, Yun Li, and Seiichi Serikawa. Underwater image de-scattering and classification by deep neural network. *Computers & Electrical Engineering*, 54:68–77, 2016.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[28] Eric Lindstrom, John Gunn, Albert Fischer, Andrea McCurdy, Linda K. Glover, and Task Team Members. A framework for ocean observing. Technical report, 2012.

[29] Huimin Lu, Yujie Li, Tomoki Uemura, Zongyuan Ge, Xing Xu, Li He, Seiichi Serikawa, and Hyoungseop Kim. FDCNet: filtering deep convolutional network for marine organism classification. *Multimedia Tools and Applications*, 77(17):21847–21860, 2017.

[30] Ammar Mahmood, Mohammed Bennamoun, Senjian An, Ferdous A. Sohel, Farid Boussaid, Renae Hovey, Gary A. Kendrick, and Robert B. Fisher. Automatic annotation of coral reefs using deep learning. In *OCEANS 2016 MTS/IEEE Monterey*. IEEE, 2016.

[31] Navid Nourani-Vatani Alon Friedman Oscar Pizarro Stefan B. Williams Michael Bewley, Bertrand Douillard. Automated species detection: An experimental approach to kelp detection from sea-floor auv images. In *Proceedings of Australasian Conference on Robotics and Automation*, 2012.

[32] Curtis D. Mobley. *Light and water: radiative transfer in natural waters*. Academic press, 1994.

[33] Linda B. Preskitt, Peter S. Vroom, and Celia Marie Smith. A rapid ecological assessment (REA) quantitative survey method for benthic algae using photoquadrats with scuba. *Pacific Science*, 58(2):201–209, 2004.

[34] Christopher Rasmussen, Jiayi Zhao, Danielle Ferraro, and Arthur Trembanis. Deep census: AUV-based scallop population monitoring. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. IEEE, 2017.

[35] Joseph Redmon. Darknet: Open source neural networks in c. `http://pjreddie.com/darknet/`, 2013–2016.

[36] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[37] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

[38] Ahmad Salman, Ahsan Jalal, Faisal Shafait, Ajmal Mian, Mark Shortis, James Seager, and Euan Harvey. Fish species classification in unconstrained underwater environments based on deep learning. *Limnology and Oceanography: Methods*, 14(9):570–585, 2016.

[39] Ahmad Salman, Shoaib Ahmad Siddiqui, Faisal Shafait, Ajmal Mian, Mark R Shortis, Khawar Khurshid, Adrian Ulges, and Ulrich Schwanecke. Automatic fish detection in underwater videos by a deep neural network-based hybrid motion learning system. *ICES Journal of Marine Science*, 2019.

[40] Shoaib Ahmed Siddiqui, Ahmad Salman, Muhammad Imran Malik, Faisal Shafait, Ajmal Mian, Mark R Shortis, and Euan S Harvey and. Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data. *ICES Journal of Marine Science*, 75(1):374–389, 2017.

[41] Richard Taylor, Norman Vine, Amber York, Steve Lerner, Dvora Hart, Jonathan Howland, Lakshman Prasad, Larry Mayer, and Scott Gallager. Evolution of a benthic imaging system from a towed camera to an automated habitat characterization system. In *OCEANS 2008*. IEEE, 2008.

[42] Sébastien Villon, Marc Chaumont, Gérard Subsol, Sébastien Villéger, Thomas Claverie, and David Mouillot. Coral reef fish detection and recognition in underwater videos by supervised machine learning: Comparison between deep learning and hog+svm methods. In *Advanced Concepts for Intelligent Vision Systems*, pages 160–171, Cham, 2016. Springer International Publishing.

[43] CVPR 2018 Workshop and Challenge: Automated Analysis of Marine Video for Environmental Monitoring. Data challenge description, 2018. `http://www.viametoolkit.org/cvpr-2018-workshop-data-challenge/challenge-data-description/`.