Marine Animal Detection and Recognition with Advanced Deep Learning Models

Peiqin Zhuang, Linjie Xing, Yanlin Liu, Sheng Guo, Yu Qiao

Shenzhen key Lab. Of CVPR, Shenzhen Institutes of Advanced Technology
Chinese Academy of Sciences, P.R. China
{pq.zhuang, lj.xing, yl.liu, sheng.guo, yu.qiao}@siat.ac.cn

Abstract. This paper summarizes SIATMMLAB's contributions in SEACLEF-2017 task [1]. We took part in three subtasks with advanced deep learning models. In Automatic Fish Identification and Species Recognition task, we exploited different frameworks to detect the proposal boxes of foreground fish, then fine-tuned a pre-trained neural network to classify the fish. In Automatic Frame-level Salmon Identification task, we utilized the BN-Inception [2] network to identify whether a video frame contains salmons or not. In Marine Animal Recognition task, we examined different neural networks to make classification based on weakly-labelled images. Our methods achieve good results in both task1 and task3.

Keywords: Deep Learning, Fish Detection, Weakly-labelled, Image Classification.

1 Introduction

Driven by the increasing demand of ecological surveillance and biodiversity monitoring under the water, more sea-related multimedia data were collected with the aid of advanced imaging systems. However, with the exponential growth of sea-related visual data, it is prohibitive to count on the manual handling by the experts to annotate these datasets. Therefore, automatic analyzing the contents of underwater image is key to make use of the exponentially increasing underwater data. SEACLEF2017 [1] have launched four tasks to explore suitable methods for handling these multimedia data.

In this paper, we elaborate our methods applied in SEACLEF2017, and the remainder of this paper is organized as follows. In Section 2, we present our approach for task 1, including the method of proposing foreground boxes, reducing background boxes and classifying the boxes we got. In Section 3, we report our approach for frame-level salmon identification in task 2. In Section 4, we describe the procedure for handling weakly-labeled fish data in task 3. Finally, we provide a comprehensive analysis of our works and discuss the directions for further research.

2 Species Recognition on Coral Reef Videos

In this task, we automatically identify and recognize fish species by giving a bounding box and a corresponding label. To reach this goal, we employed a traditional pipeline [3,4,5] for detecting fish. Firstly, we used different detection architectures based on deep neural networks to generate potential bounding boxes, and then fine-tuned a pre-trained neural network with training data to classify bounding boxes. Section 2.1 will present the two methods we used to detect bounding boxes, while Section 2.2 describes the procedure of classification. We show our final result in Section 2.3. At the end, we make some discussions on how to improve detection performance in the future.

2.1 Foreground Detection

Inspired by the past participants of SEACLEF [8,9], we separate species recognition process into detection step and classification step. In the past three years, end-to-end [3,4,5] detection methods have prevailed in most of detection task because of their dramatic performance, so we selected two latest models as our detector.

Firstly, we chose SSD [5] to differentiate the regions between foreground fish and background. With the advance of producing predictions from different features maps of different scales [5,6], we got acceptable results by using SSD to generate potential bounding boxes. Besides, we also used another detection architecture PVANET [7] to detect foreground fish. At detection step, we extracted all true positive frames with annotations and then chose one tenth of positive frames at each video to formulate validation dataset. The rest of positive frames were used for training. A detection result example is shown in Fig. 1.

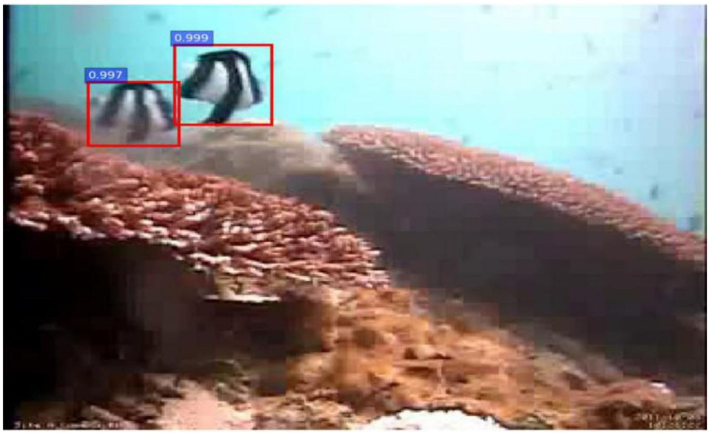

Fig. 1. A detection result example, fish were enclosed by red bounding boxes with their corresponding scores, which shows the confidence of being detected as foreground.

Afterwards to remove false positives, we adopted Sungbin Chois's [8] method to compute a background image (Fig. 2) by selecting the median value of at each pixel position for every video, then used background subtraction and erosion to create a mask (Fig. 3) for each frame. If the area of background in bounding box is greater than a threshold we set, we consider it as background, and consequently discard it. Compared to other background subtraction methods [8,9], we adopted it as a post-processing means after detection, not directly for detecting bounding boxes.

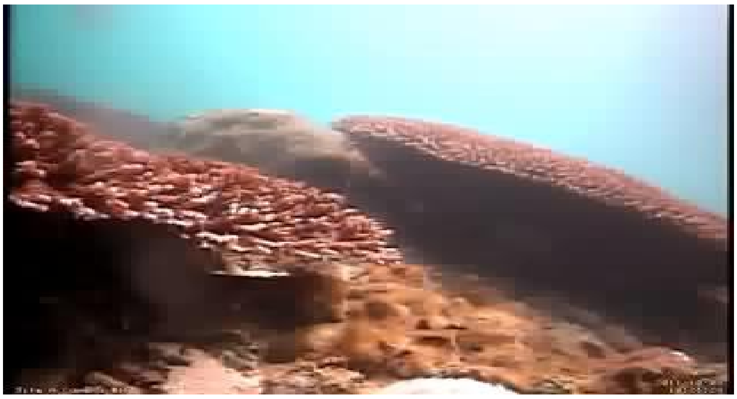

Fig. 2. An example background computed by selecting the median value at each pixel position

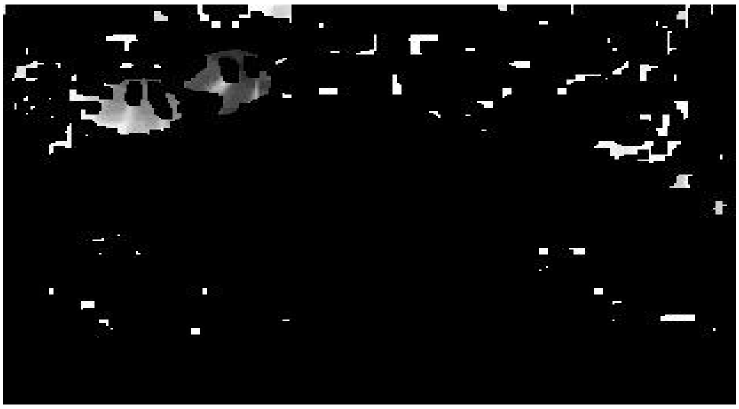

Fig. 3. An example mask computed by background subtraction and erosion indicates the region of background

2.2 Species Classification

In the classification step, we extracted true positives from the training dataset with the annotation information. Besides, we also added some background as a new class to further reduce the amounts of false positives in our final results. Given species imbalance in each video, we extracted all videos together and then separated those patches into training set and validation set according to the labels. The ratio between training dataset and validation dataset is the same as in the detection step.

For the sake of attaining a high performance of classification, we chose the ResNet-10 [10] as our classifier. According to the relevant information of this task, we set the number of output classes as 16, and used the training dataset to fine-tune this ResNet-10 network.

2.3 Results

The test dataset contains 73 videos, similar scenario to the training dataset. By employing the procedure of detection and classification, we finally submitted two results. SIATMMLAB_run1 used the SSD framework to detect foreground fish, while SIATMMLAB_run2 utilized the PVANET to generate potential bounding boxes as well. Both classification of two above results used ResNet-10 network as classifier. Our normalized scores of two results are respectively 0.66 and 0.71 (Table 1), nearly equal to the best result [8] in 2015.

Table 1. Our final results on task1 measured by counting score, precision and normalized score

	Counting score	Precision	Normalized score
SIATMMLAB_run1	0.87	0.76	0.66
SIATMMLAB_run2	0.88	0.80	0.71

2.4 Discussion

Although our approaches achieve a good performance on task 1, there still exist some aspects need to be further improved. By analyzing the results, we found that several fish were not detected in some particular video clips , which influences the counting score and precision. Besides, our detector often mistook some background regions with abundant texture for foreground fish. Since the video naturally has context information between frames, it will help us to fix the discontinuity of video sequence in detection results and achieve a higher performance in the future.

In task 1, we combined both detection and classification methods to recognize fish species on coral reef videos, and achieved an exciting results in the end. In the future, we will try more fusions of detection and classification, use some pre-processing methods to augment video resolution, reduce noise influence caused by the illumination changes, and utilize more video context information. We believe that effective incorporation of these methods will provide an opportunity for ecologist to monitor biodiversity more accurately than manual handling in the future.

3 Frame-level Salmon Identification

In this task, we need identify the appearance of salmon in each frame [1]. Since the ratio of frame which salmon appeared is rare and the salmons are often small, this task appears to be challenging for us. In Section 3.1, we will analyze some relevant information about the training data; Section 3.2 presents the method we used and our final result. Finally, we will make a discussion about task 2 in Section 3.3.

3.1 Data Analysis

The training data contains 8 videos about salmon. By analyzing the frames extracted from all videos, it is notable that the amounts of positives are much less than the counterpart of negatives. There are nearly 1300 positives, while the negatives numbers are almost 59k, which creates a large data imbalance between positive samples and negative samples. Besides, most of the frames are only filled with black background, which means information about salmon in the fame is rare (Fig. 4).. Some frames even contain a green static water turbine in it.

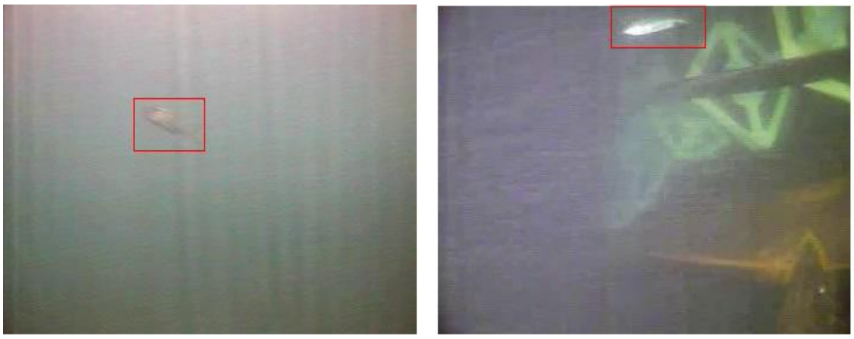

Fig. 4. Two different images show the appearance of salmon in the frame with red bounding box

3.2 Experiments and Results

In order to solve the imbalance between positives and negatives, we used all positives and randomly selected 1500 negatives to formulate our training dataset and validation dataset. We chose nine tenths of those as training set , while the rest were used as validation set.

We regarded frame-level salmon identification as a binary classification problem, so we use the BN-Inception [2] network as our classifier. Through fine-tuning our network with training dataset, we achieved 97% accuracy in our validation dataset. We also tested the model at the whole dataset, which showed only 2723 images were incorrectly classified within 59956 images.

The test dataset contains 8 videos. Compared to the training dataset, the test dataset seems to be a little different from training dataset. The green water turbine in most of frames is always revolving, which causes illumination changes and increases much noise. In this case, our model often mistook many frames with illumination changes for positives and created many false positives. We submitted one run as our final result. The precision for our result is 0.04 (Table 2), which means most of our positives are incorrect. Our model performance suffers a huge decline in test dataset.

Table 2. Our results on task2 measured by precision, recall and F-measure metric

	Precision	Recall	F-measure
SIATMMLAB_run1	0.04	0.82	0.07

3.3 Discussion

We adopted BN-Inception network to identify whether a video frame contains the salmons or not. Although our model successfully reached the goal of this task on the training dataset, the accuracy of test dataset performed poorly with challenging constraints. Firstly, since the salmon is so small and most of frames are filled with black background, it is challenging for us to choose frames, which essentially represent the properties of positives and negatives. Next, our model finally mistook many negatives for positives because of the illumination changes brought by the water turbine revolving.

Given these constraints, more work will be carried out to help improve models performance. Strengthening feature representation and introducing data-mining tricks may further solve the problem of frames selection. Besides, more pre-processing technologies will be employed to improve the quality of frames and reduce influence of illumination changes. According to the two above aspects, we will incorporate more methods to increase our model generalization and accuracy in the future.

4 Marine Animal Species Recognition

Task 3 aims to classify marine animals from weakly-labelled images collected by keyword queries on the internet [1]. With the difficulty of high similarity between two species and weakly-labelled annotations, making exact recognition seems to be challenging for us. Section 4.1 will analyze the training dataset and its existing difficulties. Section 4.2 shows the details of our experiments and final results. Finally, we will make a slight discussion about our work in Section 4.3.

4.1 Data Analysis

The training dataset includes nearly 13k images, but some of species have little numbers of images compared to other species. During the period of our experiment, we also found some other flaws in the training dataset: some maps describing the distribution region of corresponding species were added to the training dataset and sometimes two identical images both appeared in different species, which often caused classification mistakes. In this case, we made a slight examination of training dataset to pick out the bad data. Besides, we also added nearly 2k images extracted from YouTube videos to increase the generalization of our model. We split all the data into training data and validation data and selected 10 images per species as validation set.

4.2 Experiments and Results

Driven by the high performance of deep neural networks [2,10,11,12], we decided to use two latest architectures of neural network: BN-Inception [2] and ResNet-50 [10]. We did many experiments on training data, and evaluated our performance on the validation set with the top-1 and top-5 metric. Some detail information are shown in Table 3.

Table 3. Our experiment description on task 3 by using different neural networks,changing input size and crop size

	Model	Input size	Crop size	Top-1 Acc.
Experiment1	BN-Inception	360×260	224	0.840
Experiment2	BN-Inception	600×400	224	0.818
Experiment3	BN-Inception	600×400	336	0.845
Experiment4	ResNet-50	360×260	224	0.837
Experiment5	ResNet-50	600×400	224	0.800

We did such 5 experiments on the training dataset by changing the deep neural network model [2,10], resizing the input image size, using different crop size [13,14] to augment the training dataset. All our experiments achieved high performance in the validation data in spite of the challenge this task brings. The accuracies measured by top-1 metric were all higher than 0.8, and the counterparts measured by Top-5 metric were also higher than 0.93.

We finally submitted three runs. According to the source of test dataset, we separated test dataset into two parts. Those images cropped from videos, which have similar scenario with task 1's data, were tested by a new RestNet-50 network trained with task 1's video data. The rest of test data were tested by the ablation of different models [15] we have described before. SIATMMLAB_run1 used the ablation of experiment 1 and experiment 4. The ablation of experiment 1, experiment 3 and experiment 4 provided the result of SIATMMLAB_run2. SIATMMLAB_run3 includes the result of experiment 1 and experiment 3. The metric used to measure final runs is average precision, the result of our runs are shown in Table 4.

Table 4. Our final result measured by the metric of average precision

	Pr@1	Pr@2	Pr@3
SIATMMLAB_run1	0.61	0.71	0.74
SIATMMLAB_run2	0.61	0.71	0.75
SIATMMLAB_run3	0.61	0.72	0.76

4.3 Discussion

Although our model achieved more than 0.8 accuracy with top-1 metric on the validation dataset, it seems its performance slightly decreased in the test dataset. Recalling the whole procedure of our experiment, we have not used the relevance ranking information when we trained our model. Since the images are weakly-labelled and the metric used to measure our results is average precision, using the relevance information may help us to recognize the species more accurately in the future. Furthermore, some species have high degree of similarity to each other, even our human beings can not classify them easily without professional knowledge. How to find an effective method to essentially represent these species is a key problem we want to solve in our future work. Besides, some fish species were so small compared to the whole image. In this case, we will employ salience map to primarily locate the positions of fish, preparing for the recognition step.

Automatically recognizing fish species is essential for handling sea-related multimedia data.We hope to further improve classifier with more advanced models ablation and auxiliary methods so as to recognize fish species more accurately when the ecologist use these methods to monitor biodiversity.

5 Conclusions and Perspectives

This paper has described our participation in SEACLEF2017. All our approaches are based on the architectures of deep neural network, and achieved high performance both in task 1 and task 3. Although the methods using with deep neural network have good effect in this competition, there still exist some constrains, like low-resolution, illumination changes and complicated background, when we handle these underwater multimedia data. Based on these constraints, we will respectively investigate more effective methods to solve these aspects and try to raise our performance to a new level. More related work will be carried out to improve our model performance and help to promote these methods to be used in the real-world application.

References

1. Alexis Joly, Hervé Goëau, Hervé Glotin, Concetto Spampinato, Pierre Bonnet, Willem-Pier Vellinga, Jean-Christophe Lombardo, Robert Planqué, Simone Palazzo, Henning Müller. LifeCLEF 2017 Lab Overview: multimedia species identification challenges, *Proceedings of CLEF 2017*, 2017.

2. Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network by reducing internal covariate shift, In *ICML*, pages 448-456, 2015. 3.

3. Ross Girshick. Fast-rcnn, In *ICCV*, 2015. 2,9.

4. Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks, In *NIPS*, 2015.1,2,6,7,8,9.

5. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg. SSD: single shot multibox detector, In *ECCV*, 2016. 1,3,4,6,7,8.

6. Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie. Feature Pyramid Networks for Object Detection, arXiv preprint arXiv: 1612.03144, 2016.

7. Kye-Hyeon Kim, Sanghoon Hong, Byungseok Roh, Yeongjae Cheon, Minjie Park, PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection, arXiv preprint arXiv: 1611.08588, 2016.

8. Sungbin Choi, Fish identification in underwater video with deep convolutional neural nerwork : Snumedinfo at lifeclef fish task 2015. In Working Notes of the 6th International Conference of the CLEF Initiative. *CEUR Workshop Proceedings*, 2015. Vol-1391, urn:nbn:de:0074-1391-8.

9. Jonas Jäger, Erik Rodner, Joachim Denzler, Viviane Wolff,Klaus Fricke-Neuderth. SeaCLEF 2016: Object Proposal Classification for Fish Detection in Underwater Videos. In Working Notes of the 7th International Conference of the CLEF Initiative. *CEUR Workshop Proceedings*, 2016. Vol-1609, urn:nbn:de:0074-1609-5.

10. Kaiming He, Xiaoyu Zhang, Shaoqing Ren, Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770-778, 2016. 2,3.

11. Krizhevsky Alex, Sutskever Ilya, Hinton Geoffrey. Image classification with deep convolutional neural networks. In *NIPS* ,2012.

12. Karen Simonyan, Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *NIPS*, 2015.

13. Wei Liu, Andrew Rabinovich, Alexander C. Berg. ParseNet: Looking wider to see better. In *ILCR*, 2016. 2.

14. Limin Wang, Sheng Guo, Weilin Huang, Yuanjun Xiong, Yu Qiao. Knowledge Guided Disambiguation for Large-Scale Scene Classification with Multi-Resolution CNNS. *IEEE Transactions on Image Processing* 26(4), 2055-2068.

15. Sheng Guo, Weilin Huang, Limin Wang, Yu Qiao. Locally-Supervised Deep Hybrid Model for Scene Recognition. *IEEE Transactions on Image Processing* 26(2), 808-820.