SeaCLEF 2016: Object proposal classification for fish detection in underwater videos

Jonas Jäger[1,2], Erik Rodner[2], Joachim Denzler[2], Viviane Wolff[1], and Klaus Fricke-Neuderth[1]

[1] Department of Electrical Engineering and Information Technology,
Fulda University of Applied Sciences, Germany
[2] Computer Vision Group, Friedrich Schiller University Jena, Germany
{Jonas.Jaeger,Viviane.Wolff,Klaus.Fricke-Neuderth}@et.hs-fulda.de
{Erik.Rodner,Joachim.Denzler}@uni-jena.de

Abstract. This working note describes the results of CVG Jena Fulda team for the fish recognition task in SeaCLEF 2016. Our method is based on convolutional neural networks applied to object proposals for detection as well as species classification. We are using background subtraction proposals that are filtered by a binary SVM classifier for fish detection and a multiclass SVM for species classification. Both SVM's utilize CNN features extracted from AlexNet. With this pipeline we achieve a recognition precision of 66% and a normalized counting score of 58% on the provided test dataset. We also show that classification of background subtraction proposals works much better for fish detection than background subtraction on its own.

Keywords: Object proposals, R-CNN, CNN features, Fine-grained classification, Fish detection

1 Introduction

This paper presents the participation of the CVG Jena Fulda team in the *SeaCLEF 2016 Task 1*. This task deals with automatic fish recognition of coral reef species in low resolution videos. All fishes are presented in their natural unrestricted habitat. See Fig. 1 for example frames.

This task is important to enhance computer vision methods for biodiversity applications. Many scientists in the field of ecology collect large amounts of video data to monitor biodiversity in their specific applications. But manual analysis of this data is time consuming and requires knowledge of rear human experts, which makes it impossible to evaluate data in a large scale. However, this large scale analysis is essential to obtain the knowledge to save ecosystems that have a large impact on the human population. Therefore, tools for automatic video analysis need to be developed to support the work of ecologists.

We have a special interest in this task because our team works on a closely related problem [7]. In our application we deal with high resolution underwater video analysis of fish species at the Adriatic sea in Croatia.

We noticed that detection is a crucial part in a fish classification and counting system. But we also experienced that fish detection is a difficult problem, due to lighting changes and the complex background in a natural environment. Therefore, we focus in the following paper on robust fish detection.

Last years participants [2, 3] in this task used median image background subtraction for fish detection. Boom et al. [1] also utilized background subtraction methods and post processed detection results with an objectness filter to remove bad detections. In contrast to that, we classify fish proposals by CNN features [4].

In this work we propose the use of object proposal classification for fish detection. Object proposals are obtained by background subtraction and then classified into *fish* and *background* by a binary support vector machine (SVM). For fish recognition we utilize a multiclass SVM trained for the 15 considered species. Both SVM's are using the same CNN features, extracted from AlexNet [8], for prediction.

Our detection approach is very similar to the idea of region-based convolutional neural networks (R-CNN) presented by Girshick et al. [6]. In contrast to [6] we are using the background subtraction method of Stauffer and Grimson [12] instead of selective search [14] for proposal generation, since we can exploit time information in the video data. Another difference is that we do not apply domain specific fine tuning to the CNN.

2 Fish Dataset

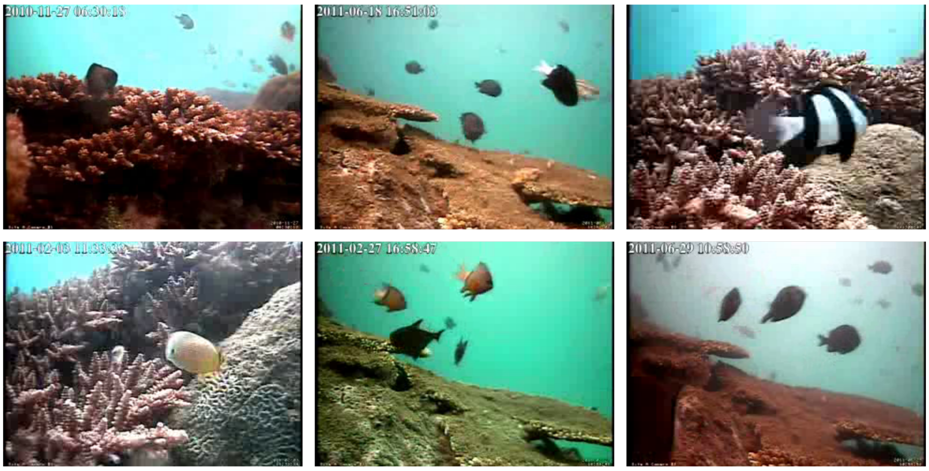


Fig. 1: Example frames from six different videos of the SeaCLEF 2016 dataset.

The provided dataset: The dataset contains videos and images of fish species in their natural coral reef environment. It is divided into a training and a test set. Example frames from six different videos are shown in Fig. 1.

The provided training set consists of 20 low resolution videos and more than 20000 sample images of 15 fish species. There are 5 videos with a resolution of 640 × 480 pixels and 15 videos with 320 × 240 pixels. All videos are annotated by two human experts with bounding boxes and species names.

The test set contains 73 videos with a resolution of 320 × 240 pixels.

Dataset preparation: We split the given training videos into two parts with 10 videos each. One part will be used as *validation set*. The other 10 videos and all sample images will be used for training and will be called *training set* in the rest of this paper.

3 Method

3.1 Overview

Our main idea is to build a fish detector and to use detections for species classification. Since the application of background subtraction methods on its own leads to a large number of false detections, we use background subtraction to get fish proposals and classify each proposal as as *fish* or *background*. Then, all fish detections are classified as one of 15 species or rejected. Both classifieres, for detection and species recognition, are using the same features.

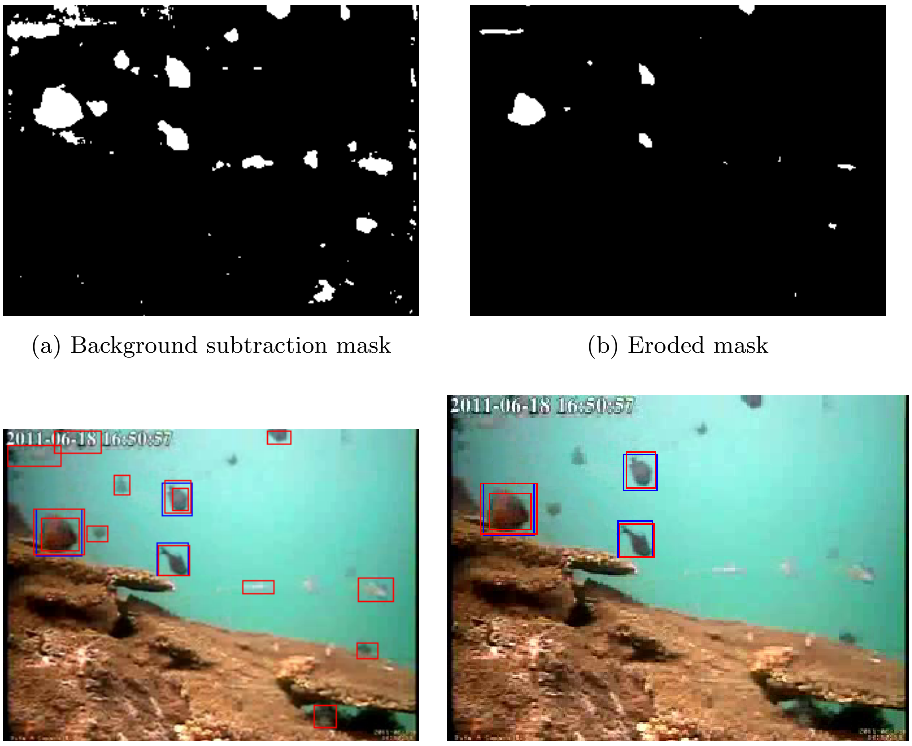
3.2 Object proposal classification for fish detection

Our fish detection approach consists of three steps. (1.) Generation of bounding box proposals. (2.) Extraction of CNN features for each proposal. (3.) Classification of each bounding box proposal as *fish* or *background*. Please see Fig. 2 for illustration of these steps.

In step (1.) we use the background subtraction algorithm of Stauffer and Grimson [12], which uses a probabilistic background model that represents each pixel as a mixture of Gaussians. The result of this algorithm is a binary mask that indicates which pixels are background (see Fig. 2a). This mask is further used to obtain a second background mask (see Fig. 2b) by applying an erosion filter to it, which allows us to separate nearby fishes.

After that we apply the blob detection method of Suzuki and Abe [13] to both masks to get bounding box proposals (see Fig. 2c). Bounding boxes that have a smaller area than 100 pixels are removed, since these proposals are to small for species classification.

(2.) Now we use the generated proposals to extract CNN features from *AlexNet* [8], that was pretrained on *ILSVRC 2012* [11]. As features we choose the activations of the 7th hidden layer (*relu7*) in the convolutional network. Note that we did not fine tune the convolutional net by training it with fish images.



(a) Background subtraction mask

(b) Eroded mask

(c) Object proposals

(d) Boxes that where classified as fish

Fig. 2: Result images of different stages in our fish detection pipeline. Red boxes are predicted and blue boxes are ground truth. This figure is best viewed in color.

(3.) Based on these features we utilize a binary SVM for classifying each bounding box proposal as *fish* or *background* (see Fig. 2d). Then we choose from all fish detections the boxes with a confidence level that is higher or equal to 0.5. In oder to obtain a probability measure from SVM scores we used Platt scaling [10] as implemented in scikit-learn [9].

As a post processing step we apply non-maximum suppression to remove duplicate boxes for all fishes.

Detector training: To train our detector we extract CNN features (see step (2.)) and fit the SVM classifier to the classes *background* and *fish*. As training data we utilize the fish sample images of the training set (see section 2) and extract all annotated fishes from the 10 training videos. As background examples we generate object proposals from training set videos and extract those boxes that have no intersection with a ground truth fish box.

3.3 Species classification using CNN features

As in our previous work [7] we use CNN features [4] and a multiclass SVM for species prediction. We utilize the same CNN features that where extracted in detection step (2.) from AlexNet [8], which was pretrained on ImageNet. As features we choose the activations of the 7th hidden layer (relu7) in the network.

When the confidence level for a classification is lower than 0.5 we consider it as an *unknown fish* and reject it. In order to get probabilities from SVM scores we use the method of Wu et al. [15].

The SVM is trained with a one-vs-rest strategy for the 15 considered fish species. The training data are composed of the provided species sample images and all annotated fishes cropped from training set (see section 2) videos.

4 Results

4.1 Fish detection results

One of our main interest is how well object proposal classification (OPC) works for fish detection compared to background subtraction on its own. For that purpose we will first describe the methods listed in results Tab. 1 and then define our Pascal VOC [5] like evaluation process. Finally we will discuss our fish detection results presented in Fig. 3 and Tab. 1.

Table 1: Average precision for OPC (object proposal classification) detection and background substraction on our validation dataset. All values are in percent.

Method	**average precision**
BgsMedian	0.34
BgsGMM	4.35
OPC (BgsMedian)	18.28
OPC (BgsGMM)	**40.66**

Methods: The first method, called *BgsMedian* in Tab. 1, computes a median background image for all frames in a video and subtracts the current frame from that background image. A specific pixel in the median image is calculated by using the median value of all pixels at same position in the video. This method was also used by last year participants [3, 2].

The second method, referenced as *BgsGMM*, was developed by Stauffer and Grimson [12] and uses a probabilistic background model that represents each pixel as a mixture of Gaussians.

To obtain bounding boxes from these background subtraction methods we applied blob detection proposed by Suzuki and Abe [13].

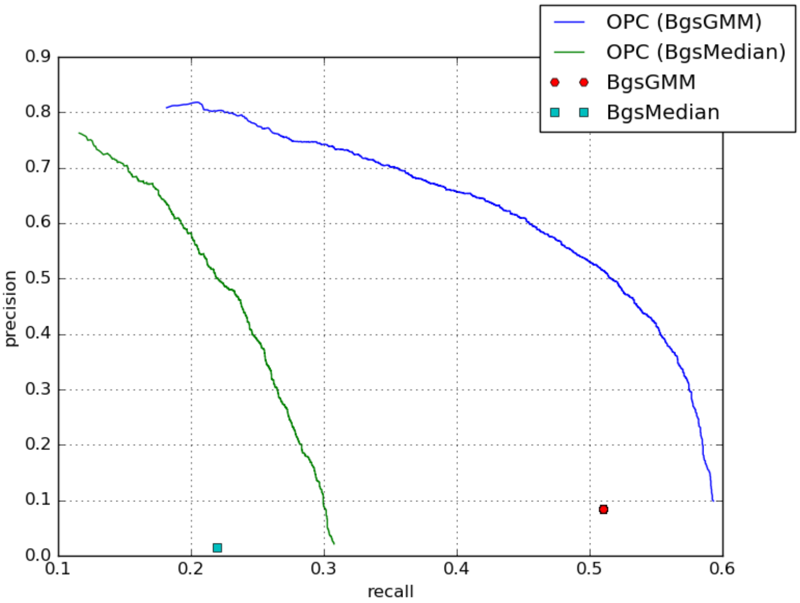


Fig. 3: Precision-Recall graph for fish detection with OPC (object proposal classification) and background subtraction on its own.

OPC (BgsMedian) and *OPC (BgsGMM)* are using the pipeline described in section 3.2 with the exception that *BgsMedian* is used for bounding box proposal generation in *OPC (BgsMedian)*.

In our Experiments we fine tuned the parameter of the background subtraction methods for fish detection when used on its own. When we used these methods for proposal generation the parameter have been adjusted to get many fish proposals.

Evaluation process: As in Pascal VOC [5], we consider a fish detection as correct (true positive) if the intersection over union ratio (iou) for a ground truth box with a predicted box is greater or equal to 0.5. If there is more than one predicted box that satisfies this condition for a specific ground truth box: Then one predicted box is considered as true positive and the remaining boxes are counted as false positives.

Discussion: Fig. 3 and Tab. 1 present detection results for the above mentioned methods. The *OPC (BgsGMM)* approach works best in our setup. For detection by background subtraction *BgsGMM* has a higher average precision score than *BgsMedain*. Whereby average precision of *BgsGMM* is 36.35% lower than *OPC (BgsGMM)*.

In general it can be observed that OPC detection approaches work better than background subtraction in our setup, although the CNN was not fine tuned to fish images.

4.2 Species classification

For species classification we used the detections of *OPC (BgsGMM)* and extracted CNN features to classify each detection as one of the 15 considered fish species. If the confidence level for a classification was lower than 0.5 it was rejected.

With this pipeline we achieve a counting score (**CS**) of **83%**, a **precision** of **66%** and a normalized counting score (**NCS**) of **58%** (see Fig. 4). CS and NCS are used as scoring functions in SeaCLEF 2016 and are defined as:

$$CS = e^{-\frac{d}{N_{gt}}} \tag{1}$$

with d as the difference between the number of ground truth occurrences N_{gt} and the predicted occurrences per species.

$$NCS = CS \cdot precision \tag{2}$$

Fig. 4 shows the results for SeaCLEF 2016. From this year participants we have achieved the best results in the fish species identification task. Compared to last years winner [3] who achieved a CS of 89%, a precision of 81% and a NCS of 72% there is still a little work to do. In future we plan to incorporate tracking, which will improve fish detection and classification results. We further plan to use larger CNN models and want to fine tune these models for fishes, since this worked really well for Choi [3].

5 Conclusion

This paper described our participation in SeaCLEF 2016 fish species recognition task. We focused on robust fish detection, since the simple application of background subtraction methods leads to a large number of false detections. Therefore we compared traditional background subtraction methods, mainly used for fish detection so far, with object proposal classification (OPC) for detection. We show that OPC fish detection (Fig. 2) works much better than background subtraction (Fig. 3) in our setup.

For species recognition we use the same CNN features as for detection and classified each fish with a multiclass SVM. Using this pipeline we achieve a normalized counting score of 58% and a precision of 66% (see Fig. 4) on the provided test dataset.

For the future we plan to incorporate fish tracking. We also want to use larger CNN models and fine tune these models to fish data.

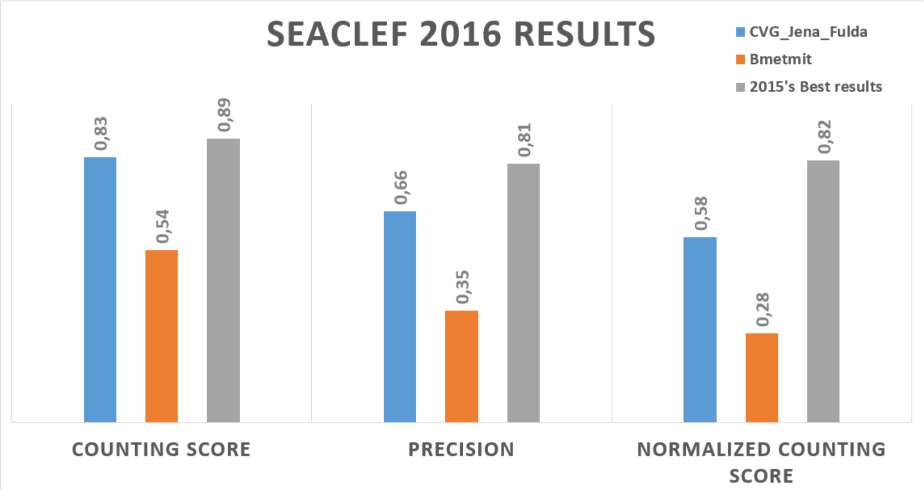


Fig. 4: Results, published on the SeaCLEF 2016 website. Our team is shown in blue.

References

1. Bastiaan J. Boom, Jiyin He, Simone Palazzo, Phoenix X. Huang, Cigdem Beyan, Hsiu-Mei Chou, Fang-Pang Lin, Concetto Spampinato, and Robert B. Fisher. A research tool for long-term and continuous analysis of fish assemblage in coral-reefs using underwater camera footage. *Ecological Informatics*, 23:83–97, 2014.
2. Jorge Cabrera-Gmez, Modesto Castrilln Santana, Antonio Domnguez-Brito, Daniel Hernandez-Sosa, Josep Isern-Gonzlez, and Javier Lorenzo-Navarro. Exploring the use of local descriptors for fish recognition in lifeclef 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.
3. Sungbin Choi. Fish identification in underwater video with deep convolutional neural network: 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.
4. Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *CoRR*, abs/1310.1531, 2013.
5. Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.
6. Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.
7. 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 T. Pltz S. McKenna T. Amaral, S. Matthews and R. Fisher, editors, *Proceedings of the Machine Vision of Animals and their Behaviour (MVAB)*, pages 6.1–6.7. BMVA Press, September 2015.

8. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
9. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
10. John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.
11. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
12. Chris Stauffer and W. Eric L. Grimson. Adaptive background mixture models for real-time tracking. In *1999 Conference on Computer Vision and Pattern Recognition (CVPR '99), 23-25 June 1999, Ft. Collins, CO, USA*, pages 2246–2252, 1999.
13. Satoshi Suzuki and Keiichi Abe. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1):32–46, 1985.
14. J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 2013.
15. Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2003.