Fish identification in underwater video with deep convolutional neural network: SNUMedinfo at LifeCLEF fish task 2015

Sungbin Choi

Department of Biomedical Engineering, Seoul National University, Republic of Korea

wakeup06@empas.com

Abstract. This paper describes our participation at the LifeCLEF Fish task 2015. The task is about video-based fish identification. Firstly, we applied foreground detection method with selective search to extract candidate fish object window. Then deep convolutional neural network is used to classify fish species per window. Classification results are post-processed to produce final identification output. Experimental results showed effective performance in spite of challenging task condition. Our approach achieved best performance in this task.

Keywords: Object detection, Image classification, Deep convolutional neural network

1 Introduction

In this paper, we describe the participation of the SNUMedinfo team at the LifeCLEF Fish task 2015. The purpose of task is automatically counting separate fish species in video segments. Training data includes video clips with annotation and sample images of 15 fish species. For a detailed introduction of the task, please see the overview paper of this task (1).

In recent years, deep Convolutional Neural Network (CNN) has improved automatic image classification performance dramatically (2). In this study, we experimented with GoogLeNet (3) which has shown effective performance in a recent ImageNet Challenge (4).

Firstly, we applied foreground detection method with selective search to extract candidate fish object window (Section 2.1). CNN is trained and used to identify fish species in candidate window (Section 2.2). Then CNN classification results are further refined to produce final identification output (Section 2.3). Our experimental methods are detailed in the next section.

2 Methods

2.1 Candidate fish object window extraction

Foreground detection

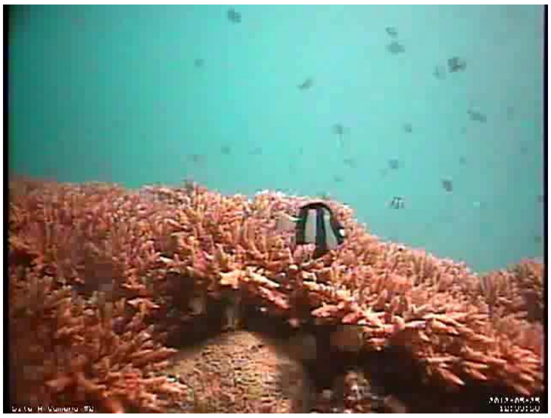

Fig. 1. Video segment image example

Firstly, we tried to identify background region per each video clip. If a video clip has S temporal segments, each pixel location has corresponding S pixel values. Per each pixel location in video clip, we took median value as background pixel value (Fig 2).

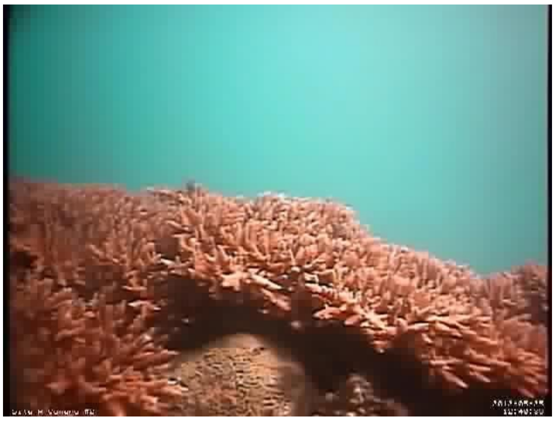

Fig. 2. Background image example

Pixels having pixel values different from this background more than predefined threshold, is considered as foreground pixel. Bilateral filter is applied to smooth foreground image (Fig 3).

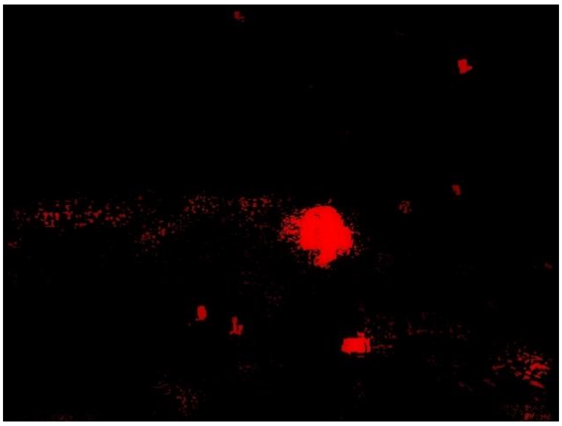

Fig. 3. Foreground image example

Then, we applied selective search (5) to extract candidate fish object window.

2.2 Fish species identification

Preparing training set for CNN

In fish task training set, there are 20 video clips with bounding box annotation, and samples images of 15 considered species. We formulated training set and validation set as follows.

Training set: Samples images of 15 fish species + 10 video clips
Validation set: Other 10 video clips

Per each video clip, among candidate fish object windows extracted from section 2.1, windows having intersection over union area (IoU) over 0.7 with ground truth bounding box annotation is considered as target fish species positive example. Candidate fish object windows having IoU less than 0.2 is considered as negative example (No fish inside window). So we have 16 labels for image classification (15 fish species + 'No fish')

Training CNN

We utilized GoogLeNet for image classification. GoogLeNet incorporates Inception module with the intention of increasing network depth with computational efficiency. Training CNN for fish identification started from GoogLeNet pretrained on ImageNet dataset. We finetuned CNN on fish identification training set (initial learning rate 0.001; batch_size:40).

2.3 Post-processing classification results

Filtering CNN output within each video segment

CNN classified results from Section 2.2 contains lots of image windows overlapped to each other, so we need to select best matching window for final output. Firstly, among all target positive windows, we selected maximum 20 windows having top classification score from CNN. Secondly, windows having IoU more than 0.3 is considered as duplicate, so it is removed.

Refining classification output by utilizing temporally connected video segment
Video segments are temporally connected, so existing fish object in previous frame is expected to be located within nearby region in next frame. Based on this idea, we applied following two rules.
Rule 1 (Adding): If video segment (k-1) and (k+1) has target positive fish object window in nearby geographic location, but video segment (k) does not have target fish object window in that location, then fish is expected to be in segment (k) also.
Rule 2 (Removing): If video segment (k) has target positive fish object, but both video segment (k-1) and (k+1) does not have target fish object window in nearby location, then it is expected that fish is expected to be not in segment (k).

3 Results

In fish task test set, 73 video clips are given. We submitted three different runs. In SNUMedinfo1 and SNUMedinfo2, assigned 10 video clips in training set and validation set is switched (Section 2.2). SNUMedinfo3 is same as SNUMedinfo1, but Filtering CNN output within each video segment step (Section 2.3) is not applied.
Evaluation metric for this task was counting score, precision and normalized counting score (For a detailed introduction to these evaluation metric, please see the overview paper of this task). Counting score is calculated based on the difference between the number of occurrences in the submitted run and the ground truth. Precision is calculated as number of true positive divided by number of true positive plus false positive. Normalized counting score is calculated as multiplication of counting score with precision. Evaluation results on test set is described in following table.

Table 1. Evaluation results of submitted runs

	Counting score	**Precision**	**Normalized counting score**
SNUMedinfo1	0.89	0.81	0.72
SNUMedinfo2	0.89	0.80	0.71
SNUMedinfo3	0.85	0.71	0.60

4 Discussion

Compared to other image recognition task such as ImageNet or LifeCLEF Plant task, this task deals with low quality underwater video. So our experiments involved additional pre-processing and post-processing step besides deep convolutional neural network training for image recognition. To further analyze contributions of each step with regard to the final performance, we need to experiment with various combinations of method options. We postpone thorough analysis of each step to future study when test set ground truth becomes available.

But generally, our overall fish identification performance was very effective in spite of challenging task conditions of varying video images in underwater scene. Our counting score approached near 0.9 and precision exceed 0.8 (Table 1).

Our post-processing step utilizing temporal neighborhood segment (Section 2.3) clearly improved performance when we compare run SNUMedinfo1 (with temporal post-processing step) to SNUMedinfo3 (without temporal post-processing step). Technically, this method is simple compared to more advanced techniques such as (6), but it was very helpful for improving precision.

5 Conclusion

This fish tasks deals with underwater video image, so it was more challenging than general image classification task and additional steps were needed for pre-processing and post-processing. We combined foreground detection method with selective search for candidate fish object window detection. Then CNN pretrained on other general object classification task is trained to classify fish species. Outputs from CNN classification results are further refined to produce final identification results. In our future work, we'll explore other methodological options to find more effective method.

6 References

1. Cappellato L, Ferro, N., Jones, G., and San Juan, E. CLEF 2015 Labs and Workshops. CEUR Workshop Proceedings (CEUR-WS.org); 2015.
2. Krizhevsky A, Sutskever I, Hinton GE, editors. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems; 2012.
3. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. arXiv preprint arXiv:14094842. 2014.
4. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. arXiv preprint arXiv:14090575. 2014.
5. Uijlings JRR, van de Sande KEA, Gevers T, Smeulders AWM. Selective Search for Object Recognition. Int J Comput Vis. 2013;104(2):154-71.
6. Kae A, Marlin B, Learned-Miller E, editors. The Shape-Time Random Field for Semantic Video Labeling. Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on; 2014 23-28 June 2014.