Fish species recognition from video using SVM classifier

Katy Blanc, Diane Lingrand, Frédéric Precioso

Univ. Nice Sophia Antipolis, I3S, UMR 7271, 06900 Sophia Antipolis, France
CNRS, I3S, UMR 7271, 06900 Sophia Antipolis, France
precioso@i3s.unice.fr

Abstract. For this first edition of LifeCLEF fish identification, we have built a processing chain based on background extraction, selection and description of keypoints with an adaptive scale and learning of each species by a binary linear SVM classifier. From the foreground segmentation, we have extracted several groups of blobs, representing a fish each. We have submitted three runs for different blob sizes and associations.

1 Introduction

The Fish task of the LifeCLEF2014 [4, 3] is organized in 4 subtasks such as video based fish identification, image based fish identification, image based fish identification and species recognition, and image based fish species recognition. We consider the subtask 3: we have to detect a fish and recognize his species inside a video. The training set is made of 285 videos labeled with 19868 bounding boxes of fishes and their species annotated. Our method is mainly based on motion detection and a set of binary Support Vector Machines trained with OpponentSift descriptors. The next section will describe our method in details. Results are discussed in Section 3, followed by conclusions and perspectives.

2 Detection and Identification Method

Our processing chain starts with a background-foreground segmentation from motion detection (see figure 1).

2.1 Background and motion detection

For this first step, an adaptive background mixture model [7] is built. This method consists in assuming that each pixel in the scene is modeled by a mixture of many Gaussian distributions and then each pixel of the background is computed with the distributions with the smallest fitness value. Finally a background subtraction is performed by marking as a foreground pixel, any pixel that is over 2.5 times any standard deviation of any background distribution. We end up with a mask for the motion detection that we first erode and then dilate to define blobs of detected moving objects.

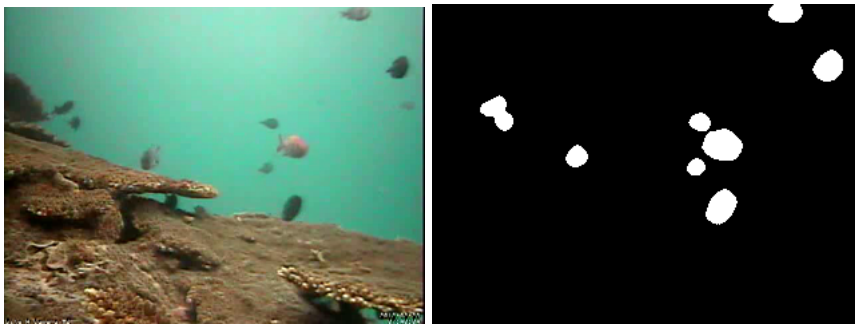

Fig. 1. An original frame and its motion mask

2.2 Points of interest and their descriptions

The XML metadata provided with the training set define the bounding boxes of annotated fishes, so we extract keypoints in these bounding boxes (figure 2). Specifically, these keypoints are located at the center of the given bounding box and several scales are set for the keypoint descriptor so that the resulting set of keypoints describe the whole fish. The initial scale is set to fit the size of the bounding box, the other scales are sequentially incremented from the initial center. To detect a fish with a part of it, the head or the tail, we also slightly shift the central initial keypoint to the left and the right of the bounding box in a spatial-pyramid-like technique. We adopt this strategy of large keypoint descriptors owing to the low resolution of the videos.

Fig. 2. Keypoints on a learning annotated fish

The OpponentSift schema [5] is used to describe these keypoints. The offline part of our processing chain ends here and we will now explain the online part of the system: analyzing a test video.

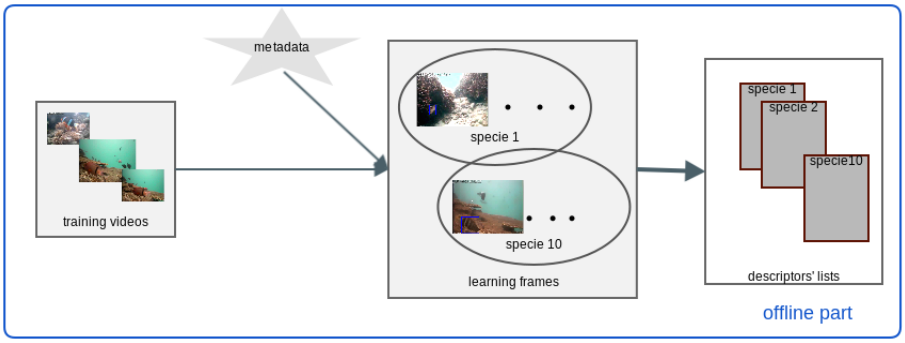

Fig. 3. Extraction of positive descriptors from training data set with metadata

2.3 Background description, filtering and learning

Since the training distribution must be the same as the test distribution in a supervised learning process, we decide to train a SVM classifier with some OpponentSift descriptors from the video background as negative training samples. The SIFT detector did not give us enough keypoints owing to the low resolution. Thus we choose to densely extract keypoints with approximately same scales as the scales of the training keypoints. Specifically, we have extracted keypoints with fixed scales from 30 to 110 pixels of diameter with respect to the size of the video.
To ensure the definition of the background, these keypoints were located in the still areas of the motion mask (black areas on figure 1).

For most of the training bounding boxes, part of the background is inside. Thus, before training, we filter the keypoints inside the bounding boxes (which should then be considered as positive training samples of the associated fish): For each one of these keypoints we look for its 10 nearest neighbors and remove any keypoint from the bounding box with more negative neighbors (keypoints from the background not in bounding boxes from the same video) than positive ones (keypoints from other bounding boxes on the same video).

Then, for each species, we train a linear SVM classifier with the species keypoints descriptors as positive samples and the aforementioned filtered background descriptors as negative examples.

We have hence done the first row of our processing chain during the query video classification.

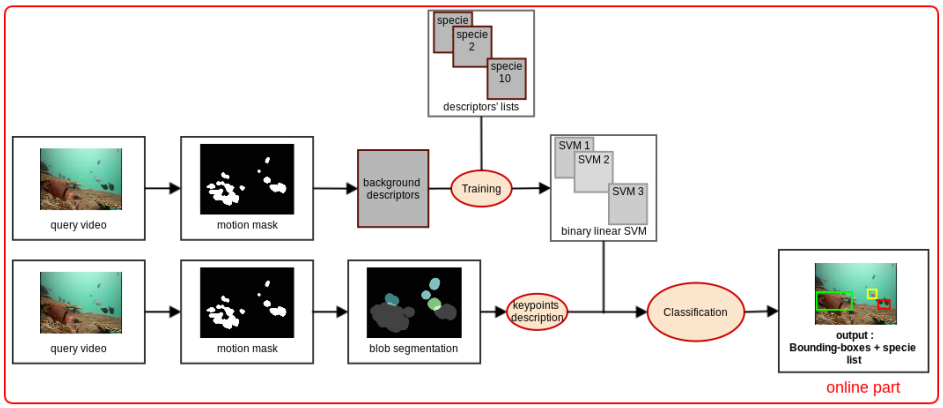

Fig. 4. Our processing chain from acquiring test video to detection and identification

2.4 Species and bounding box

First of all, we segment the motion mask to differentiate each blob. Then, for each blob, the centered and shifted keypoints are extracted with several scales (fixed scales from 30 to 110) and their scores are computed by each species SVM classifier. This score represents the distance between the current descriptor and the SVM decision boundary with the sign of the selected class. Only positively classified points with a distance larger than 0.5 are considered. For a blob, we sum up the score over each keypoint associated with each species in order to obtain a global score per species. Then, the list of potential species is given by decreasing scores. For our bounding box construction, we frame every keypoint with respect to its scale and associate the species classified with the highest score (first species from the list).

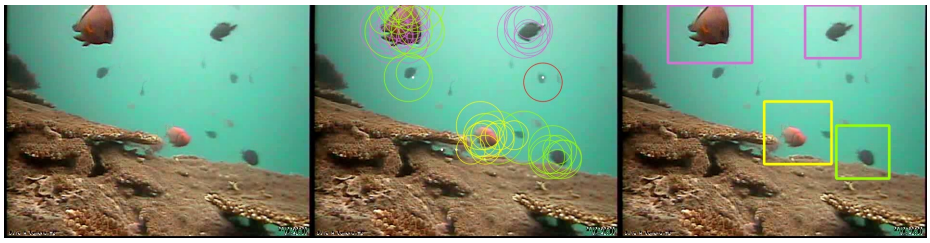

Fig. 5. From left to right:(a) original frame, (b) the detected and positively classified keypoints with the color corresponding to the species with highest score and (c) the final bounding boxes with the color corresponding to species with the highest sum over scores

Finally, we submitted three runs:

1. The first one with the blobs computed from the mask segmentation.
2. For the second run, we wanted to detect if there were several fishes in connected blobs. Thus we have computed a lighter dilation (than in the first run) on the initial motion mask and detected the dominant color in each blob. Small blobs with same dominant color were merged into one blob. Then each blob was dilated to obtain the maximum numbers of keypoints on the fishes. You can see on figure 5 that even if the fish is separated in several blobs, these ones are then merged together. Finally the same processing is applied to each group of blobs as it is in run 1.

Fig. 6. Fish segmentation with dilatation and dominant color

3. The third run has computed groups of blobs as in the run 2 except that we have provided as many answers as little connected blobs regardless their dominant color.

3 Experimentation and results

At the beginning, we tried to track fishes in order to obtain more descriptors with some standard tracking methods as CamShift and segmentation of optical flow for instance. But the tracking was not effective enough owing to the resolution. We also encountered difficulties with moving seaweed and changing brightness as warned by the organizers since the dataset contains videos recorded from sunrise to sunset. We have built up our processing chain to deal with these difficulties.

As scoring functions, the organizers of LifeCLEF have computed both average precision vs recall and precision vs recall for each fish species for the subtask3 (figure 6 and 7). Only bounding boxes matching with a bounding box from test set with a PASCAL score above a certain threshold are considered. The organizers have computed a baseline program with the ViBe background modeling approach [1, 2] for fish detection and VLFeat [6] for fish species recognition to compare against our method.

Compared to the baseline, we have a better precision and worse recall. It is worth noting that the performance in terms of species recognition (based only on the correctly detected bounding boxes) was comparable to the baseline, but our bounding boxes were often too large.

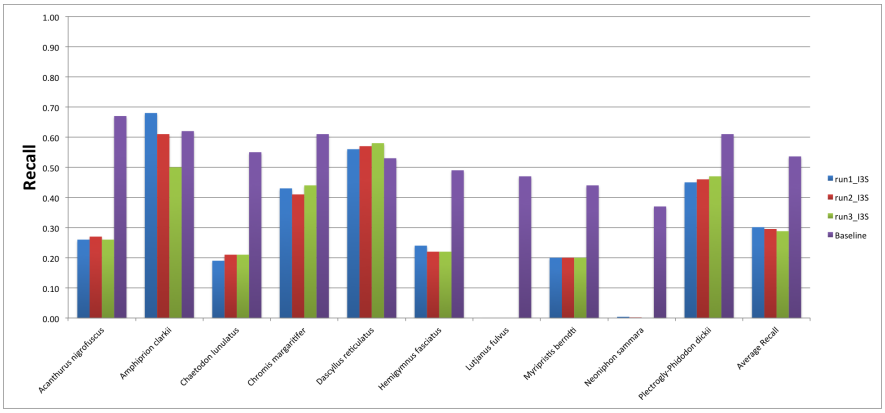

Fig. 7. Recall on each species and average recall of our three runs and the baseline of the organizers

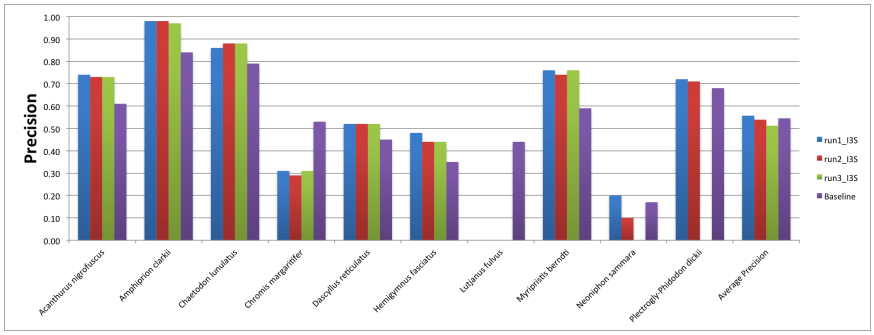

Fig. 8. Precision on each species and average precision of our three runs and the baseline of the organizers

4 Conclusion

This competition was a real challenge since the video data set was difficult regarding video resolution, natural phenomena (e.g. murky water, algae on camera lens, ...) and the huge amount of data to be processed in a limited time. Our results are so pretty encouraging.

To improve our processing chain, a good tracking from the annotated fishes would allow us obtaining more positive descriptors for each species. Moreover, some metadata could be used as the GPS coordinates. Finally, with the tracking, we could study the movement of each species to see if this is specific to species and we could try analyzing the interactions between species.

References

1. O. Barnich and M. Van Droogenbroeck. ViBe: a powerful random technique to estimate the background in video sequences. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009)*, pages 945–948, April 2009. PDF available on the University site or at the IEEE.
2. O. Barnich and M. Van Droogenbroeck. ViBe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image Processing*, 20(6):1709–1724, June 2011.
3. Spampinato Concetto, Bob Fisher, and Bas Boom. Lifeclef fish identification task 2014. In *CLEF working notes 2014*, 2014.
4. Alexis Joly, Henning Müller, Hervé Goëau, Hervé Glotin, Concetto Spampinato, Andreas Rauber, Pierre Bonnet, Willem-Pier Vellinga, and Bob Fisher. Lifeclef 2014: multimedia life species identification challenges. In *Proceedings of CLEF 2014*, 2014.
5. Koen E. A. van de Sande, Theo Gevers, and Cees G. M. Snoek. Evaluation of color descriptors for object and scene recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, USA, June 2008.
6. A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.
7. Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 2 - Volume 02*, ICPR '04, pages 28–31, Washington, DC, USA, 2004. IEEE Computer Society.