Fish Species Recognition from Video using SVM Classifier

Katy Blanc
Univ. Nice Sophia Antipolis,
I3S, UMR UNS-CNRS 7271
06900 Sophia Antipolis,
France
kblanc@i3s.unice.fr

Diane Lingrand
Univ. Nice Sophia Antipolis,
I3S, UMR UNS-CNRS 7271
06900 Sophia Antipolis,
France
lingrand@i3s.unice.fr

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

ABSTRACT

To build a detailed knowledge of the biodiversity, the geographical distribution and the evolution of the alive species is essential for a sustainable development and the preservation of this biodiversity. Massive databases of underwater video surveillance have been recently made available for supporting designing algorithms targeting the identification of fishes. However these video datasets are rather poor in terms of video resolution, pretty challenging regarding both the natural phenomena to be considered such as murky water, seaweed moving the water current, etc, and the huge amount of data to be processed.

We have designed a processing chain based on background segmentation, selection keypoints with an adaptive scale, description with OpponentSift and learning of each species by a binary linear Support Vector Machines classifier.

Our algorithm has been evaluated in the context of our participation to the Fish task of the LifeCLEF2014 challenge. Compared to the baseline designed by the LifeCLEF challenge organizers, our approach reaches a better precision but a worse recall. Our performances in terms of species recognition (based only on the correctly detected bounding boxes) is comparable to the baseline, but our bounding boxes are often too large and our score is so penalized. Our results are thus really encouraging.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Information retrieval—*Specialized information retrieval,Multimedia and multimodal retrieval, Video search*

General Terms

Video analysis

Keywords

video processing; image recognition; machine learning

MAED '14 Orlando, Florida USA
Copyright 2014 ACM 978-1-4503-3123-4/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661821.2661827.

1. INTRODUCTION

For the purpose of supporting the conception of automatic fish identification systems, the Fish task of the LifeCLEF2014 [7, 4] is organized in 4 subtasks: video-based fish identification (subtask-1), image-based fish identification (subtask-2), image-based fish identification and species recognition (subtask-3), and image-based fish species recognition (subtask-4). In this paper, we consider the subtask-3: we have to detect a fish and recognize its species inside a video. The training set is made of 285 videos, acquired with underwater still camera, and labeled with 19868 bounding boxes of fishes and their species annotated. The testing set is made of 116 videos.

The next section will review recent works on object recognition in video. In section 3, our method will be described in details. Results are discussed in Section 4, followed by conclusions and perspectives.

2. RELATED WORK

When considering video classification, the first processing step is extracting relevant features to input them in a learning machine.

Among standard video content analysis approaches, optical flow is probably the most widely considered. In [12], the authors used for the first time optical flow for motion recognition, using temporal textures (i.e. 1st and 2nd order statistics based on the direction and the magnitude of the flow). For periodic human actions (i.e. walking, running, skiing, swimming), the same authors proposed to compute optical flow, then to accumulate its magnitude in a regular spatio-temporal grid, this grid defining then a flow-based motion descriptor. This idea of extracting optical flow information with respect to a spatio-temporal grid was extended by Rodriguez *et al* [13] with flow features computed in spatio-temporal cubes over regularity flow information. These features were further classified through a template matching process based on correlation from Fourier transform.

Efros *et al* [6] also splitted the optical flow but rather than following a 2D+t regular grid, they divided the optical flow informations into 4 groups of vectors according to the flow directions. Ahad *et al* [1] recently used also this same four flow channel description in order to solve the problem of self-occlusion in a Motion Histogram Image approach.

In [2], Ali and Shah extracted more complex kinematic features from optical flow such as divergences, vorticity, symmetric and anti-symmetric flow fields, second and third principal invariants of flow gradient, rate of strain tensor and

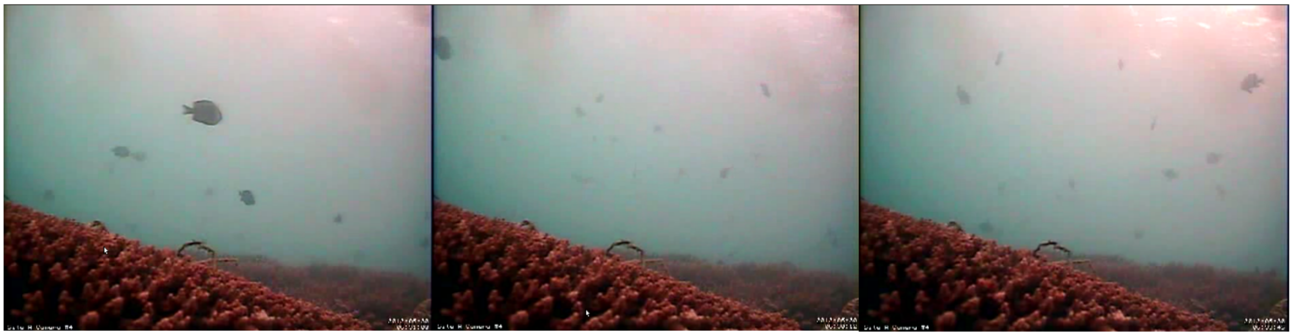

Figure 1: Variations of illumination (sun light)

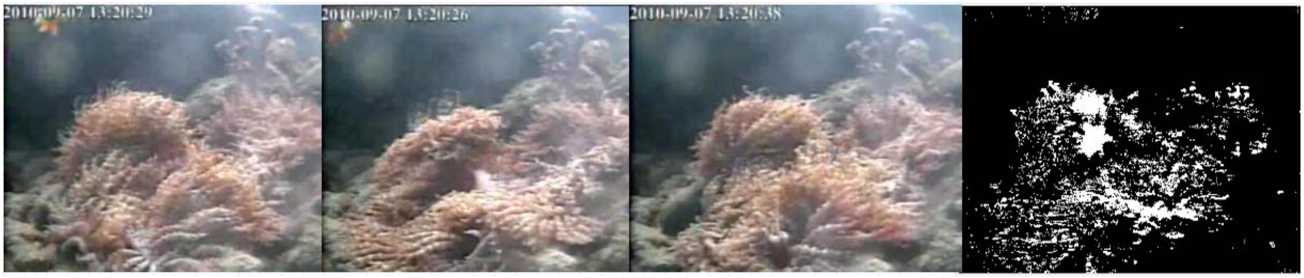

Figure 2: Seaweed moving in the background

third principal invariant of rate of rotation tensor. After a PCA on these features, the classification step is based on a nearest neighbor algorithm for Multiple Instance Learning (MIL).

On FishClef dataset, all the methods we have tried to extract a relevant optical flow were useless owing to illumination variations (figure 1), large seaweed moving (figure 2), etc. We have thus decided to extract not anymore only motion features but 2D+t features since these approaches have shown to be pretty powerful.

Indeed, in [19], after blurring and sub-sampling video sequence along temporal direction to build temporal pyramid (3 levels), the authors extract space-time gradient at each space-time point in each of the 3 cubes and then classify them by comparing gradient measurements. In [15], Thurau and Hlavac extend HOG for videos (dividing video volume into regularly spaced overlapping blocks) and classify these 2D+t HOG descriptors with a nearest neighbor algorithm.

Several authors have studied the repeatability and accuracy of detected features under changes in spatial and temporal video resolution and under camera motion [9] or under scale changes, in-plane rotations, video compression and camera motion [18]. Comparison of local space descriptors including higher order descriptors, image gradients and optical flow [10], space-time HOG, HOG and HOF [8] and in terms of image brightness, gradient ans optical flow [5] have been done.

However, the resolution and the quality of the FishClef dataset were too low for achieving good results with any of these methods. We thus came back to extract less semantic information from motion by computing only background detection.

Such an approach has already been considered by Piccardi [11]. Difference of frames are considered, optic flow is thresholded and a model of the background is estimated. Tanase [14] compensates the motion of the camera by estimating an homography from features (method based on RANSAC). Then, full frame optical flow is computed using Farneback's method. Finally, foreground objects are detected using the displacement between the synthetic background motion field and the actual motion field.

We now detail our approach.

3. THE PROCESSING CHAIN

Our method starts with an extraction and a description of keypoints on each specie.

3.1 Points of interest and their descriptions

The XML metadata provided with the training set define the bounding boxes of annotated fishes. We consider three keypoints located on the central horizontal axis and at one third and two third of the horizontal length (see figure 3). These three keypoints are described using the Opponent SIFT color descriptor [16] at different scales. Scales are computed starting from the bounding box size (around 60 pixels in diameter). Some part of a fish may not belong to the bounding box: four increasing sizes are thus considered allowing a keypoint to describe the whole fish. It may also happen that the bounding box is too large: four decreasing sizes are thus considered enabling to describe approximately three parts of a fish: head, body and tail. We adopt this strategy of large keypoint descriptors owing to the low resolution of the videos.

The offline part of the processing chain ends here (the training phase is summarized in figure 4).

We now describe the online part of the system which consists in analysing a query video.

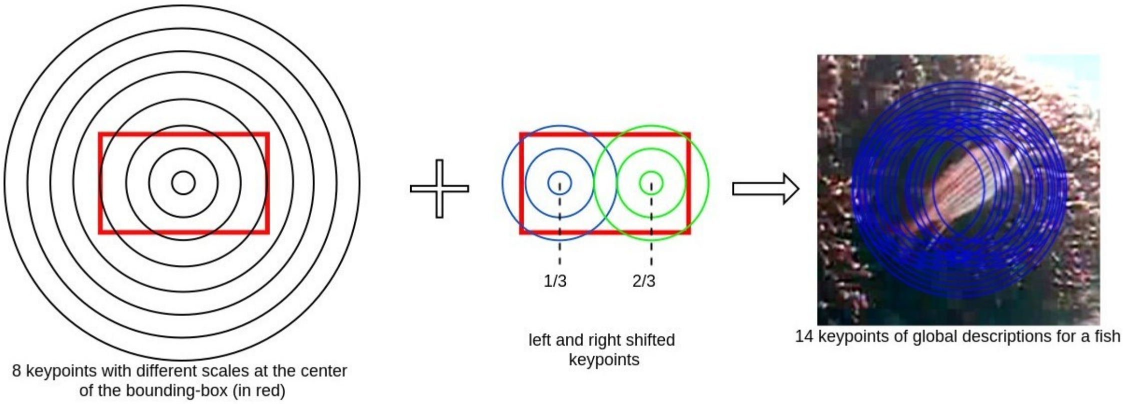
8 keypoints with different scales at the center of the bounding-box (in red)
1/3
2/3
left and right shifted keypoints
14 keypoints of global descriptions for a fish

Figure 3: Extraction of points of interest at different locations and scales with respect to the given bounding box.

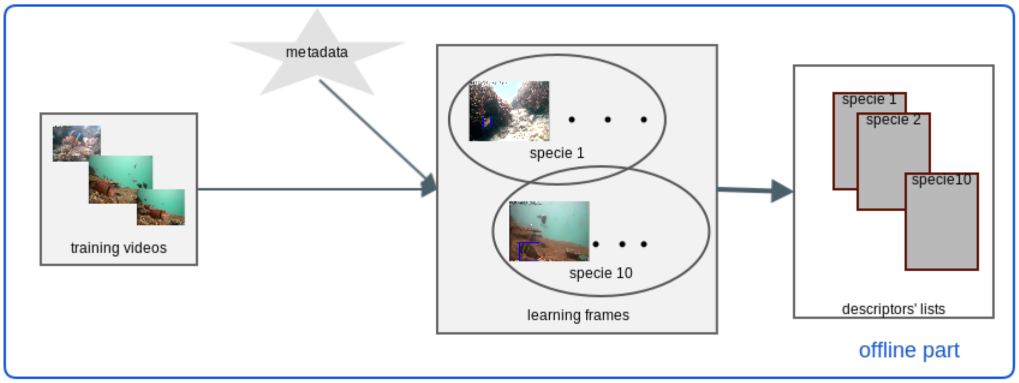

Figure 4: Extraction of positive descriptors from training data set with metadata

3.2 Background and motion detection

The first step of analysing is to detect motion from a background-foreground segmentation obtained thanks to an adaptive background mixture model [20]. 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.

3.3 Background description, filtering and learning

Since the training distribution must be the same as the test distribution in a supervised learning process and in order to distinguish specie from background and identify them,for each specie, a SVM classifier is trained with the descriptors of this specie as positive samples and some OpponentSift descriptors of the current video background as negative samples. Specifically, keypoints are densely extracted from the background with fixed scales from 30 to 110 pixels of diameter with respect to the size of the video.

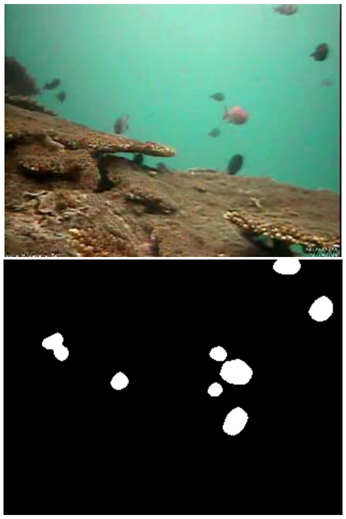

Figure 5: An original frame and its motion mask

To ensure the definition of the background, these keypoints were located in the still areas of the motion mask (black areas on figure 5). Because of the large size of bounding boxes, some descriptors of specie describe the background. Thus, before training, we filter the positive keypoints: 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) than positive ones.

Then, for each species, a linear SVM classifier is trained with the aforementioned filtered species keypoints descriptors as positive samples and the background descriptors as negative examples in similar proportion. We have hence done the first row of our processing chain during the query video classification (Figure 6).

3.4 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 (see figure 7). Then, the list of potential species is given by decreasing scores. For our bounding box construction, we frame every keypoints with respect to its scale and associate the species classified with the highest score (first species from the list).

4. EXPERIMENTATION AND RESULTS

The Fish task of the LifeCLEF2014 [7, 4] is organized in 4 subtasks:

- *Subtask 1- Video-based Fish Identification*, four videos fully labeled, 21106 annotations corresponding to 9852 different fish instances.
- *Subtask 2- Image-based Fish Identification*, 957 videos labeled with 112078 fish annotated.
- *Subtask 3- Image-based Fish Identification and Species Recognition*, 285 videos labeled with 19868 fish (and their species) annotated.
- *Subtask 4- Image-based Fish Species Recognition*, 19868 fish images annotated.

We consider the subtask 3: we have to detect a fish and recognize its species inside a video. At first, 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. Furthermore, underwater cameras are still with a fixed focal which does not provide much advantages owing to the natural phenomena aforementioned while bringing other difficulties since without specific fish focalization, any fish of any size is a possible target (as can be seen on figure 1, figure 5 or figure 8). 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 subtask-3 (figure 9 and 10). 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 [3] for fish detection and VLFeat [17] for fish species recognition to compare against our method.

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 8 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.

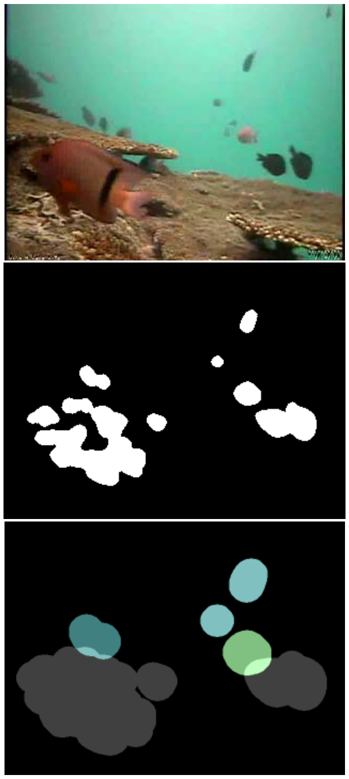

Figure 8: Fish segmentation with dilatation and dominant color

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.

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.

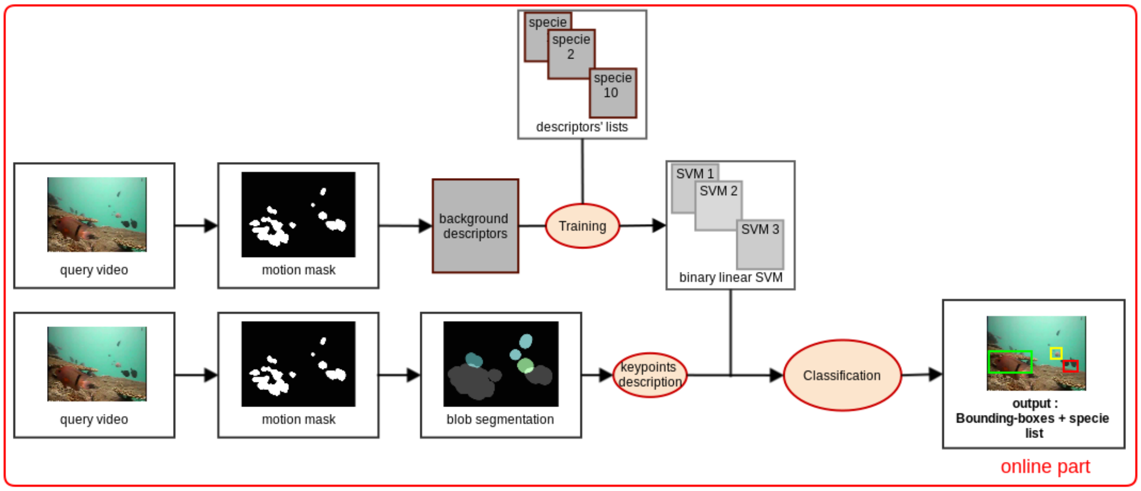

Figure 6: Our processing chain from acquiring test video to detection and identification

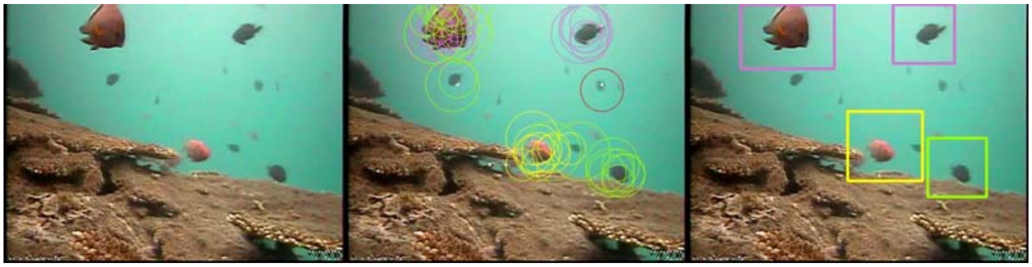

Figure 7: From left to right:(a) original frame, (b) the detected and positively classified keypoints with the color corresponding to the species with highest score and (c) the final bounding boxes with the color corresponding to species with the highest sum over scores

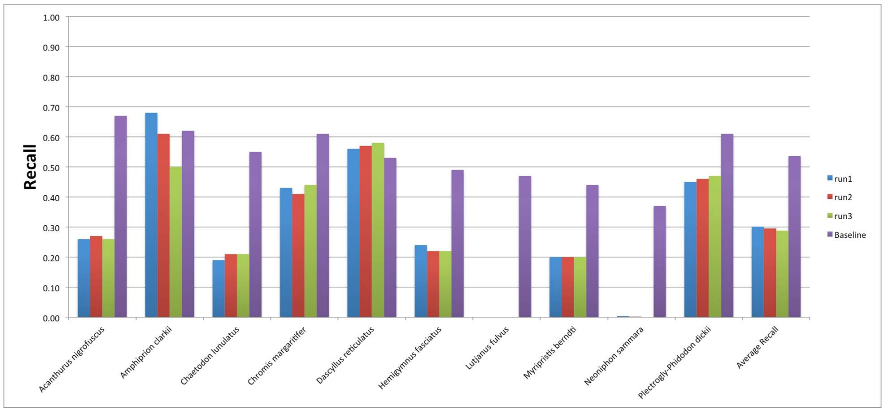

Figure 9: Recall on each species and average recall of our three runs and the baseline of the organizers

5. CONCLUSION

For the purpose of supporting the conception of automatic fish identification systems, the Fish task of the LifeCLEF2014 Challenge [7, 4] provides 285 videos for training, acquired with underwater still camera, labeled with 19868 bounding boxes of fishes and their species annotated, and 116 videos for testing. This dataset is a real challenge since the video are poor owing to both video resolution and natural phenomena (e.g. murky water, seaweed moving with the sea, ...) and the huge amount of data to be processed.

We have designed a processing chain based on background segmentation, selection keypoints with an adaptive scale, description with OpponentSift and learning of each species by a binary linear Support Vector Machines classifier. Our results are really encouraging but could be improved.

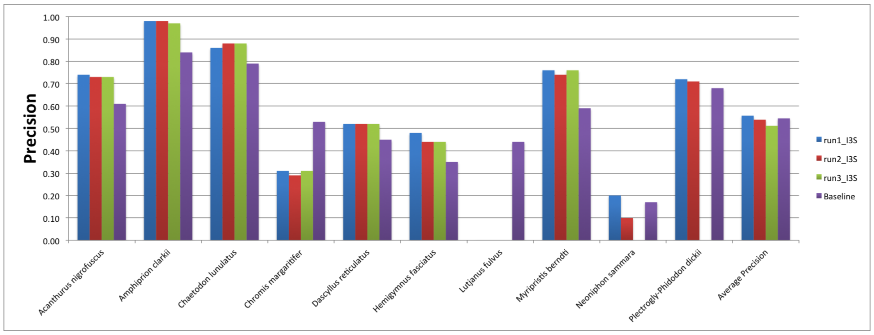

Figure 10: Precision on each species and average precision of our three runs and the baseline of the organizers

For instance in order 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 could be a significant feature to identify species and we could also try analyzing the interactions between species.

6. REFERENCES

[1] M. A. R. Ahad, T. Ogata, J. K. Tan, H. Kim, and S. Ishikawa. Motion recognition approach to solve overwriting in complex actions. In *FG*, pages 1–6. IEEE, 2008.

[2] S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(2):288–303, Feb. 2010.

[3] 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.

[4] S. Concetto, B. Fisher, and B. Boom. Lifeclef fish identification task 2014. In *CLEF working notes*, 2014.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 14th International Conference on Computer Communications and Networks*, ICCCN '05, pages 65–72, Washington, DC, USA, 2005. IEEE Computer Society.

[6] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pages 726–, Washington, DC, USA, 2003. IEEE Computer Society.

[7] A. Joly, H. Müller, H. Goëau, H. Glotin, C. Spampinato, A. Rauber, P. Bonnet, W.-P. Vellinga, and B. Fisher. Lifeclef 2014: multimedia life species identification challenges. In *Proceedings of CLEF*, 2014.

[8] A. Klaeser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *Proceedings of the British Machine Vision Conference*, pages 99.1–99.10. BMVA Press, 2008. doi:10.5244/C.22.99.

[9] I. Laptev. *Local Spatio-Temporal Image Features for Motion Interpretation.* PhD thesis, Department of Numerical Analysis and Computer Science (NADA), KTH, 2004.

[10] I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. In *In First International Workshop on Spatial Coherence for Visual Motion Analysis*, 2004.

[11] M. Piccardi. Background subtraction techniques: a review. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 3099–3104. IEEE, 2004.

[12] R. Polana and R. Nelson. Low level recognition of human motion (or how to get your man without finding his body parts). In *In Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 77–82. Press, 1994.

[13] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*. IEEE Computer Society, 2008.

[14] C. Tanase. *Towards effective spatio-temporal analysis for content-based video retrieval.* PhD thesis 2014.

[15] C. Thurau and V. Hlavác. Pose primitive based human action recognition in videos or still images. In *CVPR'08*, pages –1–1, 2008.

[16] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluation of color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, Sept. 2010.

[17] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.

[18] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of the 10th European Conference on Computer Vision: Part II*, ECCV'08, pages 650–663, Berlin, Heidelberg, 2008.

[19] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *Proc. of CVPR*, pages 123–130, 2001.

[20] Z. 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.