Efficient instance-based fish species visual identification by global representation

Pierre-Hugues Joalland[1,2], Sébastien Paris[2], and Hervé Glotin[1,2,3]

1 Aix-Marseille Université, CNRS, ENSAM, LSIS UMR 7296, 13397 Marseille, France
2 Université de Toulon, CNRS, LSIS UMR 7296, 83957 La Garde, France
3 Institut Universitaire de France, 103 Bd St-Michel, 75005 Paris, France

joalland@univ-tln.fr
sebastien.paris@lsis.org
glotin@univ-tln.fr

Abstract. This paper presents the participation of the LSIS/DYNI team for the ImageCLEF 2014 Fish identification challenge. ImageCLEF's Fish identification task provides a testbed for the system-oriented evaluation of fish species identification based on still images. The goal is to investigate image retrieval approaches in the context of images extracted from collected videos. The LSIS/DYNI team submitted three runs,, won the challenge with results that sensibly outperform the baseline (both recall and precision of 0.99) for the image-based fish recognition category with a fully automatic method. Our approach is based on a computer vision framework involving local, highly discriminative visual descriptors, sophisticated visual-patches encoder and large-scale supervised classification. The paper presents the three procedures employed, and provides an analysis of the obtained evaluation results.

Keywords: ImageCLEF, fish species identification, underwater video monitoring, images, identification, classification, Fisher Vectors, Local Ternary Patterns, late fusion, encoding/pooling.

1 Introduction

This paper presents the contribution of the LSIS/DYNI group for the LifeClef Fish identification task[1][2][3] that was organized within ImageCLEF 2014 for the fish species recognition based on still images containing only one fish instance (Subtask 4). This challenge was organized as a classification task over 10 fish species with visual content being the main available information. Considered images are extracted from underwater fish videos acquired with natural background (see Fig. 1). The LSIS/DYNI team submitted three runs, all of them based on local feature extraction and large-scale supervised classification. Our automatic methods won the challenge and sensibly outperformed the baseline for the image-based fish recognition task (both recall and precision of 0.99).

2 Task description

The task has been evaluated as a fish species retrieval task.

2.1 Training and Test data

The images dataset was built from the Fish4Knowledge (www.fish4knowledge.eu) videos in charge of monitoring Taiwan coral reefs in the past five years. The dataset contains videos recorded from sunrise to sunset showing several phenomena, e.g. murky water, algae on camera lens, etc., which makes the fish identification task more complex.
Each video has a resolution of either 320x240 or 640x480 with 5 to 8 fps. Only the 10 main species were considered.

- The training data is comprised of 9868 images. The groundtruth consists in 10 directories (10 species), each one containing the images according to the species.
- The test data is comprised of 6956 to-be-predicated images.

2.2 Task objective and evaluation metric

The goal of the task was to retrieve the correct species among the 10 possible ones for each test image.
Each participant was allowed to submit up to 3 runs. As many species as possible can be associated to each test image, sorted by decreasing confidence score. However, we chose to only provide the best ranked one by our system.

Fig. 1. Fish species identification dataset.

3 Description of used methods

For all submitted runs, we followed the same unsupervised pipeline [4][5]:

i. local feature extraction
ii. patch encoding
iii. pooling with spatial pyramid for local analysis and a linear large-scale supervised classification
iv. supervised classification using Linear SVM

For all used methods, global representation is retrieved on 1x1 plus 2x2 pooling windows.
No image specific pre-processing was performed, in particular illumination correction, background substraction. The posterior probabilities are retrieved from the SVM outputs by linear regression. Late fusion is performed by averaging posterior probabilities.

3.1 Local Ternary Patterns (LTP) → LSIS DYNI run 1

The first run corresponds to a one layer architecture based on LTP features [8], where dictionary is fixed by the LTP framework and local feature linearly encoded with a single dictionary element. We fixed t = 10. The basic idea of LTP is to approximate ternary code by concatenating two binary codes (Local Binary Patterns).
Our method is based on a multiscale version where block size is selected from 1 pixel up to 3 pixels (3 scales in total). Final features are obtained by average pooling on 1x1 + 2x2 spatial pyramid (5 windows).
Thus, features size is 7680 due to 256 x 2 codes x 3 scales x 5 windows

3.2 Late fusion of LTP and improved FV → LSIS DYNI run 2

The second run is using LTP features as in run1, coupled with improved Fisher Vectors [7], where the spatial compound of the local features were added. As local features, we chose SIFT vectors densely sampled and decorrelated by PCA in a 80 dimension space.
We first compute 25x25 SIFT patches sized 24x24 pixels per image and repeat this for 3 scales. Fisher vectors are obtained with the same spatial pyramid as in 3.1, by estimating a Gaussian Mixture Model (GMM) with 16 Gaussians. Fisher Vectors are derived from the mean values and the variances of the fitted GMM.
Thus, features size is 38400 due to 2 x 80 x 16 x 3 scales x 5 windows.

3.3 Late fusion of LTP + improved FV + Sparse Coding → LSIS DYNI run 3

Here, we took as local features some LTP patches densely sampled (25x25 per image) [6].
As in 3.2, we first compute 25x25 LTP patches sized 24x24 pixels per image and repeat this for 3 scales with the same spatial pyramid of 5 windows.
Learning of dictionary is performed by using sparse coding, with a positivity constraint for both sparse codes and dictionary elements. The dictionary finally contains 1024 elements and lp-norm pooling on the 1x1+2x2 spatial pyramid where we fixed p = 3.
Thus, features size is 15360 due to 1024 x 3 scales x 5 windows.

4 Results

4.1 Baseline

The baseline for this task is VLFeat for fish species recognition [9].

4.2 Recall score

Our runs outperformed sensibly the baseline by obtaining an average recall of 0.99 vs. Baseline 0.91 (see Fig. 2).

4.3 Precision score

Precision of 1 is obtained for almost all species expect for Chromis margaritifer, Dascyllus reticulatus and Plectrogly-Phidodon dickii species :

Run	Chromis margaritifer	Dascyllus reticulatus	Plectrogly-Phidodon dickii
1	0.95	0.96	0.96
2	0.94	0.98	0.98
3	0.97	0.97	0.98

5 Conclusions

Our methods sensibly outperformed the baseline and were ranked first of this first ImageCLEF Fish identification challenge our framework is well adapted for this easy challenge by outperforming sensibly the baseline :

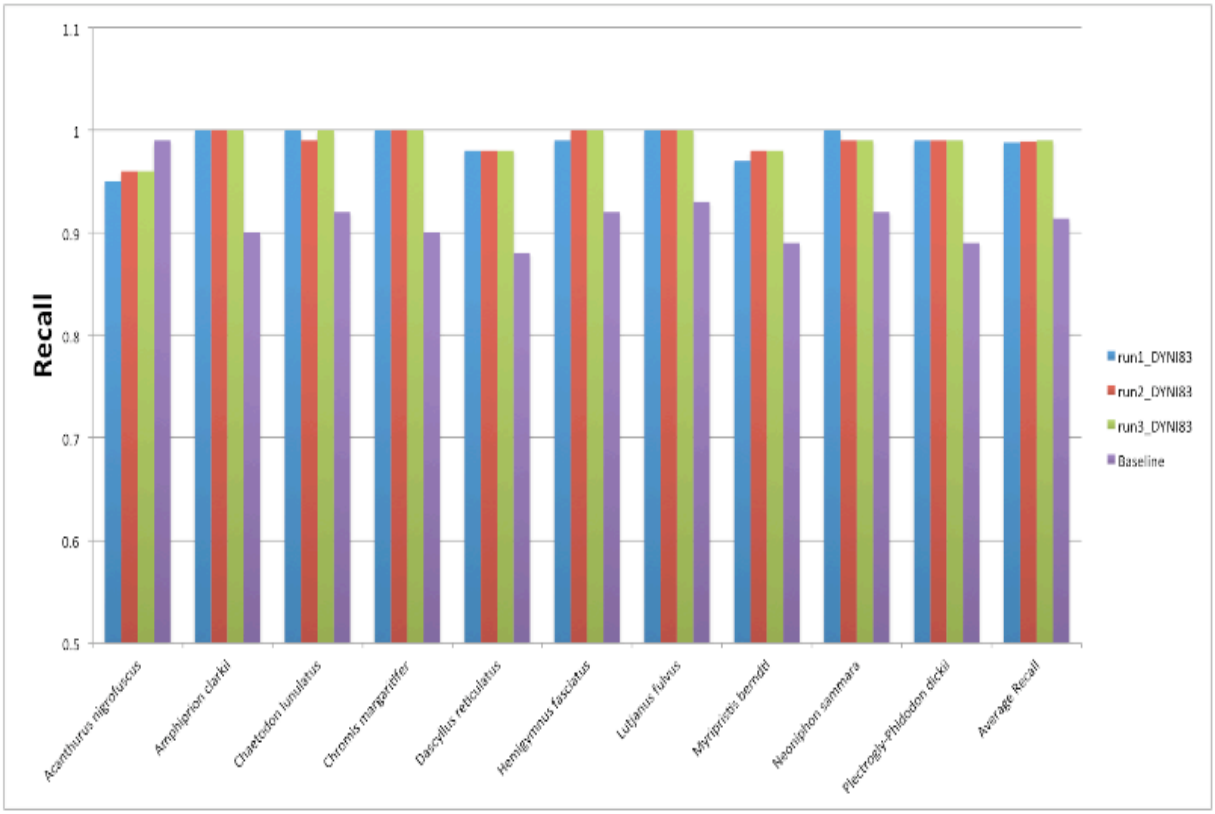

Fig. 2. Recall for the image-based fish identification

References

1. B. J. Boom, J. He, S. Palazzo, P. X. Huang, H.-M. Chou, F.-P. Lin, C. Spampinato, R. B. Fisher; A research tool for long-term and continuous analysis of fish assemblage in coral-reefs using underwater camera footage, Ecological Informatics (2014)
2. A. Joly, H. Müller, H. Goëau, H. Glotin, C. Spampinato, A. Rauber, P. Bonnet, W.P. Vellinga, B. Fisher ; multimedia life species identification challenges, LifeCLEF 2014 proceedings (2014)
3. C. Spampinato, B. Fisher, B. J. Boom ; CLEF working notes 2014, LifeCLEF Fish Identification Task 2014, FishClef 2014 proceedings (2014)
4. Paris, S., Halkias, X., Glotin, H.: Sparse coding for histograms of local binary patterns applied for image categorization: Toward a bag-of-scenes analysis. In: ICPR' 12 (2012)
5. Paris, S., Halkias, X., Glotin, H.: Participation of LSIS/DYNI to ImageCLEF 2012 – working keynotes. In: ImageCLEF'12 (2012)
6. Paris, S., Halkias, X., Glotin, H.: Efficient Bag of Scenes Analysis Categorization. In: ICPRAM'13 (2013)
7. Sanchez J., Perronnin F., Mensink T., Verbee J.: Image Classification with the Fisher Vector: Theory and Practice. In: International Journal of Computer Vision 105, 3, 222-245 (2013)
8. Tan X., Triggs B.: Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions. In: AMFG '07 - 3rd International Workshop Analysis and Modelling of Faces and Gestures 4778, 168-182 (2007)
9. Vedaldi A., Fulkerson B.: VLFeat - an open and portable library of computer vision algorithms. In: ACM International Conference on Multimedia. (2010)

Acknowledgements

This work is supported by RAPID PHRASE project with Prolexia SA.
This work is also supported by the SABIOD CNRS MI MASTODONS Big Data project on automatic species identification and will next be completed by joint bioacoustic and visual identification.