Multimed Tools Appl (2016) 75:1701–1720
DOI 10.1007/s11042-015-2601-x

# Fine-grained object recognition in underwater visual data

**C. Spampinato[1] · S. Palazzo[1] · P. H. Joalland[2,3] · S. Paris[2,3,4] · H. Glotin[2,3] · K. Blanc[5] · D. Lingrand[5] · F. Precioso[5]**

Received: 1 August 2014 / Revised: 23 March 2015 / Accepted: 1 April 2015 /
Published online: 24 May 2015


**Abstract** In this paper we investigate the fine-grained object categorization problem of determining fish species in low-quality visual data (images and videos) recorded in real-life settings. We first describe a new annotated dataset of about 35,000 fish images (*MA-35K*

✉ C. Spampinato
cspampin@dieei.unict.it

S. Palazzo
simone.palazzo@dieei.unict.it

P. H. Joalland
joalland@univ-tln.fr

S. Paris
sebastien.paris@lsis.org

H. Glotin
glotin@univ-tln.fr

K. Blanc
kblanc@i3s.unice.fr

D. Lingrand
lingrand@i3s.unice.fr

F. Precioso
precioso@i3s.unice.fr

[1] Department of Electrical, Electronics and Computer Engineering, University of Catania, Catania, Italy

[2] Aix-Marseille Université, CNRS, ENSAM, LSIS UMR 7296, 13397 Marseille, France

[3] Université de Toulon, CNRS, LSIS UMR 7296, 83957 La Garde, France

[4] Institut Universitaire de France (IUF), 75005 Paris, France

[5] I3S, UMR UNS-CNRS 7271, University of Nice Sophia Antipolis, Nice, France

dataset), derived from the Fish4Knowledge project, covering 10 fish species from the Eastern Indo-Pacific bio-geographic zone. We then resort to a label propagation method able to transfer the labels from the *MA-35K* to a set of 20 million fish images in order to achieve variability in fish appearance. The resulting annotated dataset, containing over one million annotations (*AA-1M*), was then manually checked by removing false positives as well as images with occlusions between fish or showing partially fish. Finally, we randomly picked more than 30,000 fish images distributed among ten fish species and extracted from about 400 10-minute videos, and used this data (both images and videos) for the fish task of the LifeCLEF 2014 contest. Together with the fine-grained visual dataset release, we also present two approaches for fish species classification in, respectively, still images and videos. Both approaches showed high performance (for some fish species the precision and recall were close to one) in object classification and outperformed state-of-the-art methods. In addition, despite the fact that dataset is unbalanced in the number of images per species, both methods (especially the one operating on still images) appear to be rather robust against the long-tail curse of data, showing the best performance on the less populated object classes.

**Keywords** Object classification · Marine ecosystem analysis · Environmental monitoring

## 1 Introduction

In the last decades, the extensive research on object recognition has mainly focused on distinguishing object classes which are visually dissimilar [8, 10]. Recently, computer vision researchers have put a lot of effort into recognizing sub-ordinate object classes, a.k.a. *fine-grained object recognition*, as this poses several challenges because of the lack of strongly-discriminative features.

Automated visual systems performing such tasks might provide significant support to many applications, especially those requiring specialized domain knowledge (e.g. ecology): indeed, most people can easily distinguish between a person *playing a clarinet* from one *holding a clarinet* [18], while it is much more difficult to distinguish between plant types or animal species, where inter-class similarity might be very high. Moreover, especially for the ecological context, the need for such automatic tools has become even greater due to technological advances (remotely-operated vehicles, stationary non-invasive cameras, as well as the overall reduction of device costs) which led to the collection of massive datasets, whose analysis requires automated methods as it cannot be done by human operators [3].

However, to perform automatic or semi-automatic visual tasks, it is first necessary to have large annotated datasets, whose collection requires human efforts, specialized domain knowledge (especially for fine-grained object recognition) and often money, as shown by several recent methods exploiting platforms such as Amazon Mechanical Turk.

The proposed work addresses both the problem of automatically annotating and performing fine-grained visual tasks in domain-specific datasets. In particular, we will focus on underwater visual data using a large fish image dataset built by processing [28] over one million 10-minute video clips taken for fish biodiversity monitoring within the *Fish4Knowledge project*.[1]

[1] www.fish4knowledge.eu

The difficulties of fish-species classification in our dataset lies in the fact that fish species, especially within the same family, differ by few phenotypic features and even images of different fish species may look very similar because of 1) fish high deformability, 2) light propagation in water, and 3) low resolution of the recorded videos (due to network limitations in the monitoring stations). Fig. 1 shows three samples of *Acanthurus nigrofuscus*, *Chromis margaritifer* and *Dascyllus reticulatus* fish species, comparing how they appear in high-resolution images and in our dataset.

The main contributions of our work are:

– Two fine-grained object classification approaches operating, respectively, on still images and on videos;
– The introduction and the release of a new fine-grained fish image dataset, which complement the existing fine-recognition benchmarks [26, 37] and provides a useful basis to support marine life and its biodiversity investigation.

The remainder of the paper is organized as follows: in Section 2 a reviews of the most recent fine-grained object classification methods is carried out. Section 3 describes the employed underwater visual dataset, while fine-grained object recognition methods are reported in Section 4 with the performance described and discussed in Section 5. Concluding remarks and future developments are given in Section 6.

## 2 Related work

Recently, the computer vision, machine learning and multimedia scientific communities have addressed with increasing interest the problem of *fine-grained recognition*, i.e. categorizing objects which look very similar and differ only for subtle details, such as recognizing different species of animals (e.g dogs [26], birds [5]) and plants [19]. Of course, this task represents is rather challenging, because the discriminative features among the object classes are more difficult to identify; besides, one of the consequences of the novelty (and

**Fig. 1** From left to right, examples of *Acanthurus nigrofuscus*, *Chromis margaritifer* and *Dascyllus reticulatus* fish species: how they appear in an encyclopedia (*first row*) and how they appear in our dataset (*second row*)

the complexity) of this task is the lack of datasets to be used for training and testing machine learning approaches. One example of fine-grained recognition is fish species identification at the species level. This task, *per se*, is not more difficult than other animal-based recognition ones, but their application in "real-life" context makes it very challenging because of the high variability of fish appearance and the limitation of the employed underwater imaging devices [12, 13].

Several techniques have been specifically devised to deal with fine-grained recognition [5, 9, 17, 35, 36], mostly focusing on the discovery of visual features which are more discriminative at the subordinate level. In [17], a combination of visual cues extracted from training images is used to build discriminative compound words. In [7], image patches are considered as discriminative attributes and a Conditional Random Field (CRF) framework is used to learn the attributes on a training set with humans in the loop.

Dense-sampling techniques have also been explored which decompose an image into patches and then extract low-level features from these regions to identify fine image statistics. In [38] the authors introduce "Grouplet", a set of generative local dense features, which work reliably for human activity recognition. As follow-up, the same authors, in [18], improved the performance of their former method by fusing global and local information and using a random forest–based approach to extract dense spatial information. The limitation of these methods is their efficiency, as dense sampling feature spaces are often huge and increase many-fold when employing multiple image patches of arbitrary size. In addition, they are not able to operate with low-quality images, where the finer details necessary to pick the subtle differences between fine-grained object classes are missing.

An added difficulty for the existing methods is the lack of extensive training data: for example, the Caltech-UCSD Birds (CUB-200) [5] dataset contains only 6000 images of 200 different bird species, i.e. 30 images per bird species on average. It is proven [33] that large training datasets allow for effective nonparametric object recognition methods. However, the manual collection of large scale annotated datasets is a complex, tedious and expensive task, and although disparate techniques have been proposed for automatic labeling in the case of basic object recognition [14], the line of research that seems to be more effective for fine-grained recognition still involves the presence of human operators in the annotation loop: fine-grained labels are much more difficult to acquire and the automatic selection of discriminative features often results in the selection of irrelevant features. A promising direction is resorting to crowdsourcing solutions, as in [6], whose application has proved to be effective in discriminative feature selection, avoiding overfitting issues. However, in the case of fine-grained object recognition, the annotation process may require a deep understanding of the specific domain by the human operator (it is not feasible to ask non-experts to distinguish fish or birds or plants of the same family), thus limiting the applicability of crowdsourcing approaches.

## 3 Dataset

### 3.1 Underwater image and video dataset

For the experiments described in this paper, we employed the following datasets:

– The *U-20M* dataset contains about 20 million unannotated (hence the *U* in the name) underwater images. This dataset represents the main body of images employed in our experiments.

- The *MA-35K* dataset is a subset of *U-20M* containing about 35,000 fish images belonging to 10 chosen species. Each image was manually annotated (*MA*) with the corresponding species.
- The *AA-1M* dataset contains about 1 million automatically-annotated (*AA*) images. This dataset is a subset of the images in *U-20M* which belong to the classes annotated in *MA-35K*.
- The *FishCLEF-31K* dataset consists of about 30,000 images extracted from the *AA-1M* one and was used for the fish task [29] of the LifeCLEF 2014 benchmarking initiative [16]. Together with the fish images, the videos (about 400 10-minute videoclips) these images were extracted from, were also released.

### 3.2 Dataset collection

The *U-20M* dataset represents a randomly-selected fraction of the data collected in the context of the *Fish4Knowledge* project, amounting to more than a billion fish images extracted by processing over one million videos. In particular, these images contain generally one fish per image and represent the bounding boxes of the fish as detected by background modeling approaches [1]. The number of fish species in *U-20M* is unknown as the dataset is unlabeled, however, during the Fisk4Knowledge project it was noted that the 99.9 % of the observed fish belong to only 10 species (and for the remaining species it was very difficult to gather a significant number of image samples).

To create the *U-20M* dataset we, first, selected from the one-billion dataset only one image per trajectory (extracted via object tracking [30]) in order to avoid near duplicate images. On average each trajectory consisted of about ten fish detections and, by selecting only one image per trajectory we reduced the original dataset from one billion to about 150M images. The U-20M dataset was generated by randomly (uniform pdf) selecting 20M images from this last set. It is also important to notice that the random selection keeps the fish species distribution approximately constant.

The *MA-35K* dataset is a subset of the *U-20M* dataset, containing only (but not all) fish images belonging to the 10 most common species. Image annotation was carried out manually and validated by two expert marine biologists. In the original dataset, some species were more common than others showing the long-tail issue of data; although we tried to make the dataset as uniformly distributed across species as possible, for some of them (most evidently, *Lutjanus fulvus*) it was quite difficult to find a large number of adequate images, which resulted in a lower presence in the dataset. Figure 2 shows sample pictures of the chosen 10 species.

The *AA-1M* dataset is also a subset of *U-20M*, obtained by the semi-automatic annotation approach described in [11]: briefly, images from *MA-35K* are used as queries to a similarity-based search in *U-20M*; after a check for false positives, achieved by means of mutual similarity in the retrieved images, the resulting images are then assigned the same species label as the query image.

The *FishCLEF-31K* dataset is obtained by selecting randomly about 31K images from *AA-1M*. We did not use either the *AA-1M* dataset since training classifiers on such a big dataset is impractical and computationally expensive and or the *MA-35K* one because it contains many near duplicate images. Table 1 shows the distribution of the images for the ten fish species and its partitioning as training and test set for the LIFECLEF 2014 contest.

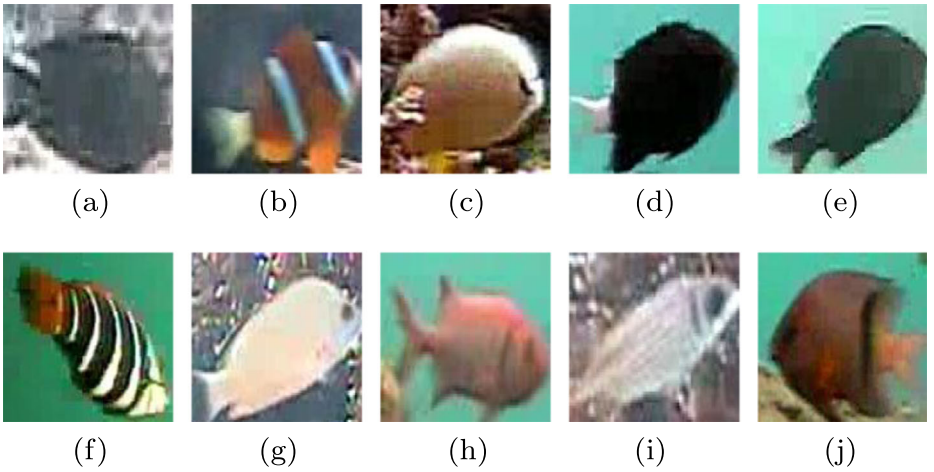



**Fig. 2** The 10 fish species analyzed in this work (images taken directly from the *MA-35K* dataset). From left to right and top to bottom: *Acanthurus nigrofuscus*, *Amphiprion clarkii*, *Chaetodon lunulatus*, *Chromis margaritifer*, *Dascyllus reticulatus*, *Hemigymnus fasciatus*, *Lutjanus fulvus*, *Myripristis kuntee*, *Neoniphon sammara*, *Plectrogly-Phidodon dickii*

For video-based fish species recognition, we employed the videos the 31K images derive from, i.e. 401 10-minute videos.

All the datasets used in this paper are publicly available at www.perceive.dieei.unict.it/datasets/fish_recognition.

## 4 Fine-grained fish species recognition

In this section we present two approaches for fish species recognition in still images and in videos. In the case of videos, before classification, fish identification was performed by means of a background modeling approach.

**Table 1** Distribution of fish species in the dataset used for the fish task of the LifeCLEF 2014 contest

These fish images were extracted from 401 10-minute videos, which were also provided as part of the video-based fish identification task

| Species | Training Set | Test Set |
|---|---|---|
| *Acanthurus nigrofoscus* | 2,511 | 725 |
| *Amphiprion clarkii* | 2,985 | 878 |
| *Chaetodon lunulatus* | 2,494 | 917 |
| *Chromis margaritifer* | 3,282 | 371 |
| *Dascyllus reticulatus* | 3,196 | 681 |
| *Hemygimnus fasciatus* | 2,224 | 852 |
| *Lutjanus fulvus* | 720 | 146 |
| *Myripristis berndti* | 2,554 | 840 |
| *Neoniphon sammara* | 2,019 | 969 |
| *Plectrogly-Phidodon dickii* | 2,456 | 577 |
| Total | 24,441 | 6,956 |

### 4.1 Image-based fish species identification

Conversely to other fish-recognition approaches [12, 13], our fine-grained fish recognition performs a multi-scale analysis: from a fast one layer encoding-pooling scheme based on MB-LTP robust to illumination/color variation transformations to a fine-grained analysis via Fisher vectors and sparse coding trained on SIFT local features. The multi-scale aggregation via late fusion permits to reach a high level of classification rate (see Section 5). More in detail, our method is based on the classic unsupervised pipeline (see [20, 24, 25]), which consists of the following three steps:

- local feature extraction
- patch encoding
- pooling operation (with a potentially spatial pyramid to improve local analysis)

These three steps constitutes a layer. Layers can be stacked together to obtain a more global representation. On the one hand, the more layers are stacked, the more the recognition system is invariant to complex image transformations, but on the other hand, if the number of layer is high, discriminability between classes can be lost. In practice, the number of layer is tuned to capture main image transformations common to all classes. On top of this architecture, a large-scale supervised classification is used, typically *via* linear-SVM or logistic regression algorithms.[2] In most computer vision approaches, only the encoding part is trained unsupervised from a random subsample set of local features and pooling parameter is fixed *ad-hoc*. In our work, we fixed the spatial pyramid of the last layer to $1 \times 1 + 2 \times 2$ representing a total of $1 + 4$ ROI to pool codes. In most of cases, the first layer is not trained but fixed according to some neuro-vision considerations. For example, SIFT patches [21] can be seen as a fixed sparse coding scheme of local gradient orientations where the sparsity level is fixed to 2 (see [4]) followed by a weighted pooling over local $4 \times 4$ windows. Local Binary Patterns [23] can be also considered to the sparse coding of local binary patterns where sparsity level is fixed to 1 and pooling is performed by histogram processing. For the image-based fish recognition, we devised three approaches: the first one using the 1-layer approach, and the other two exploiting $2-$layer hybrid architecture.

#### *4.1.1 Direct 1-layer approach*

The first approach (corresponding to the first run in the results presented in the next section) is associated with a 1-layer architecture based on the approximated Local Ternary Patterns (LTP) propsed by Tan *and al* (see [32]). In many computer-vision applications, especially in face detection/recognition, Local Binary Patterns (LBP) and derivatives are known to offer very good results for a modest price. In [32], LTP codes are approximated by the aggregation of two LBP codes. We extended the direct LTP formulation to a multi-scale version where binary codes are computed over block of $s$ pixels square instead of one unique pixel. We performed analysis with 3 block'size $s \in \{1, 2, 3\}$. We also compute LTP codes for each RGB color channels. The total feature size is $2 \times 256 \times 3 \times 3 \times 5 = 23,040$. As mentioned earlier, this architecture is fixed and no training phase is required. The processing takes less than $0.05s$ per image on a modern laptop.

[2] If main image's transformations are captured during the stacked/deep feature extraction pipeline, a non-linear classification is not improving results in practice.

#### 4.1.2 Stacked 2-layer approach

The second and third approaches are based on a 2−layer hybrid architecture where the first one employs either SIFT or LTP patches densely grid sampled and the second layer uses Fisher Vector (FV) [27] or Sparse Coding (SC) [39] framework as encoder.

In particular, the second method consists of a fusion between the approach describe above and Fisher Vector representations. Specifically, we sampled $N = \Delta_x \times \Delta_y = 25 \times 25 = 625$ SIFT patches per scale and per color channel, each of them computed over a $M = 24$ pixels square block. 3 scales is used $\sigma \in \{0.5, 0.75, 1\}$ and the 3 RGB color channel. In order satisfy the diagonal assumption of the inverse of the Fisher Information Matrix (FIM), a PCA is performed on extracted SIFT patches for each scale and for each color channel reducing dimension from 128 to 80. Following [27], we aggregated to each obtained patch, their normalized coordinates of the middle of the patch. The encoding part associated FV consists in computing local gradient *wrt.* means and variances of the log-likelihood with Gaussian Mixture modelling. We fixed $G = 32$ gaussians and used 300,000 local patches to train the GMM. The global FV representation is obtained by average pooling on each earlier local representation over each windows of the spatial pyramid and for each scale and color channel. The total feature size is $(80 + 2) \times 2 \times 32 \times 3 \times 3 \times 5 = 236,160$. Late fusion is performed by averaging posterior probabilities obtained during linear SVM training.

The last method instead fuses LTP, FV and SC representations. For SC, LTP patches instead of SIFT are chosen given their better representation capacity (see [25]). LTP patches are extracted densely in the same way as for the SIFT ones over ROI of 24 pixel square for 3 different scales and for the 3 RGB channels. Dictionaries for SC are trained for each scale and color channel. We used the classic dictionary learning procedure alternating dictionary codebook optimization with sparse codes updates. We employed block-coordinates descent (see [22]) approach for the codebook optimization adding an extra positivity constraints for each dictionary's elements. Sparse codes are updated by LARS algorithm given the current estimated dictionary adding an extra positivity constraints on sparse codes. The number of dictionary elements was fixed to $K = 1,024$ and used 300,000 local features to train each dictionary. The total feature size is $1,024 \times 3 \times 3 \times 5 = 46,080$. For the pooling stage, we used $\ell_p$-norm pooling fixing $p = 3$. It showed that $p = 3$ is offering best results over average pooling ($p = 1$) and max-pooling ($p = \infty$).

Figure 3 describes the pipelines employed for the image-based fish species recognition task.

### 4.2 Video-based fish species identification

Our video-based fish identification consists of two steps: 1) key-points are extracted from the fish images provided in the training set in an off-line fashion, and 2) the precomputed key-points are matched against dense key-points extracted from candidate fish (extracted by a background modeling approach) for fish species classification. The flowchart of this step is shown in Fig. 4.

**Off-line extraction of key-points from fish images** For each video frame in the training set, we compute three groups of key-points, using the Opponent SIFT color descriptor [40] at different scales, on the central horizontal axis and located at the 1/3, 1/2 and 2/3 of the horizontal length. Scales are computed starting from fish bounding box size (provided in

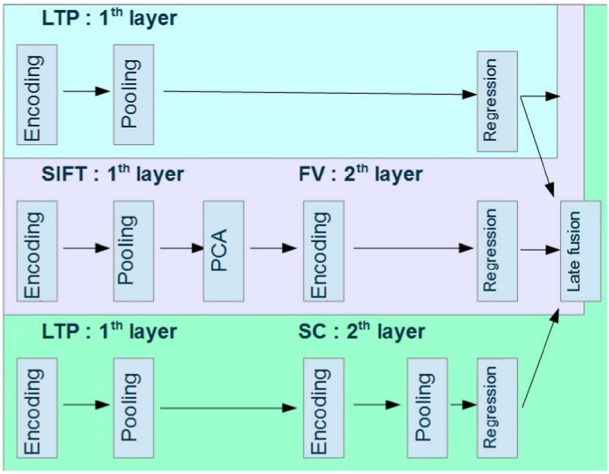



**Fig. 3** Summary of the approaches devised for fish species recognition: 1) Approach 1 provides one pipeline results (*light blue* block), b) Approach 2 aggregates two pipeline results by late fusion (*light blue* + *purple* blocks) and 3) Approaches 3 aggregates three pipeline results (*light blue* + *purple* and *green* blocks) by late fusion

the ground truth) suitably increased or decreased in order to exactly contain the whole fish and its main parts: i.e. head, body and tail. We adopt this strategy of hand-crafted large key-point descriptors owing to the low resolution of the videos and because classical detectors give generaly smaller key-points. The detection of key-points is described in Fig. 5.

**On-line video-based fish species identification** The first step of our online module detects moving objects from background-foreground segmentation maps obtained by the adaptive background mixture model described in [40]. These masks are then post-processed with morphological operators: an circular erosion of radius 3 then a circular dilation of same radius.

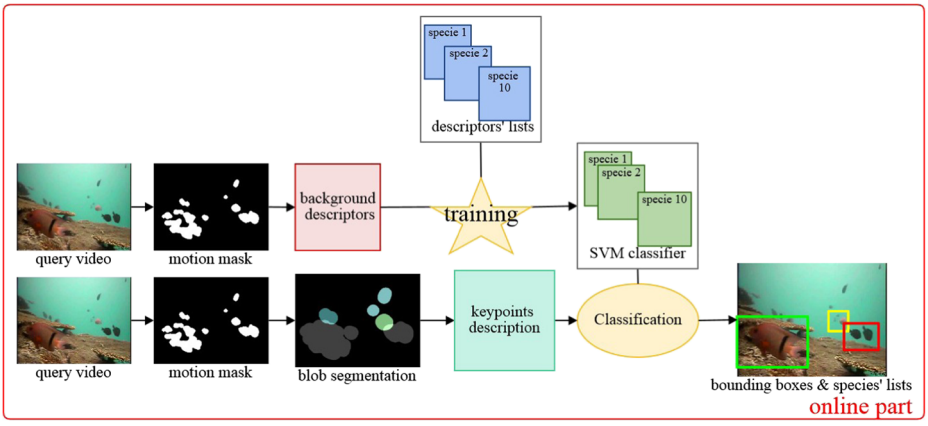



**Fig. 4** The video-based fish identification flowchart: on-line module

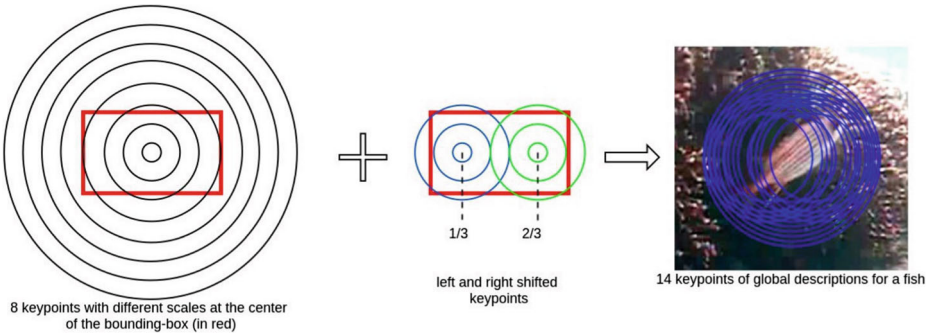



**Fig. 5** Key-point detection in fish images

In the training phase, starting from the background/foreground masks we build bounding boxes for the detected fish and then we train one SVM classifier for each fish species with the descriptors of detected fish as positive samples and Opponent SIFT descriptors of the background as negative samples. Specifically, key-points are densely extracted from the background with fixed scales from 30 to 110 pixels of diameter with respect to the size of the video. In order to avoid that background key-points (due to large bounding boxes) are used for the SVMs training, we filter out positive key-points: for each key-point inside the bounding boxes, we look for its 10 nearest neighbors and keep the best key-points (with more positive neighbors) to reach 50 percent of positive extracted descriptors.

In the test phase, for each detected blob from a video, key-points are densely extracted at several scales (fixed scales from 30 to 110), and likelihood scores are computed as the distances between the descriptors and the SVM decision boundary of each fish class. Only positively classified points with a distance larger than a threshold 1 are considered and the scores associated with each fish species are summed up in order to obtain a global likelihood score per species per blob: the fish species with the highest score is then assigned to the blob.

## 5 Performance evaluation

The approaches described in Section 4 were tested on the *FishCLEF-31K* dataset within the fish task of the LifeCLEF 2014 benchmarking initiative. In particular, the LifeCLEF 2014 Fish task (a.k.a. FishCLEF 2014) aimed at benchmarking automatic fish detection and recognition methods by processing underwater visual data. It, basically, consisted of two tasks:

**A video-based task** – detecting fish instances in key video frames and recognizing their species;
**An image-based task** – to identify fish species using only still images containing only one fish instance.

**Baseline** In order to provide a baseline for our fish species classification methods we tested 1) the VLfeat BoW [34] classification method (generally used as baseline for fine-grained recognition tasks [26]) over our *FishCLEF-31K* dataset for the task of recognizing fish in

still images and 2) the ViBe background modeling approach [1] for fish detection in videos (it proved to operate effectively with underwater videos [3]) followed by the VLFeat+BoW for fish species recognition. In both cases, performance evaluation was carried out by computing average precision and recall, and precision and recall for each fish species and the results are shown in Table 2. In particular, for the image-based task only the recall was assessed (since precision was one for all the considered species), while for the video-based task we also computed the precision as the probability of identifying background areas as fish. When computing precision and recall for both tasks we taken into account only the most probable class for each image.

**Image-based fish species recognition** [15] This section reports the performance achieved by the three methods described in Section 4.1 on the image-based task. Each method employed different features, namely: 1) only LTP 2) LTP + Improved Fisher Vectors (IFV) and 3) LTP + IFV + SC. For all the three methods, the last global representation is pooled on $1 \times 1 + 2 \times 2$ spatial pyramid. The results, in terms of recall, of the three approaches are reported and compared to our baseline in Fig. 6. As for the precision, the approach yielded a precision of 1 for almost all species expect for *Chromis margaritifer* (0.96), *Dascyllus reticulatus* (0.97) and *Plectrogly-Phidodon dickii* (0.98) species. Form recall performance, it is possible to notice that the fast LTP approach (1-layer architecture) outperforms the baseline by a large margin (except for the first specie), i.e. The other two approaches (2 and 3) slightly improved results because only few examples were misclassified by the first one.

**Video-based Fish Species Recognition** [2] The results, in terms of precision and recall, achieved by the video-based fish species recognition approach, described in Section 4.2, are shown in Figs. 7 and 8. For this task, the approach first discriminates fish from background and then assigns to each detected fish a species. A detection was considered as a true positive if the PASCAL score between it and the corresponding object in the ground truth was over 0.3. The employed parameters were $T = 0.5$ and $M = 10$ (see Section 4.2). Also for this task, we used three different settings based on how blobs detected by the fish detection

**Table 2** Fine grained classification baselines on the *FishCLEF 31K* dataset

| Fish Species | Precision—Recall Image | Precision—Recall Video |
|---|---|---|
| *Acanthurus nigrofoscus* | 0.96 – 0.99 | 0.61 – 0.67 |
| *Amphiprion clarkii* | 0.98 – 0.90 | 0.83 – 0.62 |
| *Chaetodon lunulatus* | 0.94 – 0.92 | 0.79 – 0.55 |
| *Chromis margaritifer* | 0.96 – 0.90 | 0.52 – 0.61 |
| *Dascyllus reticulatus* | 0.91 – 0.88 | 0.46 – 0.53 |
| *Hemygimnus fasciatus* | 0.99 – 0.92 | 0.34 – 0.49 |
| *Lutjanus fulvus* | 0.99 – 0.93 | 0.43 – 0.47 |
| *Myripristis berndti* | 0.99 – 0.89 | 0.58 – 0.44 |
| *Neoniphon sammara* | 0.98 – 0.92 | 0.17 – 0.34 |
| *Plectrogly-Phidodon dickii* | 0.97 – 0.89 | 0.68 – 0.61 |
| Average | 0.97 – 0.91 | 0.54 – 0.54 |

Performance, in terms of average and per species precision and recall, of our baseline approaches, respectively, on the image-based (second column) and video-based (third column) task

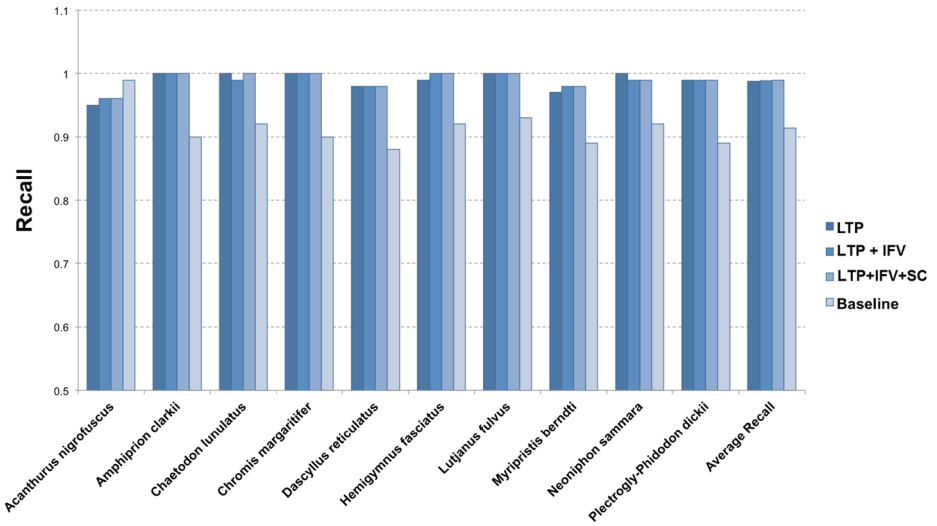



**Fig. 6** Image-based fish species recognition results [15] in terms of recall on the *FishCLEF 31K* dataset

modules were treated: 1) blobs as coming out from the background/foreground segmentation mask; 2) fish occlusion (more fish in one bounding box) separated by resorting to color features and 3) small blobs with a spatial and color coherence merged.

While the average recall obtained by our method was lower than the baseline's recall, the precision was much improved, thus implying that our classification approach based on key-points was more reliable than the fish detection baseline [1]. The reason behind the low recall may be found in the size of the computed bounding boxes when processing a video (see Fig. 9): bigger bounding boxes may contain also background objects and/or other fish instances, thus affecting the performance of the species classification method.

To demonstrate the effectiveness of this method in discriminating fish species, we adapted it to work with still images and compared (see Fig. 8) its performance to the

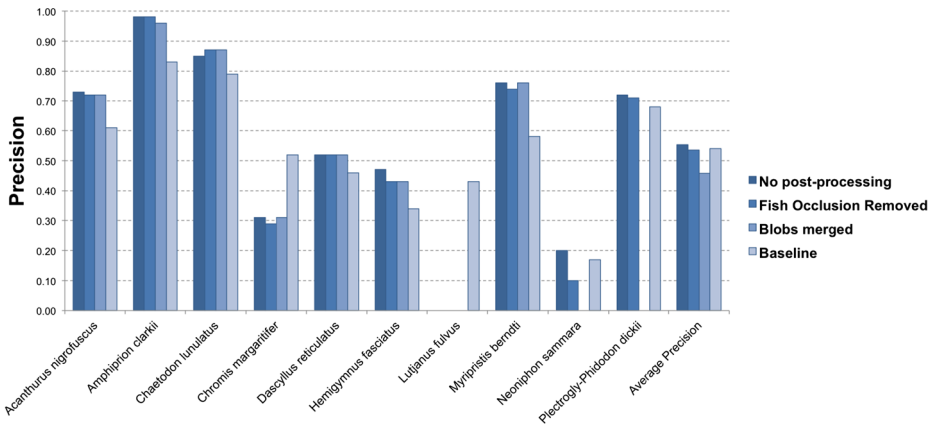



**Fig. 7** Precision of our video-based fish species recognition results [2] on the *FishCLEF 31K* dataset

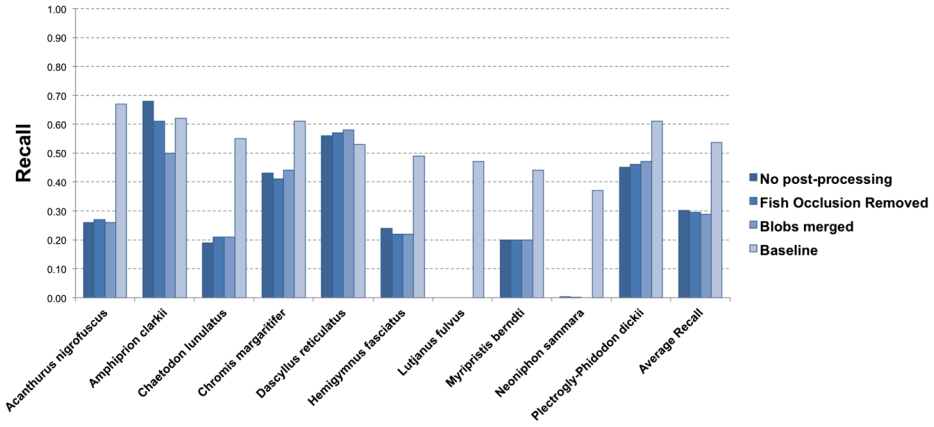



**Fig. 8** Recall of our video-based fish species recognition results [2] on the *FishCLEF 31K* dataset

ones achieved by, respectively, the method devised for the image-based task and the baseline. The results are shown in Table 3: the approach employed for the image-based task is still the most performing one, followed by the one developed for the video-based task, which was able to outperform a powerful approach such as VLFeat+BoW. However, the approach based on the extraction of keypoints is more efficient, though less accurate, than the approaches based on LPB, and this makes it particularly suitable for applications where efficiency is a key requirement.

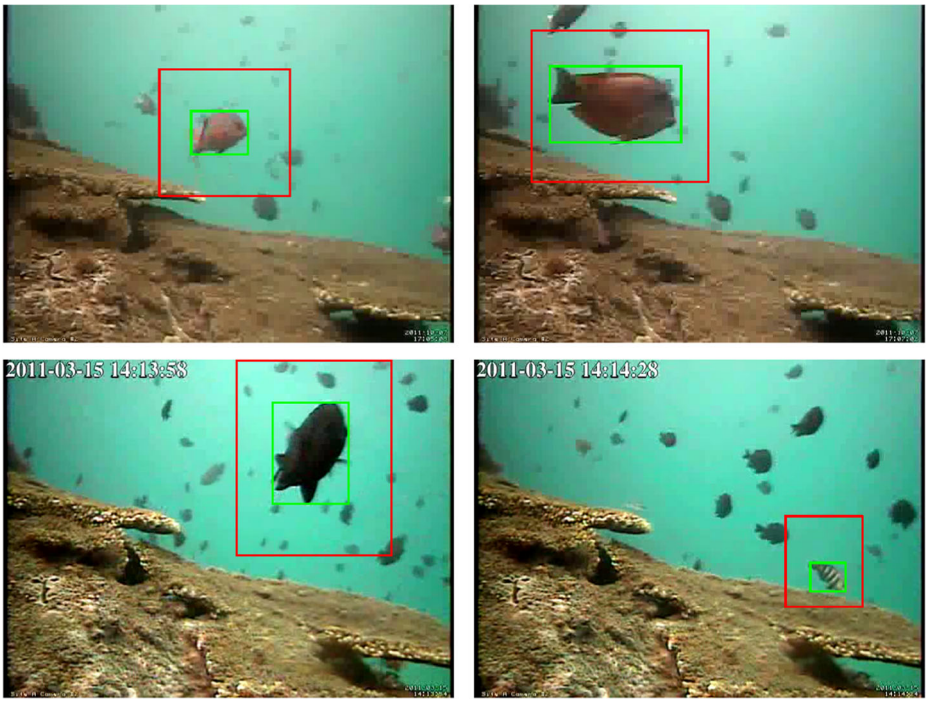



**Fig. 9** Matching between the bounding boxes computed by our video-based fish identification method [2] (in *red*) and the ones (in *green*) provided in the *FishCLEF 31K* dataset

**Table 3** Overall fine grained classification accuracy

| Fish Species | Image-based approach [15] | Video-based approach [2] | Baseline |
|---|---|---|---|
| *Acanthurus nigrofoscus* | 0.96 | 0.85 | 0.99 |
| *Amphiprion clarkii* | 1 | 0.88 | 0.90 |
| *Chaetodon lunulatus* | 1 | 0.91 | 0.92 |
| *Chromis margaritifer* | 1 | 0.99 | 0.90 |
| *Dascyllus reticulatus* | 0.98 | 0.94 | 0.88 |
| *Hemygimnus fasciatus* | 0.99 | 0.92 | 0.92 |
| *Lutjanus fulvus* | 1 | 0.91 | 0.93 |
| *Myripristis berndti* | 0.97 | 0.98 | 0.89 |
| *Neoniphon sammara* | 1 | 0.98 | 0.92 |
| *Plectrogly-Phidodon dickii* | 0.99 | 0.98 | 0.89 |
| Average | 0.99 | 0.93 | 0.91 |

Comparison (in terms of recall) between the approach devised for dealing with still images [15] (second column), the video-based fish identification approach (third column) [2] adapted to operate with still images and the baseline (VLFeat+BoW) (forth column)

## 6 Concluding remarks

In this paper we have introduced a large dataset for the fine-grained recognition problem of identifying fish species from images and videos. We have developed an effective nonparametric approach for automatic label propagation. The automatically-labeled dataset (suitably filtered) was used for benchmarking fish species recognition approaches within the fish task of the LifeCLEF 2014 initiative. Two methods for fish species recognition were also described: the first one dealing with still images and the second one with videos. Both methods achieved very high performance, although dealing with videos is much more complex as it needs reliable methods for detecting moving objects [31]. Our dataset at the moment contains only 10 fish species representing the most observed species in our underwater visual dataset. This, of course, simplifies our fine-grained recognition problem compared to other fine-grained classification benchmarks that contain much more classes, e.g. plant species [19] or bird recognition [5]. However, the extremely low image quality (the aforementioned benchmarks for fine-grained recognition contain only high resolution images) makes the classification task not trivial, especially in the case of videos. We are currently working on increasing the number of species up to 100 in order to see how the proposed methods behave in case of long-tail data distribution and to have fish families sharing common phenotypes, thus making the fine-grained classification task much more complex .

**Acknowledgments** We thank the Ministére du Redressement Productif (DGCIS) for the support to the RAPID PHRASE project, and the BPI, PACA, TPM for the FUI14 SYCIE project.

## References

1. Barnich O, Van Droogenbroeck M (June 2011) Vibe: A universal background subtraction algorithm for video sequences. IEEE Trans Image Process 20(6):1709–1724

2. Blanc FPK, Lingrand D (2014) Fish species recognition from video using SVM classifier, in *LifeClef'14 - Proceedings*, http://www.imageclef.org/2014/lifeclef/fish
3. Boom BJ, He J, Palazzo S, Huang PX, Beyan C, Chou H-M, Lin F-P, Spampinato C, Fisher RB (2014) A research tool for long-term and continuous analysis of fish assemblage in coral-reefs using underwater camera footage. Ecological Informatics 23(0):83–97
4. Boureau Y (2012) Learning hierarchical feature extractors for image recognition, Ph.D. dissertation, New York University
5. Branson S, Wah C, Schroff F, Babenko B, Welinder P, Perona P, Belongie S (2010) Visual recognition with humans in the loop. In: 11th European Conference on Computer Vision, vol 6314. Springer, pp 438–451
6. Deng J, Krause J, Fei-Fei L (2013) Fine-grained crowdsourcing for fine-grained recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 580–587
7. Duan K, Parikh D, Crandall D, Grauman K (2012) Discovering localized attributes for fine-grained recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3474–3481
8. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338
9. Farrell R, Oza O, Zhang N, Morariu V, Darrell T, Davis L (2011) Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp 161–168
10. Fei-Fei L, Fergus R, Perona P (2003) A bayesian approach to unsupervised one-shot learning of object categories. In: Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2, ser. ICCV '03, pp 1134–1141
11. Giordano D, Kavasidis I, Palazzo S, Spampinato C (2015) Nonparametric label propagation using mutual local similarity in nearest neighbors. Comp Vision Image Underst 131:116–127
12. Huang P, Boom B, Fisher R (2013) Underwater live fish recognition using a balance-guaranteed optimized tree, in Computer Vision ACCV 2012, ser. Lecture Notes in Computer Science. In: Lee K, Matsushita Y, Rehg J, Hu Z (eds), vol 7724. Springer, Berlin Heidelberg, pp 422–433. [Online]. Available:, doi:10.1007/978-3-642-37331-2_32
13. Huang P, Boom B, Fisher R (2015) Hierarchical classification with reject option for live fish recognition. Mach Vis Appl 26(1):89–102
14. Jeon J, Lavrenko V, Manmatha R (2003) Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR '03), pp 119–126
15. Joalland P, Paris S, Glotin H (2014) Efficient instance-based fish species visual identification by global representation, in *LifeClef'14 - Proceedings*, http://www.imageclef.org/2014/lifeclef/fish
16. Joly A, Muller H, Goeau H, Glotin H, Spampinato C, Rauber A, Bonnet P, Vellinga W, Fisher B (2014) Multimedia life species identification challenges. In: Proceedings of CLEF 2014, vol 1
17. Khan FS, van de Weijer J, Bagdanov AD, Vanrell M (2011) Portmanteau vocabularies for multi-cue image representation. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger K (eds) Advances in Neural Information Processing Systems (NIPS 2011), pp 1323–1331
18. Khosla A, Yao B, Fei-Fei L (2014) Integrating randomization and discrimination for classifying human-object interaction activities, in Human-Centered Social Media Analytics
19. Kumar N, Belhumeur PN, Biswas A, Jacobs DW, Kress WJ, Lopez I, Soares JVB (2012) Leafsnap: A computer vision system for automatic plant species identification. In: The 12th European Conference on Computer Vision (ECCV)
20. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol 2, pp 2169–2178
21. Lowe D (1999) Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol 2, pp 1150–1157
22. Mairal J, Bach F, Ponce J, Sapiro G (2009) Online dictionary learning for sparse coding. In: ICML '09
23. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7):971–987
24. Paris S, Halkias X, Glotin H (2012) Sparse coding for histograms of local binary patterns applied for image categorization: Toward a bag-of-scenes analysis. In: 21st International Conference on Pattern Recognition (ICPR), pp 2817–2820
25. Paris S, Halkias X, Glotin H (2013) Efficient bag of scenes analysis for image categorization. In: ICPRAM, pp 335–344

26. Parkhi OM, Vedaldi A, Zisserman A, Jawahar CV (2012) Cats and dogs. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp 3498–3505
27. Snchez J, Perronnin F, de Campos T (2012) Modeling the spatial layout of images beyond spatial pyramids. Pattern Recogn Lett 33(16):2216–2223
28. Spampinato C, Beauxis-Aussalet E, Palazzo S, Beyan C, Ossenbruggen J, He J, Boom B, Huang X (2014) A rule-based event detection system for real-life underwater domain. Mach Vis Appl 25(1):99–117
29. Spampinato C, Fisher R, Boom BJ (2014) CLEF working notes 2014, LifeCLEF Fish Identification Task 2014. In: Proceedings of CLEF 2014, vol 1
30. Spampinato C, Palazzo S, Giordano D, Kavasidis I, Lin F, Lin Y (2012) Covariance based fish tracking in real-life underwater environment. In: VISAPP 2012 - Proceedings of the International Conference on Computer Vision Theory and Applications, Volume 2, Rome, Italy, 24–26 February, 2012, pp 409–414
31. Spampinato C, Palazzo S, Kavasidis I (2014) A texton-based kernel density estimation approach for background modeling under extreme conditions. Comp Vision Image Underst 122(0):74–83
32. Tan X, Triggs B (2010) Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans Image Process 19(6):1635–1650
33. Torralba A, Fergus R, Freeman WT (2008) 80 million tiny images: a large data set for nonparametric object and scene recognition. IEEE Transactions of Pattern Analysis and Machine Intelligence 30(11):1958–1970
34. Vedaldi A, Fulkerson B (2010) VLFeat - an open and portable library of computer vision algorithms. In: ACM International Conference on Multimedia
35. Wah C, Branson S, Perona P, Belongie S (2011) Interactive localization and recognition of fine-grained visual categories. In: 2011 IEEE International Conference on Computer Vision (ICCV)
36. Yao B, Bradski GR, Li F-F (2012) A codebook-free and annotation-free approach for fine-grained image categorization. In: CVPR, pp 3466–3473
37. Yao B, Khosla A, Fei-Fei L (2011) Combining randomization and discrimination for fine-grained image categorization. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition, pp 1577–1584
38. Yao B, Li F-F (2010) Grouplet: A structured image representation for recognizing human and object interactions. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp 9–16
39. Yang J, Yu K, Gong Y, Huang TS (2009) Linear spatial pyramid matching using sparse coding for image classification. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 1794–1801. [Online]. Available: doi:10.1109/CVPRW.2009.5206757
40. Zivkovic Z (2004) Improved adaptive gaussian mixture model for background subtraction. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, vol 2, pp 28–31

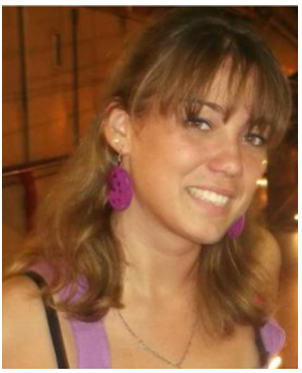
**Concetto Spampinato** received the Laurea degree in Computer Engineering in 2004, grade 110/110 cum laude, and the Ph.D. in 2008 from the University of Catania in Italy, where he is currently Research Assistant. His research interests include computer vision, pattern recognition and content based multimedia retrieval, with a particular focus on environmental applications. He has co-authored over 100 publications in international refereed journals and conference proceedings and he is member of the International Association for Pattern Recognition (IAPR).

**Simone Palazzo** received the Laurea degree in Computer Engineering in 2010, grade 110/110 cum laude from the University of Catania, Italy where he is currently doing his Ph.D. His interest activities include image and signal processing, image enhancement and reconstruction.

**P. H. Joalland** currently works at University of Toulon as research engineer. His research interests include computer vision, pattern recognition, indexation but also underwater acoustic, telecom systems, inertial navigation and industrial property. Previously, he worked at Mitsubishi Electric and Jouve as project manager and is author of two patents. He received his M.D. in Computer Engineering in 1993 from the University of Rennes.

**Sébastien Paris** University of Marseille Dr. Sébastien PARIS is associate professor at the university of Marseille (France) and LSIS laboratory - UMR CNRS 7296 since 2005. Previously, in late 2000, he got Ph.D. Thesis in signal processing, in the domain of sonar system, and more precisely about frequency line tracking for passive systems. This work was held in the SIS laboratory in Toulon (France), under the direction of Pr C. Jauffret. Then he was postdoctor in the INRIA bretagne laboratory, in the Signal Processing department from 2001 to 2002. Since 2005, he joined the LSIS lab to work in machine learning field He also participated to international evaluation about Plant identification (PLANT 2010), Fish categorization (Fish2012). He works on machine learning and computer vision for recognition/categorization tasks.

**Hervé Glotin** is professor of Computer Science at the Institut Universitaire de France (IUF) and Univ. of Toulon (UTLN) since 2010, in the Systems and Information Sciences CNRS lab. (LSIS). He is leading the DYNI team on stochastic multimodal information retrieval. He received in 2001 his PhD 'Robust adaptive multi-stream automatic speech recognition using voicing and localization cues' from Inst. of Perceptual Artificial Intelligence (IDIAP-CH) and Inst. of Spoken Communication Grenoble (INPG-FR). In 2000 he was expert at Johns Hopkins CSLP lab. with IBM human language team where he co-designed the Via-Voice audiovisual Large Vocabulary Speech Recognition system. His research today deals with multimodal information retrieval including video and scaled (bio)acoustics. He is since 10 years the general chair of ERMITES summer school / workshop on advanced multimodal information retrieval, and the general chair of several workshop in environmental data, as in NIPS, ICML, ICDM2015. He initiated in 2012 and heads the CNRS Big Data project on Scaled Acoustic Biodiversity (http://sabiod.org), involving machine learning, signal processing and bioacoustics. He is co-author of more than 100 of internat. refereed articles.



**Katy Blanc** received her M.S. degree in Modelization and Applied Math from the engineering school Polytech'Nice Sophia, at University Nice Sophia Antipolis (UNS), France, in 2014. She is currently PhD student at I3S laboratory, Joint CNRS-UNS Unit of Research 7271, France. Her research interests are mainly Computer Vision and Machine Learning for video content retrieval. Her work focuses fluid mechanics and tensors to design new video content description, and kernel design for multi-class support vector machines.

**Dr. Diane Lingrand** is associate professor at the Laboratory I3S, Joint CNRS — University of Nice-Sophia Antipolis Unit of Reseach 7271. She received her PhD in Computer Vision in 1999 from the University of Nice - Sophia Antipolis for her work on uncalibrated monocular video sequence analysis at INRIA, Sophia Antipolis. She has been postdoctoral fellow at the Montreal Neurological Institute (MNI- McGill) where she studied MNI images and EEG signals for epilepsy. Her research have focused more specially about 3D Interactions with 3D objects and 3D images, using Vision-based input devices and facilitating the Human-Computer Interactions by using information on the context. Her research interests are machine learning, mainly Support Vector Machines, for computer vision problems such as plant identification in image and metadata datasets, action recognition in videos, locations and objects recognition in 3D scenes, and perceptual vision models for image classification.

**Pr. Frédéric Precioso** is a Full Professor (since 2011) at the Engineering school Polytech'Nice, part of University of Nice-Sophia Antipolis. He is the head of MinD Group (Mining Data) in the Laboratory I3S, Joint CNRS — University of Nice-Sophia Antipolis Unit of Reseach 7271. He received his Ph.D. in Signal and Image Processing, from University of Nice-Sophia Antipolis, France, in 2004, for his work on parametric active contour segmentation methods. After a year as post-doctorate Marie-Curie Fellow at CERTH-Informatics and Telematics Institute, (Thessaloniki, Greece), where he worked on semantic methods for multimedia content understanding, he has become associate professor at Ecole Nationale Supérieure de l'Electronique et de ses Applications (ENSEA), Cergy, in 2005. He is interested, for ten years now, in video and image segmentation using active contours more recently focusing on bio-medical applications. Since 2005, his main topics of interest are image and video object detection and classification, interactive learning in this context, content-based image and video retrieval systems and scalability of such systems. He served for two years as associated member of the Technical Committee of IEEE Bio Inspired Signal Processing and for two years he is a full member of this same technical committee.