Overview of the LifeCLEF 2014 Fish Task

Concetto Spampinato[1], Simone Palazzo[1], Bas Boom[2], and Robert B. Fisher[2]

[1] Department of Electrical, Electronics and Computer Engineering
University of Catania, Italy
{cspampin, simone.palazzo}@dieei.unict.it
[2] School of Informatics, University of Edinburgh, UK
{bboom, rbf}@inf.ed.ac.uk

Abstract. This paper describes the LifeCLEF 2014 fish task, which aimed at benchmarking automatic fish detection and recognition methods by processing underwater visual data. The task consisted of video-based subtasks for fish detection and fish species recognition in videos and one image-based task for fish species classification in still images.
Our underwater visual datasets consisted of about 2,000 videos taken from the Fish4Knowledge video repository and more than 200,000 annotations automatically obtained and manually validated.
About fifty teams registered to the fish task, but only two teams submitted runs: the *I3S team* for subtask 3 and *LSIS/DYNI team* for subtask 4. The results achieved by both teams are satisfactory.

Keywords: Underwater video analysis, Object classification, Fine-grained visual categorisation

1 Introduction

Underwater video and imaging systems are used increasingly in a range of monitoring or exploratory applications, in particular for biological (e.g. benthic community structure, habitat classification), fisheries (e.g. stock assessment, species richness), geological (e.g. seabed type, mineral deposits) and physical surveys (e.g. pipelines, cables, oil industry infrastructure). Their usage has benefitted from the increasing miniaturisation and cost-effectiveness of submersible ROVs (remotely operated vehicles) and advances in underwater digital cameras. These technologies have revolutionised our ability to capture high-resolution images in challenging aquatic environments and are also greatly improving our ability to effectively manage natural resources, increasing our competitiveness and reducing operational risks in industries that operate in both marine and freshwater systems. Despite these advances in data collection technologies, the analysis of video data usually requires very time-consuming and expensive input by human observers. This is particularly true for ecological and fishery video data, which often requires laborious visual analysis. This analytical "bottleneck" greatly restricts the use of these otherwise powerful video technologies and demands effective methods for automatic content analysis to enable proactive provision of analytical information.

The development of automatic video analysis tools is, however, particularly challenging because of the complexities of underwater video recordings in terms of the variability of scenarios and factors that may degrade the video quality such as water clarity and/or depth. A recent advance in this direction was developed as part of the Fish4Knowledge (F4K)[3] project (funded by The European Union 7th Framework Programme), where computer vision and machine learning techniques were developed to extract information about fish density and richness from videos taken by underwater cameras installed at coral reefs in Taiwan [1].

However, the underwater is a rather complex environment because of several factors that make the study of it particularly complex and challenging. In particular, dynamic or multi-modal backgrounds, abrupt lighting changes (due also to water caustics), and radical and instant water turbidity changes affect the ability to perform visual tasks also for humans. To complicate even more the situation, the underwater environment shows two almost exclusive characteristics with respect to other domains: three degrees of freedom and erratic and extremely fast movements of objects (i.e. fish). This makes fish less predictable than people or vehicles, as fish may move in all three directions changing suddenly their size and their shape in the video. As a consequence of this, although in the F4K project reliable approaches for video-based fish detection and species identification were devised, the problem of automatic analysis of underwater visual data remains still open.

In this paper we describe the "Fish Task" organised as part of the LifeCLEF 2014, where video and image based approaches for fish detection and recognition were tested on underwater video/image datasets achieving very good results.

The remainder of the paper is as follows: Sect. 2 provides an overview of the task as well as the underwater video and image datasets used, Sect. 3 presents the participants to the fish task and their approaches whose results, compared to our baselines, are given in Sect. 4. Sect. 5 concludes this report.

2 Fish Task Description

The LifeCLEF 2014 Fish task aimed at benchmarking automatic fish detection and recognition methods by processing underwater visual data. It basically consisted of three video-based subtasks and one image-based task. The video-based subtasks were:

Subtask 1 – detecting moving objects in videos by either background modeling or object detection methods;
Subtask 2 – detecting fish instances in video frames, thus discriminating fish instances from non-fish ones;
Subtask 3 – detecting fish instances in video frames and recognising their species;

The image-based subtask, instead, had the goal:

[3] www.fish4knowledge.eu

Subtask 4 – to identify fish species using only still images containing only one fish instance.

The participants had to submit at most three runs for a subtask. The run file must have had the same format as the ground truth xml file (see below), i.e. it must contain the frame where the fish was detected together with the bounding box (for all the subtasks 1, 2, and 3), contours (only task 1) and species name (for subtasks 3 and 4) of the fish.

2.1 Dataset

The underwater video dataset was derived from the Fish4Knowledge video repository, which contains about 700,000 10-minute video clips that were taken in the past five years to monitor Taiwan coral reefs. The Taiwan area is particularly interesting for studying the marine ecosystem, as it holds one of the largest fish biodiversities of the world with more than 3,000 different fish species whose taxonomy is available at *http://fishdb.sinica.edu.tw*. 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. As the LifeCLEF 2014 fish task included four subtasks, we employed different datasets for each subtask. In particular, for subtask 1 we used eight videos (four for training and four for testing) fully labeled (each single fish instance was annotated) using the tool in [2] resulting in 21,106 annotations corresponding to 9,852 different fish for the training dataset and in 14,829 annotations corresponding to 899 fish for the test set. For the other three subtasks, we used the (partially overlapping) following datasets:

- The *D20M* dataset, which contains about 20 million underwater images randomly selected from the whole F4K dataset.
- The *D35K* dataset, which is a subset of *D20M* containing about 35000 fish images belonging to 10 fish species. Each image of this dataset was manually annotated with the corresponding species.
- The *D1M* dataset, which contains about 1 million automatically-annotated images by using the method in [3] and further manually checked. This dataset is meant to be the subset of the images in *D20M* which belong to the classes annotated in *D35K*.

The *D20M* dataset represents a randomly-selected fraction of the data collected within the Fish4Knowledge project, amounting to more than a billion fish images. The *D35K* dataset is a subset of the *D20M* one, containing only (but not all) fish images belonging to the 10 most common species. Image annotation was carried out manually and validated by expert marine biologists. Figure 1 shows sample pictures of the chosen 10 species.

The *D1M* dataset is also a subset of *D20M*, obtained by the automatic annotation technique described in [3] and further manually checked: briefly, images

Fig. 1. The 10 fish species analysed in this work. Please note that these images are taken directly from the *D35K* 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*.

from *D35K* are used as queries to a similarity-based search in *D20M*; after a check for false positives, the resulting images are then assigned to the same species label as the query image. We did not use directly the *D35K* dataset fish species classification, because it contained several near duplicate images.

The datasets used in the LifeCLEF 2014 fish task (subtasks 2, 3 and 4) were generated from the *D1M* dataset. In particular for subtask 2, which was meant to identify only some fish instances (unlike the subtask 1 which asked for identifying all fish in a video) in video frames, we used 112,078 annotations taken from 957 videos for the training set and 15,245 annotations from 89 videos for the test set.

For subtasks 3 and 4, we, instead, used 24,441 (belonging to 285 videos) annotated fish (and their species) in the training set and 6,956 annotations (belonging to 116 videos) in the test set. Details about the species distribution in the training and test datasets are shown in Table 1.

It is important to note that some species were more common than others; 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.

The datasets (both the training and the test ones) were provided as zipped folders containing:

- A ground truth folder where the ground truth for all subtasks were given as XML files (Fig. 2):
- A video folder where all the videos of the entire dataset were contained
- A image folder which contains the images corresponding to the bounding boxes given in the XML files. These images were provided only for subtasks 2, 3, 4 in the training phase and only for subtask 4 in the test phase.

Table 1. Distribution of fish species in the training and test sets for subtasks 3 and 4.

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

3 Participants

About 50 teams registered to the fish task, but only two teams submitted runs for the fish task: one, the *I3S team*, for subtask 3 and one, the *LSIS/DYNI team* for subtask 4.

The strategy employed by the *I3S team* for fish identification and recognition (subtask 3) consisted of, first, applying a background modeling approach based on Mixture of Gaussian for moving object segmentation. SVM learning using keyframes of species as positive entries and background of current video as negative entries was used for fish species classification.

The *LSIS/DYNI team* submitted three runs for subtask 4. Each run followed the strategy proposed in [4] which, basically, consisted of extracting low level features, patch encoding, pooling with spatial pyramid for local analysis and a linear large-scale supervised classication by averaging posterior probabilities estimated through linear regression of linear SVM's outputs. No image specific pre-processing regarding illumination correction or background subtraction was performed.

```
<video id=VIDEO_ID location=LOCATION_CAMERA date_time=DATE width=W height=H frame_rate=FPS gps=GPS>
    <frame number=NUM_FRAME>
        <object id=OBJECT_ID filename=FILE_NAME_OF_THE_IMAGE species=SPECIES_NAME> <!-- SPECIES_NAME is provided for subtask 3 only -->
            <bounding_box>COORDINATES (X,Y) OF BOUNDING BOX </bounding_box> <!-- For subtasks 2 and 3 -->
            <contour>COORDINATES (X,Y) OF CONTOUR </contour>  <!-- For subtask 1 only -->
        </object>
    </frame>
</video>
```

Fig. 2. Example of a ground truth XML file

4 Results

Performance evaluation was carried out on the released test sets by computing: average precision and recall, and precision and recall for each fish species for subtask 3; average recall and recall for each fish species for subtask 4. The baselines for subtasks 3 and 4 were, respectively:

- **Subtask 3**: The ViBe [5] background modeling approach for fish detection combined to VLFeat BoW [6] for fish species recognition
- **Subtask 4**: VLFeat BoW for fish species recognition

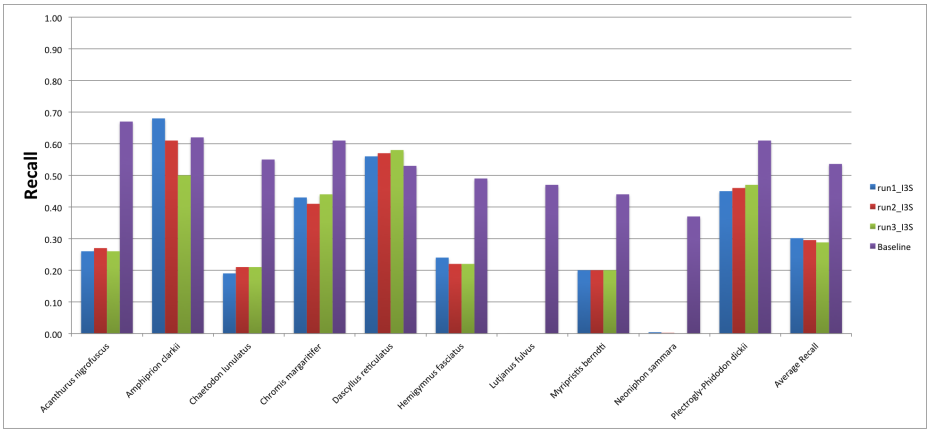

Fig. 3. Average recall and recall for fish species achieved by the *I3S team* on subtask 3 compared to our baseline

The results achieved by the *I3S team* for subtask 3 are reported in Fig. 3 and compared to our baseline. When computing the performance in terms of fish detection, we relaxed the constraint on the PASCAL score as the *I3S team* often detected correctly a fish but the bounding box' size was much bigger than the one provided in our dataset (see Fig. 4).

While the average recall obtained by the *I3S team* was lower than the baseline's recall, the precision was improved (see Fig. 5), thus implying that their fish species classification approach was reliable more than the fish detection approach. Furthermore, we allowed participants to provide a ranked list (top three species) of fish species for each fish instance in the test set and we treated a recognition as a true positive if the correct species was in the top three species. Considering only the most probable class provided in the results submitted by the *I3S team* makes the performance drop considerably. The reason behind this may be found in the size of the bounding boxes extracted by the *I3S team*. In fact bigger bounding boxes may contain also background objects and/or other

fish instances, thus affecting the performance of the fish species classification approach.

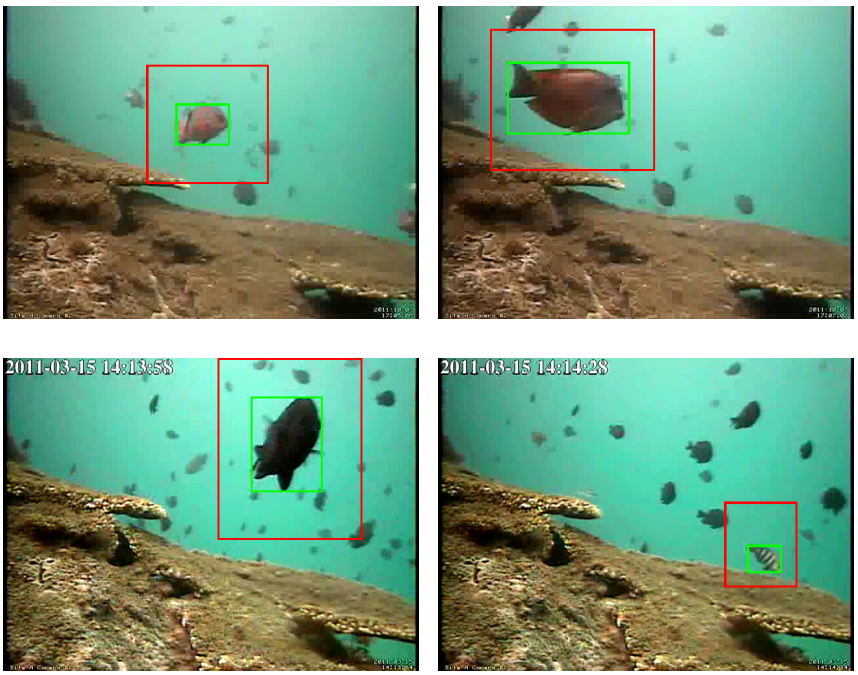

Fig. 4. Matching between the bounding boxes computed by the *I3S team* (in red) and the ones (in green) provided in our dataset

The results obtained by the *LSIS/DYNI team* for subtask 4 are shown in Fig. 6. In this case, the performance were computed by taking into account only the most probable fish species (i.e. the first class in the provided ranked list).

Please note the image-based recognition task (subtask 4) was easier than subtask 3 since it does need any fish identification module (which is the most complex part in video-based fish identification) and we had only ten fish species with very distinctive features. However, *LSIS/DYNI team* did a great job outperforming our baseline.

5 Concluding remarks

In this report, we described the LifeCLEF 2014 fish task, which aimed at benchmarking machine learning and computer vision methods for fish detection and recognition in underwater "real-life" video footage. Although the fifty teams registered for the fish task, only two teams submitted runs: the *I3S team* for subtask 3 and *LSIS/DYNI team* for subtask 4. The main challenge of our underwater

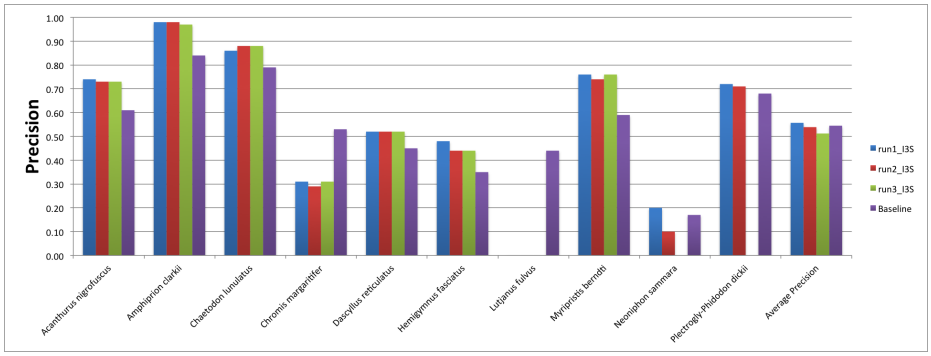

Fig. 5. Average precision and precision for fish species achieved by the *I3S team* on subtask 3 compared to our baseline

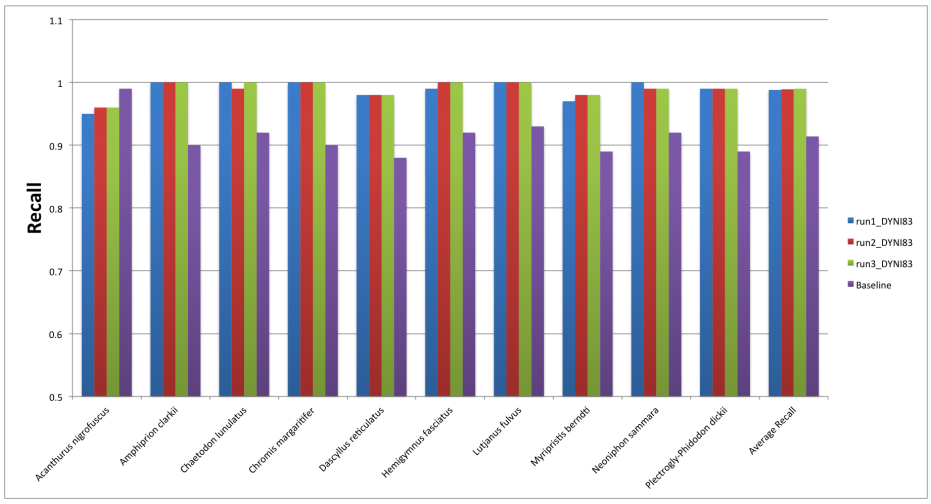

Fig. 6. Average recall and recall for fish species achieved by the *LSIS/DYNI team* on subtask 4 compared to our baseline

dataset was to identify fish correctly (as moving objects) by processing the entire video sequences. This appears evident by looking the results achieved by the *I3S team* (Fig. 3 and 5): for instance, the *I3S team* was not able to detect any of the Lutjanus fulvus instances, as it is a rather tiny fish which often tends to hide behind rocks. The subtask 4 was easier (as shown by the results achieved by our baseline) as the fish were already segmented from the background and the considered fish species have very distinctive features that make their identification in the feature space almost trivial. Nevertheless, the results achieved by the *LSIS/DYNI team* are excellent, outperforming our baseline for almost all species.

We have just completed the labelling of other 25 species and we are working on removing the near duplicates from the dataset to make the fish identification and recognition task more challenging.

We would like to thank to all teams who participated to the task and the ImageCLEF and LifeCLEF organisation who made the first edition of fish task possible.

References

1. Boom, B.J., He, J., Palazzo, S., Huang, P.X., Beyan, C., Chou, H.M., Lin, F.P., Spampinato, C., Fisher, R.B.: A research tool for long-term and continuous analysis of fish assemblage in coral-reefs using underwater camera footage. Ecological Informatics **http://dx.doi.org/10.1016/j.ecoinf.2013.10.006** (2013)
2. Kavasidis, I., Palazzo, S., Salvo, R., Giordano, D., Spampinato, C.: An innovative web-based collaborative platform for video annotation. Multimedia Tools and Applications **70**(1) (2014) 413–432
3. Giordano, D., Palazzo, S., Spampinato, C.: Nonparametric label propagation using mutual local similarity in nearest neighbors. To appear on Computer Vision and Image Understanding (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: 2012 21st International Conference on Pattern Recognition (ICPR). (Nov 2012) 2817–2820
5. Barnich, O., Van Droogenbroeck, M.: Vibe: A universal background subtraction algorithm for video sequences. Image Processing, IEEE Transactions on **20**(6) (June 2011) 1709–1724
6. Vedaldi, A., Fulkerson, B.: VLFeat - an open and portable library of computer vision algorithms. In: ACM International Conference on Multimedia. (2010)