LifeCLEF 2016: Multimedia Life Species Identification Challenges

Alexis Joly[1], Hervé Goëau[2], Hervé Glotin[3], Concetto Spampinato[4], Pierre Bonnet[5], Willem-Pier Vellinga[6], Julien Champ[1], Robert Planqué[6], Simone Palazzo[4], Henning Müller[7]

[1] Inria, LIRMM, Montpellier, France
[2] IRD, UMR AMAP, France
[3] AMU, CNRS LSIS, ENSAM, Univ. Toulon, IUF, France
[4] University of Catania, Italy
[5] CIRAD, UMR AMAP, France
[6] Xeno-canto foundation, The Netherlands
[7] HES-SO, Sierre, Switzerland

Abstract. Using multimedia identification tools is considered as one of the most promising solutions to help bridge the taxonomic gap and build accurate knowledge of the identity, the geographic distribution and the evolution of living species. Large and structured communities of nature observers (e.g., iSpot, Xeno-canto, Tela Botanica, etc.) as well as big monitoring equipment have actually started to produce outstanding collections of multimedia records. Unfortunately, the performance of the state-of-the-art analysis techniques on such data is still not well understood and is far from reaching real world requirements. The LifeCLEF lab proposes to evaluate these challenges around 3 tasks related to multimedia information retrieval and fine-grained classification problems in 3 domains. Each task is based on large volumes of real-world data and the measured challenges are defined in collaboration with biologists and environmental stakeholders to reflect realistic usage scenarios. For each task, we report the methodology, the data sets as well as the results and the main outcomes.

1 LifeCLEF Lab Overview

Identifying organisms is a key for accessing information related to the ecology of species. This is an essential step in recording any specimen on earth to be used in ecological studies. But unfortunately, this is difficult to achieve due to the level of expertise necessary to correctly record and identify living organisms (for instance plants are one of the most difficult group to identify with more than 300.000 species). This *taxonomic gap* has been recognized since the Rio Conference of 1992, as one of the major obstacles to the global implementation of the Convention on Biological Diversity. Among the diversity of methods used for species identification, Gaston and O'Neill [21] discussed in 2004 the potential of automated approaches typically based on machine learning and multimedia data analysis methods. They suggested that, if the scientific community is able to (i) overcome the production of large training datasets, (ii) more precisely identify and evaluate the error rates, (iii) scale up automated approaches, and (iv) detect novel species, it will then be possible to initiate the development of a generic automated species identification system that could open up vistas of new opportunities for pure and applied work in biological and related fields.

Since the question raised in [21], "automated species identification: why not?", a lot of work has been done on the topic [46,9,69,62,1,68,38,20,17] and it is still attracting much research today, in particular on deep learning techniques. In parallel to the emergence of automated identification tools, large social networks dedicated to the production, sharing and identification of multimedia biodiversity records have increased in recent years. Some of the most active ones like iNaturalist[8], iSpot [58], Xeno-Canto[9] or Tela Botanica[10] (respectively initiated in the US for the two first and in Europe for the two last), federate tens of thousands of active members, producing hundreds of thousands of observations each year. Noticeably, the Pl@ntNet initiative was the first one attempting to combine the force of social networks with that of automated identification tools [38] through the release of a mobile application and collaborative validation tools. As a proof of their increasing reliability, most of these networks have started to contribute to global initiatives on biodiversity, such as the Global Biodiversity Information Facility (GBIF[11]) which is the largest and most recognized one. Nevertheless, this explicitly shared and validated data is only the tip of the iceberg. The real potential lies in the automatic analysis of the millions of raw observations collected every year through a growing number of devices but for which there is no human validation at all.

The performance of state-of-the-art multimedia analysis and machine learning techniques on such raw data (e.g., mobile search logs, soundscape audio recordings, wild life webcams, etc.) is still not well understood and is far from reaching the requirements of an accurate generic biodiversity monitoring system. Most existing research before LifeCLEF has actually considered only a few douzen or up to hundreds of species, often acquired in well-controlled environments [28,50,43]. On the other hand, the total number of living species on earth is estimated to be around 10K for birds, 30K for fish, 300K for flowering plants (cf. ThePlantlist[12]) and more than 1.2M for invertebrates [3]. To bridge this gap, it is required to boost research on large-scale datasets and real-world scenarios.

In order to evaluate the performance of automated identification technologies in a sustainable and repeatable way, the LifeCLEF[13] research platform was created in 2014 as a continuation of the plant identification task [39] that was

[8] http://www.inaturalist.org/
[9] http://www.xeno-canto.org/
[10] http://www.tela-botanica.org/
[11] http://www.gbif.org/
[12] http://www.theplantlist.org/
[13] www.lifeclef.org

run within the ImageCLEF lab [14] the three years before [28,29,27]. LifeCLEF enlarged the evaluated challenge by considering birds and fishes in addition to plants, and audio and video contents in addition to images. In this way, it aims at pushing the boundaries of the state-of-the-art in several research directions at the frontier of information retrieval, machine learning and knowledge engineering including (i) large scale classification, (ii) scene understanding, (iii) weakly-supervised and open-set classification, (iv) transfer learning and fine-grained classification and (v), humanly-assisted or crowdsourcing-based classification. More concretely, the lab is organized around three tasks, each based :

PlantCLEF: an image-based plant identification task making use of Pl@ntNet collaborative data

BirdCLEF: an audio recordings-based bird identification task making use of Xeno-canto collaborative data

SeaCLEF: a video and image-based identification task dedicated to sea organisms (making use of submarine videos and aerial pictures).

As described in more detail in the following sections, each task is based on big and real-world data and the measured challenges are defined in collaboration with biologists and environmental stakeholders so as to reflect realistic usage scenarios. The main novelties of the 2016th edition of LifeCLEF compared to the previous years are the following:

1. **Introduction of new contents types**: Both the plant and the bird tasks introduced new types of contents in their respective test sets so as to focus on more automated biodiversity monitoring scenarios. The test set of the plant task was composed of the raw image search logs of the Pl@ntNet mobile application (whereas previous editions were based on explicitly shared and collaboratively validated citizen sciences data). For the bird task, the novelty was the inclusion of *soundscape recordings*, i.e. continuous recordings of a specific environment over a long period.
2. **Identification of the individual level**: Previous editions of LifeCLEF were only concerned with species identification, i.e. retrieving the taxonomic name of an observed living plant or animal. The sea task conducted in 2016, however, included an identification challenge at the individual level. For some groups, notably whales, it is actually preferable to monitor the organisms at the individual level rather than at the species level. This problem is much less studied than species recognition and, to the best of our knowledge, WhaleCLEF is the first system-oriented evaluation dedicated to this challenge in the literature.

Overall, more than 130 research groups from around the world registered to at least one task of the lab. Fourteen of them finally crossed the finish line by participating in the collaborative evaluation and by writing technical reports describing in details their evaluated system.

[14] `http://www.imageclef.org/`

2 Task1: PlantCLEF

Image-based plant identification is the most promising solution towards bridging the botanical taxonomic gap, as illustrated by the proliferation of research work on the topic [33,10,41,35,2] as well as the emergence of dedicated mobile applications such as LeafSnap [43] or Pl@ntNet [38]. As promising as these applications are, their performance is still far from the requirements of a real-world's ecological surveillance scenario. Allowing the mass of citizens to produce accurate plant observations requires to equip them with much more effective identification tools. As an illustration, in 2015, 2,328,502 millions queries have been submitted by the users of the Pl@ntNet mobile apps but only less than 1% of them were finally shared and collaboratively validated. Allowing the exploitation of the unvalidated observations could scale up the world-wide collection of plant records by several orders of magnitude. Measuring and boosting the performance of automated identification tools is therefore crucial. As a first step towards evaluating the feasibility of such an automated biodiversity monitoring paradigm, we created a new testbed entirely composed of image search logs of the Pl@ntNet mobile application (contrary to the previous editions of the PlantCLEF benchmark that were only based on explicitly shared and validated observations).

As a concrete scenario, we focused on the monitoring of invasive exotic plant species. These species represent today a major economic cost to our society (estimated at nearly 12 billion euros a year in Europe) and one of the main threats to biodiversity conservation [71]. This cost can even be more important at the country level, such as in China where it is evaluated to be about 15 billion US dollars annually [72], and more than 34 billion US dollars in the US [52]. The early detection of the appearance of these species, as well as the monitoring of changes in their distribution and phenology, are key elements to manage them, and reduce the cost of their management. The analysis of Pl@ntNet search logs can provide a highly valuable response to this problem because the presence of these species is highly correlated with that of humans (and thus to the density of data occurrences produced through the mobile application).

2.1 Dataset and evaluation protocol

For the training set, we provided the PlantCLEF 2015 dataset enriched with the ground truth annotations of the test images (that were kept secret during the 2015 campaign). In total, this data set contains 113,205 pictures of herb, tree and fern specimens belonging to 1,000 species (living in France and neighboring countries). Each image is associated with an XML file containing the taxonomic ground truth (species, genus, family), as well as other meta-data such as the type (fruit, flower, entire plant, etc.), the quality rating (social-based), the author name, the observation Id, the date and the geo-loc (for some of the observations).

For the test set, we created a new annotated dataset based on the image queries that were submitted by authenticated users of the Pl@ntNet mobile application in 2015 (unauthenticated queries had to be removed for copyright issues). A fraction of that queries were already associated to a valid species name because they were explicitly shared by their authors and collaboratively revised. We included in the test set the 4633 ones that were associated to a species belonging to the 1000 species of the training set (populating the known classes). Remaining pictures were distributed to a pool of botanists in charge of manually annotating them either with a valid species name or with newly created tags of their choice (and shared between them). In the period of time devoted to this process, they were able to manually annotate 1821 pictures that were included in the test set. Therefore, 144 new tags were created to qualify the unknown classes such as for instance *non-plant objects*, *legs* or *hands*, *UVO* (Unidentified Vegetal Object), *artificial plants*, *cactaceae*, *mushrooms*, *animals*, *food*, *vegetables* or more precise names of horticultural plants such as roses, geraniums, ficus, etc. For privacy reasons, we had to remove all images tagged as *people* (about 1.1% of the tagged queries). Finally, to complete the number of test images belonging to unknown classes, we randomly selected a set of 1546 image queries that were associated to a valid species name that do not belong to the France flora (and thus, that do not belong to the 1000 species of the training set or to potentially highly similar species). In the end, the test set was composed of 8,000 pictures, 4633 labeled with one of the 1000 known classes of the training set, and 3367 labeled as new unknown classes. Among the 4633 images of known species, 366 were tagged as *invasive* according to a selected list of 26 potentially invasive species. This list was defined by aggregating several sources (such as the National Botanical conservatory, and the Global Invasive Species Programme) and by computing the intersection with the 1000 species of the training set. Based on the previously described testbed, we conducted a system-oriented evaluation involving different research groups who downloaded the data and ran their system. To avoid participants tuning their algorithms on the invasive species scenario and keep our evaluation generalizable to other ones, we did not provide the list of species to be detected. Participants only knew that the targeted species were included in a larger set of 1000 species for which we provided the training set. Participants were also aware that (i) most of the test data does not belong to the targeted list of species (ii) a large fraction of them does not belong to the training set of the 1000 species, and (iii) a fraction of them might not even be plants. In essence, the task to be addressed is related to what is sometimes called *open-set* or *open-world* recognition problems [5,56], i.e., problems in which the recognition system has to be robust to unknown and never seen categories. Beyond the brute-force classification across the known classes of the training set, a big challenge is thus to automatically reject the false positive classification hits that are caused by the unknown classes i.e., by the distractors). To measure this ability of the evaluated systems, each prediction had to be associated with a confidence score in $p \in [0, 1]$ quantifying the probability that this prediction is true (independently from the other predictions).

The metric used to evaluate the performance of the systems is the classification Mean Average Precision (MAP-open), considering each class c_i of the training set as a query. More concretely, for each class c_i, we extract from the run file all predictions with $PredictedClassId = c_i$, rank them by decreasing probability $p \in [0, 1]$ and compute the average precision for that class. The mean is then computed across all classes. Distractors associated to high probability values (i.e., false alarms) are likely to highly degrade the MAP, it is thus crucial to try rejecting them. To evaluate more specifically the targeted usage scenario (i.e., invasive species), a secondary MAP was computed by considering as queries only a subset of the species that belong to a black list of invasive species.

2.2 Participants and results

94 research groups registered to LifeCLEF plant challenge 2016 and downloaded the dataset. Among this large raw audience, 8 research groups succeeded in submitting *runs*, i.e., files containing the predictions of the system(s) their ran. Details of the methods and systems used in the runs are synthesised in the overview working note of the task [26] and further developed in the individual working notes of the participants (Bluefield [34], Sabanci [22], CMP [64], LIIR, Floristic [30], UM [47], QUT [48], BME [4]). We give hereafter a few more details of the 3 systems that performed the best:

Bluefield system: A VGGNet [59] based system with the addition of Spatial Pyramid Pooling, Parametric ReLU and unknown class rejection based on the minimal prediction score of training data (Run 1). Run 2 is the same as run 1 but with a slightly different rejection making use of a validation set. Run 3 and 4 are respectively the same as Run 1 and 2 but the scores of the images belonging to the same observation were summed and normalised.

Sabanci system: Also a CNN-based system with 2 main configurations. Run 1: An ensemble of GoogleLeNet [66] and VGGNet [59] fine-tuned on both LifeCLEF 2015 data (for recognizing the targeted species) and on 70K images of the ILSCVR dataset (for rejecting unknown classes). Run 2 is the same than Run 1 but without rejection.

CMP system: A ResNet [36] based system with the use of bagging in Run 1 (3 networks) and without bagging (in Run 2).

We report in Figure 1 the scores achieved by the 29 collected runs for the two official evaluation metrics (MAP-open and MAP-open-invasive). To better assess the impact of the distractors i.e., the images in the test set belonging to unknown classes), we also report the MAP obtained when removing them (and denoted as MAP-closed). As a first noticeable remark, the top-26 runs which performed the best were based on Convolutional Neural Networks (CNN). This definitely confirms the supremacy of deep learning approaches over previous methods, in particular the one bases on hand-crafted features (such as BME TMIT Run 2). The different CNN-based systems mainly differed in (i) the architecture of the used CNN, (ii) the way in which the rejection of the unknown classes was managed and (iii), various system design improvements such as classifier ensembles, bagging or observation-level pooling. An impressive MAP of 0.718 (for the targeted invasive species monitoring scenario) was achieved by the best system configuration of Bluefield (run 3). The gain achieved by this run is however more related to the use of the observation-level pooling (looking at Bluefield run 1 for comparison) than to a good rejection of the distractors. Comparing the metric

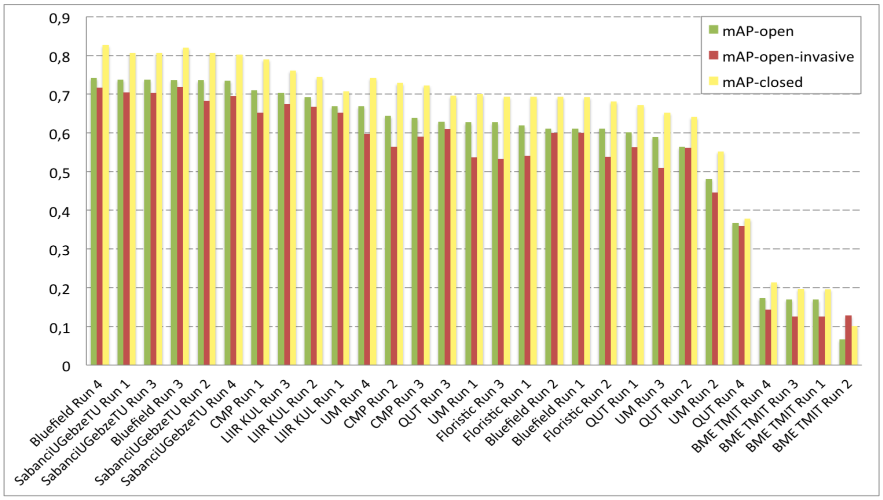

Fig. 1. Scores achieved by all systems evaluated within the plant identification task of LifeCLEF 2016, **MAP-open**: mean Average Precision on the 1000 species of the training set and distractors in the test set, **MAP-open-invasive**: mean Average Precision with distractors but restricted to 26 invasive species only, **MAP-closed**: mean Average Precision on the 1000 species but without distractors in the test set

MAP-open with MAP-closed, the figure actually shows that the presence of the unknown classes degrades the performance of all systems in a roughly similar way. This difficulty of rejecting the unknown classes is confirmed by the very low difference between the runs of the participants who experimented their system with or without rejection (e.g., Sabanci Run 1 vs. Run 2 or FlorisTic Run 1 vs. Run 2). On the other side, it is noticeable that all systems are quite robust to the presence of unknown classes since the drop in performance is not too high. Actually, as the CNNs were pre-trained on a large generalist data set beforehand, it is likely that they have learned a diverse enough set of visual patterns to avoid underfiting. Now it is important to notice that the proportion of unknown classes in the test set was still reasonable (actually only 42%) because of the procedure used to create it. In further work, we will attempt to build a test set closer to the true statistics of the queries. This is however a hard problem. Even experts are actually doubtful of the true label of many images that do not not contain enough visual evidences. Thus, they tend to annotate only the contents they are sure of i.e., the less confused ones. To build a more complete ground truth, it is required to take into account this doubt, during the annotation process, but also when measuring the accuracy of the evaluated systems.

3 Task2: BirdCLEF

The general public as well as professionals like park rangers, ecological consultants and of course the ornithologists themselves are potential users of an automated bird identifying system, typically in the context of wider initiatives related to ecological surveillance or biodiversity conservation. Using audio records rather than bird pictures is justified by current practices [9,69,68,8]. Birds are actually not easy to photograph as they are most of the time hidden, perched high in a tree or frightened by human presence, and they can fly very quickly, whereas audio calls and songs have proved to be easier to collect and very discriminant.

Before LifeCLEF started in 2014, three previous initiatives on the evaluation of acoustic bird species identification took place, including two from the SABIOD[15] group [25,24,7]. In collaboration with the organizers of these previous challenges, BirdCLEF 2014, 2015 and 2016 challenges went one step further by (i) significantly increasing the species number by an order of magnitude, (ii) working on real-world social data built from thousands of recordists, and (iii) moving to a more usage-driven and system-oriented benchmark by allowing the use of meta-data and defining information retrieval oriented metrics. Overall, the task is much more difficult than previous benchmarks because of the higher confusion risk between the classes, the higher background noise and the higher diversity in the acquisition conditions (different recording devices, contexts diversity, etc.). It therefore produces substantially lower scores and offers a better progression margin towards building real-world general identification tools.

The main novelty of the 2016 edition of the task with respect to the two previous years was the inclusion of *soundscape recordings* in addition to the usual xeno-canto recordings that focus on a single foreground species (usually thanks to mono-directional recording devices). Soundscapes, on the other hand, are generally based on omnidirectional recording devices that continuously monitor a specific environment over a long period. This new kind of recording fits better to the (possibly crowdsourced) passive acoustic monitoring scenario that could augment the number of collected records by several orders of magnitude.

3.1 Data and task description

The training and test data of the challenge consists of audio recordings collected by Xeno-canto (XC)[16]. Xeno-canto is a web-based community of bird sound recordists worldwide with about 3,000 active contributors that have already collected more than 300,000 recordings of about 9550 species (numbers for June 2016). Nearly 1000 (in fact 999) species were used in the BirdCLEF dataset, representing the 999 species with the highest number of recordings in October 2014 (14 or more) from the combined area of Brazil, French Guiana, Suriname, Guyana, Venezuela and Colombia, totalling 33,203 recordings produced by thousands of users. This dataset includes the entire dataset from the 2015 BirdCLEF challenge [32], which contained about 33,000 recordings.

The newly introduced test data in 2016 contains 925 soundscapes provided by 7 recordists, sometimes working in pairs. Most of the soundscapes have a length of (more or less) 10 minutes, each coming often from a set of 10-12 successive

[15] Scaled Acoustic Biodiversity http://sabiod.univ-tln.fr

[16] `http://www.xeno-canto.org/`

recordings collected from one location. The total duration of new testing data to process and analyse is thus equivalent to approximately 6 days of continuous sound recording. The number of known species (i.e belonging to the 999 species in the training dataset) varies from 1 to 25 species, with an average of 10.1 species per soundscape.

To avoid any bias in the evaluation related to the used audio devices, each audio file has been normalized to a constant bandwidth of 44.1 kHz and coded in 16 bits in wav mono format (the right channel was selected by default). The conversion from the original Xeno-canto data set was done using ffmpeg, sox and matlab scripts. The optimized 16 Mel Filter Cepstrum Coefficients for bird identification (according to an extended benchmark [15]) were computed, together with their first and second temporal derivatives on the whole set. They were used in the best systems run in ICML4B and NIPS4B challenges. However, due to technical limitations, the soundscapes were not normalized and directly provided to the participants in mp3 format (shared on the xeno-canto website, the original raw files being not available).

All audio records are associated with various meta-data including the species of the most active singing bird, the species of the other birds audible in the background, the type of sound (call, song, alarm, flight, etc.), the date and location of the observations (from which rich statistics on species distribution can be derived), some text comments of the authors, multilingual common names and collaborative quality ratings. All of them were produced collaboratively by the Xeno-canto community.

Participants were asked to determine all the active singing birds species in each query file. It was forbidden to correlate the test set of the challenge with the original annotated Xeno-canto data base (or with any external content as many of them are circulating on the web). The whole data was split in two parts, one for training (and/or indexing) and one for testing. The test set was composed of (i) all the newly introduced soundscape recordings and (ii), the entire test set used in 2015 (equal to about 1/3 of the observations in the whole 2015 dataset). The training set was exactly the same as the one used in 2015 (i.e., the remaining 2/3 of the observations). Note that recordings of the same species done by the same person on the same day are considered as being part of the same observation and cannot be split across the test and training set. The XML files containing the meta-data of the *query* recordings were purged so as to erase the taxon name (the ground truth), the vernacular name (common name of the bird) and the collaborative quality ratings (that would not be available at query stage in a real-world mobile application). Meta-data of the recordings in the training set were kept unaltered.

The groups participating in the task were asked to produce up to 4 runs containing a ranked list of the most probable species for each query record of the test set. Each species was associated with a normalized score in the range $[0, 1]$ reflecting the likelihood that this species is singing in the sample. The primary metric used was the Mean Average Precision averaged across all queries.

3.2 Participants and results

84 research groups registered for the bird challenge and downloaded the data and 6 of them finally submitted runs. Details of the systems and the methods used in the runs are synthesised in the overview working note of the task [31] and further developed in the individual working notes of the participants ([18,45,51,67,54]). We give hereafter more details of the 3 systems that performed the best.

Cube system was based on a CNN architecture of 5 convolutional layers combined with the use of a rectify activation function followed by a max-pooling layer. Based on spectrogram analysis and some morphological operations, silent and noisy parts were first detected and separated from the birds song (or call) parts. Spectrograms were then split into chunks of 3 seconds that were used as inputs of the CNN after several data augmentation techniques. Each chunk identified as a bird song was first concatenated with 3 randomly selected chunks of background noise. Time shift, pitch shift and randomized mixes of audio files from the same species were then used as complementary data augmentation techniques. All the predictions of the distinct chunks are finally averaged to get the prediction of the entire test record. Run 1 was an intermediate result obtained after only one day of training. Run 2 differs from run 3 by using 50% smaller spectrograms in (pixel) size for doubling the batch size and thus allowing to have more iterations for the same training time (4 days). Run 4 is the average of predictions from run 2 and 3 and reaches the best performance, showing the benefit of bagging (as for the plant identification task).

TSA system: As in 2014 and 2015, this participant used two hand-crafted parametric acoustic features and probabilities of species-specific spectrogram segments in a template matching approach. Long segments extracted during BirdCLEF2015 were re-segmented with a more sensitive algorithm. The segments were then used to extract Segment-Probabilities for each file by calculating the maxima of the normalized cross-correlation between all segments and the target spectrogram image via template matching. Due to the very large amount of audio data, not all files were used as a source for segmentation (i.e., only good quality files without background species were used). The classification problem was then formulated as a multi-label regression task solved by training ensembles of randomized decision trees with probabilistic outputs. The training was performed in 2 passes, one selecting a small subset of the most discriminant features by optimizing the internal MAP score on the training set, and one training the final classifiers on the selected features. Run 1 used one single model on a small but highly optimized selection of segment-probabilities. A bagging approach was used consisting in calculating further segment-probabilities from additional segments and to combine them either by blending (24 models in Run 3). Run 4 also used blending to aggregate model predictions, but the predictions were included that after blending resulted in the highest possible MAP score calculated on the entire training set (13 models including the best model from 2015).

WUT system: like the Cube team, they used a CNN-based learning framework. Starting from denoised spectrograms, silent parts were removed with percentile thresholding, giving thus around 86.000 training segments varying in length and associated each with a single main species. As a data augmentation technique and for fitting the 5 seconds fixed input size of the CNN, segments were adjusted by either trimming or padding. The 3 first successive runs are produced by deeper and deeper or/and wider and wider filters. Run 4 is as an ensemble of neural networks averaging the predictions of the 3 first runs.

Figure 2 reports the performance measured for the 18 submitted runs. For each run (i.e., each evaluated system), we report the overall mean Average Precision (official metric) as well as the MAP for the two categories of queries: the soundscapes recordings (newly introduced) and the common observations (the same as the one used in 2015). To measure the progress over last year, we also plot on the graph the performance of last year's best system [44] (orange dotted line). The first noticeable conclusion is that, after two years of resistance of bird song identification systems based on engineering features, convolutional neural networks finally managed to outperform them (as in many other domains). The best run based on CNN (Cube Run 4) actually reached an impressive MAP of 0.69 on the 2015 testbed to be compared to respectively 0.45 and 0.58 for the best systems based on hand-crafted features evaluated in 2015 and 2016. To our knowledge, BirdCLEF is the first comparative study reporting such an important performance gap in bio-acoustic large-scale classification. A second important remark is that this performance of CNNs was achieved without any fine-tuning contrary to most computer vision challenges in which the CNN is generally pre-trained on a large training data such as ImageNet. Thus, we can hope for even better performance, e.g., by transferring knowledge from other bio-acoustic contexts or other domains. It is important to notice that the second system based on CNN (WUT) did not perform as well as the Cube system and did not outperform the system of TSA based on hand-crafted features. Looking at the detailed description of the two CNN architectures and their learning framework, it appears that the way in which audio segments extraction and data augmentation is performed does play a crucial role. The Cube system does notably include a randomized background noise addition phase which makes it much more robust to the diversity of noise encountered in the test data.

If we now look at the scores achieved by the evaluated systems on the soundscape recordings only (yellow plot), we can draw very different conclusions. First of all, we can observe that the performance on the soundscapes is much lower than on the classical queries, whatever the system. Although the classical recordings also include multiple species singing in the background, the soundscapes appear to be much more challenging. Several tens of species and even much more individual birds can actually be singing simultaneously. Separating all these sources seem to be beyond the scope of state-of-the-art audio representation learning methods. Interestingly, the best system on the soundscape queries was the one of TSA based on the extraction of very short species-specific spectrogram segments and a template matching approach. This very fine-grained approach allows the extracted audio patterns to be more robust to the species overlap problem. On the contrary, the CNN of Cube and WUT systems were optimized for the mono-species segments classification problem. The data augmentation method of the Cube system was in particular only designed for the single species case. It addressed the problem of several individual birds of the same species singing together (by mixing different segments of the same class) but it did not address the multi-label issue (i.e., several species singing simultaneously [16]), and is getting close to the simple reference MFCC model provided for comparison to the baseline [54].

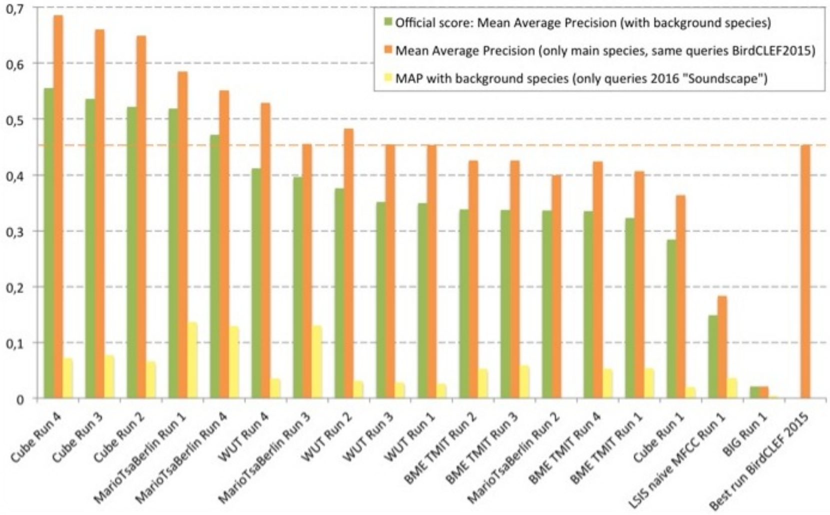

Fig. 2. Scores of the LifeCLEF 2016 Bird Identification Task

4 Task3: SeaCLEF

The SeaCLEF 2016 task originates from the previous editions of the fish identification task (in 2014 and 2015), i.e., video-based coral fish species identification for ecological surveillance and biodiversity monitoring. SeaCLEF 2016 extends the previous ones in that it does not only consider fish species, but sea organisms in general. The need of automated methods for sea-related visual data is driven by the advances in imaging systems (in particular underwater) and their employment for marine ecosystem analysis and biodiversity monitoring. Indeed in recent years we have assisted an exponential growth of sea-related visual data, in the forms of images and videos, for disparate reasoning ranging from fish biodiversity monitoring to marine resource managements to fishery to educational purposes. However, the analysis of such data is particularly expensive for human operators, thus limiting greatly the impact of that the technology may have in understanding and sustainably exploiting the sea.

The task aims at evaluating two kinds of automated identification scenarios: species recognition and individuals recognition. Whereas species recognition is the most common practice, it is preferable for some groups to monitor the organisms at the individual level rather than at the species level. This is notably the case of big animals, such as whales and elephants, whose population might be scarce and travelling for long distances. Monitoring individual animals allows gathering valuable information about population sizes, migration, health, sexual maturity and behavior patterns.

4.1 Coral Reef Species Identification in Underwater Videos

The goal of the task was to automatically detect and recognize coral reef species in underwater videos. The typical usage scenario of automated underwater video analysis tools is to support marine biologists in studying thoroughly the marine ecosystem and fish biodiversity. Also, scuba divers, marine stakeholders and other marine practitioners may benefit greatly from these kinds of tools. Recently, underwater video and imaging systems have been employed since they do not affect fish behavior and may provide large amounts of visual data at the same time. However, manual analysis as performed by human operators is largely impractical, and requires automated methods. Nevertheless, the development of automatic video analysis tools is 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.

Despite some preliminary work, mainly carried out in controlled environments (e.g., labs, cages, etc.) [49,19], the most important step in the automated visual analysis has been done in the EU-funded Fish4Knowledge (F4K)[17] project, where computer vision methods were developed to extract information about fish density and richness from videos taken by underwater cameras installed at coral reefs in Taiwan [62,63,6,61]. Since the F4K project, many researchers have directed their attention towards underwater video analysis [53,55], including some recent initiatives by the National Oceanographic and Atmospheric Administration (NOAA) [57] and the fish identification task at LifeCLEF 2014 and 2015 [12,13,60]. Although there are recent advances in the underwater computer vision field, the problem is still open and needs several (joint) efforts to devise robust methods able to provide reliable measures on fish populations.

Data. The training dataset consists of 20 videos manually annotated, a list of fish species (15) and for each species, a set of sample images to support the learning of fish appearance models. Each video is manually labelled and agreed by two expert annotators and the ground truth consists of a set of bounding boxes (one for each instance of the given fish species list) together with the fish species. In total the training dataset contains more than 9,000 annotations (bounding boxes + species) and more than 20000 sample images. However, it is not a statistical significant estimation of the test dataset rather its purpose is as a familiarization pack for designing the identification methods. The training dataset is unbalanced in the number of instances of fish species: for instance it

[17] `http://www.fish4knowledge.eu/`

contains 3165 instances of “Dascyllus Reticulates” and only 72 instances of “Zebrasoma Scopas”. This was done not to favour nonparametric methods against model-based methods. For each considered fish species, its *fishbase.org* link is also given so as to give access to more detailed information about fish species including complementary high quality images. In order to evaluate the identification process independently from the tracking process, temporal information was not be exploited. This means that the annotators only labelled fish for which the species was clearly identifiable regardless from previous identifications. Each video is accompanied by an XML file containing instances of the provided list species. For each video, information on the location and the camera recording the video is also given.

The test dataset consists of 73 underwater videos. The list of considered fish species is the same than the one released with the training dataset (i.e., 15 coral reef fish species). The number of occurrences per fish species is provided in Table 4.1. It is noticeable, that for three fish species there were no occurrences in the test set, and also that in some video segments there were no fish at all. This was done to evaluate the method’s capability to reject false positives.

Task Description. The main goal of the video-based fish identification task is to automatically count fish per species in video segments (e.g., video X contains $N1$ instances of fish of species 1, ..., N_n instances of fish species N). However, participants were also asked to identify fish bounding boxes. The ground truth for each video (provided as an XML file) contains information on fish species and location. The participants were asked to provide up to three runs. Each run had to contain all the videos included in the set and for each video the frame where the fish was detected together with the bounding box, and species name (only the most confident species) for each detected fish.

Table 1: Fish species occurrences in the test set.

Fish Species Name	**Occurrences**	**Fish Species Name**	**Occurrences**
Abudefduf vaigiensis	93	Acanthurus nigrofuscus	129
Amphirion clarkii	517	Chaetodon lunulatus	1876
Chaetodo speculum	0	Chaetodon trifascialis	1317
Chromis chrysura	24	Dacyllus aruanus	1985
Dascyllus reticulatus	5016	Hemigymnus melapterus	0
Myripristis kuntee	118	Neoglyphidodon nigroris	1531
Pempheris vanicolensis	0	Plectrogly-phidodon dickii	700
Zebrasoma scopas	187		

Metrics. As metrics, we used the “**Counting Score (CS)**” and the “**Normalized Counting Score (NCS)**”, defined as:

$$CS = e^{-}\frac{d}{N_{gt}} \tag{1}$$

with d being the difference between the number of occurrences in the run (per species) and, N_{gt}, the number of occurrences in the ground truth. The Normalised Counting S instead depends on precision Pr:

$$NCS = CS \times Pr = CS \times \frac{TP}{TP + FP} \tag{2}$$

with TP and FP being the True Positive and the False Positive. As detection was considered a true positive if the intersection over union score of its bounding box and the ground truth was over 0.5 and the species was correctly identified.

Participants and Results Figure 3 shows, respectively, the average (per video and species) normalized counting score, precision and counting score obtained by the two participating teams (CVG [37] and BMETMIT [14]) who submitted one run each.

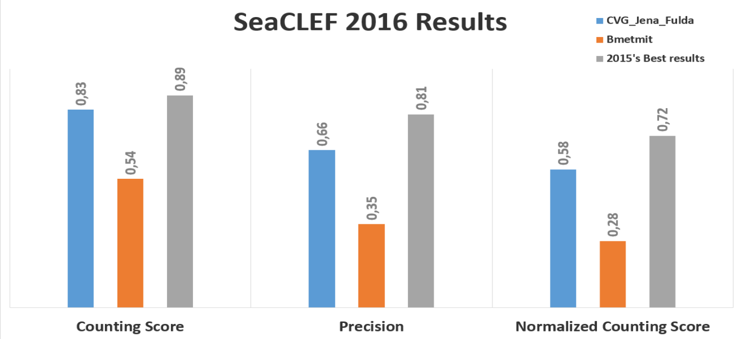

Fig. 3. SeaCLEF 2016 Results.

Figure 4 gives the detailed normalized counting scores per fish species. In addition to the results obtained in 2016, the graphs also show the best performance achieved on the same dataset in 2015. This comparison shows that, unfortunately, none of the 2016 approaches outperformed the one by SNUMED INFO, which performed the best in 2015 (described in details in [11]). This system was based on the GoogLeNet [65] Convolutional Neural Network (CNN). Potential fish instances were previously segmented from the video through a stationary foreground detection using background subtraction and a selective search strategy [70]. Producing the final output counts was finally achieved by grouping the temporally connected video segments classified by the CNN. The system used in 2016 by CVG [37] was inspired by a region-based convolutional neural network (R-CNN [23]), with the difference that it employed background subtraction instead of selective search for bounding box proposal generation. More specifically, CVG's method used off-the-shelf AxelNet CNN [42] for feature extraction (7th hidden layer relu7), and then trained a multiclass support vector machine (MSVMs) for species classification. Its achieved performance, in terms of counting score of 0.83, over the 15 considered fish species was fairly good. Its lower value with respect to SNUMED INFO's one (0.89) can be explained with the fact that CVG did not apply any domain-specific fine tuning of CNN. In the case of normalised counting score, the gap between CVG and SNUMED INFO was higher, and this is due to the fact that CVG employed background subtraction for proposal generation, which is known to be prone to false positives, instead of the more effective selective search used by SNUMED INFO. For a similar reason BMETMIT achieved the lowest normalised counting score, while its lower counting score can be ascribed to the used shallow classifier operating on SURF features, while the other two methods (CVG and SNUMED INFO) resorted on deep-learning methods.

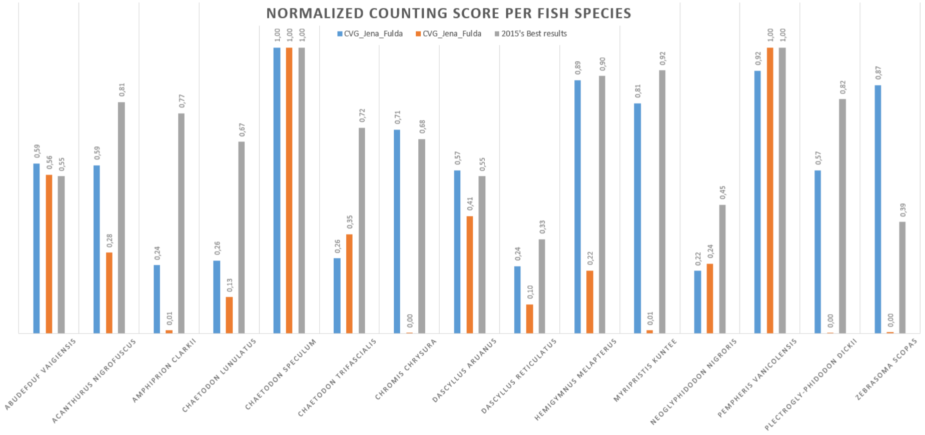

Fig. 4. Normalised Counting Score detailed by fish species.

4.2 Individual Humpback Whale Identification

Using natural markings to identify individual animals over time is usually known as *photo-identification*. This research technique is used on many species of marine mammals. Initially, scientists used artificial tags to identify individual whales, but with limited success (most tagged whales were actually lost or died). In the 1970s, scientists discovered that individuals of many species could be recognized by their natural markings. These scientists began taking photographs of individual animals and comparing these photos against each other to identify individual animal's movements and behavior over time. Since its development, photo-identification has proven to be a useful tool for learning about many marine mammal species including humpbacks, right whales, finbacks, killer whales, sperm whales, bottlenose dolphins and other species to a lesser degree. Nowadays, this process is still mostly done manually making it impossible to get an accurate count of all the individuals in a given large collection of observations.

Researchers usually survey a portion of the population, and then use statistical formulae to determine population estimates. To limit the variance and bias of such an estimator, it is however required to use large-enough samples which still makes it a very time-consuming process. Automating the *photo-identification* process could drastically scale-up such surveys and open brave new research opportunities for the future.

Data and related challenges The dataset used for the evaluation consisted of 2005 images of humbpack whales caudals collected by the CetaMada NGO between 2009 and 2014 in the Madagascar area. Each photograph was manually cropped so as to focus only on the caudal fin that is the most discriminant pattern for distinguishing an individual whale from another. Figure 5 displays six of such cropped images, each line corresponding to two images of the same individual. As one can see, the individual whales can be distinguished thanks to their natural markings and/or the scars that appear along the years. Automatically finding such matches in the whole dataset and rejecting the false alarms is difficult for three main reasons. The first reason is that the number of individuals in the dataset is high, around $1,200$, so that the proportion of true matches is actually very low (around 0.05% of the total number of potential matches).

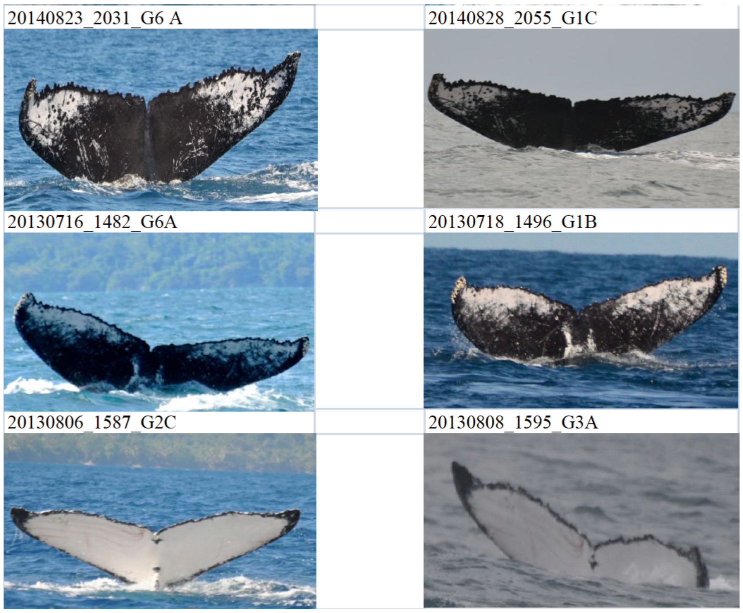

Fig. 5. 3 good matches (each line corresponds to 2 images of the same individual whale)

The second difficulty is that distinct individuals can be very similar at a first glance as illustrated by the false positive examples displayed in Figure 6. To discriminate the true matches from such false positives, it is required to detect very small and fine-grained visual variations such as in a spot-the-difference game. The third difficulty is that all images have a similar water background of which the texture generates quantities of local mismatches.

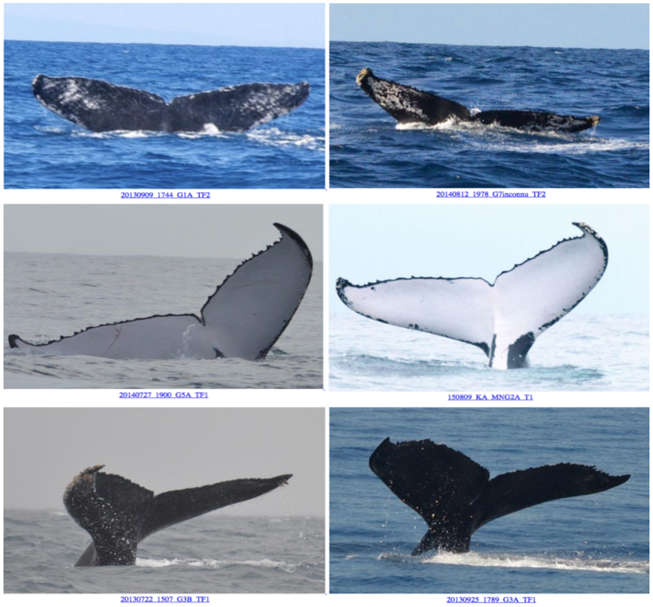

Fig. 6. 3 false positives (each line corresponds to 2 distinct individual whales)

Task Description The task was simply to detect as many true matches as possible from the whole dataset, in a fully unsupervised way. Each evaluated system had to return a *run file* (i.e., a raw text file) containing as much lines as the number of discovered matches, each match being a triplet of the form *imageX.jpg;imageY.jpg;score* where *score* is a confidence score in $[0, 1]$ (1 for highly confident matches). The retrieved matches had to be sorted by decreasing confidence score. A run should not contain any duplicate match (e.g., *image1.jpg;image2.jpg;score* and *image2.jpg;image1.jpg;score* should not appear in the same run). The metric used to evaluate each run is the Average Precision:

$$AveP = \frac{\sum_{k=1}^{K} P(k) \times rel(k)}{M}$$

where M is the total number of true matches in the groundtruth, k is the rank in the sequence of returned matches, K is the number of retrieved matches, $P(k)$ is the precision at cut-off k in the list, and $rel(k)$ is an indicator function equaling 1 if the match at rank k is a relevant match, 0 otherwise. The average is over all true matches and the true matches not retrieved get a precision score of 0.

Participants and Results Two research groups participated to the evaluation and submitted a total of 6 run files. Table 4.2 provides the scores achieved by the six runs. Details of the systems and methods used can be found in the individual working notes of the participants (INRIA [40], BME-MIT [14]). We give hereafter a synthetic description of the evaluated systems/configurations:

INRIA system: This group used a large-scale matching system based on local visual features, approximate k-nn search of each individual local feature via multi-probe hashing, and a RANSAC-like spatial consistency refinement step used to reject false positives (based on a rotation-and-scale transformation model). The run named *ZenithINRIA_SiftGeo* used affine SIFT features whereas the one named *ZenithINRIA_GoogleNet_3layers_borda* used off-the-shelf local features extracted at three different layers of GoogLeNet [65] (layer *conv2-3x3*: 3136 local features per image, layer *inception_3b_output*: 784 local features par image, layer *inception_4c_output*: 196 local features per image). The matches found using the 3 distinct layers were merged through a late-fusion approach based on Borda. Finally, the last run *ZenithINRIA_SiftGeo_QueryExpansion* differs from *ZenithINRIA_SiftGeo* in that a query expansion strategy was used to re-issue the regions matched with a sufficient degree of confidence as new queries.

BME-MIT system: This group used aggregation-based image representations based on SIFT features (extracted either on a dense grid or around Laplace-Harris points), a GMM-based visual codebook learning (256 visual words),and Fisher Vectors (FVs) for the global image representation. A RBF kernel was used to measure the similarity between image pairs. Runs *bmetmit_whalerun_2* and *bmetmit_whalerun_3* differ from *bmetmit_whalerun_1* in that segmentation propagation was used beforehand so as to separate the background (the water) from the whale's caudal fin. In *bmetmit_whalerun_3* the segmentation mask was applied only for filtering the features during the codebook learning phase. In run 2 the mask was also used to when computing the FVs of each image.

Table 2: Individual whale identification results: AP of the 6 evaluated systems

Run name	**AP**
ZenithInria SiftGeo	0.49
ZenithInria SiftGeo QueryExpansion	0.43
ZenithInria GoogleNet 3layers borda	0.33
bmetmit whalerun 1	0.25
bmetmit whalerun 3	0.10
bmetmit whalerun 2	0.03

The main conclusion we can draw from the results of the evaluation (cf. table 4.2) is that spatial consistency of the local features is crucial for rejecting the false positives (as proved by the much higher performance of INRIA system). As powerful as aggregation-based methods such as Fisher Vectors are for fine-grained classification, they do not capture the spatial arrangement of the local features which is a precious information for rejecting the mismatches without supervision. Another reason explaining the good performance of the best run *ZenithINRIA_SiftGeo* is that it is based on affine invariant local features contrary to *ZenithINRIA_GoogleNet_3layers_borda* and **BME-MIT** runs that use grid-based local features. Such features are more sensitive to small shifts and local affine deformations even when learned through a powerful CNN. Finally, neither segmentation nor query expansion succeeded in improving the results. Segmentation is always risky because of the risk of over segmentation which might remove the useful information from the image. Query expansion is also a risky solution in that it is highly sensitive to the decision threshold used for selecting the re-issued matched regions. It can be considerably increase recall when the decision threshold is well estimated but at the opposite, it can also boost the false positives when the threshold is too low.

5 Conclusions and Perspectives

With more than 130 research groups who downloaded LifeCLEF 2016 datasets and 14 of them who submitted runs, the third edition of the LifeCLEF evaluation did confirm a high interest in the evaluated challenges. The main outcome of this collaborative effort is a snapshot of the performance of state-of-the-art computer vision, bio-acoustic and machine learning techniques towards building real-world biodiversity monitoring systems. The results did show that very high identification success rates can be reached by the evaluated systems, even on large number of species (up to 1000 species). The most noticeable progress came from the deployment of deep Convolutional Neural Networks for the bird songs identification challenge. We observed a similar performance gap to the one observed in many domains beforehand (in particular the LifeCLEF plant identification task two years ago). Interestingly, this was achieved without any fine-tuning which means that the xeno-canto dataset is sufficiently rich to allow the CNN learning relevant audio features. This opens the door to transfer learning opportunities in other bio-acoustic domains for which training data are sparser. Regarding the plant task, the main conclusion was that CNNs appeared to be quite robust to the presence of unknown classes in the test set. The proportion of novelty was however still moderate, near 50% and might be increased in further evaluations so as to better fit reality. Finally, the two newly introduced scenarios, i.e., soundscape-based monitoring of birds and unsupervised identification of individual whales appeared to be quite challenging. Bird soundscapes, in particular, seem to be out of reach for current audio representation learning methods because of the very large number of overlapping sound sources in single recordings. The identification of individual whales was more effective (thanks to the use of spatial verification) but there is still room for improvement before fully automating the *Photo-identification* process used by biologists.

Acknowledgements The organization of the PlantCLEF task is supported by the French project Floris'Tic (Tela Botanica, INRIA, CIRAD, INRA, IRD) funded in the context of the national investment program PIA. The organization of the BirdCLEF task is supported by the Xeno-Canto foundation for nature sounds as well as the French CNRS project SABIOD.ORG and Floris'Tic.The organization of the SeaCLEF task is supported by the Ceta-mada NGO and the French project Floris'Tic.

References

1. MAED '12: Proceedings of the 1st ACM International Workshop on Multimedia Analysis for Ecological Data. ACM, New York, NY, USA (2012), 433127
2. Aptoula, E., Yanikoglu, B.: Morphological features for leaf based plant recognition. In: Proc. IEEE Int. Conf. Image Process., Melbourne, Australia (2013)
3. Baillie, J.E.M., H.T.C., Stuart, S.: 2004 iucn red list of threatened species. a global species assessment. IUCN, Gland, Switzerland and Cambridge, UK (2004)
4. Bálint Pál Tóth, Márton Tóth, D.P., Szúcs, G.: Deep learning and svm classification for plant recognition in content-based large scale image retrieval. In: Working notes of CLEF 2016 conference (2016)
5. Bendale, A., Boult, T.E.: Towards open world recognition. CoRR (2014)
6. 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 23, 83 – 97 (2014)
7. Briggs, F., Huang, Y., Raich, R., Eftaxias, K., et al., Z.L.: The 9th mlsp competition: New methods for acoustic classification of multiple simultaneous bird species in noisy environment. In: IEEE Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–8 (2013)
8. Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X.Z., Raich, R., Hadley, S.J., Hadley, A.S., Betts, M.G.: Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. The Journal of the Acoustical Society of America 131, 4640 (2012)
9. Cai, J., Ee, D., Pham, B., Roe, P., Zhang, J.: Sensor network for the monitoring of ecosystem: Bird species recognition. In: Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Conference on (2007)
10. Cerutti, G., Tougne, L., Vacavant, A., Coquin, D.: A parametric active polygon for leaf segmentation and shape estimation. In: International Symposium on Visual Computing. pp. 202–213 (2011)
11. Choi, S.: Fish identification in underwater video with deep convolutional neural network: Snumedinfo at lifeclef fish task 2015. In: Working notes of CLEF (2015)
12. Concetto, S., Palazzo, S., Fisher, B., Boom, B.: Lifeclef fish identification task 2014. In: CLEF working notes 2014 (2014)
13. Concetto, S., Palazzo, S., Fisher, B., Boom, B.: Lifeclef fish identification task 2015. In: CLEF working notes 2015 (2015)
14. Dävid Papp, D.L., Szücs, G.: Object detection, classification, tracking and individual recognition for sea images and videos. In: Working notes of CLEF (2016)
15. Dufour, O., Artieres, T., Glotin, H., Giraudet, P.: Clusterized mfcc and svm for bird song. In: Identification, Soundscape Semiotics, Localization, Categorization (2014)
16. Dufour, O., Glotin, H., Artieres, T., Bas, Y., Giraudet, P.: Multi-instance multi-label acoustic classification of plurality of animals: birds, insects & amphibian. In: 1st Workshop on Neural Infor. Proc. Scaled for Bioacoustics. pp. 164–174. in conj. with NIPS (2013)

17. Dugan, P., Zollweg, J., Popescu, M., Risch, D., Glotin, H., LeCun, Y., Clark, C.: High performance computer acoustic data accelerator: A new system for exploring marine mammal acoustics for big data applications (2015)
18. Elias Sprengel, Martin Jaggi, Y.K., Hofmann, T.: Audio based bird species identification using deep learning techniques. In: Working notes of CLEF (2016)
19. Evans, F.: Detecting fish in underwater video using the em algorithm. In: Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on. vol. 3, pp. III–1029–32 vol.2 (2003)
20. Farnsworth, E.J., Chu, M., Kress, W.J., Neill, A.K., Best, J.H., Pickering, J., Stevenson, R.D., Courtney, G.W., VanDyk, J.K., Ellison, A.M.: Next-generation field guides. BioScience 63(11), 891–899 (2013)
21. Gaston, K.J., O'Neill, M.A.: Automated species identification: why not? 359(1444), 655–667 (2004)
22. Ghazi, M.M., Yanikoglu, B., Aptoula, E.: Open-set plant identification using an ensemble of deep convolutional neural networks. In: Working notes of CLEF (2016)
23. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. CoRR abs/1311.2524 (2013)
24. Glotin, H., Clark, C., LeCun, Y., Dugan, P., Halkias, X., Sueur, J.: Bioacoustic challenges in icml4b. In: in Proc. of 1st workshop on Machine Learning for Bioacoustics. No. USA, ISSN 979-10-90821-02-6 (2013), `http://sabiod.org/ICML4B2013_proceedings.pdf`
25. Glotin, H., Dufour, O., Bas, Y.: Overview of the 2nd challenge on acoustic bird classification. In: Proc. Neural Information Processing Scaled for Bioacoustics. NIPS Int. Conf., Ed. Glotin H., LeCun Y., Artières T., Mallat S., Tchernichovski O., Halkias X., USA (2013), `http://sabiod.univ-tln.fr/nips4b`
26. Goëau, H., Bonnet, P., Joly, A.: Plant identification in an open-world (lifeclef 2016). In: CLEF working notes 2016 (2016)
27. Goëau, H., Bonnet, P., Joly, A., Bakic, V., Barthélémy, D., Boujemaa, N., Molino, J.F.: The imageclef 2013 plant identification task. In: CLEF. Valencia (2013)
28. Goëau, H., Bonnet, P., Joly, A., Boujemaa, N., Barthélémy, D., Molino, J.F., Birnbaum, P., Mouysset, E., Picard, M.: The imageclef 2011 plant images classification task. In: CLEF 2011 (2011)
29. Goëau, H., Bonnet, P., Joly, A., Yahiaoui, I., Barthélémy, D., Boujemaa, N., Molino, J.F.: Imageclef2012 plant images identification task. In: CLEF12. Rome
30. Goëau, H., Champ, J., Joly, A.: Floristic participation at lifeclef 2016 plant identification task. In: Working notes of CLEF 2016 conference (2016)
31. Goëau, H., Glotin, H., Planqué, R., Vellinga, W.P., Joly, A.: Lifeclef bird identification task 2016. In: CLEF working notes 2016 (2016)
32. Goëau, H., Glotin, H., Vellinga, W.P., Planque, R., Rauber, A., Joly, A.: Lifeclef bird identification task 2015. In: CLEF working notes 2015 (2015)
33. Goëau, H., Joly, A., Selmi, S., Bonnet, P., Mouysset, E., Joyeux, L., Molino, J.F., Birnbaum, P., Bathelemy, D., Boujemaa, N.: Visual-based plant species identification from crowdsourced data. In: ACM conference on Multimedia (2011)
34. Hang, S.T., Tatsuma, A., Aono, M.: Bluefield (kde tut) at lifeclef 2016 plant identification task. In: Working notes of CLEF 2016 conference (2016)
35. Hazra, A., Deb, K., Kundu, S., Hazra, P., et al.: Shape oriented feature selection for tomato plant identification. International Journal of Computer Applications Technology and Research 2(4), 449–meta (2013)
36. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)

37. Jäger, J., Rodner, E., Denzler, J., Wolff, V., Fricke-Neuderth, K.: Seaclef 2016: Object proposal classification for fish detection in underwater videos. In: Working notes of CLEF 2016 conference (2016)
38. Joly, A., Goëau, H., Bonnet, P., Bakić, V., Barbe, J., Selmi, S., Yahiaoui, I., Carré, J., Mouysset, E., Molino, J.F., et al.: Interactive plant identification based on social image data. Ecological Informatics 23, 22–34 (2014)
39. Joly, A., Goëau, H., Bonnet, P., Bakic, V., Molino, J.F., Barthélémy, D., Boujemaa, N.: The imageclef plant identification task 2013. In: International workshop on Multimedia analysis for ecological data (2013)
40. Joly, A., Lombardo, J.C., Champ, J., Saloma, A.: Unsupervised individual whales identification: spot the difference in the ocean. In: Working notes of CLEF (2016)
41. Kebapci, H., Yanikoglu, B., Unal, G.: Plant image retrieval using color, shape and texture features. The Computer Journal 54(9), 1475–1490 (2011)
42. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
43. Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I.C., Soares, J.V.B.: Leafsnap: A computer vision system for automatic plant species identification. In: European Conference on Computer Vision. pp. 502–516 (2012)
44. Lasseck, M.: Improved automatic bird identification through decision tree based feature selection and bagging. In: Working notes of CLEF 2015 conference (2015)
45. Lasseck, M.: Improving bird identification using multiresolution template matching and feature selection during training. In: Working notes of CLEF conference (2016)
46. Lee, D.J., Schoenberger, R.B., Shiozawa, D., Xu, X., Zhan, P.: Contour matching for a fish recognition and migration-monitoring system. In: Optics East. pp. 37–48. International Society for Optics and Photonics (2004)
47. Lee, S.H., Chang, Y.L., Chan, C.S., Remagnino, P.: Plant identification system based on a convolutional neural network for the lifeclef 2016 plant classification task. In: Working notes of CLEF 2016 conference (2016)
48. McCool, C., Ge, Z., Corke, P.: Feature learning via mixtures of dcnns for fine-grained plant classification. In: Working notes of CLEF 2016 conference (2016)
49. Morais, E., Campos, M., Padua, F., Carceroni, R.: Particle filter-based predictive tracking for robust fish counting. In: Computer Graphics and Image Processing, 2005. SIBGRAPI 2005. 18th Brazilian Symposium on. pp. 367–374 (2005)
50. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (Dec 2008)
51. Piczak, K.: Recognizing bird species in audio recordings using deep convolutional neural networks. In: Working notes of CLEF 2016 conference (2016)
52. Pimentel, D., Zuniga, R., Morrison, D.: Update on the environmental and economic costs associated with alien-invasive species in the united states. Ecological economics 52(3), 273–288 (2005)
53. Ravanbakhsh, M., Shortis, M.R., Shafait, F., Mian, A., Harvey, E.S., Seager, J.W.: Automated fish detection in underwater images using shape-based level sets. The Photogrammetric Record 30(149), 46–62 (2015)
54. Ricard, J., Glotin, H.: Bag of mfcc-based words for bird identification. In: Working notes of CLEF 2016 conference (2016)
55. Rodriguez, A., Rico-Diaz, A., Rabuñal, J., Puertas, J., Pena, L.: Fish monitoring and sizing using computer vision. In: Ferrández Vicente, J.M., Álvarez Sánchez,

J.R., de la Paz López, F., Toledo-Moreo, F.J., Adeli, H. (eds.) Bioinspired Computation in Artificial Systems, Lecture Notes in Computer Science, vol. 9108, pp. 419–428. Springer International Publishing (2015)

56. Scheirer, W.J., Jain, L.P., Boult, T.E.: Probability models for open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2014)
57. Sigler, M., DeMaster, D., Boveng, P., Cameron, M., Moreland, E., Williams, K., Towler, R.: Advances in methods for marine mammal and fish stock assessments: Thermal imagery and camtrawl. Marine Technology Society Journal 49(2) (2015)
58. Silvertown, J., Harvey, M., Greenwood, R., Dodd, M., Rosewell, J., Rebelo, T., Ansine, J., McConway, K.: Crowdsourcing the identification of organisms: A case-study of ispot. ZooKeys (480), 125 (2015)
59. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)
60. Spampinato, C., Palazzo, S., Joalland, P., Paris, S., Glotin, H., Blanc, K., Lingrand, D., Precioso, F.: Fine-grained object recognition in underwater visual data. Multimedia Tools and Applcations (MTAP-D-14-00618) (2014)
61. Spampinato, C., Beauxis-Aussalet, E., Palazzo, S., Beyan, C., van Ossenbruggen, J., He, J., Boom, B., Huang, X.: A rule-based event detection system for real-life underwater domain. Machine Vision and Applications 25(1), 99–117 (2014)
62. Spampinato, C., Chen-Burger, Y.H., Nadarajan, G., Fisher, R.B.: Detecting, tracking and counting fish in low quality unconstrained underwater videos. In: VISAPP (2). pp. 514–519. Citeseer (2008)
63. Spampinato, C., Palazzo, S., Boom, B., van Ossenbruggen, J., Kavasidis, I., Di Salvo, R., Lin, F.P., Giordano, D., Hardman, L., Fisher, R.B.: Understanding fish behavior during typhoon events in real-life underwater environments. Multimedia Tools and Applications 70(1), 199–236 (2014)
64. Śulc, M., Mishkin, D., Matas, J.: Very deep residual networks with maxout for plant identification in the wild. In: Working notes of CLEF 2016 conference (2016)
65. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. CoRR (2014)
66. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
67. Tóth, B.P., Czeba, B.: Convolutional neural networks for large-scale bird song classification in noisy environment. In: Working notes of CLEF conference (2016)
68. Towsey, M., Planitz, B., Nantes, A., Wimmer, J., Roe, P.: A toolbox for animal call recognition. Bioacoustics 21(2), 107–125 (2012)
69. Trifa, V.M., Kirschel, A.N., Taylor, C.E., Vallejo, E.E.: Automated species recognition of antbirds in a mexican rainforest using hidden markov models. The Journal of the Acoustical Society of America 123, 2424 (2008)
70. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision 104 (2013)
71. Weber, E., Gut, D.: Assessing the risk of potentially invasive plant species in central europe. Journal for Nature Conservation 12(3), 171–179 (2004)
72. Weber, E., Sun, S.G., Li, B.: Invasive alien plants in china: diversity and ecological insights. Biological Invasions 10(8), 1411–1429 (2008)