HAL
open science

LifeCLEF 2017 Lab Overview: Multimedia Species Identification Challenges

Alexis Joly, Hervé Goëau, Hervé Glotin, Concetto Spampinato, Pierre Bonnet, Willem-Pier Vellinga, Jean-Christophe Lombardo, Robert Planque, Simone Palazzo, Henning Müller

To cite this version:

Alexis Joly, Hervé Goëau, Hervé Glotin, Concetto Spampinato, Pierre Bonnet, et al.. LifeCLEF 2017 Lab Overview: Multimedia Species Identification Challenges. CLEF: Cross-Language Evaluation Forum, Sep 2017, Dublin, Ireland. pp.255-274, 10.1007/978-3-319-65813-1_24 . hal-01629191

HAL Id: hal-01629191

https://hal.science/hal-01629191v1

Submitted on 16 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LifeCLEF 2017 Lab Overview: multimedia species identification challenges

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

[1]Inria, LIRMM, Montpellier, France
[2]CIRAD, UMR AMAP, France
[3]AMU, Univ. Toulon, CNRS, ENSAM, LSIS UMR 7296, IUF, France
[4]University of Catania, Italy
[5]Xeno-canto foundation, The Netherlands
[6]ES-SO, Sierre, Switzerland

May 2017

Abstract

Automated multimedia identification tools are an emerging solution towards building accurate knowledge of the identity, the geographic distribution and the evolution of living plants and animals. Large and structured communities of nature observers 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 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 uses and ecology of species. This is an essential step in recording any specimen on

earth to be used in ecological studies. 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 an estimated number of 400,000 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 [11] 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 theoretical and applied work in biological and related fields.

Since the question raised in Gaston and O'Neill [11], *automated species identification: why not?*, a lot of work has been done on the topic (*e.g.* [33, 5, 46, 45, 24]) 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 eBird[1] [41], iNaturalist[2], iSpot [38], Xeno-Canto[3] or Tela Botanica[4] (respectively initiated in the US for the two first ones and in Europe for the three last one), 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 [24] 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[5]) 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 dozen or up to hundreds of species, often acquired in well-controlled envi-

[1] `http://ebird.org/content/ebird/`
[2] `http://www.inaturalist.org/`
[3] `http://www.xeno-canto.org/`
[4] `http://www.tela-botanica.org/`
[5] `http://www.gbif.org/`

ronments [14, 36, 31]. On the other hand, the total number of living species on earth is estimated to be around 10K for birds, 30K for fish, 400K for flowering plants (cf. State of the World's Plants 2017[6]) 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[7] research platform was created in 2014 as a continuation of the plant identification task [25] that was run within the ImageCLEF lab [8] the three years before [14, 15, 13]. LifeCLEF enlarged the evaluated challenge by considering birds and marine animals 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 :

PlantCLEF: an image-based plant identification task making use of Pl@ntNet collaborative data, Encyclopedia of Life' data, and Web 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 2017th edition of LifeCLEF compared to the previous years are the following:

1. **Scalability**: To fully reach its objective, an evaluation campaign such as LifeCLEF requires a long term research effort so as to (i) encourage non incremental contributions, (ii) measure consistent performance gaps and (iii), progressively scale up the problem. Therefore, the number of species was increased considerably between the 2016-th and the 2017-th edition. The plant task, in particular, made a big jump with 10,000 species instead of 1,000 species in the training set. This makes it one of the largest image classification benchmark.
2. **Noisy + clean data**: The focus of the plant task this year was to study the impact of training identification systems on noisy Web data rather then clean data. Collecting clean data massively is actually prohibitive in terms of human cost whereas noisy Web data can be collected at a very cheap cost. Therefore, we built two large-scale datasets illustrating the same 10K species:

[6] `https://stateoftheworldsplants.com/`
[7] `http://www.lifeclef.org`
[8] `http://www.imageclef.org/`

one with clean labels coming from the Web platform Encyclopedia Of Life[9] [47], and one with a high degree of noise - domain noise as well as category noise - crawled from the Web without any filtering.

3. **Time-coded soundscapes**: As the soundscapes data appeared to be very challenging in 2016 (with an accuracy below 15%), we introduced in 2017 new soundscape recordings containing time-coded bird species annotations thanks to the involvement of expert ornithologists. In total, 4,5 hours of audio recordings were collected and annotated manually with more than 2000 identified segments.
4. **New organisms and identification scenarios**: The SeaCLEF task was extended with novel scenarios involving new organisms, *i.e* (i) salmons detection for the monitoring of water turbine, and (ii), marine animal species recognition using weakly-labeled images and relevance ranking.

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

2 Task1: PlantCLEF

The 2017-th edition of PlantCLEF is an important milestone towards working at the scale of continental floras. Thanks to the long term efforts made by the biodiversity informatics, it is actually now possible to aggregate clean data about tens of thousands species world wide. The international initiative Encyclopedia of Life (EoL) in particular is one of the biggest resource of plant pictures. However, the majority of plant species are still very poorly illustrated in such expert databases (or often not illustrated at all). A much larger number of plant pictures are spread on the Web through botanist blogs, plant lovers web-pages, image hosting websites and on-line plant retailers. The LifeCLEF 2017 plant identification challenge proposes to study to what extent a huge but very noisy training set collected through the Web is competitive compared to a relatively smaller but trusted training set checked by experts. As a motivation, a previous study conducted by Krause et al. [29] concluded that training deep neural networks on noisy data was unreasonably effective for fine-grained recognition. The PlantCLEF challenge completes their work in several points:

1. it extends their result to the plant domain. The specificity of the plant domain is that it involves much more species than birds. As a consequence the degree of noise might be much higher due to scarcer available data and higher confusion risks.
2. it scales the comparison between clean and noisy training data to 10K of species. The clean training sets used in their study were actually limited to few hundreds of species.
3. it uses a third-party test dataset that is not a subset of either the noisy dataset or the clean dataset. More precisely, it is composed of images

[9] http://eol.org/

submitted by the crowd of users of the mobile application Pl@ntNet[23]. Consequently, it exhibits different properties in terms of species distribution, pictures quality, etc.

In the following subsections, we synthesize the resources and assessments of the challenge, summarize the approaches and systems employed by the participating research groups, and provide an analysis of the main outcomes. A more detailed description of the challenge and a deeper analysis of the results can be found in the CEUR-WS proceedings of the task [12].

2.1 Dataset and evaluation protocol

To evaluate the above mentioned scenario at a large scale and in realistic conditions, we built and shared three datasets coming from different sources. As training data, in addition to the data of the previous years, we provided two new large data sets both based on the same list of 10,000 plant species (living mainly in Europe and North America):

Trusted Training Set *EoL10K*: a trusted training set based on the online collaborative Encyclopedia Of Life (EoL). The 10K species were selected as the most populated species in EoL data after a curation pipeline (taxonomic alignment, duplicates removal, herbaria sheets removal, etc.). The training set has a massive class imbalance with a minimum of 1 picture for *Achillea filipendulina* and a maximum of 1245 pictures for *Taraxacum laeticolor*.

Noisy Training Set *Web10K*: a noisy training set built through Web crawlers (Google and Bing image search engines) and containing 1.1M images. This training set is also imbalanced with a minimum of 4 pcitures for *Plectranthus sanguineus* and a maximum of 1732 pictures for *Fagus grandifolia*.
The main idea of providing both datasets is to evaluate to what extent machine learning and computer vision techniques can learn from noisy data compared to trusted data (as usually done in supervised classification). Pictures of EoL are themselves coming from several public databases (such as Wikimedia, Flickr, iNaturalist) or from some institutions or less formal websites dedicated to botany. All the pictures can be potentially revised and rated on the EoL website. On the other side, the noisy training set will contain more images for a lot of species, but with several type and level of noises which are basically impossible to automatically filter: a picture can be associated to the wrong species but the correct genus or family, a picture can be a portrait of a botanist working on the species, the pictures can be associated to the correct species but be a drawing or an herbarium sheet of a dry specimen, etc.

Mobile search test set: the test data to be analyzed within the proposed challenge is a large sample of the query images submitted by the users of the mobile application Pl@ntNet (iPhone[10] & Androïd[11]). It contains covering a

[10] `https://itunes.apple.com/fr/app/plantnet/id600547573?mt=8`
[11] `https://play.google.com/store/apps/details?id=org.plantnet`

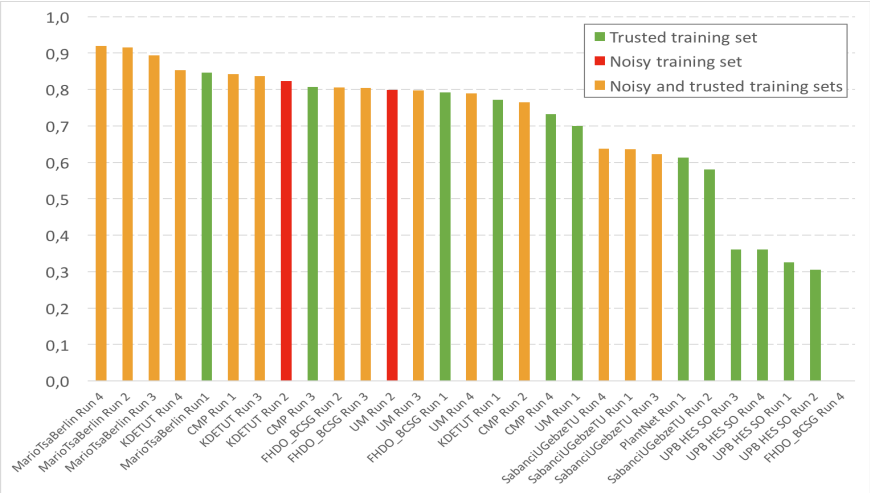

Figure 1: Performance achieved by all systems evaluated within the plant identification task of LifeCLEF 2017

large number of wild plant species mostly coming from the Western Europe Flora and the North American Flora, but also plant species used all around the world as cultivated or ornamental plants, or even endangered species precisely because of their non-regulated commerce.

2.2 Participants and results

80 research groups registered to LifeCLEF plant challenge 2017 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) they ran. Details of the methods and systems used in the runs are synthesised in the overview working note of the task [12] and further developed in the individual working notes of the participants (CMP [40], FHDO BCSG [35], KDE TUT [18], Mario MNB [32], Sabanci Gebze[2], UM [34] and UPB HES SO [44]). We report in Figure 1 the performance achieved by the 29 collected runs. The PlantNet team provides a baseline for the task with the system used in Pl@ntNet app, based on inception model Szegedy [43] and described [1].

Trusted or noisy ? As a first noticeable remark, the measured performances are very high despite the difficulty of the task with a median Mean Reciprocal Rank (MRR) around 0.8, and a highest MRR of 0.92 for the best system Mario MNB Run 4. A second important remark is that the best results are obtained mostly by systems that learned on both the trusted and the noisy datasets. Only two runs (KDE TUT Run 2 and UM Run 2) used exclusively the noisy dataset but gave better results than most of the methods using only the

trusted dataset. Several teams also tried to filter the noisy dataset, based on the prediction of a preliminary system trained only on the trusted dataset (*i.e.* by rejecting pictures whose label is contradictory with the prediction). However, this strategy did not improve the final predictor and even degraded the results. For instance Mario MNB Run 2 (using the raw Web dataset) performed better than Mario MNB Run 3 (using the filtered Web dataset).

Succeeding strategies with CNN models: Regarding the used methods, all submitted runs were based on Convolutional Neural Networks (CNN) confirming definitively the supremacy of this kind of approach over previous methods. A wide variety of popular architectures were trained from scratch or fine-tuned from pre-trained weights on the ImageNet dataset: GoogLeNet[43] and its improved inception v2[21] and v4 [42] versions, inception-resnet-v2[42], ResNet-50 and ResNet-152 [19], ResNeXT[19], VGGNet[39] and even the AlexNet[30]. One can notice that inception v3 was not experimented despite the fact it is a recent model giving state of art performances in other image classification benchmarks. It is important to note that the best results were obtained with ensemble classifiers as for the KDE TUT team who learned and combined predictions from the ResNet-50 and two declinations of this architecture, as the CMP and FHDO BCSG teams with the inception-resnet-v2. Bootstrap aggregating (bagging) was also very efficient to extend the number of classifiers by learning several models with the same architecture but on different training and validation subsets. This is the case of the best run Mario MNB Run 4 for instance where at the end a total of 12 CNNs were learned and combined (7 GoogLeNet, 2 ResNet-152, 3 ResNeXT). CMP team combined also numerous models, a total of 17 models for instance for the CMP Run 1 with various sub-training datasets and bagging strategies, but all with the same inception-resnet-v2 architecture. Another key for succeeding the task was the use of data augmentation with usual transformations such as random cropping, horizontal flipping, rotation, for increasing artificially the number of training samples and helping the CNNs to generalize better. Mario MNB team added two more interesting transformations with slight modifications of the color saturation and lightness, and they correlated the intensity of these transformations with the diminution of the learning rate during training to let the CNNs see patches closer to the original image at the end of each training process. Last but not least, Mario MNB is the only team who extended the test images with similar transformations: each image from a given test observation was augmented with 4 more transformed images.

The most recent model race: We can make more comments about the choice of the CNN architectures: one can suppose that the most recent models such as inception-resnet-v2 or inception-v4 should lead to better results than older ones such as AlexNet, VGGNet and GoogleNet. For instance, the runs with GoogleNet and VGGNet by Sabanci [2], or with a PReLU version of inception-v1 by the PlantNet team, or with the historical AlexNet architecture by the UPB HES SO team [44] performed the worst results. However, one can

notice that the "winning" team used also numerous GoogLeNet models, while the old VGGNet used in UM run 2 gave quite high and intermediate results around a MRR of 0.8. This highlights how much the training strategies are important and how ensemble classifiers, bagging and data augmentation can greatly improve the performance even without the most recent architectures from the state of the art. Besides the use of ensemble of classifiers, some teams also tried to propose modifications of existing models. KDE TUT, in particular, modified the architecture of the first convolutional layers of ResNet-50 and report consistent improvements in their validation experiments [18]. CMP also reported slight improvements on the inception-resnet-v2 by using a maxout activation function instead of RELU. The UM team proposed an original architecture called Hybrid Generic-Organ learned on the trusted dataset (UM Run 1). Unfortunately, it performed worst than a standard VGGNet model learned on the noisy dataset (UM Run 2). This can be partially explained by the fact that the HBO-CNN model need tagged images (flower, fruit, leaf,...), a missing information for the noisy dataset and partially available for the trusted dataset.

The GPU race: Like discussed above, best performances were obtained with ensembles of very deep networks (up to 17 learned models for the CMP team) learned over millions of images produced with data augmentation techniques. In the case of the best run Mario MNB Run 4, test images were also augmented so that the prediction of a single image finally relies on the combination of 60 probability distributions (5 patches x 12 models). Overall, the best performing system requires a huge GPU consumption so that their use in data intensive contexts is limited by cost issues (*e.g.* the Pl@ntNet mobile application accounts for millions of users). A promising solution towards this issue could be to rely on knowledge distilling [20]. Knowledge distilling consists in transferring the generalization ability of a cumbersome model to a small model by using the class probabilities produced by the cumbersome model as soft targets for training the small model. Alternatively, more efficient architectures and learning procedures should be devised.

3 Task2: BirdCLEF

The general public as well as professionals like park rangers, ecological consultants and of course ornithologists are potential users of an automated bird song identifying system. A typical professional use would be in the context of wider initiatives related to ecological surveillance or biodiversity conservation. Using audio records rather than bird pictures is justified [5, 46, 45, 4]since birds are in fact not that easy to photograph and calls and songs have proven to be easier to collect and have been found to be species specific.

The 2017 edition of the task shares similar objectives and scenarios with the previous edition: (i) the identification of a particular bird species from a recording of one of its sounds, and (ii) the recognition of all species vocalising in so-called "soundscapes" that can contain up to several tens of birds vocalising.

The first scenario is aimed at developing new automatic and interactive identification tools, to help users and experts to assess species and populations from field recordings obtained with directional microphones. The soundscapes, on the other side, correspond to a much more passive monitoring scenario in which any multi-directional audio recording device could be used without or with very light user's involvement. These (possibly crowdsourced) passive acoustic monitoring scenarios could scale the amount of annotated acoustic biodiversity records by several orders of magnitude.

3.1 Data and task description

As the soundscapes appeared to be very challenging in 2015 and 2016 (with an accuracy below 15%), new soundscape recordings containing time-coded bird species annotations were integrated in the test set (so as to better understand what makes state-of-the-art methods fail on such contents). This new data was specifically created for BirdCLEF thanks to the work of three people: Paula Caycedo Rosales (ornithologist from the Biodiversa Foundation of Colombia and Instituto Alexander von Humboldt, Xeno-Canto member), Hervé Glotin (bio-accoustician, co-author of this paper) and Lucio Pando (field guide and ornithologist). In total, about 6,5 hours of audio recordings were collected and annotated in the form of time-coded segments with associated species name. This data is composed of two main subsets:

Peru soundscapes, about 2 hours (1:57:08) 32 annotated segments: recorded in the summer of 2016 with the support of Amazon Explorama Lodges and the SABIOD[12] project. These recordings have been realized in the jungle canopy at 35 meters high (the highest point of the area), and at the level of the Amazon river, in the Peruvian basin. The recordings are sampled at 96 kHz, 24 bits PCM, stereo, dual -12 dB, using multiple systems: TASCAM DR, SONY M10, Zoom H1.

Colombia soundscapes, about 4,5 hours (4:25:55), 1990 annotated segments: These documents were annotated by Paula Caycedo Rosales, ornithologist from the Biodiversa Foundation of Colombia and an active Xeno-Canto member.

In addition to these newly introduced records, the test set still contained the 925 soundscapes and 8,596 single species recordings of BirdCLEF 2016 (collected by the members of Xeno-Canto[13] network, see [16] for more details).

As for the training data, we consistently enriched the training set of the 2016 edition of the task, in particular to cover the species represented in the newly introduced time-coded soundscapes. Therefore, we extended the covered geographical area to the union of Brazil, Colombia, Venezuela, Guyana, Suriname,

[12] http://sabiod.org
[13] `http://www.xeno-canto.org/contributors`

French Guiana, Bolivia, Ecuador and Peru, and collected all Xeno-Canto records in these countries. We then kept only the 1500 species having the most recordings so as to get sufficient training samples per species (48,843 recordings in total). The training set has a massive class imbalance with a minimum of four recordings for *Laniocera rufescens* and a maximum of 160 recordings for *Henicorhina leucophrys*. Recordings are associated to various metadata such as the type of sound (call, song, alarm, flight, etc.), the date, the location, textual comments of the authors, multilingual common names and collaborative quality ratings.

Participants were asked to run their system so as to identify all the actively vocalising birds species in each test recording (or in each test segment of 5 seconds for the soundscapes). The submission *run files* had to contain as many lines as the total number of identifications, with a maximum of 100 identifications per recording or per test segment). Each prediction had to be composed of a species name belonging to the training set and a normalized score in the range $[0, 1]$ reflecting the likelihood that this species is singing in the segment. The used evaluation metric used was the Mean Average Precision. Up to 4 *run files* per participant could be submitted to allow evaluating different systems or system configurations.

3.2 Participants and results

78 research groups registered for the BirdCLEF 2017 challenge and downloaded the data. Only 5 of them finally submitted run files and technical reports. Details of the systems and the methods used in the runs are synthesized in the overview working note of the task [17] and further developed in the individual working notes of the participants ([28, 10, 37, 9]). Below we give more details about the 3 systems that performed the best runs:

DYNI UTLN system (Soundception) [37]: This system is based on an adaptation of the image classification model Inception V4 [42] extended with a time-frequency attention mechanism. The main steps of the processing pipeline are (i) the construction of multi-scaled time-frequency representations to be passed as RGB images to the Inception model, (ii) data augmentation (random hue, contrast, brightness, saturation, random crop in time and frequency domain) and (iii) the training phase relying on transfer learning from the initial weights of the Inception V4 model (learned in the visual domain using the ImageNet dataset).

TUCMI system [28]: This system is also based on convolutional neural networks (CNN) but using more classical architectures than the Inception model used by DYNI UTLN. The main steps of the processing pipeline are (i) the construction of magnitude spectrograms with a resolution of 512x256 pixels, which represent five-second chunks of audio signal, (ii) data augmentation (vertical

roll, Gaussian noise, Batch Augmentation) and (iii) the training phase relying on either a classical categorical loss with a softmax activation (TUCMI Run 1), or on a set of binary cross entropy losses with sigmoid activations as an attempt to better handle the multi-labeling scenario of the soundscapes (TUCMI Run 2). TUCMI Run 3 is an ensemble of 7 CNN models including the ones of Run 1 and Run 2. TUCMI Run 4 was an attempt to use geo-coordinates and time as a way to reduce the list of species to be recognized in the soundscapes recordings. Therefore, the occurrences of the eBird initiative were used complementary to the data provided within BirdCLEF. More precisely, only the 100 species having the most occurrences in the Loreto/Peru area for the months of June, July and August were kept in the training set.

Cynapse system [9]: This system is based on a multi-modal deep neural network taking audio samples and metadata as input. The audio is fed into a convolutional neural network using four convolutional layers. The additionally provided metadata is processed using fully connected layers. The flattened convolutional layers and the fully connected layer of the metadata were joined and put into a large dense layer. For the sound pre-processing and data augmentation, they used a similar pipeline as the best system of BirdCLEF 2016 described in [8]. The two runs Cynapse Run 2 and 3 mainly differ in the FFT window size used for constructing the time-frequency representation passed as input to the CNN (respectively 512 and 256). Cynapse Run 4 is an average of Cynapse Run 2 and 3.
Figure 2 reports the performance measured for the 18 submitted runs. For each run (*i.e.* each evaluated system), we report the Mean Average Precision for the three categories of queries: traditional mono-directional recordings (the same as the one used in 2016), non time-coded soundscape recordings (the same as the one used in 2016) and the newly introduced time-coded soundscape recordings. To measure the progress over last year, we also plot on the graph the performance of last year's best system [8]

It is remarked that all submitted runs were based on Convolutional Neural Networks (CNN) confirming the supremacy of this approach over previous methods (in particular the ones based on hand-crafted features which were performing the best until 2015). The best MAP of 0.71 (for the single species recordings) was achieved by the best system configuration of DYNI UTLN (Run 1).That rather similar to the MAP of 0.68 achieved last year by [8] but with 50% more species in the training set. Regarding the newly introduced time-coded soundscapes, the best system was also the one of DYNI UTLN (Run 1) whereas it did not introduce any specific features towards solving the multi-labeling issue. The main conclusions we can draw from the results are the following:
The network architecture plays a crucial role: Inception V4 that was known to be the state of the art in computer vision [42] also performed the best within the BirdCLEF 2017 challenge that is much different (time-frequency representations instead of images, a very imbalanced training set, mono- and multi-labeling scenarios, etc.). This shows that its architecture is intrinsically

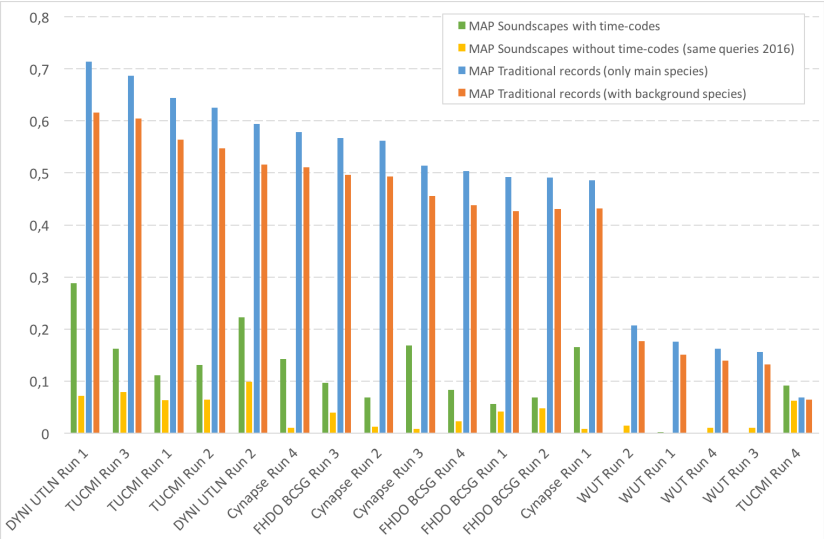

Figure 2: BirdCLEF 2017 results overview - Mean Average Precision

well-suited for a variety of machine-learning tasks across different domains.

The use of ensembles of networks improves the performance consistently: This can be seen through Cynapse Run 4 and TUCMI Run 3 that outperform the other respective runs of these participants.

The use of a multi-labeling training:. The use of the binary cross-entropy losses in TUCMI Run 2 did allow a slight performance gain compared to the classical softmax loss in TUCMI Run 1. Unfortunately, this was not enough to compensate the gains due to other factors in DYNI UTLN Run 1 (in particular the network architecture).

The use of metadata was not successful: The attempt of Cynapse did not allow a sufficient performance improvement to compensate the gains due to other factors (in particular the network architecture). TUCMI Run 4, working on the restricted list of the 100 most likely species according to eBird data, did not outperform the other TUCMI Runs on the soundscape test data.

4 Task3: SeaCLEF

The SeaCLEF 2017 task originates from the previous editions (2014 and 2015,2016) of marine organism identification in visual data for ecological surveillance and biodiversity monitoring. SeaCLEF 2017 significantly extends past editions in the tackled marine organisms species as well in the application tasks. The need

of automated methods for sea-related multimedia data and to extend the originally tasks is driven by the recent sprout of marine and ocean observation approaches (mainly imaging - including thermal - systems) and their employment for marine ecosystem analysis and biodiversity monitoring. Indeed in recent years we have assisted an exponential growth of sea-related multimedia data in the forms of images/videos/sounds, 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 that the technology may have in understanding and sustainably exploiting the sea/ocean.

4.1 Data and task description

The SeaCLEF 2017 challenge was composed of four subtasks and related datasets:

Subtask 1 - Automated Fish Identification and Species Recognition on Coral Reef Videos: The participants have access to a training set consisting of twenty underwater videos in which bounding boxes and fish species labels are provided. Then, providing a test set of 20 videos, the goal of the task is to automatically detect and recognize fish species. The evaluation metrics are the precision, the counting score (CS), and the normalized counting score (NCS). More information about these metrics and the whole set up of the subtask can be found in the LifeCLEF 2016 overview [26].

Subtask 2 - Automated Frame-level Salmon Identification in Videos for Monitoring Water Turbine: The participants have access to a training set consisting of eight underwater videos with frame-level annotations indicating the presence of salmons. Then, providing a test set of 8 videos, the goal of the task is to identify in which frames salmon appear. Such events are pretty rare and salmons are often very small, thus the task mainly pertains detection of rare events involving unclear objects (salmons).

Subtask 3 - Marine Animal Species Recognition using Weakly-Labelled Images and Relevance Ranking: Contrary to the previous subtasks, this one aims at classifying marine animals from 2D images. The main difficulties of the task are: 1) high similarity between species and 2) weak annotations, for training, gathered automatically from the Web and filtered by non-experts. In particular, the training dataset consists of up to 100 images for each considered species (in total 148 fish species). Training images are weakly labelled, i.e., Web images have been retrieved automatically from the Web using marine animal scientific names as query. The retrieved images were then filtered by non-experts who were instructed to only remove images not showing fish/marine animals. Furthermore, the relevance ranking to the query is provided for each crawled image and can be used during training.

Subtask 4 - Whale Individual Recognition: This subtask aims at au-

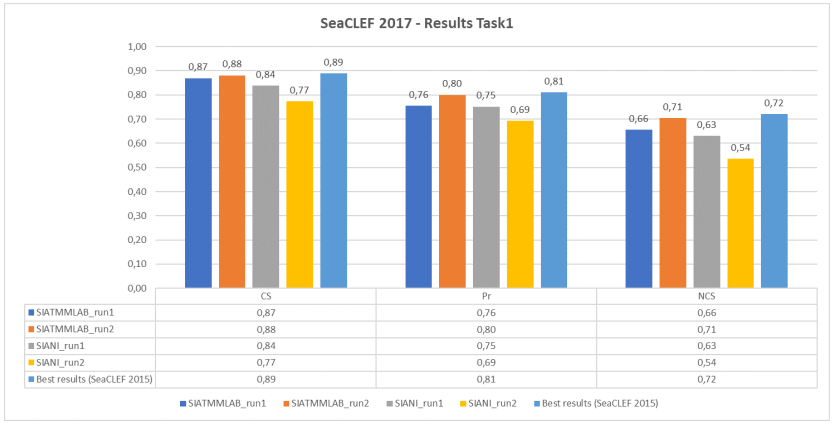

Figure 3: Results of SeaCLEF subtask 1 - Automated Fish Identification and Species Recognition on Coral Reef Videos

tomatically matching image pairs, over a large set of images, of same individual whales through the analysis of their caudal fins. Indeed, the caudal fin is the most discriminant pattern for distinguishing an individual whale from another. Finding the images that correspond to the same individual whale is a crucial step for further biological analysis (e.g. for monitoring population displacement) and it is currently done manually by human operators (hence a painful, error prone and unscalable process usually known as *photo-identification*). The metric used to evaluate each run is the Average Precision. More information about this metric and the whole set up of the subtask can be found in the LifeCLEF 2016 overview [26].

4.2 Participants and Results

Over 40 research teams registered and downloaded data for the SeaCLEF 2017 challenge. Only seven of the registered participants submitted runs. No one team participated to all the four challenge subtasks and only one team (SIATMMLAB) submitted runs for more than one task (subtask 1, 2 and 3).

Subtask 1 results: Figure 3 displays the results obtained by the 2 participating groups who submitted a total of 4 runs. Details of the methods of the team SIATMMLAB can be found in [48]. Unfortunately, none of them was able to outperform the baseline method that obtained the best results in LifeCLEF 2015 (by SNUMED [6]).

Subtask 2 results: For this task, only one participant (SIATMMLAB [48]) submitted only one run, achieving the following results: Precision = 0.04, Recall = 0.82 and F-measure = 0.07. The low performance, especially due to false positives, demonstrates the complexity of this task.

Subtask 3 results: Figure 4 reports the performance in terms of P@1, P@3 and

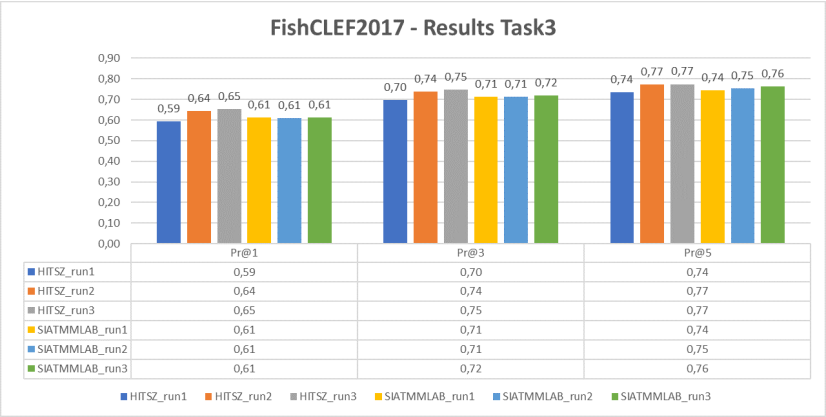

Figure 4: Results of SeaCLEF subtask 3 - Marine Animal Species Recognition using Weakly-Labelled Images and Relevance Ranking

P@5 of the two participant teams (HITSZ and SIATMMLAB) that submitted three runs each. Details of the methods used by the SIATMMLAB can be found in [48].

Subtask 4 results: Results achieved by the three groups who participated to subtask 4 are summarized in Table 1[14]. A first conclusion is that, as last year, using a RANSAC-like spatial consistency checking step is crucial for reaching good performance. The spatial arrangement of the local features is actually a precious information for rejecting the masses of mismatches obtained by simply matching the low level (SIFT) features. The two other elements explaining the big performance gap achieved by BME_DCLab_Run3 was (i) the use of a preliminary segmentation to separate the fin from the background and (ii), the use of a clustering algorithm on top of the image matching graph (to recover many pairs that were missed by the matching process but that can be infer by transitivity).

For all the first three subtasks, Convolutional Neural Networks based on either Inception or Res-Net were employed. The good performance achieved by CNNs over the subtasks 1 and 3 combined to the low one in subtask 3 confirm, what is already known in the literature, that CNNs work best with large data and that, instead, show limitations in case of detection of rare instances (despite no one of the employed methods discussed any data augmentation technique). Another observation is related to the fact that all methods exploit only visual cues to perform the tasks, which is expected given the nature of the released data. To overcome this limitation, this year we added in a task (subtask 3) another dimension related to image search ranking to support image classifica-

[14]We precise that there was probably a bug in the runfile MLRG_Run2 that performed abnormally low with regard to the used technique

Table 1: Individual whale identification results (SeaCLEF subtask 4)

Run name	**AP**	**method description**	paper
BME_DCLab_Run3	0.51	SIFT features matching + RANSAC spatial consistency filtering + random walks on the matches to discover clusters	[7]
ZenithINRIA_Run1	0.39	SIFT features matching (through multi-probe hashing) + RANSAC-like spatial consistency filtering	[27]
BME_DCLab_Run2	0.30	SIFT features matching + re-ranking	[7]
BME_DCLab_Run1	0.30	SIFT features matching	[7]
MLRG_Run2	0.01	Preprocessing using Grabcut Segmentation and Clustering Techniques + SIFT features matching (through FLANN) + re-ranking	[22]

tion using multimedia data, which, however, was not used by the participants. Finally, the low performance in subtask 3 and 4 clearly demonstrate that some problems, especially in the underwater domain, cannot be tackled merely by brute force learning (or fine-tuning) of low and middle-level visual features.

5 Conclusions and Perspectives

With about 130 research groups who downloaded LifeCLEF 2017 data and 18 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 rates can be reached by the evaluated systems, even on large number of species (up to 10,000 species). The most noticeable progress came from the deployment of new convolutional neural network architectures, confirming the fast growing progress of that techniques. Interestingly, the best performing system on the bird sounds recognition task was based on an the architecture of the image-based CNN Google model (Inception V4). This shows the convergence of the best performing technique whatever the targeted domain. Another important outcome was about the use of noisy Web data to train such deep learning models. The plant task confirmed that doing so allows achieving very good performance, even better than the one obtained with validated data. The combination of both noisy and clean data enabled even better performance gains, up to an amazing accuracy of 92% for the 10K species plant challenge. Despite these impressive results, there is still a large room of improvements for several of the evaluated challenges including: (i) the soundscapes for birds monitoring, (ii)

the underwater imagery for fish monitoring, and (iii) the photo-identification of whale individuals.

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 EADM GDR CNRS MADICS, BRILAAM STIC-AmSud, and Floris'Tic. The annotations of some soundscape were prepared with regreted wonderful Lucio Pando at Explorama Lodges, with the support of Pam Bucur, Marie Trone and H. Glotin. The organization of the SeaCLEF task is supported by the Ceta-mada NGO and the French project Floris'Tic.

References

[1] Affouard, A., Goeau, H., Bonnet, P., Lombardo, J.C., Joly, A.: Pl@ntnet app in the era of deep learning. In: 5th International Conference on Learning Representations (ICLR 2017), April 24-26 2017, Toulon, France (2017)

[2] Atito, S., Yanikoglu, B., Aptoula, E.: Plant identification with large number of classes: Sabanciu-gebzetu system in plantclef 2017. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)

[3] Baillie, J., Hilton-Taylor, C., Stuart, S.N.: 2004 IUCN red list of threatened species: a global species assessment. Iucn (2004)

[4] 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)

[5] 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)

[6] Choi, S.: Fish identification in underwater video with deep convolutional neural network: Snumedinfo at lifeclef fish task 2015. In: Working Notes of CLEF 2015 (Cross Language Evaluation Forum) (2015)

[7] Dávid Papp, F.M., Szűcs, G.: Image matching for individual recognition with sift, ransac and mcl. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)

[8] Elias Sprengel, Martin Jaggi, Y.K., Hofmann, T.: Audio based bird species identification using deep learning techniques. In: Working Notes of CLEF 2016 (Cross Language Evaluation Forum) (2016)

[9] Fazekas, B., Schindler, A., Lidy, T.: A multi-modal deep neural network approach to bird-song identication. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)

[10] Fritzler, A., Koitka, S., Friedrich, C.M.: Recognizing bird species in audio files using transfer learning. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)

[11] Gaston, K.J., O'Neill, M.A.: Automated species identification: why not? Philosophical Transactions of the Royal Society of London B: Biological Sciences 359(1444), 655–667 (2004)

[12] Goëau, H., Bonnet, P., Joly, A.: Plant identification based on noisy web data: the amazing performance of deep learning (lifeclef 2017). In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)

[13] Goëau, H., Bonnet, P., Joly, A., Bakic, V., Barthélémy, D., Boujemaa, N., Molino, J.F.: The imageclef 2013 plant identification task. In: CLEF 2013. Valencia (2013)

[14] 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)

[15] Goëau, H., Bonnet, P., Joly, A., Yahiaoui, I., Barthélémy, D., Boujemaa, N., Molino, J.F.: Imageclef2012 plant images identification task. In: CLEF 2012. Rome (2012)

[16] Goëau, H., Glotin, H., Planqué, R., Vellinga, W.P., Joly, A.: Lifeclef bird identification task 2016. In: CLEF 2016 (2016)

[17] Goëau, H., Glotin, H., Planqué, R., Vellinga, W.P., Joly, A.: Lifeclef bird identification task 2017. In: CLEF 2017 (2017)

[18] Hang, S.T., Aono, M.: Residual network with delayed max pooling for very large scale plant identification. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)

[19] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)

[20] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)

[21] Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. CoRR abs/1502.03167 (2015), `http://arxiv.org/abs/1502.03167`

[22] Jaisakthi, S., Mirunalini, P., Jadhav, R.: Automatic whale matching system using feature descriptor. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)

[23] Joly, A., Bonnet, P., Goëau, H., Barbe, J., Selmi, S., Champ, J., Dufour-Kowalski, S., Affouard, A., Carré, J., Molino, J.F., et al.: A look inside the pl@ ntnet experience. Multimedia Systems 22(6), 751–766 (2016)

[24] 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)

[25] 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)

[26] Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Champ, J., Planqué, R., Palazzo, S., Müller, H.: Lifeclef 2016: multimedia life species identification challenges. In: International Conference of the Cross-Language Evaluation Forum for European Languages 2016 (2016)

[27] 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 (Cross Language Evaluation Forum) (2016)

[28] Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M., Eibl, M.: Large-scale bird sound classification using convolutional neural networks. In: CLEF 2017 (2017)

[29] Krause, J., Sapp, B., Howard, A., Zhou, H., Toshev, A., Duerig, T., Philbin, J., Fei-Fei, L.: The unreasonable effectiveness of noisy data for fine-grained recognition. In: European Conference on Computer Vision. pp. 301–320. Springer (2016)

[30] 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)

[31] 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)

[32] Lasseck, M.: Image-based plant species identification with deep convolutional neural networks. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)

[33] 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)

[34] Lee, S.H., Chang, Y.L., Chan, C.S.: Lifeclef 2017 plant identification challenge: Classifying plants using generic-organ correlation features. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)

[35] Ludwig, A.R., Piorek, H., Kelch, A.H., Rex, D., Koitka, S., Friedrich, C.M.: Improving model performance for plant image classification with filtered noisy images. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)

[36] 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)

[37] Sevilla, A., Glotin, H.: Audio bird classification with inception v4 joint to an attention mechanism. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)

[38] 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)

[39] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)

[40] Šulc, M., Matas, J.: Learning with noisy and trusted labels for fine-grained plant recognition. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)

[41] Sullivan, B.L., Aycrigg, J.L., Barry, J.H., Bonney, R.E., Bruns, N., Cooper, C.B., Damoulas, T., Dhondt, A.A., Dietterich, T., Farnsworth, A., et al.: The ebird enterprise: an integrated approach to development and application of citizen science. Biological Conservation 169, 31–40 (2014)

[42] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv preprint arXiv:1602.07261 (2016)

[43] 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. pp. 1–9 (2015)

[44] Toma, A., Stefan, L.D., Ionescu, B.: Upb hes so @ plantclef 2017: Automatic plant image identification using transfer learning via convolutional neural networks. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)

[45] Towsey, M., Planitz, B., Nantes, A., Wimmer, J., Roe, P.: A toolbox for animal call recognition. Bioacoustics 21(2), 107–125 (2012)

[46] 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)

[47] Wilson, E.O.: The encyclopedia of life. Trends in Ecology & Evolution 18(2), 77–80 (2003)

[48] Zhuang, P., Xing, L., Liu, Y., Guo, S., Qiao, Y.: Marine animal detection and recognition with advanced deep learning models. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017)