



Plant Identification: Experts vs. Machines in the Era of Deep Learning

Pierre Bonnet, Hervé Goëau, Siang Thye Hang, Mario Lasseck, Milaň Sulc, Valéry Malécot, Philippe Jauzein, Jean-Claude Melet, Christian You, Alexis Joly

► To cite this version:

Pierre Bonnet, Hervé Goëau, Siang Thye Hang, Mario Lasseck, Milaň Sulc, et al.. Plant Identification: Experts vs. Machines in the Era of Deep Learning: Deep learning techniques challenge flora experts. Multimedia Tools and Applications for Environmental & Biodiversity Informatics, Chapter 8, Editions Springer, pp.131-149, 2018, Multimedia Systems and Applications Series, 978-3-319-76444-3. 10.1007/978-3-319-76445-0_8 . hal-01913277

HAL Id: hal-01913277

<https://hal.science/hal-01913277>

Submitted on 6 Nov 2018

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.

Chapter 8

Plant Identification: Experts vs. Machines in the Era of Deep Learning

Deep learning techniques challenge flora experts

Pierre Bonnet, Hervé Goëau, Siang Thye Hang, Mario Lasseck, Milan Šulc, Valéry Malécot, Philippe Jauzein, Jean-Claude Melet, Christian You, and Alexis Joly

Abstract Automated identification of plants and animals have improved considerably in the last few years, in particular thanks to the recent advances in deep learning. The next big question is how far such automated systems are from the human expertise. Indeed, even the best experts are sometimes confused and/or disagree between each others when validating visual or audio observations of living organism. A picture or a sound actually contains only a partial information that is usually not sufficient to determine the right species with certainty. Quantifying this uncertainty and comparing it to the performance of automated systems is of high interest for both computer scientists and expert naturalists. This chapter reports an experimental study following this idea in the plant domain. In total, 9 deep-learning systems

A. Joly,
Inria ZENITH team, France, e-mail: alexis.joly@inria.fr

P. Bonnet, H. Goëau
CIRAD, UMR AMAP, F-34398 Montpellier, France — AMAP, Univ Montpellier, CIRAD, CNRS, INRA, IRD, Montpellier, France, e-mail: pierre.bonnet@cirad.fr & herve.goeau@cirad.fr

S. T. Hang
Toyohashi University of Technology, Japan, e-mail: hang@kde.cs.tut.ac.jp

M. Lasseck
Museum fuer Naturkunde Berlin, Leibniz Institute for Evolution and Biodiversity Science, Germany, e-mail: Mario.Lasseck@mfn-berlin.de

M. Šulc
Czech Technical University in Prague, Czech Republic, e-mail: sulcmila@cmp.felk.cvut.cz

V. Malécot
IRHS, Agrocampus-Ouest, INRA, Universit d'Angers, Angers, France, e-mail: valery.malecot@agrocampus-ouest.fr

P. Jauzein
AgroParisTech UFR Ecologie Adaptations Interactions, Thiverval-Grignon, e-mail: p.jauzein@free.fr

J.-C. Melet, e-mail: jcd.melet@wanadoo.fr · C. You
Socit Botanique Centre Ouest, Nercillac, France e-mail: you.christian@neuf.fr

implemented by 3 different research teams were evaluated with regard to 9 expert botanists of the French flora. Therefore, we created a small set of plant observations that were identified in the field and revised by experts in order to have a near-perfect golden standard. The main outcome of this work is that the performance of state-of-the-art deep learning models is now close to the most advanced human expertise. This shows that automated plant identification systems are now mature enough for several routine tasks, and can offer very promising tools for autonomous ecological surveillance systems.

8.1 Introduction

Automated species identification was presented 15 years ago as a challenging but very promising solution for the development of new research activities in Taxonomy, Biology or Ecology [17]. With the development of an increasing number of web and mobile applications based on visual data analysis, the civil society was able in the recent years to evaluate the progress in this domain, and to provide new data for the development of large-scale systems. To evaluate the performance of automated plant identification technologies in a sustainable and repeatable way, a dedicated system-oriented benchmark was setup in 2011 in the context of the CLEF evaluation forum [21]. A challenge called PlantCLEF was organized in this context using datasets co-produced with actors of the civil society (such as educators, nature lovers, hikers). Years after years, the complexity and size of this testbed was increasing and allowed dozens of research teams to evaluate the progress and limits of the machine learning systems they developed. In 2017, the PlantCLEF challenge was organized on a dataset covering 10,000 plant species. This was the first evaluation at this scale in the world, and results were promising and impressive with accuracies reaching 90% of correct identification for the best system. This amazingly high performance raises the question of how far automated systems are from the human expertise and of whether there is a upper bound that can not be exceeded. A picture (or a set of pictures) actually contains only a partial information about the observed plant and it is often not sufficient to determine the right species with certainty. For instance, a decisive organ such as the flower or the fruit, might be not visible at the time the plant was observed. Or some of the discriminant patterns might be very hard or unlikely to be observed in a picture such as the presence of hairs or latex, or the morphology of the underground parts. As a consequence, even the best experts can be confused and/or disagree between each others when attempting to identify a plant from a set of pictures. Estimating this intrinsic data uncertainty according to human experts and comparing it to the performance of the best automated systems is of high interest for both computer scientists and expert naturalists.

A first step in that direction had been taken in 2014 through a first *Man vs. Machine experiment* conducted by some of the authors of this chapter [1]. At that time, it was concluded that machines were still far from performing as well as expert botanists. The best methods were only able to outperform the participants that de-

clared themselves as amateurs of botany. Computer vision has made great progress since that time, in particular thanks to the advances in deep learning. Thus, this chapter presents an upgraded *Human vs. Machine* experiment in the continuity of the previous study but using state-of-the-art deep learning systems. For a fair comparison, we also extended the evaluation dataset to more challenging species and we involved expert botanists with a much higher expertise on the targeted flora. In total, 9 deep-learning systems implemented by 3 different research teams were evaluated with regard to 9 expert botanists among the most renowned in Europe. The rest of this chapter is organized as follows. In section 8.2, we first return to the process of identifying a plant by an expert in order to fully understand its mechanisms and issues. Then, in section 8.3, we give an overview of the state-of-the-art in automated plant identification by synthesizing the results of the international evaluation campaign LifeCLEF 2017 co-organized by some of the authors of this chapter. Finally, in section 8.4.2, we report the results and analysis of our new *experts vs. machines* experiment.

8.2 Understanding the plant identification process by botanists

For a botanist, identifying a plant means associating a scientific name to an individual plant. More precisely, that means assigning that individual plant to a group, called a taxon. Such taxon had a name selected according to a set of rules. The delimitation of taxa and the scientific names applying to them are the result of a process called taxonomy (or systematics). This process is in the hands of a relatively low number of scientists. During that process, hundreds of herbarium sheets (i.e. dry plants collected during the past centuries and mounted on a large piece of paper together with annotations such as date, place, collector name) and usually a lower number of living plants are compared. Such comparison may be based on macromorphological, micromorphological or molecular data, manually or computationally analyzed. This comparison allows delimiting groups on the basis of certain features. This is a step where the taxonomist should tell apart variability in the morphology of the various parts of the individuals assigned to a peculiar taxa and features shared by all the specimens assigned to that taxa. The obtained groups are hierarchically organized, in a classification. The most common rank in such classifications is the species, but other ranks are used such as genus, family. Thus, identifying a plant is commonly treated as giving the scientific name at the specific rank. To do this, botanists relies on various methods involving memory and observation. As the result of a more or less long learning, botanist may have an implicit knowledge of the appearance and the variability of a species. Botanists may also rely on diagnostic characters, i.e. features (morphological) that tell apart individual of a peculiar species from any other species in an area. For example, any fan-like leaf, with a median sinus, collected on a tree, may be assigned to *Ginkgo biloba* (among living plants). Diagnostic characters may also correspond to some higher ranked taxa, for example, umbel-like inflorescences of *Apiaceae*. Additionally, botanists

may also use identification keys. Such tools consist in a set of alternatives, usually a pair of morphological characters (for example "leaf less than 10 cm long" versus "leaf equal or more than 10 cm long"). At each set of alternatives the botanist should select the morphological character best applying to his sample. This drives him toward another set of alternative or to the name of his material. Production of identification keys is a complex process, and, when allowing the identification of the plants of an area or a large taxonomic group (such as a family or a genus), are assembled in books called Floras or Monographs. Such published paper material is generally used by professional botanists, students, land managers or nature observers in general.

In the field, expert botanists may apply more or less simultaneously the three above-listed methods, i.e. implicit knowledge, diagnostic characters and keys. Further elements may also be involved in the identification process.

1. According to the period of the year, the location, the altitude, and the local environment (such as the level of sun exposure, the distance to a river stream or a disturbed area, the soil quality, etc.), the botanist will have in mind a selection of potential plant species that occur in the prospected area. The size and the quality of this potential species list will be directly related to his/her expertise and experience on this flora.
2. When a botanist sees one or several specimens of the same species to be identified, he/she will first select the one(s) that appear(s) to be the most informative, *e.g.* the most healthy, the one with the higher number of reproductive organs (flowers and fruits), or vegetative parts (stems, leaves). Due to this selection, he/she will access to the plant that will have the most higher volume of information, and that gives the best chances to lead to a correct identification.
3. Whether or not he/she uses a key, he/she may look attentively at several parts of the plants. The habit, *i.e.* the shape of the whole plant, will then usually be the first morphological attribute analyzed by the botanist, simply because it can be seen the farthest. The flowers and the fruits, if present, are also very regularly observed, as they are the most informative parts of the plant. Several attributes will be analyzed such as their position and insertion on the plants, their number, density, size, shape, structure, etc. Unfortunately, most plants are in flowers and fruits only a small fraction of the year (from few days to few weeks). In such situation, it is often necessary to analyze dry or dead flowers or fruits, if present. Regarding vegetative parts, most of the time, leaves are the first part to be analyzed. The botanist may examine their position and distribution along the stem, their shape, color, vein network, pubescence, etc. He/she will also try to observe uncommon particularities on the plant such as the presence of spines, of swollen parts, if some latex is flowing from the stem, or if the plant has a specific smell, etc.
4. The number of observed attributes is very variable from one plant to another. It depends on its growing stage, on the number of its morphological similar relatives for the considered flora, and of the expertise of the botanist. For example, in Europe, if a botanist identifies a specimen as belonging to the *Moraceae* family (based on the analysis of the leaf, fruit, and latex), he already knows that the num-

ber of potential species is very small. He/she doesn't have to look to many more characters for its species identification. On the other hand, if he/she identifies a specimen as a representative of the *Poaceae* family (based on the analysis of the fruits), he will have to look to many different characters as this family is one of the most rich in temperate regions (with hundreds or thousands of species).

5. If using a key, the botanist will look more precisely on the features considered at each set of alternatives, following the order used in the key (thus going from one part to another and back to the first for example). If he knows and recognizes on his sample diagnostic features applying to a group of species (genus, family for example), he may go directly to the part of the key dealing with that group. If he had implicit knowledge of the plant at hand, he may use the key in a reverse way. In such situation he will go to the set of alternatives that ends with the species' name he had in mind, and look at the characters that are used in the few previous sets of alternatives. Whatever the botanist selects himself the characters to look at or follows the order imposed by the key, for the same character (for example number of petals) the botanist will look at several relevant parts of the plant (in the example, several flowers), or even to several individuals, in order to prevent him looking at an anomaly.
6. During the whole identification process, botanists often use micro-lens. This allows them observing very small plant parts such as the inner parts of the flowers, or the hair shape on the leaf surface.
7. They may bring back to their offices specimens who are not easily identifiable in the field either because of lack of some characters or because of the size of such characters. They may also bring back specimens which are the most interesting for their research subject for further comparison with previously identified material.

The identification process in the field allows to better understand the assets and limits of an image-based identification. A picture (or a set of pictures) only provides a partial view of all the attributes that can be observed in the field. Indeed, the degree of informativeness of an observation is itself highly dependent on the botanical expertise of the photographer. Observations made by novices might for instance be restricted to the habit view which makes the identification impossible in some cases. Furthermore, the image-based identification process cannot be as iterative and dynamic as in the field. If the botanist realizes that an attribute is missing when following a dichotomous key, he cannot return to the observation of the plant.

8.3 State-of-the-art of automated plant identification

To evaluate the performance of automated plant identification technologies in a sustainable and repeatable way, a dedicated system-oriented benchmark was setup in 2011 in the context of ImageCLEF¹. Between 2011 and 2017, about 10 research groups participated yearly to this large collaborative evaluation by benchmarking

¹ www.imageclef.org

their image-based plant identification systems (see [21, 19, 20, 13, 12, 18, 10] for more details). The last edition, in 2017, was an important milestone towards building systems working at the scale of a continental flora [10]. To overcome the scarcity of expert training data for many species, the objective was 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. [11] concluded that training deep neural networks on noisy data was very effective for fine-grained recognition tasks. The PlantCLEF 2017 challenge completed their work in two main points: (i) it extended it to the plant domain and (ii), it scaled the comparison between clean and noisy training data to 10K of species. In the following subsections, we synthesize the methodology and main outcomes of this study. A more detailed description and a deeper analysis of the results can be found in [10].

8.3.1 Dataset and evaluation protocol

Two large training data sets both based on the same list of 10.000 plant species (living mainly in Europe and North America) were provided:

Trusted Training Set *EoL10K*: a trusted training set based on the online collaborative Encyclopedia Of Life (EoL)². The 10K species were selected as the most populated species in EoL data after a curation pipeline (taxonomic alignment, duplicates removal, herbarium sheets removal, etc.).

Noisy Training Set *Web10K*: a noisy training set built through Web crawlers (Google and Bing image search engines) and containing 1.1M images.

The main idea of providing both datasets was 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 that pictures can be potentially revised and rated on the EoL website. On the other side, the noisy set contained more images for a lot of species, but with several types and levels of noise 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 was a large sample of the

² <http://eol.org/>

query images submitted by the users of the mobile application PI@ntNet (iPhone³ & Android⁴). It contained a large number of wild plant species mostly coming from the Western Europe Flora and the North American Flora, but also species used all around the world as cultivated or ornamental plants.

8.3.2 Evaluated systems

Eight research groups participated to the evaluation. Details of the methods and systems they used are synthesized in the overview of the task [10] and further developed in the individual working notes of the participants (CMP [2], FHDO BCSG [3], KDE TUT [4], Mario MNB [5], Sabanci Gebze[6], UM [7] and UPB HES SO [8]). Participants were allowed to run up to 4 systems or 4 different configurations of their system. In total, 29 systems were evaluated. We give hereafter more details of the techniques and methods used by the 3 participants who developed the best performing systems:

Mario TSA Berlin, Germany, 4 runs, [5]: this participant used ensembles of fine-tuned CNNs pre-trained on ImageNet based on 3 architectures (GoogLeNet, ResNet-152 and ResNeXT) each trained with bagging techniques. Intensive data augmentation was used to train the models with random cropping, horizontal flipping, variations of saturation, lightness and rotation. Test images were also augmented and the resulting predictions averaged. *MarioTsaBerlin Run 1* results from the combination of the 3 architectures trained on the trusted datasets only (EOL and PlantCLEF2016). Run 2 exploited both the trusted and the noisy dataset to train four GoogLeNet's, one ResNet-152 and one ResNeXT. In Run 3, two additional GoogLeNet's and one ResNeXT were trained using a filtered version of the web dataset and images of the test set that received a probability higher than 0.98 in Run 1. The last and "winning" run *MarioTsaBerlin Run 4* finally combined all the 12 trained models.

KDE TUT, Japan, 4 runs, [4]: this participant introduced a modified version of the ResNet-50 model. Three of the intermediate convolutional layers used for downsampling were modified by changing the stride value from 2 to 1 and preceding it by max-pooling with a stride of 2, to optimize the coverage of the inputs. Additionally, they switched the downsampling operation with the convolution for delaying the downsampling operation. This has been shown to improve performance by the authors of the ResNet architecture themselves. During the training they used data augmentation based on random crops, rotations and optional horizontal flipping. Test images were also augmented through a single flip operation and the resulting predictions averaged. Since the original ResNet-50 architecture was modified,

³ <https://itunes.apple.com/fr/app/plantnet/id600547573?mt=8>

⁴ <https://play.google.com/store/apps/details?id=org.plantnet>

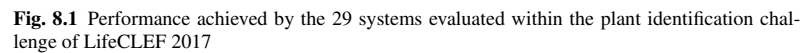
no fine-tuning was used and the weights were learned from scratch starting with a big learning rate value of 0.1. The learning rates were multiplied by 0.1 twice, throughout the training process, over 100 epochs according to a schedule ratio 4:2:1 indicating the number of iterations using the same learning rate (limited to a total number of 350 000 iterations in the case of the big noisy dataset due to technical limitations). Run 1, 2, 3 were trained respectively on the trusted dataset, noisy dataset, and both datasets. The final run 4 is a combination of the the outputs of the 3 runs.

CMP, Czech Republic, 4 runs, [2]: this participant based his work on the Inception-ResNet-v2 architecture [29] which introduces inception modules with residual connections. An additional maxout fully-connected layer with batch normalization was added on top of the network, before the classification fully-connected layer. Hard bootstrapping was used for training with noisy labels. A total of 17 models were trained using different training strategies such as: with or without maxout, with or without pre-training on ImageNet, with or without bootstrapping, with and without filtering of the noisy web dataset. CMP Run 1 is the combination of all the 17 networks by averaging their results. CMP Run 3 is the combination of the 8 networks that were trained on the trusted EOL data solely. CMP Run2 and CMP Run 4 are post-processings of CMP Run1 and CMP Run 3 aimed at compensating the asymmetry of class distributions between the test set and the training sets.

8.3.3 Results

We report in Figure 8.1 the performance achieved by the 29 evaluated systems. The used evaluation metric is the Mean Reciprocal Rank (MRR), *i.e.* the mean of the inverse of the rank of the correct species in the predictions returned by the evaluated system.

The first main outcome of that experiment was that the identification performance of state-of-the-art machine learning systems is impressive (with a median MRR around 0.8 and a maximal MRR of 0.92 for the best evaluated system *Mario MNB Run 4*). A second important conclusion was that the best results were obtained by the systems that were trained on both the trusted and the noisy dataset. Nevertheless, the systems that were trained exclusively on the noisy data (KDE TUT Run 2 and UM Run 2) performed better than the ones using the trusted data solely. This demonstrates that crawling the web without any filtering is a very effective way of creating large-scale training sets of plant observations. It opens the door to the possibility of building even larger systems working at the scale of the world's flora (or at least on 100K species). Regarding the machine learning methods used by the participants, it is noticeable that all evaluated systems 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 popular ImageNet dataset:



8.4 Human vs. Machine experiment

The amazingly high performance of machine learning techniques measured within the LifeCLEF 2017 challenge raises several questions regarding automated species identification: Is there still a margin of progression ? Are machine learning algorithms becoming as effective as human experts ? What is the maximum reachable performance when using only images as the main source of information ? As discussed above, a picture actually contains only a partial information about the observed plant and it is often not sufficient to determine the right species with certainty.

Estimating this intrinsic uncertainty, thanks to human experts, is thus of crucial interest to answer the question of whether the problem is solved from a computer science perspective. Therefore, we conducted two experiments described in the two following subsections. The first one (section 8.4.1) extends the results of the previous *Human vs. Machine experiment* that we conducted in 2014. It aims at measuring the progress that were made by automated identification systems since that time. The second experiment is based on a new testbed involving more challenging species and a panel of botanists with a much higher expertise on the targeted flora. It aims at answering the main questions asked in this paper. In the aim to start to response to these answers, we conducted several experiments with the some of the most state-of-the-art automated plant identification methods.

8.4.1 Progress made since 2014

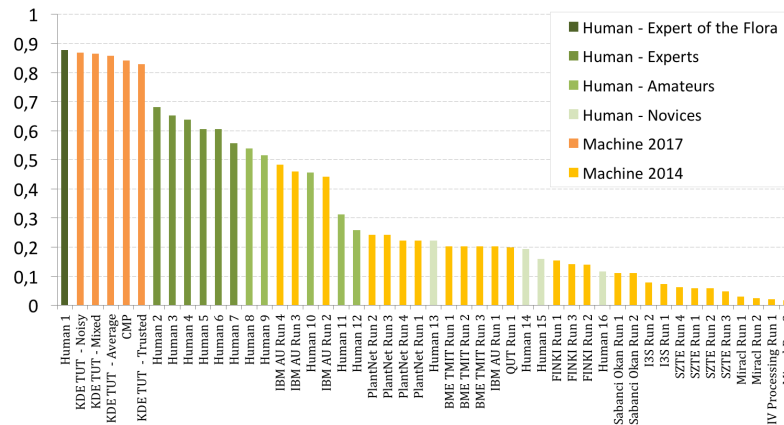
As discussed above, a first human vs. machine experiment [1] was conducted in 2014 based on 100 botanical observations that were identified by a panel of people with various expertise as well as by the systems evaluated within the LifeCLEF 2014 challenge. The 100 plants were selected at random from the whole set of observations of the PlantCLEF 2014 dataset [13]. This reduced test set was then shared with a large audience of potential volunteers composed of four target groups: **expert of the Flora** (highly skilled people such as taxonomists, expert botanists of the considered flora), **expert** (skilled people like botanists, naturalists, teachers, but not necessarily specialized on the considered Flora), **amateur** (people interested by plants in parallel of their professional activity and having a knowledge at different expertise levels), and **novice** (inexperienced users). The identification propositions were collected through a user interface presenting the 100 observations one by one (with one or several pictures of the different organs) and allowing the user to select up to three species for each observation thanks to a drop-down menu covering the 500 species of the PlantCLEF 2014 dataset. The most popular common names were also displayed in addition to the scientific name of the taxon to facilitate the participation of amateurs and novices. If the user didn't provide any species proposition for a given observation, the rank of the correct species was considered as infinite in the evaluation metric. We restricted the evaluation to the knowledge-based identification of plants, without any additional information or tools during the test. Concretely, the participants were not allowed to use external resources like field guides or Flora books. Among all contacted people, 20 of them finally accepted to participate: 1 expert of the French flora, 7 from the expert group, 7 from the amateur group, 5 from the novice group.

The performance of the 27 systems evaluated within LifeCLEF 2014 were computed on the same 100 observations than the ones identified by the human participants. To allow a fair comparison with human-powered identifications, the number of propositions was also limited to 3 (*i.e.* to the 3 species with the highest score for each test observation). To measure the progress since 2014, we did propose to the research

groups who participated to the 2017-th edition of LifeCLEF to run their system on the same testbed. The three research groups who developed the best performing systems accepted to do so but only two of them (CMP and KDE TUT) were eligible for that experiment (the systems of Mario MNB were actually trained on a dataset that contained the 100 observations of the test set). Figure 8.2 reports the Mean Reciprocal Rank scores obtained by all human participants and all automated identification systems ("machines"). The description of the systems that were evaluated in 2014 ("Machine 2014") can be found in [13]). The description of the systems that were evaluated in 2017 ("Machine 2017") can be found in section 8.3.2.

The main outcome of Figure 8.2 is the impressive progress that was made by machines between 2014 and 2017. This progress is mostly due to the use of recent deep convolutional neural network architectures but also to the use of a much larger training data. Actually, the systems experimented in 2014 were trained on 60.962 images, while the systems experimented in 2017 were trained on respectively 256,287 pictures (EOL data) for CMP Run3 and CMP Run4, KDE TUT Run1, and on 1.1M pictures (EOL + Web) for the other ones. Interestingly, the fact that the 2017 systems were trained on 10K species rather than 500 species did not affect their performance to much (this might even have increased their performance).

To conclude this first experiment with regard to our central question, one can notice that the quality of the identifications made by the best evaluated system is very close to the one of the only highly skilled botanist (qualified as "Expert of the flora" in Figure 8.2). All other participants, including the botanists who were not directly specialists of the targeted flora, were outperformed by the five systems experimented in 2017.



8.4.2 Experts vs. Machines experiment (2017)

In the aim to evaluate more precisely the capacities of state-of-the-art plant identification systems compared to human expertise, we did set up a new evaluation with (i) a more difficult test set and (ii), a group of highly skilled experts composed of the most renowned botanists of the considered flora.

8.4.2.1 Test set description

The new test set was created according to the following procedure. First, 125 plants were photographed between May and June 2017, a suitable period for the observation of flowers in Europe, in a botanical garden called the "Parc floral de Paris", and in a natural area located in the north of Montpellier city (southern part of France, close to the Mediterranean sea). The photos have been done with two smartphone models, an iPhone 5 and a Samsung S5 G930F, by a botanist and an amateur under his supervision. The selection of the species has been motivated by several criteria including (i) their membership to a difficult plant group (*i.e.* a group known as being the source of many confusions), (ii) the availability of well developed specimens with well visible organs on the spot and (iii), the diversity of the selected set of species in terms of taxonomy and morphology. About fifteen pictures of each specimen were acquired in order to cover all the informative parts of the plant. However, all pictures were not included in the final test set in order to deliberately hide a part of the information and increase the difficulty of the identification. Therefore, a random selection of only 1 to 5 pictures was operated for each specimen. In the end, a subset of 75 plants illustrated by a total of 216 images related to 33 families and 58 genera was selected. This test set is available online ⁵ under an open data license (CC0) in order to foster further evaluations by other research teams.

8.4.2.2 Experiment description

The test set was sent to 20 expert botanists, working part-time or full-time as taxonomist, botanist, or research scientist specialist of the considered flora. Few of them were recognized as non-professional expert botanists. Most of them are or were involved (i) in the conception of renowned books or tools dedicated to the French flora (ii) or in the study of large plant groups such as: Mediterranean flora[31]; Flora of ile-de-France[25]; Flora of cultivated fields[24]; author of the French national reference checklist[16]; author of the study of traits of Mediterranean species[27], publication on FloreNum⁶, etc. In addition to the test set, we provided to the experts an exhaustive list of 2,567 possible species, which is basically the subpart of the 10.000 species used in PlantCLEF2017 related to the French

⁵ <http://otmedia.lirmm.fr/LifeCLEF/mvsm2017/>

⁶ <http://www.florenum.fr/>

flora exclusively. Regarding the difficulty of the task and contrary to the previous human vs. machine experiment done in 2014, each participant was allowed to use any external resource (book, herbarium material, computational tool, web app, etc.), excepted automated plant identification tools such as Pl@ntNet. For each plant, the experts were allowed to propose up to 3 species names ranked by decreasing confidence. Among the 20 contacted experts, 9 of them finally completed the task on time and returned their propositions.

In parallel, we did propose to the research groups who participated to the 2017-th edition of LifeCLEF to run their system on the same testbed than the one sent to the experts. The three research groups who developed the best performing systems accepted to do so and provided a total of 9 run files containing the species predictions of their systems with different configurations (see section 8.3.2 for more details).

8.4.2.3 Results

Figure 8.3 displays the top-1 identification accuracy achieved by both the experts and the automated systems. Table 8.1 reports additional evaluation metrics namely the Mean Reciprocal Rank score (MRR), the top-2 accuracy and the top-3 accuracy. As a first noticeable outcome, none of the botanist correctly identified all observations. The top-1 accuracy of the experts is in the range 0.613 – 0.96, with a median value of 0.8. This illustrates the high difficulty of the task, especially when reminding that the experts were authorized to use any external resource to complete the task, Flora books in particular. It shows that a large part of the observations in the test set do not contain enough information to be surely identified when using classical identification keys. Only the four experts with an exceptional field expertise were able to correctly identify more than 80% of the observations.

Besides, Figure 8.3 shows that the top-1 accuracy of the evaluated systems is in the range 0.56-0.733 with a median value of 0.66. This is globally lower than the experts but it is noticeable that the best systems were able to perform similarly or slightly better than three of the highly skilled participating experts. Moreover, if we look at the top-3 accuracy values provided in Table 8.1, we can see that the best evaluated system returned the correct species within its top-3 predictions for more than 89% of the test observations. Only the two best experts obtained a higher top-3 accuracy. This illustrates one of the strength of the automated identification systems. They can return an exhaustive ranked list of the most probable predictions over all species whereas this is a very difficult and painful task for human experts. Figure 8.5 displays the further top-K accuracy values as a function of K for all the evaluated systems. It shows that the performance of all systems continues to increase significantly for values of K higher than 3 and then becomes more stable for values of K in the range [20-50]. Interestingly, the best system reaches a top-11 accuracy of 0.973%, *i.e.* the same value of the top-1 accuracy of the best expert, and a 100% top-K accuracy for $K = 39$. In view of the thousands of species in the whole check list, it is likely that such a system would be very useful even for the experts them-

selves. By providing an exhaustive short list of all the possible species, it would help them to not exclude any candidate species that they might have missed otherwise.

Run	RunType	MRR	Top1	Top2	Top3
Expert 1	human	0.967	0.96	0.973	0.973
Expert 2	human	0.947	0.933	0.96	0.96
Expert 3	human	0.88	0.88	0.88	0.88
Expert 4	human	0.864	0.84	0.88	0.893
Expert 5	human	0.8	0.8	0.8	0.8
Expert 6	human	0.78	0.773	0.787	0.787
Mario TSA Berlin - Noisy	machine	0.819	0.733	0.827	0.893
Mario TSA Berlin - Average	machine	0.805	0.733	0.813	0.853
Expert 7	human	0.74	0.72	0.76	0.76
KDE TUT Mixed	machine	0.786	0.707	0.8	0.827
Mario TSA Berlin - Filtered	machine	0.751	0.693	0.747	0.787
KDE TUT Average	machine	0.753	0.667	0.76	0.787
Expert 8	human	0.64	0.64	0.64	0.64
KDE TUT - Noisy	machine	0.75	0.64	0.8	0.813
Expert 9	human	0.62	0.613	0.627	0.627
CMP	machine	0.679	0.6	0.667	0.72
KDE TUT - Trusted	machine	0.656	0.573	0.613	0.72
Mario TSA Berlin - Trusted	machine	0.646	0.56	0.64	0.68

Table 8.1 Results of the human vs. machine 2017 experiments ordered by the top 1 accuracy

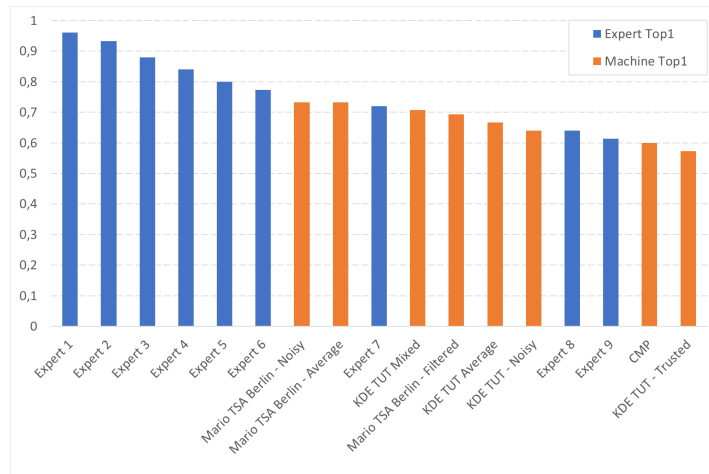


Fig. 8.3 Identification performance achieved by machines and human experts for the human vs. machine 2017 experiments

To further understand the limitations and the margin of progress of the evaluated identification systems, we did analyze more deeper which of the 75 test observations were correctly identified or missed compared to the expert's propositions. The main outcome of that analysis is that the automated systems perform as well as experts for about 86% of the observations, *i.e.* for 65 of the 75 test observations, the best evaluated system ranked the right species at a lower or equal rank than the best expert. Among the 10 remaining observations, 6 were correctly identified in the top-3 predictions of the best system and 9 in the top-5. Figure 8.4 displays 3 of the most difficult observations for the machines, *i.e.* the ones that were not identified by any system within the top-3 predictions. It is likely that the cause of the identification failure differs from an observation to another one. For the observation n74, for instance, it is likely that the main cause of failure is a mismatch between the training data and the test sample. Actually, the training samples of that species usually contain visible flowers whereas only the leaves are visible in the test sample. For the observation n29, it is more likely that the failure is due to the intrinsic difficulty of the *Carex* genus within which many species are very similar visually. Most of the proposals in machine runs are nevertheless under the *Carex* genus. For observation n43, the fact that most of images were not focused on a single leaf but dedicated to the illustration of the whole plant, which has a common aspect of a tuft of leaves, is probably at the origin of the misidentification. The small size of the discriminant organs and the cluttered background in the test sample makes the identification even more difficult.

Id	Species	Images				
74	Lathyrus vernus (L.) Bernh					
29	Carex distans L.					
43	Apium graveolens L.					

Fig. 8.4 Examples of observations well identified by experts but missed by the automated identification systems

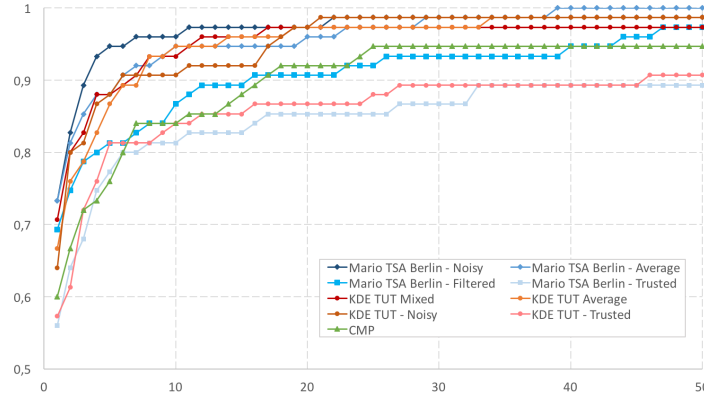


Fig. 8.5 Top-K accuracy of the evaluated system as a function of K

8.5 Conclusion and perspectives

The goal of this paper was to answer the question of whether automated plant identification systems still have a margin of progression or if they already perform as well as experts for identifying plants in images. Our study first shows that identifying plants from images solely is a difficult task, even for some of the highly skilled specialists who accepted to participate to the experiment. This confirms that pictures of plants only contain partial information and that it is often not sufficient to determine the right species with certainty. Regarding the performance of the machine learning algorithms, our study shows that there is still a margin of progression but that it is becoming tighter and tighter. Indeed, the evaluated systems were able to correctly identify as many plants as three of the experts whereas all of them were specialists of the considered flora. The best system was able to correctly classify 73.3% of the test samples including some belonging to very difficult taxonomic groups. This performance is still far from the best expert who correctly identified 96.7% of the test samples, however, as shown in our study, a strength of the automated systems is that they can return instantaneously an exhaustive list of all the possible species whereas this is a very difficult task for humans. We believe that this already makes them highly powerful tools for modern botany. Indeed, classical field guides or identification keys are much more difficult to handle and they require much more time to achieve a similar result. Furthermore, the performance of automated systems will continue to improve in the following years thanks to the quick progress of deep learning technologies. It is likely that systems capable of identifying the entire world's flora will appear in the next few years. The real question now is how to integrate them in pedagogical tools that could be used in teaching programs effectively and in a sustainable way. They have the potential to become essential tools

for teachers and students, but they should not replace an in-depth understanding of botany.

Acknowledgements Most of the work conducted in this paper was funded by the Floris’Tic initiative, especially for the support of the organization of the PlantCLEF challenge. Milan Šulc was supported by CTU student grant SGS17/185/OHK3/3T/13. Valéry Malécot was supported by ANR ReVeRIES (ref: ANR-15-CE38-0004-01). Authors would like to thank the botanists who accepted to participate to this challenge : Benoit Bock (PhotoFlora), Nicolas Georges (Cerema), Arne Saatkamp (Aix Marseille Universit, IMBE), François-Jean Rousselot, and Christophe Girod.

References

1. Bonnet, P., Joly, A., Goau, H., Champ, J., Vignau, C., Molino, J. F., Barthélémy Daniel Boujemaa, N. (2016). Plant identification: man vs. machine. *Multimedia Tools and Applications*, 75(3), 1647-1665.
2. Sulc, M., Matas, J. (2017). Learning with noisy and trusted labels for fine-grained plant recognition. *Working Notes of CLEF*, 2017.
3. Ludwig, A. R., Piorek, H., Kelch, A. H., Rex, D., Koitka, S., Friedrich, C. M. (2017). Improving model performance for plant image classification with filtered noisy images. *Working Notes of CLEF*, 2017.
4. Hang, S. T., Aono, M. (2017). Residual network with delayed max pooling for very large scale plant identification. *Working Notes of CLEF*, 2017.
5. Lasseck, M. (2017). Image-based plant species identification with deep convolutional neural networks. *Working Notes of CLEF*, 2017.
6. Atito, S., Yanikoglu, B., Aptoula, E. Plant Identification with Large Number of Species: SabanciU-GebzeTU System in PlantCLEF 2017.
7. Lee, S. H., Chang, Y. L., Chan, C. S. (2017). Lifeclef 2017 plant identification challenge: Classifying plants using generic-organ correlation features. *Working Notes of CLEF*, 2017.
8. Toma, A., Stefan, L. D., Ionescu, B. (2017). Upb hes so@ plantclef 2017: Automatic plant image identification using transfer learning via convolutional neural networks. *Working Notes of CLEF*, 2017.
9. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).
10. Goau, H., Bonnet, P., Joly, A. (2017). Plant identification based on noisy web data: the amazing performance of deep learning (lifeclef 2017). *CEUR Workshop Proceedings*.
11. Krause, J., Sapp, B., Howard, A., Zhou, H., Toshev, A., Duerig, T., ... Fei-Fei, L. (2016, October). The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision* (pp. 301-320). Springer International Publishing.
12. Goau, H., Bonnet, P., Joly, A. (2015). LifeCLEF Plant Identification Task 2015. *CEUR Workshop Proceedings*.
13. Goau, H., Joly, A., Bonnet, P., Selmi, S., Molino, J. F., Barthélémy, D., Boujemaa, N. (2014). Lifeclef plant identification task 2014. In *CLEF2014 Working Notes. Working Notes for CLEF 2014 Conference*, Sheffield, UK, September 15-18, 2014 (pp. 598-615). CEUR-WS.
14. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (pp. 248-255). IEEE.
15. Farnsworth, E. J., Chu, M., Kress, W. J., Neill, A. K., Best, J. H., Pickering, J., ... Ellison, A. M. (2013). Next-generation field guides. *BioScience*, 63(11), 891-899.

16. Bock B. (2014) Référentiel des trachéophytes de France métropolitaine réalisé dans le cadre d'une convention entre le Ministre chargé de l'Ecologie, le MNHN, la FCBN et Tela Botanica. Editeur Tela Botanica. Version 2.01 du 14 février 2014.
17. Gaston, K. J., O'Neill, M. A. (2004). Automated species identification: why not?. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 359(1444), 655-667.
18. Goau, H., Bonnet, P., Joly, A. (2016). Plant identification in an open-world (lifeclef 2016). In *Working Notes of CLEF 2016-Conference and Labs of the Evaluation forum*, Évora, Portugal, 5-8 September, 2016. (pp. 428-439).
19. Goau, H., Bonnet, P., Joly, A., Yahiaoui I., Barthelemy D., Boujemaa N., Molino J.-f. (2012). The ImageCLEF 2012 Plant Identification Task. *CEUR Workshop Proceedings*.
20. Goau, H., Joly, A., Bonnet, P., Bakic, V., Barthélémy, D., Boujemaa, N., Molino, J. F. (2013, October). The imageCLEF plant identification task 2013. In *Proceedings of the 2nd ACM international workshop on Multimedia analysis for ecological data* (pp. 23-28). ACM.
21. Goau, H., Bonnet, P., Joly, A., Boujemaa, N., Barthelemy, D., Molino, J. F., ... Picard, M. (2011, September). The ImageCLEF 2011 plant images classification task. In *ImageCLEF 2011*.
22. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
23. Ioffe, S., Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning* (pp. 448-456).
24. Jauzein P. (1995). *Flore des champs cultivés*. Num.3912, Editions Quae.
25. Jauzein P., Nawrot O. (2013). *Flore d'Ile-de-France: clés de détermination, taxonomie, statuts*, Editions Quae.
26. Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).
27. Saatkamp, A., Affre, L., Dutoit, T., Poschlod, P. (2011). Germination traits explain soil seed persistence across species: the case of Mediterranean annual plants in cereal fields. *Annals of botany*, 107(3), 415-426.
28. Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
29. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. A. (2017). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *AAAI* (pp. 4278-4284).
30. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).
31. Tison, J. M., Jauzein, P., Michaud, H., Michaud, H. (2014). *Flore de la France méditerranéenne continentale*. Turriers: Naturalia publications.