AI-based Identification of Plant Photographs from Herbarium Specimens

Hervé H.G. Goëau[‡], Pierre Bonnet[§], Alexis A.J. Joly[|]

‡ CIRAD, Montpellier, France
§ UMR AMAP, CIRAD, Montpellier, France
| Inria, Montpellier, France

Corresponding author: Alexis A.J. Joly (alexis.joly@inria.fr)

Abstract

Automated plant identification has recently improved significantly due to advances in deep learning and the availability of large amounts of field photos. As an illustration, the classification accuracy of 10K species measured in the LifeCLEF challenge (Goëau et al. 2018) reached 90%, very close to that of human experts. However, the profusion of field images only concerns a few tens of thousands of species, mainly located in North America and Western Europe. Conversely, the richest regions in terms of biodiversity, such as tropical countries, suffer from a shortage of training data (Pitman 2021). Consequently, the identification performance of the most advanced models on the flora of these regions is much lower (Goëau et al. 2019).

Nevertheless, for several centuries, botanists have systematically collected, catalogued, and stored plant specimens in herbaria. Considerable recent efforts by the biodiversity informatics community, such as DiSSCo (Addink et al. 2018) and iDigBio (Matsunaga et al. 2013), have made millions of digitized specimens from these collections available online. A key question is therefore whether these digitized specimens could be used to improve the identification performance of species for which we have very few (if any) photos. However, this is a very difficult problem from a machine learning point of view. The visual appearance of a herbarium specimen is actually very different from a field photograph because the specimens are dried and crushed on a herbarium sheet before being digitized (Fig. 1).

To advance research on this topic, we built a large dataset that we shared as one of the challenges of the LifeCLEF 2020 (Goëau et al. 2020) and 2021 evaluation campaigns (Goëau et al. 2021). It includes more than 320K herbarium specimens collected mostly from the Guiana Shield and the Northern Amazon Rainforest, focusing on about 1K plant species of the French Guiana flora. A valuable asset of this collection is that some of the specimens are accompanied by a few photos of the same specimen, allowing for more precise machine learning. In addition to this training data, we also built a test set for model evaluation, composed of 3,186 field photos collected by two of the best experts on Guyanese flora.

Based on this dataset, about ten research teams have developed deep learning methods to address the challenge (including the authors of this abstract as the organizing team). A detailed description of these methods can be found in the technical notes written by the participating teams (Goëau et al. 2020, Goëau et al. 2021). The methods can be divided into two categories:

- those based on classical convolutional neural networks (CNN) trained simply by mixing digitized specimens and photos and
- those based on advanced domain adaptation techniques with the objective of learning a joint representation space between field and herbarium representations.

The domain adaptation methods themselves were of two types, those based on

1. adversarial regularization (Motiian et al. 2017) to force herbarium specimens and photos to have the same representations,
2. metric learning to maximize inter-species distances and minimize intra-species distances in the representation space

In Table 1, we report the results achieved by the different methods evaluated during the 2020 edition of the challenge. The evaluation metric used is the mean reciprocal rank (MRR), i.e., the average of the inverse of the rank of the correct species in the list of the predicted species. In addition to this main score, a second MRR score is computed on a subset of the test set composed of the most difficult species, i.e., the ones that are the least frequently photographed in the field. The main outcomes we can derive from these results are the following:

Classical deep learning models fail to identify plant photos from digitized herbarium specimens. The best classical CNN trained on the provided data resulted in a very low MRR score (0.011). Even with the of use additional training data (e.g. photos and digitized herbarium from GBIF) the MRR score remains very low (0.039).

Domain adaptation methods provide significant improvement but the task remains challenging. The best MRR score (0.180) was achieved by using adversarial regularization (FSDA Motiian et al. 2017). This is much better than the classical CNN models but there is still a lot of progress to be made to reach the performance of a truly functional identification system (the MRR score on classical plant identification tasks can be up to 0.9).

No method fits all. As shown in Table 1, the metric learning method has a significantly better MRR score on the most difficult species (0.107). However, the performance of this method on the species with more photos is much lower than the adversarial technique.

In 2021, the challenge was run again but with additional information provided to train the models, i.e., species traits (plant life form, woodiness and plant growth form). The use of the species traits allowed slight performance improvement of the best adversarial adaptation method (with a MRR equal to 0.198).

In conclusion, the results of the experiments conducted are promising and demonstrate the potential interest of digitized herbarium data for automated plant identification. However, progress is still needed before integrating this type of approach into production applications.

Keywords

plant, identification, photos, herbarium, domain adaptation

Presenting author

Hervé Goëau

Presented at

TDWG 2021

References

- Addink W, Koureas D, Casino A (2018) DiSSCo: The physical and data infrastructure for Europe's Natural Science Collections. EGU general assembly conference abstracts.
- Goëau H, Bonnet P, Joly A (2018) Overview of ExpertLifeCLEF 2018: how far automated identification systems are from the best experts ? CLEF task overview 2018, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2018, Avignon, France.
- Goëau H, Bonnet P, Joly A (2019) Overview of LifeCLEF Plant Identification task 2019: diving into data deficient tropical countries. CLEF task overview 2019, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2019, Lugano, Switzerland.
- Goëau H, Bonnet P, Joly A (2020) Overview of LifeCLEF Plant Identification task 2020. CLEF task overview 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece.
- Goëau H, Bonnet P, Joly A (2021) Overview of PlantCLEF 2021: cross-domain plant identification. Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum.
- Matsunaga A, Figueiredo R, Thompson A, Traub G, Beaman R, Fortes JA (2013) Integrated digitized biocollections (iDigBio) cyberinfrastructure status and futures. TDWG 2013 Annual Conference.
- Motiian S, Jones Q, Iranmanesh S, Doretto G (2017) Few-shot adversarial domain adaptation. Advances in Neural Information Processing Systems.
- Pitman NA (2021) Identifying gaps in the photographic record of the vascular plant flora of the Americas. Nature Plants https://doi.org/10.1038/s41477-021-00974-2

Figure 1.

A herbarium sheet (left) and a field photo (right) of the same individual plant (*Unonopsis stipitata* Diels).

Table 1.

Synthesis of the obtained results.

	MRR	MRR on most difficult species
Best classical CNN	0.011	0.004
Best classical CNN with additional training data	0.039	0.007
Best domain adaptation method based on metric learning	0.121	**0.107**
Best domain adaptation method based on adversarial regularization	**0.180**	0.052