Automated Herbarium Specimen Identification using Deep Learning

Jose Carranza-Rojas, Alexis Joly, Pierre Bonnet, Hervé Goëau, Erick Mata-Montero

▶ **To cite this version:**

Jose Carranza-Rojas, Alexis Joly, Pierre Bonnet, Hervé Goëau, Erick Mata-Montero. Automated Herbarium Specimen Identification using Deep Learning. TDWG: Biodiversity Information Standards, Oct 2017, Ottawa, Canada. pp.e20302, 10.3897/tdwgproceedings.1.20302 . hal-01629142

HAL Id: hal-01629142
https://hal.science/hal-01629142

Submitted on 30 Jul 2019

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.

PROCEEDINGS OF TDWG

Biodiversity Information Science and Standarts 1: e20302
doi: 10.3897/tdwgproceedings.1.20302

OPEN ACCESS

Conference Abstract

Automated Herbarium Specimen Identification using Deep Learning

Jose Carranza-Rojas‡, Alexis A.J. Joly§, Pierre Bonnet|, Hervé H.G. Goëau|, Erick Mata-Montero¶

‡ Costa Rica Institute of Technology, PARMA Group, School of Computing, Cartago, Costa Rica
§ INRIA, Montpellier, France
| UMR AMAP, CIRAD, Montpellier, France
¶ Costa Rica Institute of Technology, School of Computing, Cartago, Costa Rica

Corresponding author: Jose Carranza-Rojas (jcarranza@itcr.ac.cr), Alexis A.J. Joly (alexis.joly@inria.fr)

Received: 14 Aug 2017 | Published: 16 Aug 2017

Citation: Carranza-Rojas J, Joly A, Bonnet P, Goëau H, Mata-Montero E (2017) Automated Herbarium Specimen Identification using Deep Learning. Proceedings of TDWG 1: e20302.
https://doi.org/10.3897/tdwgproceedings.1.20302

Abstract

Hundreds of herbarium collections have accumulated a valuable heritage and knowledge of plants over several centuries (Page et al. 2015). Recent initiatives, such as iDigBio (https://www.idigbio.org), aggregate data from and images of vouchered herbarium sheets (and other biocollections) and make this information available to botanists and the general public worldwide through web portals. These ambitious plans to transform and preserve these historical biodiversity data into digital format are supported by the United States National Science Foundation (NSF) Advancing the Digitization of Natural History Collections (ADBC) and the digitization is done by the Thematic Collections Networks (TCNs) funded under the ADBC program. However, thousands of herbarium sheets are still unidentified at the species level while numerous sheets should be reviewed and updated following more recent taxonomic knowledge. These annotations and revisions require an unrealistic amount of work for botanists to carry out in a reasonable time (Bebber et al. 2010). Computer vision and machine learning approaches applied to herbarium sheets are promising (Wijesingha and Marikar 2012) but are still not well studied compared to automated species identification from leaf scans or pictures of plants taken in the field.

In a recent study, we evaluate the accuracy with which herbarium images can be potentially exploited for species identification with deep learning technology (Carranza-Rojas et al.

© Carranza-Rojas J et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

2017), particularly Convolutional Neural Networks (CNN) (Szegedy et al. 2015). This type of network allows automatic learning of the most prominent visual patterns in the images since they are trainable end-to-end (thus, differentiable), as opposed to previous approaches that use custom, hand-made feature extractors. A first challenge is to use herbarium sheet images alone to automatically identify the species of plants mounted on herbarium sheets. Secondly, we propose studying if the combination of herbarium sheet images with photos of plants in the field (Joly et al. 2015, Carranza-Rojas and Mata-Montero 2016) is a viable idea to train models that provide accurate results during identification. Finally, we explore if herbarium images from one region with a specific flora can be used in transfer learning (a technique in deep learning that first allows training a model with a dataset and then once trained, uses the weighted results to train another model with that knowledge as the baseline) to another region with other species; for example, in a region under-represented in terms of collected data.

Our evaluation shows that the accuracy for species identification with deep learning technology, based on herbarium images, reaches 90.3% on a dataset of more than 1200 European plant species. This could potentially lead to the creation of a semi-, or even fully automated system to help taxonomists and experts with their annotation, classification, and revision works.

In this paper, we take a closer look at the accuracy levels achieved with respect to the first two challenges. We evaluate the accuracy levels for each species included in the dataset, which encompasses 253,733 images, 1,204 species.

Keywords

Biodiversity Informatics; Computer Vision; Deep Learning; Plant Identification; Herbaria

Presenting author

Jose Carranza-Rojas

Acknowledgements

Thanks to the National Museum of Costa Rica for their help with the collection, identification, and digitization of samples in the Costa Rican leaf-scan dataset. Special thanks to the Costa Rica Institute of Technology for partially sponsoring this research. We would also like to thank the large community that has actively engaged in iDigBio initiatives, for the valuable access to their herbarium data.

Hosting institution

Costa Rica Institute of Technology, Costa Rica

INRIA, France

CIRAD, France

References

- Bebber DP, Carine MA, Wood JRI, Wortley AH, Harris DJ, Prance GT, Davidse G, Paige J, Pennington TD, Robson NKB, Scotland RW (2010) Herbaria are a major frontier for species discovery. Proceedings of the National Academy of Sciences 107 (51): 22169-22171. https://doi.org/10.1073/pnas.1011841108
- Carranza-Rojas J, Mata-Montero E (2016) Combining leaf shape and texture for Costa Rican plant species identification. CLEI Electronic Journal 19 (1): 7:1-7:29. https://doi.org/10.19153/cleiej.19.1.7
- Carranza-Rojas J, Goeau H, Bonnet P, Mata-Montero E, Joly A (2017) Going deeper in the automated identification of herbarium specimens. BMC Evolutionary Biology 17 (1): 181. https://doi.org/10.1186/s12862-017-1014-z
- Joly A, Goëau H, Glotin H, Spampinato C, Bonnet P, Vellinga W, Planqué R, Rauber A, Palazzo S, Fisher B, Müller H (2015) LifeCLEF 2015: Multimedia life species identification challenges. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-319-24027-5_46
- Page L, MacFadden B, Fortes J, Soltis P, Riccardi G (2015) Digitization of biodiversity collections reveals biggest data on biodiversity. BioScience 65 (9): 841-842. https://doi.org/10.1093/biosci/biv104
- Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) https://doi.org/10.1109/cvpr.2015.7298594
- Wijesingha D, Marikar F (2012) Automatic detection system for the identification of plants using herbarium specimen images. Tropical Agricultural Research 23 (1): 42-50. https://doi.org/10.4038/tar.v23i1.4630