Species Detection and Segmentation of Multi-specimen Historical Herbaria

Krishna Kumar Thirukokaranam Chandrasekar‡, Kenzo Milleville‡, Steven Verstockt‡

‡ Ghent University, Ghent, Belgium

Corresponding author: Krishna Kumar Thirukokaranam Chandrasekar (krishnakumar.tc@ugent.be)

Abstract

Historically, herbarium specimens have provided users with documented occurrences of plants in specific locations over time. Herbarium collections have therefore been the basis of systematic botany for centuries (Younis et al. 2020). According to the latest summary report based on the data from Index Herbariorum, there are around 3400 active herbaria in the world containing 397 million specimens that are spread across 182 countries (Thiers 2021). Exponential growth in high quality image capturing devices induced by the enormous amount of uncovered collections has further led to rising interest in large scale digitization initiatives across the world (Le Bras et al. 2017). As herbarium specimens are increasingly becoming digitised and accessible in online repositories, an important need has also emerged to develop automated tools to process and enrich these collections to facilitate better access to the preserved archives.

This rising number of digitised herbarium sheets provides an opportunity to employ computer-based image processing techniques, such as deep learning, to automatically identify species and higher taxa (Carranza-Rojas and Joly 2018, Carranza-Rojas et al. 2017, Younis et al. 2020) or to extract other useful information from the herbaria sheets, such as detecting handwritten text, color bars, scales and barcodes. The species identification task works well for herbarium sheets that have only one species in a page. However, there are many herbarium books that have multiple species on the same page (as shown in Fig. 1) for which the complexity of the identification problem increases tremendously. It also involves a great deal of time and effort if they are to be enriched manually. In this work, we propose a pipeline that can automatically detect, identify, and enrich plant species in multi-specimen herbaria.

The core idea of the pipeline is to detect unique plant species and handwritten text around the plant species and map the text to the correct plant species. As shown in Fig. 2, the proposed pipeline begins with the pre-processing of the images. The images are rotated and aligned such that the longest edge is maintained as its height. In the case of herbarium books, the pages are detected and morphological transformations are performed to reduce occlusions (Thirukokaranam Chandrasekar and Verstockt 2020). A

YOLOv3 (You Only Look Once version 3) object detection model (Zhao and Li 2020) is trained from scratch to detect *plants* and *text*. The model was trained on a dataset of single species herbarium sheets with a mosaic augmentation technique to extend the *plants* model to detect multiple species. The first results of the training shows impressive results although it could be further improved with more labelled data. We also plan to train an object segmentation model and contrast its performance with the plant detection model for multi-specimen herbarium sheets. After detecting both the plants and the text, the text will be recognized with a state-of-the-art handwritten text recognition (HTR) model. The recognized text can then be matched with a database of specimens, to identify each detected specimen. Furthermore, additional textual metadata (e.g. date, locality, collector's name, institution) visible on the sheet will be recognized and used to enrich the collection.

Keywords

plant detection, handwritten text recognition (HTR), object segmentation

Presenting author

Krishna Kumar Thirukokaranam Chandrasekar

Presented at

TDWG 2021

Conflicts of interest

References

- Carranza-Rojas J, Goeau H, Bonnet P, Joly A, et al. (2017) Going deeper in the automated identification of Herbarium specimens. BMC Evolutionary Biology 17.
- Carranza-Rojas J, Joly A, et al. (2018) Automated Identification of Herbarium Specimens at Different Taxonomic Levels. In: Bonnet P, et al. (Ed.) Multimedia Tools and Applications for Environmental & Biodiversity Informatics. Springer
- Le Bras G, Pignal M, Jeanson M (2017) The French Muséum National d'Histoire Naturelle vascular plant herbarium collection dataset. Scientific Data 4.
- Thiers B (2021) The world's herbaria 2020: A summary report based on data from Index Herbariorum. William and Lynda Steere Herbarium, The New York Botanical Garden
- Thirukokaranam Chandrasekar KK, Verstockt S (2020) Page Boundary Extraction of Bound Historical Herbaria. Proceedings of the 12th International Conference on Agents and Artificial Intelligence https://doi.org/10.5220/0009154104760483

- Younis S, Schmidt M, Weiland C, Dressler S, Seeger B, Hickler T (2020) Detection and annotation of plant organs from digitised herbarium scans using deep learning. Biodiversity Data Journal 8.
- Zhao L, Li S (2020) Object Detection Algorithm Based on Improved YOLOv3. Electronics 9 (3). https://doi.org/10.3390/electronics9030537

Figure 1.

An example of multi-specimen herbarium page with handwritten text descriptions around it.

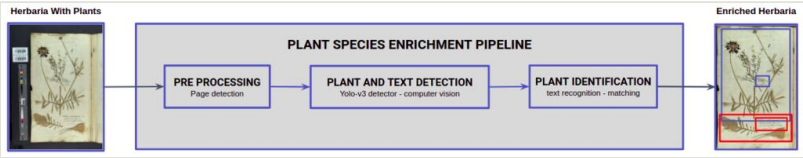

Figure 2.

Proposed enrichment pipeline for automatic identification of plant species in a multi species herbarium book.