Challenges in Curating 2D Multimedia Data in the Application of Machine Learning in Biodiversity Image Analysis

Yasin Bakış‡, Xiaojun Wang‡, Hank Bart‡

‡ Tulane University, New Orleans, United States of America

Corresponding author: Yasin Bakış (ybakis@tulane.edu)

Abstract

Over 1 billion biodiversity collection specimens ranging from fungi to fish to fossils are housed in more than 1,600 natural history collections across the United States. The digitization of these specimens has risen significantly within the last few decades and this is only likely to increase, as the use of digitized data gains more importance every day. Numerous experiments with automated image analysis have proven the practicality and usefulness of digitized biodiversity images by computational techniques such as neural networks and image processing. However, most of the computational techniques to analyze images of biodiversity collection specimens require a good curation of this data. One of the challenges in curating multimedia data of biodiversity collection specimens is the quality of the multimedia objects—in our case, two dimensional images. To tackle the image quality problem, multimedia needs to be captured in a specific format and presented with appropriate descriptors. In this study we present an analysis of two image repositories each consisting of 2D images of fish specimens from several institutions—the Integrated Digitized Biocollections (iDigBio) and the Great Lakes Invasives Network (GLIN). Approximately 70 thousand images have been processed from the GLIN repository and 450 thousand images have been processed from the iDigBio repository and their suitability assessed for use in neural network-based species identification and trait extraction applications. Our findings showed that images that came from the GLIN dataset were more successful for image processing and machine learning purposes. Almost 40% of the species have been represented with less than 10 images while only 20% have more than 100 images per species.

We identified and captured 20 metadata descriptors that define quality and usability of the image. According to the captured metadata information, 70% of the GLIN dataset images were found to be useful for further analysis according to the overall image quality score. Quality issues with the remaining images included: curved specimens, non-fish objects in the images such as tags, labels and rocks that obstructed the view of the

specimen; color, focus and brightness issues; folded or overlapping parts as well as missing parts.

We used both the web interface and the API (Application Programming Interface) for downloading images from iDigBio. We searched for all fish genera, families and classes in three different searches with the images-only option selected. Then we combined all of the search results and removed duplicates. Our search on the iDigBio database for fish taxa returned approximately 450 thousand records with images. We narrowed this down to 90 thousand fish images aided by the multimedia metadata with the downloaded search results, excluding some non-fish images, fossil samples, X-ray and CT (computed tomography) scans and several others. Only 44% of these 90 thousand images were found to be suitable for further analysis.

In this study, we discovered some of the limitations of biodiversity image datasets and built an infrastructure for assessing the quality of biodiversity images for neural network analysis. Our experience with the fish images gathered from two different image repositories has enabled describing image quality metadata features. With the help of these metadata descriptors, one can simply create a dataset for a desired image quality for the purpose of analysis. Likewise, the availability of the metadata descriptors will help advance our understanding of quality issues, while helping data technicians, curators and the other digitization staff be more aware of multimedia.

Keywords

neural networks, image processing, image quality, metadata, fish, biodiversity collection specimens

Presenting author

Yasin Bakış

Presented at

TDWG 2021

Conflicts of interest