Grace Cheung

Advisor: Professor Aaron Dollar

CPSC 490

6 February 2020

Senior Project Proposal: Tech for Conservation

Background

Image data is often a key component of conservation biology. It is used for a variety of research applications, from surveying wildlife populations and biodiversity to identifying specific species, making it crucial for conservation efforts. Currently, the standard process for collecting image data in the field is a manual and often analog method, with field assistants hand-writing annotations and digitizing them later. This process is laborious, slow, and vulnerable to human error. It can take years to annotate the millions of images needed for a single project, by which time the data might then be outdated or irrelevant. Then, because of differing methods of digitization, the resulting data can be difficult to share or to reproduce.

Currently, there are a few examples of the combination of technology with ecology, specifically in collecting image data. Seek is an augmented reality app created by iNaturalist that allows kids and families to identify wildlife through live image recognition using their phone camera, with the collected data sent to iNaturalist's global biodiversity database. While Seek demonstrates the effectiveness of technology in helping with image data, this app is not designed for research situations.

So, for my senior project, I will be working with Elizabeth Brooks (a EECS senior) to create a research-oriented web-app and desktop app that could accelerate, simplify, and standardize researchers' workflows, saving time, money, and effort.

General Project Description

The entirety of the project will be a web-app and an Electron desktop app that packages the browser version. The app will focus on plant images for now, as they are simpler to analyze. This app will have three main features: 1) image uploading with customizable annotations, 2) summary statistics, and 3) some machine-learning powered image processing.

1) Researchers will be able to create projects and customize the available annotations for each project. Field assistants can then upload images and select the appropriate annotations, including taxonomic classifications. A secondary feature is to allow for tagging of specific structures within the image.
2) The summary statistics page will help researchers visualize the data in a standardized way, which they currently do manually using mostly Excel. For now, we potentially propose a correlation matrix and heat map of traits as well as taxonomic statistics. Further features within this page will be finalized after conducting user research with biodiversity informatics specialists and ecology researchers at Yale.
3) Machine learning will be used to filter out empty images automatically, as well as to investigate the possibility of automatic annotation and identification. The Python package Pyodide will allow for integration of the machine learning with the web-app. The Python package Scikit-learn will be used to investigate automation. The proposed way to begin investigating this is to run dimensionality reduction on certain plant traits to determine how each plant varies on a certain trait for better identification.

Deliverables

The user research will be done primarily by Elizabeth. She will also be coding the majority of the frontend for both the main feature of image uploading and annotation as well as for the summary statistics.

My personal deliverables:

- Database creation and maintenance and some data processing. Elizabeth and I will be planning together and communicating about the main feature of the project (the image uploading and annotations, as well as tagging specific structures) as well as the summary statistics page. However, I will focus on coding the backend and she will focus on coding the frontend. So, I will code the database as well as the majority of the data processing algorithms.
- Machine learning for image processing. I will be investigating the majority of the machine learning for filtering and identifying images, as well as determine how viable my methods are. Training data will be provided by Yale researchers.