Data Management Plan: Machine learning Image classification of bark beetle species

Christopher Marais

Data Collection

The main aim and output of the project is to through established machine learning methods train a model that can classify various bark beetle species from images. An additional output of the project is to generate a database of images to enrich the current collection of bark beetle samples.

The images used to train the models and do experiments with will be generated in the Forest Entomology lab at the University of Florida using high resolution cameras. The initial images will be stored and backed up in their RAW format. During experimentation and model testing the images will be processed into JPEG and PNG formats. These processed images will be stored alongside the RAW images. The generated models will be stored as serialized objects alongside the Python code to interpret and use them.

We have reviewed existing datasets and are generating our own data specifically to enrich the current collection of samples and to produce images of a sufficient resolution. However, the model will be tested using images from an existing (https://barkbeetles.info/) database in collaboration with the owners of the database. These images will only be stored temporarily for the duration of the testing period. Identification of the images used during testing will be recorded and stored with instructions of where they be obtained.

All data will be stored with descriptive guides containing the metadata of where, when and how the data was obtained. Additionally, these guides will indicate how the data was processed and used. The metadata automatically generated during the capturing of images will be stored in a table in CSV format alongside the names of the images.

Ethics

The data generated and processed in this study do not carry significant ethical weight. The images will remain the intellectual property of the lab until publication. The data will be made available to other researchers on BioImage Archive when the study is published.

Storage and backup

The metadata, code and guides on the data will be stored on Github as a private repository. This repository will be updated whenever new data is generated, or code is changed. The repository will be made public after publication of the study. The Github repository will be the main backup and storage of the metadata and code. Additional copies will exist on two machines. One in the Forest Entomology lab on the University of Florida campus. Another will be on my (Christopher Marais) personal laptop.

The images in their RAW format will mainly be stored and backed up on OneDrive and on the Forest Entomology lab server with another backup on an external hard drive. Additional copies will exist on my laptop and on the high-performance server at the University of Florida (HiperGator). The backups on the lab server and OneDrive will be updated whenever new images are added. The external hard drive will be updated weekly and the copies on the HiperGator will be updated as required for processing.

Selection and preservation

All images in their RAW format, code, and models will be stored long-term. Any processed images will be stored short to medium term and relaced with guides on how to process the images for long term storage. All data on any personal computers or the HiperGator will be removed after the study is complete.

Data sharing

All data collected in this project will be held private to the Forest Entomology lab until publication of the study whereafter it will be made publicly available through GitHub and BioImage Archive. All documentation for how to access the data and the publication will be made available on GitHub. The GitHub repository will also be referenced if the study is published.

Responsibilities and resources

All responsibilities regarding the maintenance, creation and management of the data created in this study will be mine (Christopher Marais) for the duration of the study. After the completion of the study all data management responsibilities will be passed on to my supervisor (Dr Jiri Hulcr).