Markus Kangur (1007458038)

STA2453 PROJECT PROPOSAL: ZOOPLANKTON CLASSIFICATION

Introduction

Zooplankton (singular zooplankter) are a diverse class of microscopic marine organisms that drift with the tides and currents of the waters they inhabit. They play an important role in the aquatic food chain as a source of food for larger marine organisms. Importantly, zooplankton species can be highly sensitive to environmental conditions, so zooplankton species composition can be used as an indicator of ecosystem health. However manual classification is difficult, time consuming, and can be vulnerable to observation bias if multiple people are contributing to the classification effort. To address these issues, the goal of this project is to train a discriminative model that will classify images of zooplankton according to their species. In particular, we are interested in applying machine learning methods to estimate a function that maps the inputted zooplankton images to the appropriate discrete set of labels, the plankton species.

Data Structure

The Ontario Ministry of Natural Resources (MNR) has produced a labeled dataset to be used for this purpose. After completion of field sampling, they used a FlowCam to image the plankton, followed by manual labeling of 31 different classes of zooplankton. However, the MNR is primarily concerned with identifying 7 key types of zooplankton: Calanoids, Cyclopoids, Bosmina, Harpacticoids, Chironomids, Chydoridae and Daphnia.

The dataset contains almost 18000 mosaics which are each composed of approximately 80 individual plankton images. Each mosaic is stored as a 1200 x 1920 pixel 3-channel TIFF image. Figure 1 below shows a segment of a typical mosaic. After conversion to grayscale, each image will be a matrix of pixel values between 0 (black) and 255 (white).

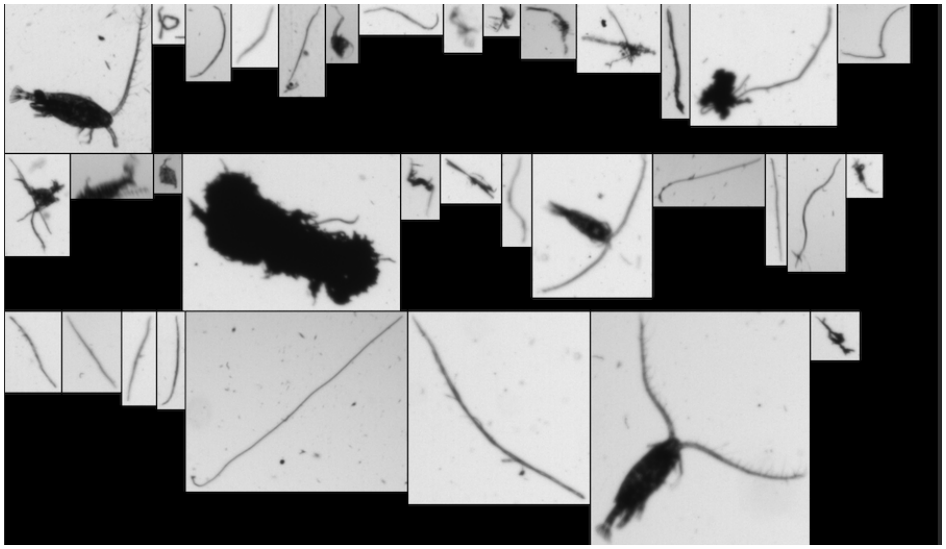

Figure 1: Example Plankton Mosaic Segment

Markus Kangur (1007458038)

The master table contains environmental features recorded during sampling for each mosaic, such as latitude, longitude, depth, and date. This master table also provides a link to other tables which contain all the geometric features computed by the FlowCam, including transparency, symmetry, compactness and volume. These other tables also contain the class labels, as well as individual image dimensions to facilitate extraction from the mosaic.

Potential Challenges

While the mosaics are a consistent size, the individual images within have very different sizes, as seen in Figure 1. This presents modelling challenges since many methods require input images of constant dimension, meaning rescaling may be necessary. Additionally, the zoo plankton species can appear in many different spatial orientations or positions in the images, meaning that within class variation is expected to be high. This may make it difficult to discriminate between different species. It was also noted that some of the tables containing labels and geometric information were missing. Finally, the modelling approaches applied in this project will be limited by the available computational power.

Potential Modelling Approaches

In all cases, our models should be evaluated based on their classification accuracy on a dedicated test set. Given that we have a large dataset of labeled training examples, it may be appropriate to apply a convolutional neural network (CNN). These models are well suited for inputs with a two-dimensional spatial structure. CNNs also exhibit translational invariance, meaning they are not very sensitive to image sizes and feature locations. Since CNNs have historically been used for image classification, there exist many openly available prebuilt architectures such AlexNet, ResNet, and DenseNet. While they have varying complexity and performance, their application will be constrained by computational power. Such a model would take individual plankton images as inputs and then output a predicted label.

There also exist opportunities to incorporate the auxiliary/geometric data into the modeling process. These features could be used independently, or a relatively simple CNN could be used to extract additional features from the images. In both cases, the features could be used in gradient boosting algorithms such as XGBoost to perform classification. Finally, it may be necessary to reduce the required computing power needed. This could be accomplished by working with a subset of the data, combining classes to reduce the number of output classes, or choosing modeling approaches that are less intensive.

STA2453 PROJECT TIMELINE

	Jan 23 – Jan 29	Jan 30 – Feb 5	Feb 6 – Feb 12	Feb 13 – Feb 19	Feb 20 – Feb 26	Feb 27 – Mar 5	Mar 6 – Mar 12	Mar 13 – Mar 19	Mar 20 – Mar 26	Mar 27 – Apr 2
Choose & Obtain Dataset										
Write Proposal										
Data Cleaning & Preparation										
Data Exploration										
Write EDA Report										
Research Modeling Options										
Train & Evaluate Models										
Update GitHub Repository										
Write Progress Report										
Write Final Report										
Prepare Final Presentation										