Author: Nadia Irwanto
Created: 09/09/2020. Updated: 16/12/2020

0 Overview

This project aims to train an instance segmentation model to identify all instances of planktonic foraminifera in a microscope slide image. The model will be trained on images from the PRISM collection of planktic foraminifera species from the Pliocene Epoch. Documenting these species will be helpful to climate and biology researchers using planktonic foraminifera data in their work. Finally, this project will contribute to the limited literature on generalizing instance segmentation to the natural history domain.

1 Background

Climate Change and the Mid-Piacenzian Age

The current phenomenon of global warming is unprecedented over the last few millennia. However, climate change is a complex process that makes it difficult to quantify each driving force as well as project future trends.

The Mid-Piacenzian Age of the Pliocene Epoch, about 3 Ma, is a potential if imperfect analogue for near-future climate conditions. The global mean temperature of the Piacenzian Earth is estimated to have been approximately 2–3°C warmer than at present, which is within the range of warming predicted for the end of the twenty-first century. At the same time, atmospheric CO_2 levels are similar to those today, at about 380-420 ppm[1]. Additionally, the similarity of the Mid-Piacenzian Earth to the modern one in terms of continental positions, land elevations and oceanic circulation patterns allows us to draw many parallels between their environmental changes. Thus, understanding the Pliocene climate is paramount to our ability to predict, adapt to, and mitigate the effects of future climate change.

The PRISM Project and Endless Forams

The Pliocene Research, Interpretation and Synoptic Mapping (PRISM) Project is a collaborative data analysis and climate modeling effort that aims to reconstruct this most recent warm period. The PRISM reconstruction is a high-resolution and multi-faceted description of the Pliocene, and is used ubiquitously to model ground-truth simulations of the past and to form the basis of future model projections.

Planktonic foraminifera are the primary source of data for deriving PRISM ocean temperatures. Planktonic foraminifera are single-celled protists that live near the surface of marine environments and have a long geologic record. Their small size and large abundance as microfossils, in combination with their widespread geographical distribution as planktonic organisms, make them ideal environmental and bio-chronological indicators. As many Piacenzian species are extant, estimating paleotemperatures based on modern calibrations is

[1] de la Vega, E., Chalk, T.B., Wilson, P.A. et al. Atmospheric CO2 during the Mid-Piacenzian Warm Period and the M2 glaciation. Sci Rep 10, 11002 (2020).https://www.nature.com/articles/s41598-020-67154-8

possible. Each species lives in a well-defined ecological niche, and hence the quantitative analysis of species assemblages is widely used to reconstruct environmental conditions.

Moreover, planktonic foraminifera data can be used to compare how species that still exist today behaved in the warmer environmental conditions of the Pliocene Epoch, or to investigate how the morphology of a species responds to time. These insights can provide useful context for understanding modern global warming.

A challenge faced by researchers using planktonic foraminifera data is that the high morphological variability within species makes it difficult to correctly identify an organism's species. To address this issue, the Hull lab at Yale previously released Endless Forams, a database of more than 34,000 planktonic foraminifera for taxonomic training. As part of the group's data collection efforts, they developed an image processing software, AutoMorph, to extract 2D and 3D shapes of planktonic foraminifera from slides images under light microscopy. The PRISM images are a valuable addition that will allow researchers to document far, far more of the Pliocene species.

However, the color, lightning and background variation of the PRISM dataset pose a challenge to AutoMorph's non-neural net approach. More specifically, AutoMorph employs RGB and gamma filters as well as user-defined threshold values that are not easily generalized over all the objects in the PRISM images. Hence, the goal of this project is to train an instance segmentation model to robustly identify and segment all instances of planktonic foraminifera in a microscope slide image. By using a machine learning approach, the model should not only perform well on the PRISM images, but also extend easily to future datasets.

Finally, existing work on using image segmentation models in the natural history domain is limited. However, the model will need to account for specific considerations; for example, images should not be flipped in data augmentation, or the chirality of an organism will be lost. This project will also serve as an example of how to apply deep learning techniques to taxonomic tasks such as species identification.

2 Project Description

For this project, I will first engage in a directed reading of the current literature on instance segmentation. This will begin with a review of foundational papers in order to attain a basic level of fluency with the key concepts and techniques used in image segmentation. The focus of my desk research will lead up to the baseline frameworks that have driven the rapid advances in instance segmentation results. I will then delve deeper into the more recent literature, including the newest state-of-the-art (SOTA) models as well as more incremental variants of model architectures. The list below gives an idea of the topics that will be reviewed:

- Fundamental concepts: R-CNNs, FCN
- Convolutional backbone architectures: VGG, ResNet, ResNeXt, FPN
- Seminal work: Mask R-CNN, UNet, DeepLab, SegNet
- Latest SOTA: DetectoRS, ResNeSt
- Model variants: Mask Scoring R-CNN; Wide UNet, UNet++
- Model training techniques: e.g. data augmentation

In parallel with the literature review, I will prepare the annotated dataset. The images have already been retrieved and preprocessed to build a composite image that stitches together in-focus pixels from images captured at different focal depths, and are ready for annotation. This entails exploring available image annotations tools, followed by using the selected tool to obtain a dataset of 2,000-3,000 segmented planktic foraminifera specimens. The annotated images will then be subjected to an exploratory data analysis to identify issues such as class imbalance, and potentially augmented as appropriate.

The bulk of the project will be spent on modelling the data. I will establish a performance baseline and choose several models to train from the landscape of instance segmentation model architectures. As model training is an iterative process, I will likely start by using only a subset of the data and adopting methods such as random search with cross-validation to make training more efficient. Models will be compared not only based on their overall performance, but the types of errors that each model makes. Once I have shortlisted which models to use, I will proceed to optimize model performance by training on the full dataset, looking into model ensembles, and adopting training procedure refinements.

3 Deliverables

1. A written report detailing i) the current state of the literature on instance segmentation and its application to the biological domain, as well as ii) the data, methodology, and results of the experiments.
2. Code for the final instance segmentation model for detecting and delineating planktonic foraminifera.
3. The dataset used to train the model.

4 Project Timeline

The first half of the semester will be dedicated to carrying out a comprehensive review of the literature pertaining to instance segmentation as well as preparing the dataset. The second half of the semester, starting November, will be spent training and optimizing the model.

Date	**Milestone(s)**
07/31/20	First meeting
08/12/20	Background reading on image segmentation, Endless Forams, and AutoMorph
08/25/20	Look into image annotation tools ● Labelbox: commonly used ● SuperAnnotate ● Segments.ai
09/09/20	First proposal draft Label a few images to gauge timeframe

09/23/20	Second proposal draft (send to Professor Zucker) Continue image annotation
10/07/20	Complete image annotation for the 11 slides. Complete literature review on fundamental concepts
10/21/20	Complete literature review on convolutional backbone architecture, seminal work, and latest SOTA
BREAK / BUFFER	
12/16/20	Complete literature review Begin training Mask R-CNN to establish performance baseline
12/30/20	Train 2-3 more models with fundamentally different architectures (SOLOv2, ___), using default parameters Evaluate models (e.g. by their accuracy and types of errors), and shortlist 1 model to optimize
01/13/21	Employ hyperparameter tuning, model ensembles, and other optimization methods
01/27/21	Propose final instance segmentation model Submit report first draft for review

Biweekly meetings will be scheduled with Dr. Allison Hsiang and Professor Pincelli Hull in order to discuss progress and subsequent direction.

5 References

1. de la Vega, E., Chalk, T.B., Wilson, P.A. et al. Atmospheric CO2 during the Mid-Piacenzian Warm Period and the M2 glaciation. Sci Rep 10, 11002 (2020).
2. Dowsett, H. J. et al. The PRISM (Pliocene Palaeoclimate) reconstruction: time for a paradigm shift. Philos. Trans. R Soc. 371, 20120524 (2013).
3. Hsiang, A. Y. et al. Endless Forams: >34,000 Modern Planktonic Foraminiferal Images for Taxonomic Training and Automated Species Recognition Using Convolutional Neural Networks. Paleoceanography and Paleoclimatology 34, 1157–1177 (2019).
4. Hsiang, A. Y., Nelson, K., Elder, L. E., Sibert, E., Kahanamoku, S. S., Burke, J. E., Kelly, A., Liu, Y., & Hull, P. M. (2017). AutoMorph: Accelerating morphometrics with automated 2D and 3D image processing and shape extraction. Methods in Ecology and Evolution, 9(3), 605–612. https://doi.org/10.1111/2041-210X.12915