Monterey Bay Aquarium Research Institute

Investigating the ecology of carbon flux in the water column using a semi-supervised particle classification approach

Sachithma Edirisinghe, University of Ruhuna, Sri Lanka

Mentor: Colleen Durkin

Summer 2022

Keywords: MorphoCluster; AyeRIS; particle classification; semi-supervised

ABSTRACT

Sinking particles in the water column are a major pathway of carbon flux to the deep ocean but they are difficult to study as particles and their associated transformations show high variability over time, depth and location. However, characterizing particles has proven to lead to more accurate estimates of carbon flux and therefore is of great significance. While autonomous in-situ imaging platforms are promising methods, the amount of data generated is a challenge. AyeRIS, a novel in-situ imaging system was used to capture images of the water column in Monterey Bay, California on 11th November 2021 and these images were run on MorphoCluster, a semi-supervised marine image classification software, to test the efficiency of the software in classifying a large marine imaging dataset and to test what particles can be resolved using MorphoCluster and AyeRIS. The major steps included data preparation, unsupervised steps; deep learning features extraction, clustering, and supervised steps; validation, and growing. 51% of the dataset was classified using MorphoCluster into 5 particle classes through three iterations, with the significant classes being fecal pellets, larvaceans, crustaceans, and sinking particles. MorphoCluster was more efficient in classifying the abundant particle classes than the rarer ones and the most abundant clusters comprised of blurry objects which may be unfocused smaller particles such as detritus and aggregates. The difficulty in classifying them may have been due to the resolution of AyeRIS images. While the feature extraction and clustering steps of MorphoCluster required high computer

performance, a significant amount of time and effort was also needed for the supervised validation and growing, especially for the more rare classes.

1 INTRODUCTION

1.1 BIOLOGICAL PUMP AND CARBON FLUX IN THE WATER COLUMN

The biological pump of the ocean refers to the downward flux of carbon from the surface of the oceans to the sea floor through a set of biological, chemical and physical processes, where it may be sequestered into the earth's interior over millennia (Boyd et al., 2019; Volk & Hoffert, 1985). An estimated value of 5 -12 $Pg^{-}C\ year^{-1}$ is sequestered by the biological pump (Boyd & Trull, 2007; Henson et al., 2011) making it an important component of the global carbon cycle and a driver of global climate through the regulation of the partial pressure of atmospheric carbon dioxide (*p*CO2) (Sarmiento & Gruber, 2006). However, there is still much uncertainty around carbon flux values due to the high variability of particles and processes involved, over time, location and depth.

At the sunlit surface of the ocean, primary producers (phytoplankton) fix carbon as organic matter through photosynthesis. This organic matter is then transported downwards via different pathways to the deep ocean, during which it may undergo different ecological transformations. Two major pathways include vertical migrations by zooplankton such as copepods and physical mixing. The more passive pathway is through the gravitational settling of particles sinking out of the euphotic zone. They may either be phytoplankton cells that sink directly, fecal pellets egested by zooplankton grazing on phytoplankton or detritus and aggregates formed from phytoplankton. Durkin et al., 2021 in a study conducted across 4 major ocean ecosystems using sediment trap observations, categorized sinking particles into 9 major classes; aggregates, dense detritus, large loose fecal pellets, long cylindrical fecal pellets, salp fecal pellets, Rhizaria (Phaeodarians), mini (spherical) fecal pellets and short (ellipsoid or oval) fecal pellets.

Resolving the particles involved in the biological pump allows for better quantification of its carbon fluxes as more accurate particle sizes and carbon content can be estimated when the identity of the particle is known as shown by Durkin et al., 2021.

1.2 STUDYING THE CARBON FLUX IN THE WATER COLUMN: IMAGING SYSTEMS

Particles in the water column has been studied mainly by using instruments that intercept particles as they sink down the water column such as sediment traps (either deep-moored, surface-tethered free-drifting, or neutrally buoyant) (Boyd et al., 2019). However, these methods may be more time-consuming and less efficient compared with in-situ imaging platforms deployed on autonomous vehicles. AyeRIS is one such imaging system, deployed on a Long Range Autonomous Underwater Vehicle (LRAUV), developed by the BioInspiration Lab at Monterey Bay Aquarium Research Institute. It has seven 23MP cameras and can capture an imaged volume of up to 2L. However, AyeRIS generates a huge amount of imaging data which can be a setback due to the amount of time needed for analysis if performed manually.

1.3 MORPHOCLUSTER

MorphoCluster is a software invented by Martin Schroeder in 2020 (Schroeder et al., 2020) that uses a semi-supervised approach to the classification of large marine imaging datasets by clustering the dataset. Being partially unsupervised, it utilizes the ability of a deep neural network to learn distinctive features of images and clusters the dataset based on similar features using an HBDSCAN* algorithm. The supervised part includes an interactive web tool, that allows the user to revise and grow the clusters, manage the hierarchy of clusters and annotate them.

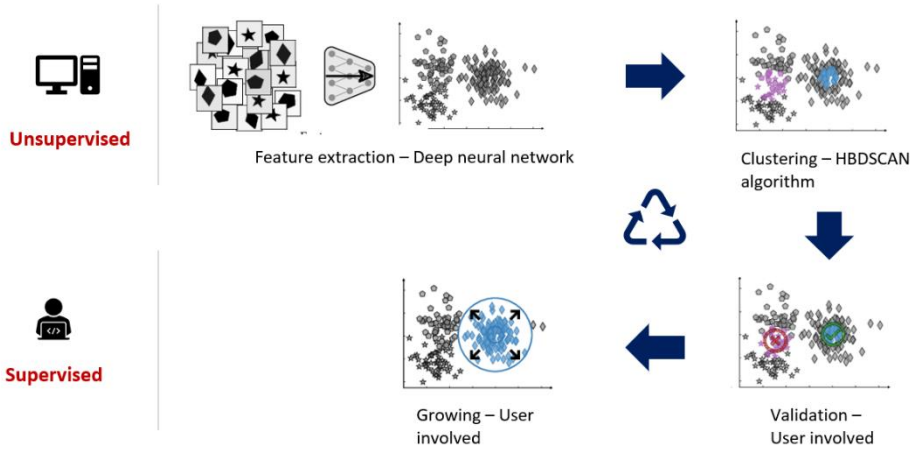

Figure 01: Overview of the feature extraction, clustering, validation and growing steps of MorphoCluster as modified from Schroeder et al., 2020. The next steps (not shown) are hierarchical arrangement and labeling of clusters

1.4 OBJECTIVES

- Evaluating MorphoCluster software as a tool to rapidly classify a large marine imaging dataset

- Exploring the extent to which particles can be resolved from AyeRIS images using MorphoCluster

2 METHODS AND MATERIALS

2.1 DATA COLLECTION

Images used were collected by the AyeRIS imaging system mounted on a LRAUV in Monterey Bay, California along the transect shown in the below figure, on 11th November 2021. This was one of the first deployments of the AyeRIS in the ocean and it reached a maximum depth of 142.7 m during this deployment.

Figure 02: Transect (in yellow) followed by LRAUV, Galene, indicating the area of Monterey Bay sampled by AyeRIS

2.2 IMAGE PROCESSING

The output data from AyeRIS consisted of grey scale images and each image contained many objects or regions of interest ie., fecal pellets, aggregates, zooplankton. For particle classification, individual regions of interest (ROIs) were required. The processing of extracting individual ROIs from the raw images involved the steps included below in Figure 03. The resulting dataset included 259 103 grey scale ROI images and these were used for analysis using MorphoCluster.

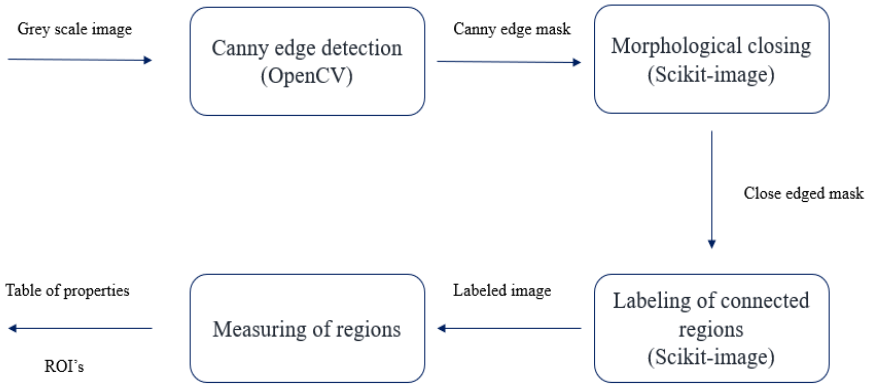

Figure 03: Steps followed in the processing of images from AyeRIS imaging system to produce regions of interest, ROI's and their table of properties.

2.3 IMAGE CLASSIFICATION USING MORPHOCLUSTER

2.3.1 DATA PREPARATION

The input data for MorphoCluster consisted of only one .zip file. hence called `archive.zip`. The image archive .zip file had to consist two files,

i. The data set of ROI's to be classified (single object images)
ii. A .csv file (hence called index.csv file) containing two columns; "object_id" which was a unique identifier for each image which in this study was the image name and "path" which was the file path for each image.

This image archive had to be placed in the data directory of the software in our computer along with the deep learning model parameters. The code for the data preparation step was written in R and is as follows,

```
files <- list.files(path = "<insert file path>")
n <- length(files)

index <- matrix(NA, nrow = n, ncol = 2)
colnames(index) <- c("object_id", "path")

for(i in 1:n){
  index[i,1] <- c(sprintf("%s", files[i]))
  index[i,2] <- c(sprintf("rois, /%s", files[i]))
  }

write.csv(index, "index.csv", row.names = FALSE, col.names = FALSE)
```

2.3.2 FEATURE EXTRACTION

A decapitated deep neural network, ResNet18 acts as the feature extractor in MorphoCluster. The network used model weights pre-trained on the ImageNet data unless specified otherwise (`--parameters-fn`). The input mean and standard deviation of the estimated mean color values of the images (R,G,B values) also had to be provided. As the output of feature extraction, it produced a 512d feature vector for each image (`features.h5`).

The command used to run the feature extraction was as follows,

```
morphocluster features [--parameters-fn model_state.pth] [--input-mean 0.9,0.9,0.9] [--input-std 1,1,1] archive.zip features.h5
```

The mean and the standard deviations of ImageNet images were used for this study (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]).

2.3.3 CLUSTERING

The clustering step is done through the HBDSCAN* algorithm which clusters the densest regions in the feature space as clusters of similar feature images. The algorithm is characterized by two parameters; minimum cluster size (m) and neighborhood size (k). The neighborhood size (k) was

set to k=8 and the minimum cluster size, the minimum number of features required to make a cluster in the feature space, was exponentially reduced (Figure 04) as recommended in Schroeder & Kiko, 2022.

An additional Principal Component Analysis step was included here to reduce dimensionality of the feature vector from 512 to 64. The command used to run the clustering initially, was as follows,

```
morphocluster cluster [--pca 64] --min-cluster-size m --min-samples k features.h5
tree.zip
```

The output was saved as the tree.zip in the data directory of the software.

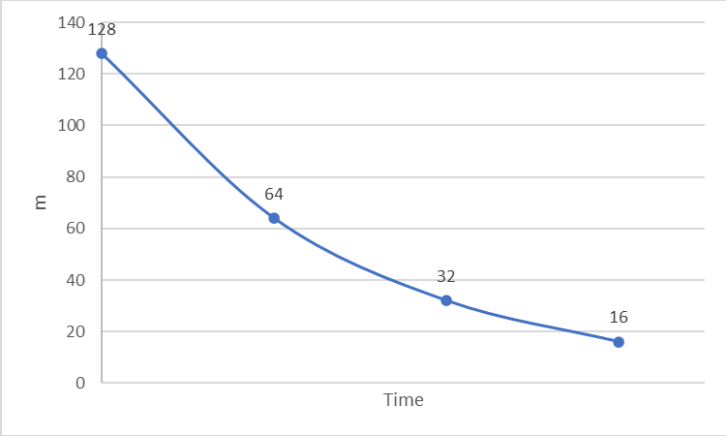

Figure 04: Minimum cluster size, k, was reduced exponentially starting from k = 128.

2.3.4 LOADING WEB APPLICATION

Next the image archive zip file, extracted features file (features.h5), and the clustering output file (tree.zip) were loaded into the web application using the following commands.

```
flask load-objects archive.zip

flask load-features [--pca 64] features.h5

flask load-project tree.zip
```

Once all the required files were loaded into the web application, the parent node (unclustered images) and the cluster nodes were shown on the interface.

2.3.5 CLUSTER VALIDATION

In the validation step, the clusters obtained were looked through manually for homogeneity. If the cluster was pure, the cluster was named and accepted (validated) and if the cluster was mixed, it was sent back to the parent node, that is the set of unclustered images (rejected). The validated clusters resulting from validation can also be termed as cluster "seeds" as they are the cores of the densest regions of the feature space.

2.3.6 CLUSTER GROWING

In this step, these cluster seeds were then "grown" into bigger clusters. This meant that more images were collected from the neighborhood size of the dense regions until a "similarity threshold", which is the boundary around a cluster seed in the feature space from which outwards are images that cannot be considered similar to the ones in the cluster seed, was identified and reached (See Figure 01 for a visual representation). Unclustered images from the neighborhood ordered from decreasing similarity to the cluster seed, were displayed as "recommended members" as pages of 50 images each. The similarity boundary was chosen by finding the image which was strictly dissimilar to the seed. This was done in 2 ways; the binary search or the turtle mode.

The binary search was the faster method and in this method the search is done through pages as a whole. If the first page has all similar images, once reviewed and accepted, the next page shown will have skipped a couple pages. And if that page was accepted the next page had skipped double the number of pages than before. Thereby the number of pages skipped between each page review doubled, increasing the number of images the user had to go through, making it faster. In the turtle mode, however, the user had to go through each object on each page individually. After all the clusters were grown, the output was saved in the export directory in the data directory of the software on the computer.

2.3.7 REITERATION

Once the grown clusters were saved, the clustering step was repeated for the unclustered pool of images with a reduced minimum cluster size (m). Here instead of the archive.zip file, the output zip file from the growing in the previous iteration saved in the export directory of the data directory was used. The resulting zip file from the clustering was given a new name (referred to as tree-2.zip here)

```
morphocluster cluster [--pca 64] --min-cluster-size 64 features.h5 tree-64.zip --tree
/data/export/2020-05-15-10-34-34--3--tree-128.zip
```

Next, the steps from 2.3.5 to 2.3.7 were repeated and along with 2.3.4 this process was repeated until no more pure clusters emerged from the clustering.

3 RESULTS

3.1 1ST ITERATION

The first iteration (m = 128) for the dataset of 259 103 images resulted in 31 clusters. The biggest cluster emerged was a group of unfocused objects termed “blobs”. With the validation step, 4 clusters were rejected due to being too mixed and 27 were validated, named and grown. At the end of growing, the clusters had 67 391 images or accounted for 26.01 % of the total data set. The clusters were mainly blobs and noise but there were three fecal pellet clusters.

Figure 05: An example of a cluster of the ‘blobs’

3.2 2ND ITERATION

In the second iteration (m = 64), there were an additional 19 clusters resulting from the clustering step. 4 clusters were rejected and 15 were validated and grown. At the end of the iteration, there the clusters included 40 % of the data set. An important new cluster that emerged was the cluster of crustaceans of different species.

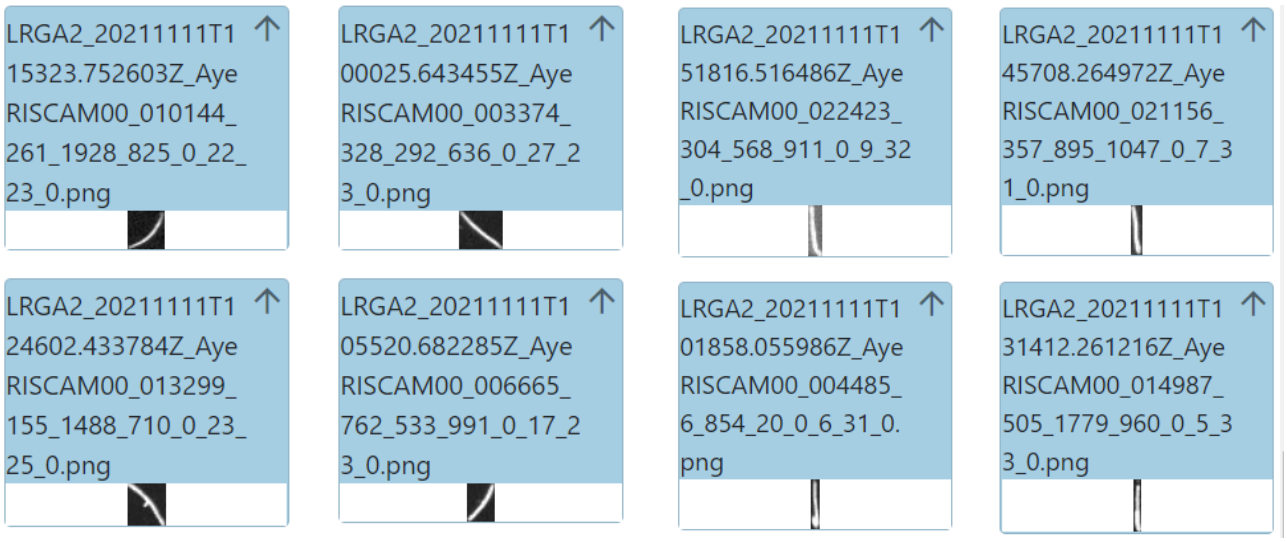

Figure 06: Two clusters of fecal pellets

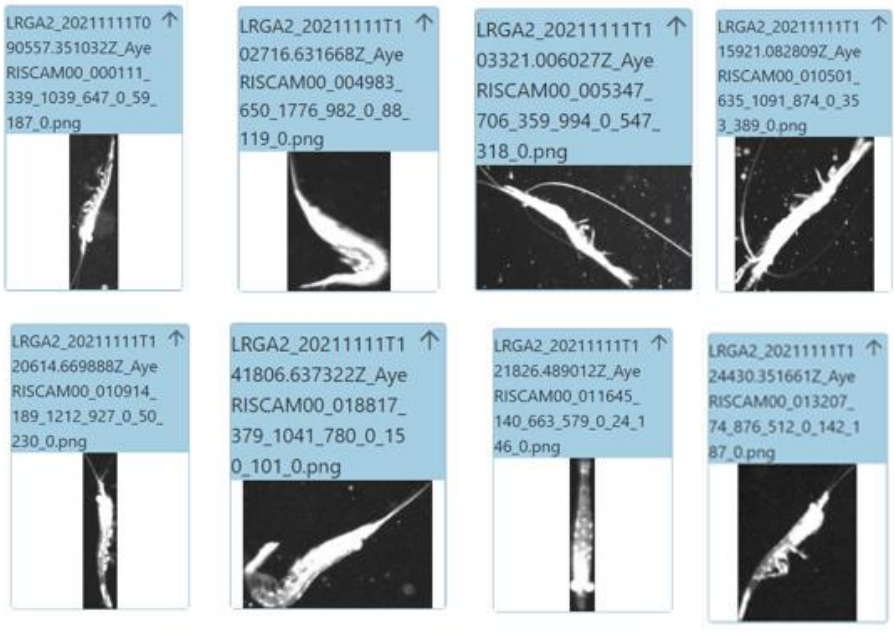

Figure 07: An example cluster of the crustaceans

3.3 3RD ITERATION

In the third iteration (m = 32), 22 new clusters emerged from the clustering step. 2 clusters were rejected and 20 were validated and grown. At the end of the iteration, 51% of the data set was clustered and there were 61 clusters in total. An important new cluster that emerged here was the cluster of objects that looked like sinking long aggregates of particles.

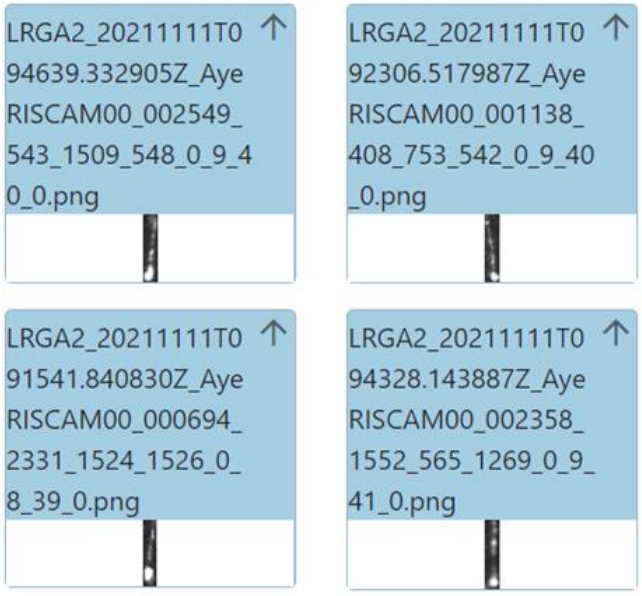

Figure 08: An example cluster of the ‘sinking particles’

3.4 4TH ITERATION

The fourth iteration (m = 16), resulted in an additional 33 clusters. However due to time constraints, the validation step could was not finished. A new cluster that emerged here was a cluster of larvaceans.

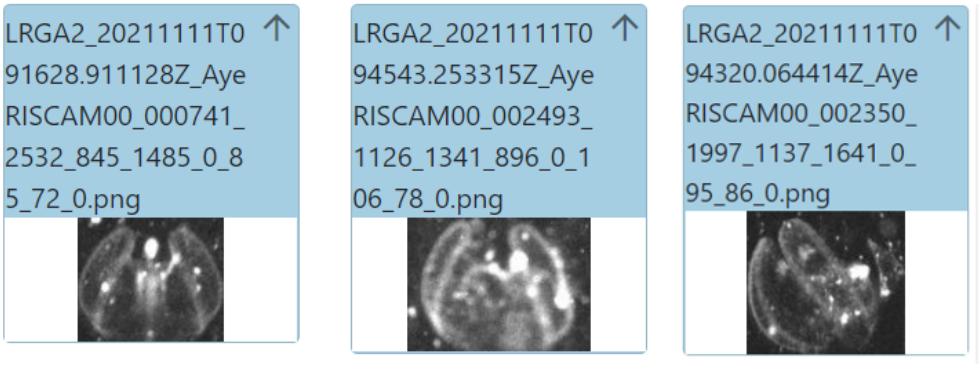

Figure 09: Cluster of the larvaceans and their mucus houses

Thereby at the finish of the project, the data set was classified up to a half or 51% ().

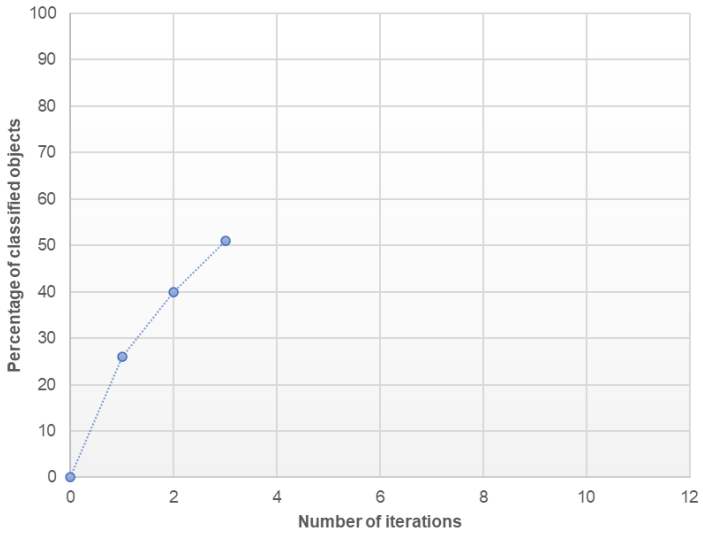

Figure 10: Percentage of classified images with the number of iterations completed (4th iteration not shown as it was not completed).

4 DISCUSSION

4.1 EVALUATION OF MORPHOCLUSTER AS A TOOL FOR RAPID ANNOTATION OF LARGE MARINE IMAGING DATASETS

Initially, the setting up and running of MorphoCluster was a lengthy process. And it was due to the instructions for running MorphoCluster were not updated with the software version in the GitHub ReadMe which was the one that was referred. The updated command syntax had several key changes that without knowing, delayed the project.

With the data preparation, during the creation of the .csv file, there were two errors made with the file paths. For example, for the images in the directory named “rois”, the format used was, “/roi/LRGA4_20220411T063917.447285Z_AyeRISCAM02_0_0_0_2100_0_172_172_0.png”, while the corrected format was, “rois/LRGA4_20220411T063917.447285Z_AyeRISCAM02_0_0_0_2100_0_172_172_0.png”.

Moreover, MorphoCluster was initially run on a virtual machine with low RAM. Therefore the feature extraction phase and clustering phase for an initial testing dataset of ~500 000 images took approximately 1 day and 4 days respectively. Since this was not viable, a switch was made to a high-performance computer (DeepRip) which was much faster.

In the validation phase, while there were a few clusters that had one or two members rejected, many of the clusters had to have several to many members rejected. When the number of additional clusters increased, the time taken for validation increased. It was mostly the insignificant clusters such as noise and ‘blobs’ that remained the same through the validation. Also, the feature of the validation step where the cluster members are arranged in such a way that the members next to each other were most dissimilar, was very helpful in rejecting members and increasing the homogeneity of the cluster.

The growing phase was the step that took up the majority of the time. Moreover, this step retained some subjectivity as the similarity threshold was determined by myself. Also for example, if the cluster was fecal pellets, while the pages should be aligned such that most similar-looking objects to fecal pellets were shown first, in the growing stages of many clusters, there were images that clearly looked like, fecal pellets in this example, after images that were clearly different. This led

to switching to turtle mode for some clusters, which was very slow. This was most common in rare clusters like the animals as they were less numerous in the data set. However since there was a lot of noise and 'blobs' in the data set, and it was needed to pull out those images as soon as possible, the growing step was done a bit leniently. But clusters of importance were handled strictly.

According to Schroeder et al., 2020, the clustering needs to be reiterated until no more pure clusters emerge or when the graph in Figure 10 reaches a steady state, which seems to have started happening. The hypothetical prediction shown in Figure 11 where it reaches a steady state around iteration 5, is based on the fact that the curve gradient had started to decrease.

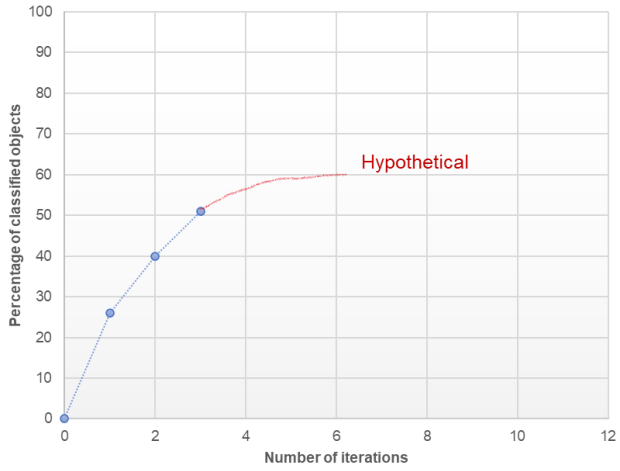

Figure 11: Prediction of the curve of percentage of classified objects with iteration number

Overall, while the clustering was mainly data-driven, fast, and able to pull out clusters that may have been unperceivable to humans, it did need a user to review and grow the clusters which were relatively slow and a lot more subjective. It did a relatively fair job of pulling out the noise, blob, and fecal pellet clusters but not so well with the others. This may be due to the high number of noise and blob images and therefore it would however be interesting to see what results would come of a dataset without a lot of noise.

4.2 PARTICLE TYPES RESOLVED USING AYERIS AND MORPHOCLUSTER

Better characterization of the particle types involved in the biological pump allows for more accurate quantification of carbon fluxes as showed by Durkin et al., 2021. The particle types in the water column identified from AyeRIS images using MorphoCluster were fecal pellets and the 'sinking particles'. The fecal pellet clusters began to emerge in the first iteration and this may indicate the high number of them in the data set, consistent with Durkin et al., 2021, where they were a main class of particles identified. While there were different sizes of fecal pellets, the clusters were not separated as such. There were mainly long, thin pellets and short, looser pellets. Moreover, the 'sinking particles' identified here were consistent with the type of particles observed in the water column during an ROV dive in Monterey Bay (personal observation) and they may be loose aggregates that are sinking vertically, giving them an elongated streamlined shape.

Figure 12: Magnified image of a member of the cluster, 'sinking particles'

The main animals identified were crustaceans and larvaceans while fish, ctenophores, and chaetognaths were spotted in the dataset when they came up as recommended members in the growing phases. The greyscale property and the resolution of the images certainly made it harder to identify particles in the validation and growing phases. Identification of these organisms simultaneously with the particles may help resolve the various ecological processes occurring in the water column as particles sink, in the longer term, such as zooplankton-particle interactions (Möller et al., 2012; Christiansen et al., 2018).

The 'blobs' clusters mainly had objects that were blurry, unfocused and hence depicted a hexagonal shape corresponding to the camera aperture or with very high brightness that masked any identifiable features. Since they were also in high abundance in the data set, they may be another major type of smaller particle such as detritus or aggregates. This may also be a result of the lower spatial resolution of the images (compared to the coverage). Since this high amount of

noise and ‘blobs’ delayed the clustering of interesting image classes, MorphoCluster would be more effective if these images could be filtered out in the image processing steps. AyeRIS images also had high temporal resolution and along with the high spatial coverage, this may be beneficial in calculating carbon fluxes more accurately. But there also may be an image overlap especially with the animals as they move around and particle counts taken from them may be needed to be treated with caution.

5 CONCLUSIONS/RECOMMENDATIONS

Overall, 51% of the dataset of 259 103 images were able to be classified using MorphoCluster into mainly 5 particle classes, ‘blobs’, ‘sinking particles’, fecal pellets, crustaceans and larvaceans. MorphoCluster seemed to be more efficient in classifying the more abundant particle classes in the data, ‘blobs’, noise and fecal pellets than the rarer yet significant classes, animals and ‘sinking particles’. Validation and growing phases required the most time and the growing phase especially was conflicting as images seemed at times not arranged in decreasing similarity in a page of recommended members and it retained some subjectivity with the user determining the similarity boundary. Large sized objects were able to be identified more clearly due to the quality of images. Moreover, due to the resolution of AyeRIS images, the smaller particles may not have been captured well and they may be the abundant ‘blobs’ in the dataset. Continuing the iterations until no new pure clusters emerge is recommended as the immediate next step and finding a method to filter out “blobs’ and noise before using MorphoCluster would be better to assess the efficiency of MorphoCluster in classifying large marine imaging datasets.

6 REFERENCES

Boyd, P.W., Claustre, H., Levy, M., Siegel, D.A. and Weber, T., 2019. Multi-faceted particle pumps drive carbon sequestration in the ocean. *Nature*, *568*(7752), pp.327-335.

Volk, T. and Hoffert, M.I., 1985. Ocean carbon pumps: Analysis of relative strengths and efficiencies in ocean-driven atmospheric CO2 changes. *The carbon cycle and atmospheric CO2: natural variations Archean to present*, *32*, pp.99-110.

Boyd, P.W. and Trull, T.W., 2007. Understanding the export of biogenic particles in oceanic waters: Is there consensus?. *Progress in Oceanography*, *72*(4), pp.276-312.

Henson, S.A., Sanders, R., Madsen, E., Morris, P.J., Le Moigne, F. and Quartly, G.D., 2011. A reduced estimate of the strength of the ocean's biological carbon pump. *Geophysical Research Letters*, *38*(4).

Durkin, C.A., Buesseler, K.O., Cetinić, I., Estapa, M.L., Kelly, R.P. and Omand, M., 2021. A visual tour of carbon export by sinking particles. *Global Biogeochemical Cycles*, *35*(10), p.e2021GB006985.

Schröder, S.M., Kiko, R. and Koch, R., 2020. MorphoCluster: efficient annotation of plankton images by clustering. *Sensors*, *20*(11), p.3060.

Schröder, S.M. and Kiko, R., 2022. Assessing Representation Learning and Clustering Algorithms for Computer-Assisted Image Annotation—Simulating and Benchmarking MorphoCluster. *Sensors*, *22*(7), p.2775.

Möller, K.O., John, M.S., Temming, A., Floeter, J., Sell, A.F., Herrmann, J.P. and Möllmann, C., 2012. Marine snow, zooplankton and thin layers: indications of a trophic link from small-scale sampling with the Video Plankton Recorder. *Marine Ecology Progress Series*, *468*, pp.57-69.