Plankton Image Identification Using Principal Components Analysis

Quantitative Engineering Analysis, Spring 2020

S. C. McAneney and A. R. Platt

1. Summary

Plankton identification is a time-consuming and tedious task for scientists, but is important for better understanding ocean ecosystems. In this paper we attempt to improve this process with an algorithm that can identify and classify plankton based on their shape, using image processing and principal components analysis (PCA) to identify and compare the key features of each species. We achieve 26% accuracy, confirming the findings of Tang et al (2006) [1] that using a single feature vector will not produce accurate enough results for use in a scientific setting. To improve this algorithm, as Tang and team do in their study, we recommend performing multiple analyses of different features and doing PCA across those features.

2. Introduction

Background

Image classification is becoming widely used in wildlife identification. One of the primary uses is being able to collect data without disrupting animals. Organizations like WildTrack and Ocean Alliance use it to noninvasively identify wildlife through footprints, features, and markings [6][7]. Classification of plankton samples is a logical extension of this technology. Classifying plankton is one of the most time-consuming jobs associated with studying plankton, and several attempts have been made to make a reliable algorithm to perform this task, using PCA [1], or multiple kernel learning [2], [3].

Algorithm

The algorithm we are reproducing is described by Tang et al. (2006) [1]. It identifies the outline of plankton in an image. It then computes the centroid of the shape and the geodesic distance from each boundary pixel to the center of the shape. When plotted, this gives an “unwrapped” version of the plankton image. By performing a discrete Fourier transform on this function, we can express the plankton shape as a series of frequencies, and then perform principal component analysis to identify the key features of the plankton. These can then be compared to the key features of unknown plankton to classify them. Using the Fourier transform method means that the orientation of the plankton in the initial image does not matter.

Ethical Implications

This technology has vast potential to help scientists in the area of marine study. Plankton are often a good indicator of the overall health of an ocean ecosystem. Being able to rapidly count and classify plankton would allow scientists to cut down on the amount of man-hours currently being put into tedious plankton classification, which would better utilize grant resources.

The harmful implications of this technology are largely negligible if a high enough accuracy can be achieved. Since the training data would be composed of plankton images, it is unlikely that there would be ethical problems with this technology with regards to humans. However, it would be important to make sure the algorithm is trained on all the plankton that could potentially be found in an area. The most immediate danger this represents is mass misclassification of a species as another species, which could lead to inaccurate conclusions about an ecosystem or species being studied, as well as comprising scientific integrity.

Question

In this paper, we investigate the question of whether plankton classification by image can reach a high enough accuracy to be useful to scientists and avoid the aforementioned concerns. Specifically, we attempt to identify plankton by performing PCA on a Fourier descriptor of the image. We aim to replicate the results of Tang et al., while using more recent plankton image data collected by the Woods Hole Oceanographic Institution.

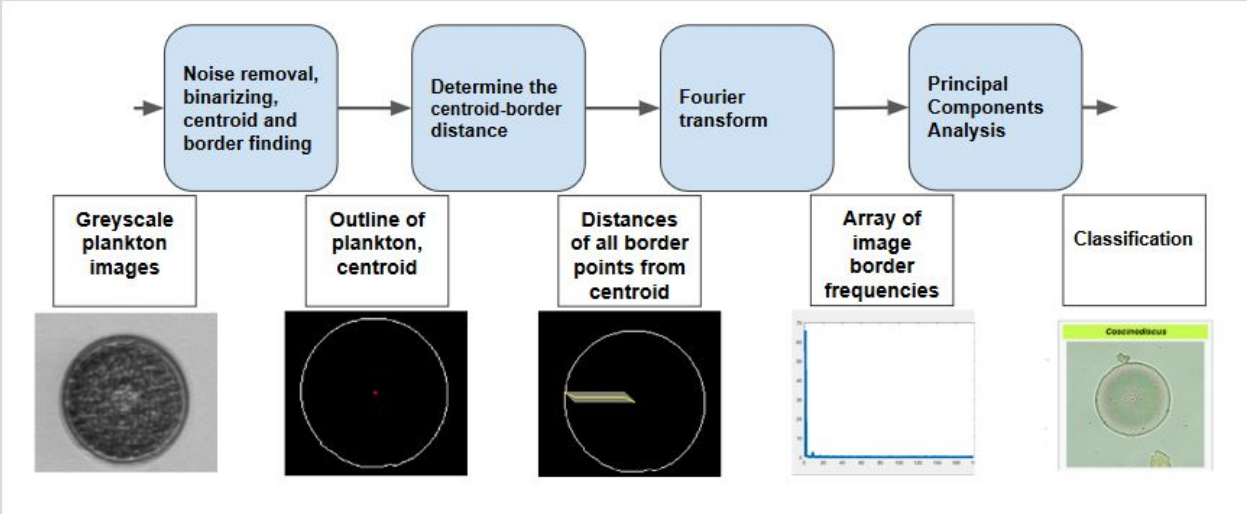
3. Methods

Our algorithm attempts to classify plankton using principal components analysis, which breaks data down into eigenvectors that represent the directions of greatest variation. We can then project the test and training data onto these vectors, which transforms them into *blobspace*. This allows us to compare the blobspace test images with the blobspace training images, and classify the test images as the same species as the nearest training images.

However, our data set requires processing before the principal components analysis can take place. The images we have contain plankton in many different rotations, positions, and sizes. This means that a pixel-to-pixel analysis of raw plankton images would not be able to accurately compare plankton as one could compare objects that are usually in the same orientation, like faces. To process the data, we first reduced each plankton image down to the outline of the plankton and its centroid, found the geodesic distance between the centroid and the border pixels, and performed a Fourier

transform. This presents each image as a frequency, which does not reflect confounding factors such as rotation or position.

An overview of our process can be seen in the diagram below:



Part 1: Image Processing

The training data for this algorithm was collected from the Woods Hole Oceanographic Institution's open source annotated image archive, specifically designated for this purpose [4]. Our image processing procedure is as follows: the program takes in grayscale images of plankton (figure 1). We apply a median filter, which replaces each pixel with the median of the surrounding pixels, eliminating some of the noise. The images are then binarized, turning them to only black and white, represented by 0s (black) and 1s (white) (figure 2). We then continue cleaning the image by determining the connected components in the image and deleting all but the largest. This leaves us with only the outline of the plankton (figure 3).

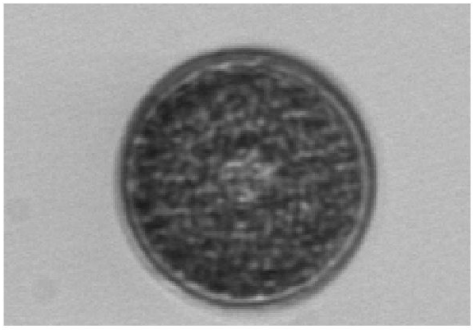


Figure 1

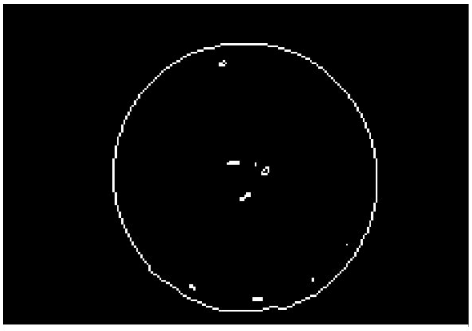


Figure 2

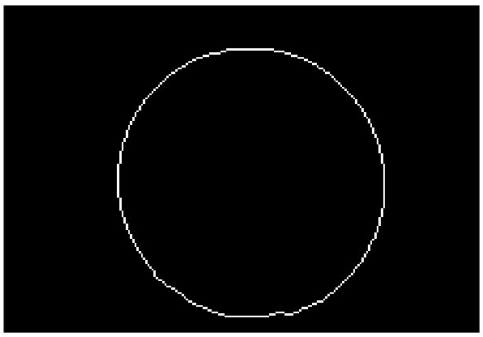


Figure 3

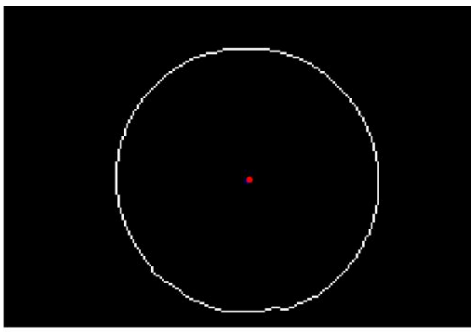


Figure 4

Part 2: Computation

Once the outline has been determined, the centroid is calculated (figure 4). The outline is traced in order to assign an index to each pixel. The indexes must be sequential along the shape outline in order to plot the distances for each point correctly and obtain the “unwrapped” shape of the plankton. The geodesic distance from each pixel in the boundary to the centroid is calculated and stored in a vector [5]. Geodesic distance was used instead of Euclidean because geodesic distance allows us to find the shortest distance within a boundary. While the example shown in the figures is roughly circular, the majority of plankton are shaped such that Euclidean distance would cut through the border of the plankton, inaccurately representing its shape. When plotted against the pixel index, these distances give us a visualization of the “unwrapped” plankton shape. The data from these unwrapped images were repeated until all images were the same length. This allowed us to compare all the images to each other.

Finally, a Discrete Fourier Transform was performed on the distances vector, allowing us to express the shape of the plankton as frequencies. By expressing the plankton as frequencies rather than pixels, we can compare plankton without interference from rotation or photograph size, but by shape alone.

Part 3: PCA

First, to perform principal component analysis (PCA), we represent the training data in a matrix A, an *n x m* matrix where each row contains the Fourier values for an image as computed in the previous section.

Next, calculate the covariance matrix, R, of A:

$$R = \frac{1}{N} * A^T * A \quad (1)$$

The covariance matrix is effectively a correlation matrix that retains the magnitude of the data. The eigenvalues, λ, and the eigenvectors, Q, can then be calculated such that λ is a diagonal matrix containing the eigenvalues $\lambda_1, \lambda_2 \ldots \lambda_N$ and Q is a square matrix where each column contains eigenvectors $v_1, v_2 \ldots v_N$ of length 1. The eigenvector corresponding to the largest eigenvalue points in the direction of the greatest variation in the data, the second largest points orthogonal to the first and in the direction of the second greatest variation in the data, and so on. These represent the principal components of the data. We can then project our data onto the principal components into what we will call blobspace using the following equation:

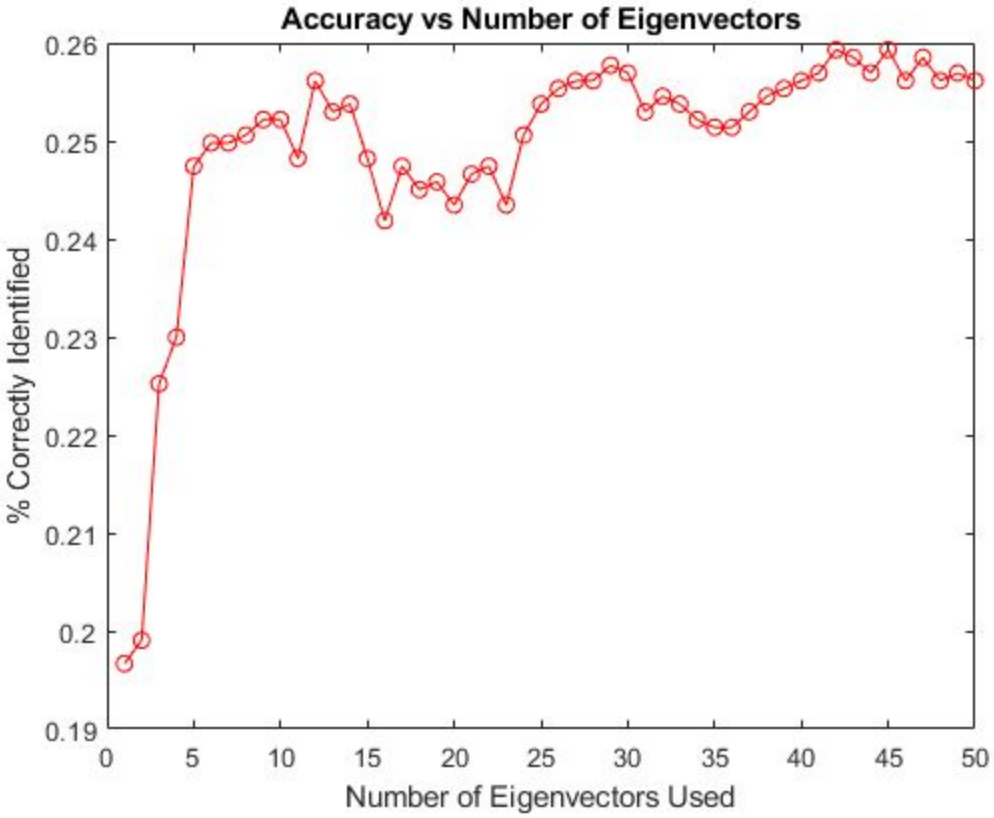
$$blobspace = Q^T * A^T . \quad (2)$$

To test the algorithm, we perform the same operations on a set of test data, but this time project it into the blobspace defined by the training data. We then use the k-nearest neighbors (kNN) algorithm to classify the test data. This means that an unknown plankton will be classified as whatever species it is closest to in blobspace. We can also test different numbers of nearest neighbors to see which produces the highest accuracy. With multiple nearest neighbors, the k nearest neighbors are considered. Whichever there are most of is what the plankton is classified as. For instance, if k=5, and 1 nearest neighbor is one species and 4 are a second species, the plankton will be classified as the second species.

4. Detailed Findings

To test our algorithm, we used 2157 training images and 1262 test images, encompassing 12 different species of plankton. These species were chosen arbitrarily from the available test images.

We determined the accuracy of our algorithm by dividing the number of images classified correctly by the number of images classified. Our algorithm obtained no higher than 26% accuracy, as shown in the figure below.



The accuracy increased significantly as the number of principal components used increased from 0 to about 5. It continued to increase from 5 to about 10, where it approximately leveled off. This requires significantly more eigenvectors to achieve the accuracy that Tang's team achieved with only a few, and its peak accuracy is considerably lower than Tang's team found, even using only the Fourier transform method. This difference in accuracy is likely due to differences in the quality of our data as compared to Tang's.

Most of this inaccuracy is due to the differing resolution between the images. Differing resolutions between the images mean that if there are two pictures of each image where one has twice the resolution of the other, the higher resolution picture will have twice the number of points in its outline. Because we needed a matrix with transforms of the same length in order to perform PCA, we took the size of the largest image and had all the other images repeat their elements until they were the same length as the longest image. Using this system would make the above hypothetical image with half the resolution and thus half the number of boundary pixels look like its frequencies were twice those of the higher resolution image. This would make the two images appear completely differently in the Fourier transform, even though they are from the same species, and would likely result in the incorrect classification of the images.

In addition, although plankton can vary greatly in size, this is not identifiable in the image without scale reference. Because the Fourier transform represents the pixel

distance from the centroid to the boundary at each frequency, this means a very small plankton would appear the same as a very large plankton with a similar shape in a similarly sized image.

Due to these limitations, we have confirmed what Tang's team found, which is that principal component analysis of Fourier descriptors alone is not an accurate enough tool to be helpful to scientists in classifying plankton. It would ultimately take more time to double check the accuracy of the algorithm than to simply classify all species by hand. The harmful implications of this technology are still minimal with regards to humans, especially as the algorithm cannot correctly classify the majority of the plankton data it was given. However, the implications with regards to studies are more serious. Our algorithm is only trained on twelve species of plankton and does not achieve an accuracy above 50%. This could lead to misclassification of any plankton the algorithm is not trained on, yielding misleading results about ecosystems being studied. Using an algorithm this limited and inaccurate would certainly compromise the scientific integrity of the study in which it was used.

A smaller thing that could improve the accuracy of our algorithm is to make a more sophisticated image pre-processing procedure. The current system is unreliable and cannot determine the difference between image noise and plankton composed of multiple pieces.

In terms of larger structural changes to improve accuracy, one of the most important changes would be to find a way to normalize the images for scale and resolution, so that those factors do not influence the Fourier function as mentioned above. In addition, we would want to consider other features such as moment invariants (parts of an image that do not change with translation, rotation, or scale), and possibly color. As suggested by Tang's work, combining vectors that contain multiple different methods of feature analysis could produce a greater accuracy than any one method of feature analysis.

5. Recommendations

Overall, though our algorithm did achieve some success, it is far from a model that could be used by plankton researchers. However, there are many ways in which the algorithm could be improved which were beyond the scope of this project. Currently, the size and coloration of the image is not accounted for when classifying, in an expansion of this algorithm, different feature vectors could be analyzed and combined to express this. Furthermore, normalizing the scale and resolution of the images would improve our results.

Although our algorithm in its current stages is not very useful in a scientific context, other more accurate plankton classification systems are being developed and will likely be used by researchers in the near future. The development of these systems has the potential to greatly help scientists studying marine ecosystems, saving them

hours of laborious classification and allowing them to allocate their resources to other areas.

6. References

[1] Tang, Xiaoou, et al. “Binary Plankton Image Classification.” IEEE Journal of Oceanic Engineering, vol. 31, no. 3, July 2006, pp. 728–735., doi:10.1109/joe.2004.836995. A system of plankton classification on which we based our algorithm

[2] Zhao, Feng, et al. “Binary SIPPER Plankton Image Classification Using Random Subspace.” Neurocomputing, vol. 73, no. 10-12, June 2010, pp. 1853–1860., doi:10.1016/j.neucom.2009.12.033. A system of plankton classification using multiple kernel learning.

[3] Zheng, Haiyong, et al. “Automatic Plankton Image Classification Combining Multiple View Features via Multiple Kernel Learning.” BMC Bioinformatics, vol. 18, no. S16, Dec. 2017, doi:10.1186/s12859-017-1954-8. A system of plankton classification using multiple kernel learning

[4] Woods Hole Oceanographic Institution. “Woods Hole Open Access Server.” Woods Hole Open Access Server, 204AD, darchive.mblwhoilibrary.org/handle/1912/7341. Classified images which were used to develop and train our algorithm.

[5] Eddins, Steve L. “Exploring Shortest Paths - Part 5 .” 13 Dec. 2011, blogs.mathworks.com/steve/2011/12/13/exploring-shortest-paths-part-5/.

[6] Li, Binbin V., et al. “Using Footprints to Identify and Sex Giant Pandas.” *Biological Conservation*, vol. 218, 2018, pp. 83–90., doi:10.1016/j.biocon.2017.11.029.

[7] “Whale Conservation- SnotBot Provides Answers.” *Intel*, www.intel.com/content/www/us/en/analytics/artificial-intelligence/technology-innovation/whales-snotbot-ocean-health.html.