Image Feature Extraction for Plankton Classification

An exploration of image feature extraction and classification on large oceanographic data

May 12, 2015

Authors

NICK HOCKENSMITH
KEVIN PARK
DANE SKINNER
Oregon State University
Corvallis

$\mathcal{KHP}$
2015
KIDDER HALL PRESS

Contents

1 Introduction

Plankton, perhaps surprisingly, form a critical link in the global ecosystem and are a fundamental source of food and energy for aquatic wildlife. As such, the population levels of plankton are an ideal metric for determining the health and viability of oceans and aquatic ecosystems. The challenge thus becomes determining the best way to classify and count the multitude of phytoplankton and zooplankton species from a sample of ocean water. Modern imaging systems can easily produce hundreds of thousands of images in a short time scale, so using human based means is daunting and often of minimal utility.

To address this challenge, we each explored a different method of image feature extraction followed by a different approach of test data classification. N. Hockensmith used image moments to extract features from images and built the classification model using the random forest algortihm. K. Park used a histogram method for feature extraction followed by random forest classification. Finally, D. Skinner extracted feature vectors using an R command from the company Indico, classifying these vectors via the K-Nearest Neighbors algorithm.

2 Feature Extraction

The most difficult task in image classification is extracting information (features) from the images. The different species of plankton come in varying shapes and sizes. Classification is further complicated by the varying quality of the images as well as the heterogeneity in pixel resoltuion. The literature on image processing is broad in scope, and the proposed methods of feature extraction can be quite complicating. Some of the more complicated methods are SIFT, SURF, Bag of Features, etc. However, for the purposes of this report, image features will be extracted using *Krawtchouk* moments and the *Histogram* method. Both methods were found to have performed better than "best guess" criteria as reported on Kaggle's National Data Science Competition board. Additionally, features were also obtained using a more complex image extraction package available on R.

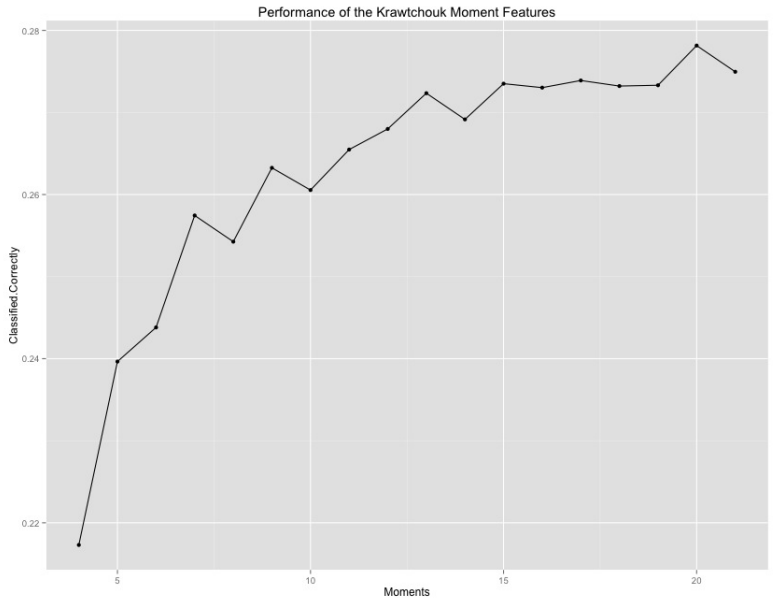
2.1 Krawtchouk Moments

Krawtchouk moments belong to the broader class of discrete, orthogonal moments. The $(m+n)$ Krawtchouk moment itself is defined as the sum, over the all x-y coordiantes, of the product of the pixel intensity function of the image $f(x, y)$ with the two weighted Krawtchouk polynomials ($K_n(a; p, N)$, $a \in \{x, y\}$ and $n \in \mathbb{N}$ is order of the moment in the x- or y-direction), where on ploynomial is specified in the x-direction and the other in the y-direction. The equation is provided below.

$$Q_{nm} = \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} \bar{K}_n(x; p_1, N-1) \bar{K}_m(y; p_2, M-1) f(x, y),$$

The motivation behind choosing to use Krawtchouk moments for the project is two-fold. First, they appear to be popular in the literture for their ability to reconstruct images at lower orders of moments. Second, their orthogonal properties make them invariant to scaling, translation,

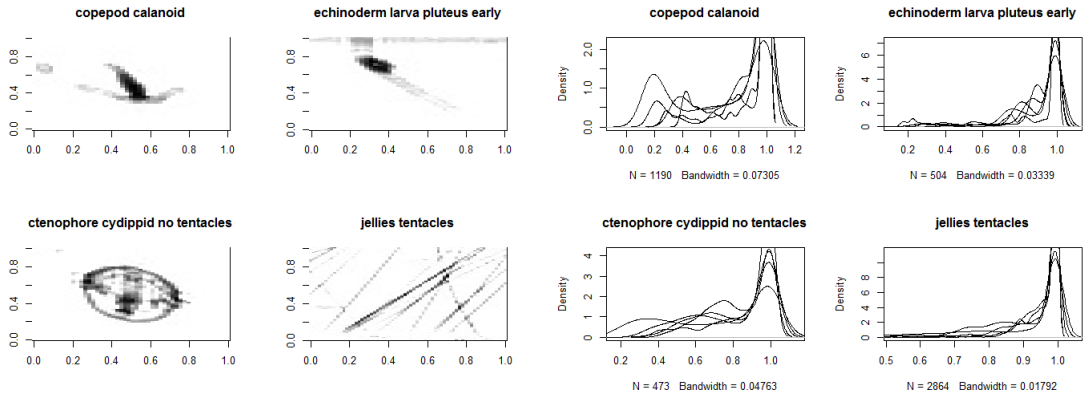
and rotation. For simplicity, only moments of order $(n + n)$ will be considered. The optimal order was determined by comparing the classification error across the different moments. As you can see below, moments of order 5 through 22 and thir respective correct classification percentages are provided. Observe that model performance begins to plateau after the 15^{th} moment.



(a) Moment

2.2 Histogram Method

The histogram method extracts information from the distribution of gray scale values that make up the image. We simply segmented the color ranges into bins and counted the frequency of values within each bin. Notice in the density plots for the species Copepod Calanoid and



(b) Plankton

(c) Grayscale Density

Ctenophore Cydipid (no tentacles) provided above that the distribution of values on the grey-scale are distinct. Note that values were scaled down to the interval $[0, 1]$. In general, this will usually be the case because of the uniqueness in the size and shape of the various species of plankton. 10 features total were extracted from each image, each feature representing one bin on the unit interval (with a bin-width of 0.1).

3 Classification Models

Two methods for classification, and their results, are provided below. The first method used the Random Forest and the second method the K-Nearest Neighbors algorithm.

3.1 Random Forest

The random forest procedure was implemented, setting the number of trees for each model fitting to 500. Kaggle scores were obtained from fitting models for the Krawtchouck and Histogram methods individually. For the *Krawtchouk* moments, the 10^{th}-order moment obtained a score of 3.67 (714 out of 1049) and the 22^{nd}-order moment received a score of 3.54 (704 out of 1049). Even though the 22^{nd}-order model performed better, the model benefited marginally from the addition of the 384 additional features. The histogram method received a score of 3.29 (660 out of 1049). Pooling features from both methods provided a significant boost in precision. The combination of the 9^{th}-order moments with the 10-bin features yields a score of 2.66 (598 out of 1049). This was our best result under the current timeframe of the project.

3.2 K-Nearest Neighbors

The K-Nearest Neighbors algorithm was a logical choice for classification of plankton because the algorithm is simple to impement and can provide as a benchmark for further classification. The algorithm works by assigning classes to images in the test set based on the Euclidean distance between images in the training set.

R does not contain much in the way of built-in image extraction features, but it does contain the command image_features in the indicoio package. This command produces a sparse, 2048 digit feature vector for each image that can then be used to calculate the Eucliden distances between different feature vectors. The drawbacks of this approach for feature extraction includes the production of possibly "too much" data to be reasonably useful and would likely require several more time intensive steps of data preparation for variable reduction.

The first few attempts at implementing the k-NN algorithm revealed some critical issues. One such issue is the choice in the number of nearest neighbors to use. Too few neighbors resulted in images being classified to different groups with equal probabilities. Kaggle consequently counts this is an incorrect classification. Another challenge is the susceptibility of the k-NN algorithm to noise and unimportant variables. As mentioned before, the feature vector is likely too large. Because of this, there is a reasonable chance that k-NN is suffering from too many variables. Currently, the best Kaggle score received for this approach is above the benchmark score.

4 Discussion and Future Work

In summary, the basic feature extraction methods decribed in this report do provide crucial information for the classification of plankton, as demonstrated with Kaggle scores above the "random chance" threshold. However, these methods are far from perfect in classifying plankton. In addition, using a large feature extraction algorithm, such as the indico.io package, doesn't necessarly mean the prouction of more features will build a better model.

Future work would include further refinements to both the classification algorithms and the extraction of image features with the goal of reducing misclassification as much as possible. Also, attempting other classification algorithms (Neuaral Networks, Naive Bayes, etc.) would also be a prime area to explore, but as always, ever expanding computation time cannot be ignored.

5 References

1. Yap, Paramesran, and Ong, *Image Analysis by Krawtchouk Moments*, IEEE Transactions on Image Processing, Vol. 12, No. 11, November 2003

2. Rani and Devaraj, *Face recognition using Krawtchouk moment*, Sadhana, Vol. 37, Part 4, August 2012, pp. 441460