# The analysis and results for the Right Whale Recognition kaggle competition by Sunday Srinivasan, Anton Rosenbloom. And Eric Dybdahl.

The competition statement is straight forward: given 11468 JPG images of right whales with 4544 set aside as a training set can you come up with a machine learning algorithm that will predict which 447 whales match with the remaining 6924 images.

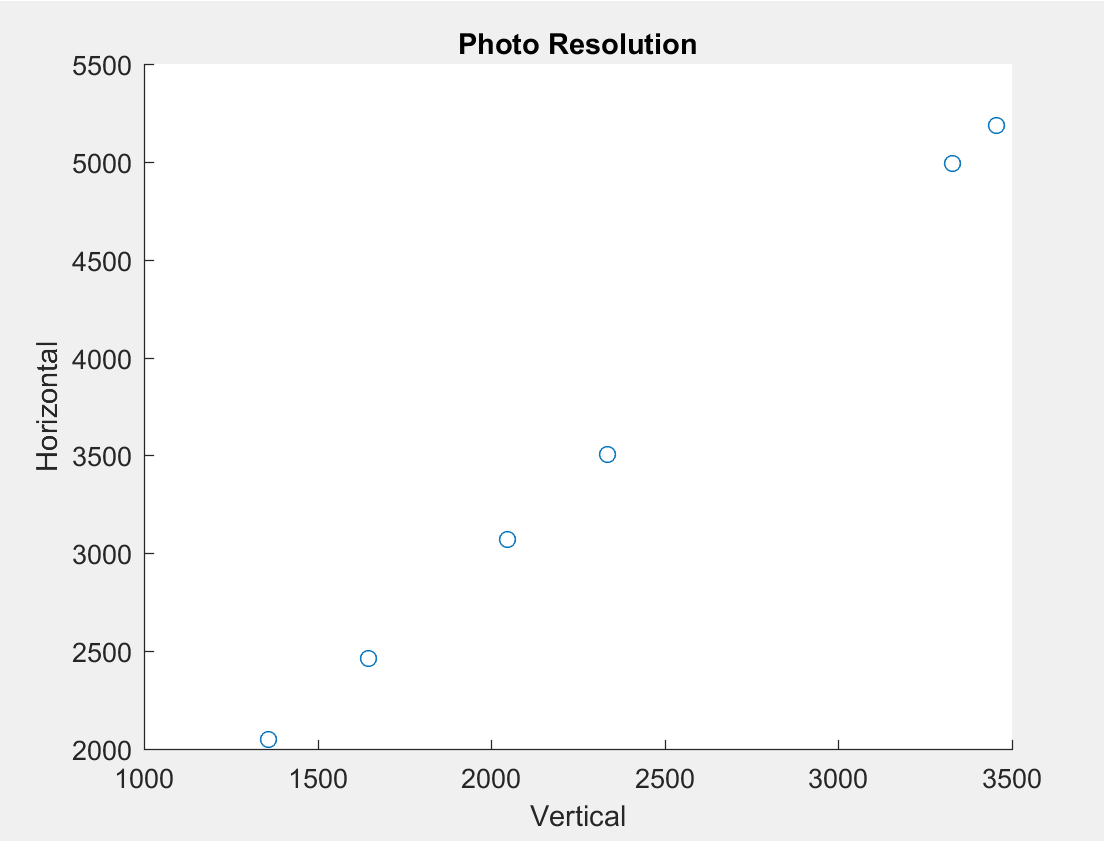
With this goal in mind, the task involved 3 sub-tasks: 1) Exploratory Analysis and Preprocessing of images (Eric), 2) Extracting features from images using the approach of Eigenfaces for recognition (Sundar) and 3) ML based classification (Anton).

These sub-tasks and lesson’s learned are detailed in the following.

# Data Analysis and image processing (Eric Dybdahl)

The Images are the dataset. There is no other attribute and the whale ids or targets are assigned to only the training set of 4544 photos. The image processing problem is to both modify and classify the photos adequately so that a recognizer can be applied to the photos.

The photos are all aerial shots taken from different angles under varied lighting and ocean conditions. The photo resolutions varied but aspect ratios remained constant as shown in the following scatter plot and histogram of the 4544:





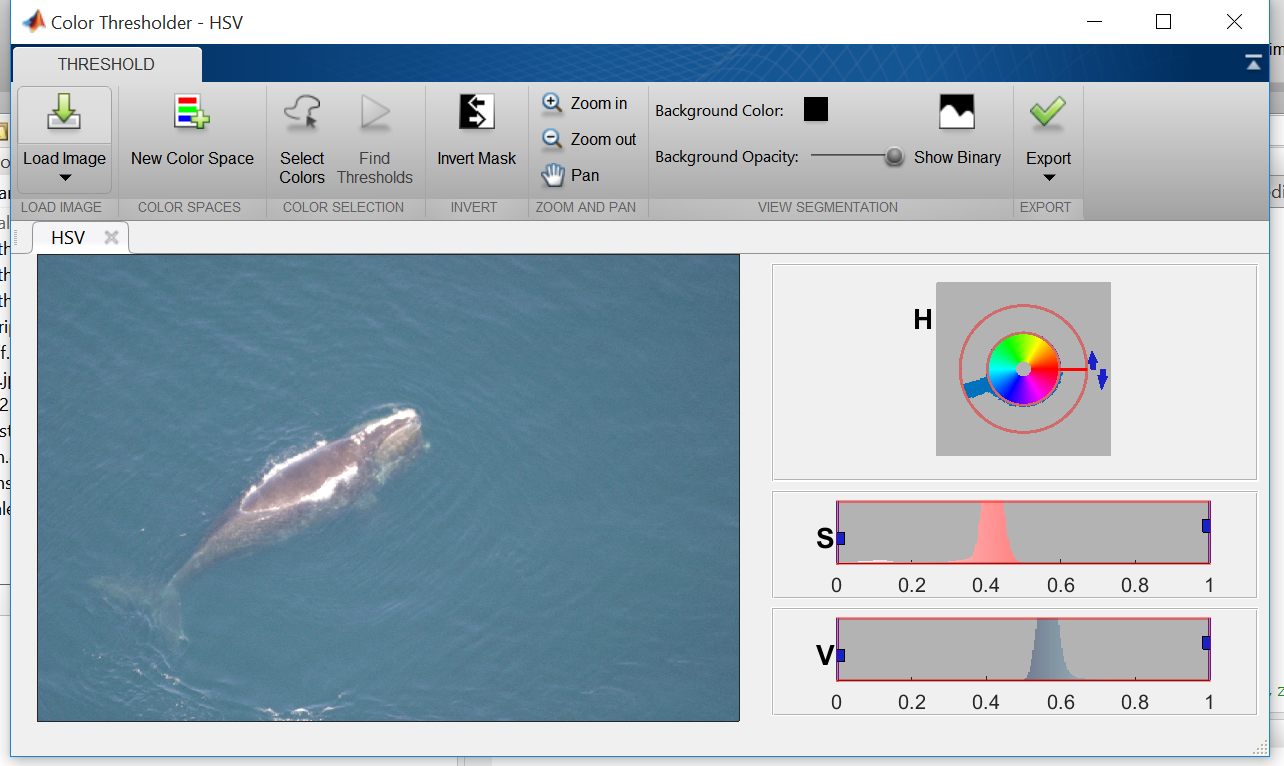
With the majority being 2048x3072.

We are trying to use a facial recognition algorithm on the whale images which means that we need to crop down to an image of the whale that is as uniform as possible. The image processing problem to be solved are as follow:

1. Orient the whale horizontally in the image.
2. Isolate the head and crop down to it.
3. Remove the obscuring spray from the whale image or find a set with minimal spray

The methods used in attempting to solve these problems were color separation, edge detection least squares fit and affine transformation.

We were able to achieve through the use of HSV color separation a distribution of the whale and a distribution of the spray for each photo. A typical HSV color distribution for a given whale is illustrated below:



Were there are five distribution regions for these photos:

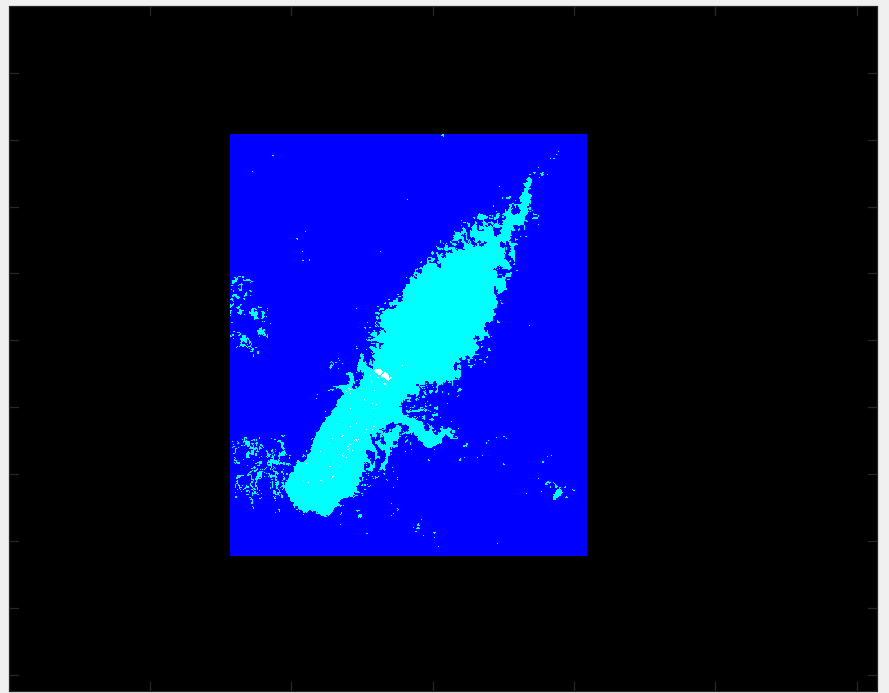
1. One in the H or hue range centered around 5.5 and describes mostly the blue green color of the water.
2. One in the S or saturation range above 0.2 and show strong in colorful photos.
3. One in the S range below 0.2 which defines most of the gray whale and the spray.
4. One in the V or value range below 0.8 which defines the most of photos light intensity.
5. One in the V range above 0.8 which defines the spray and the whales white head ornament.

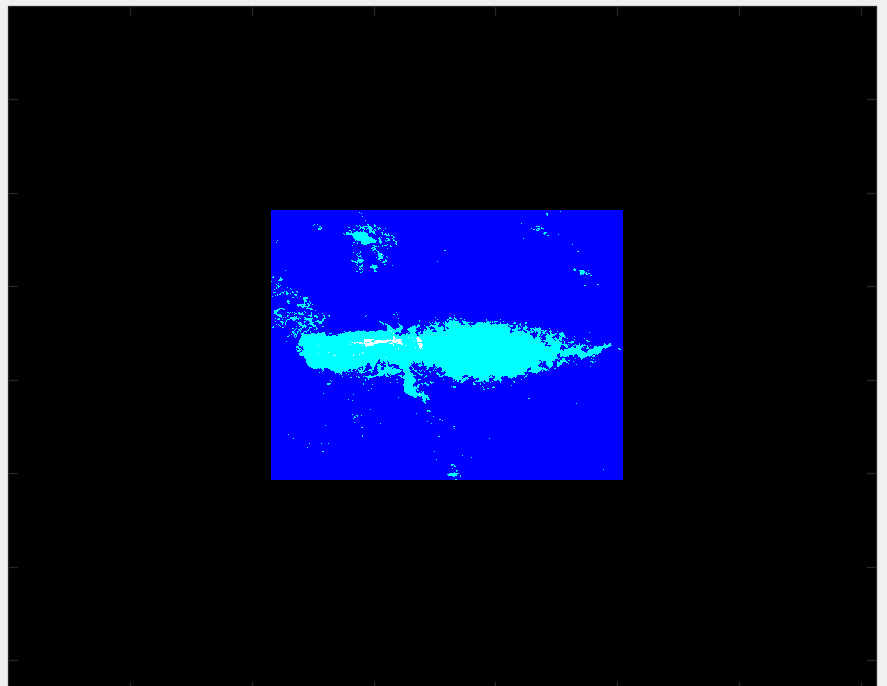
The intensity, mean and variance and existence or absence of these distribution characterize a whale photos color range.

The whale distribution can be pulled out with V < 0.8, and S < 0.2 or H > 0.6. The distribution of the spray with V > 0.8.

With the use of morphing techniques the whale distribution can be cleaned up so that we can get a center of mass and create an ROI, region of interest and zero out anything outside of it. Then run a least squares line through the center of mass. Using the center of mass and the slope of the line the photo can be translated and rotated with an affine transformation such the whale is both centered and horizontal. Cropping down on the whale distribution a new photo is created that is a centered and horizontal whale.









The direction of the head can be found by a color separation of the white head ordinate and the spray. It happens that in general the head ordinate is more yellow and slightly more color saturated than the spray. By calculating the center of mass for the head ordinate the direction of the head can be determined and rotated so that the head is facing right if necessary. Then the left half can be zeroed out so that a new center of mass and ROI can be calculated and used to crop down on the head and resize the image to 1024x1536.



With adjustments most of the whale images can be cropped down by using this method. Unfortunately only through visual examination of the results can we be able to know if the image was cropped down successfully. It would be nice to have some way of scorning the final image.

As of yet we have not attempted any adjustment for the spray.

This method was used to orient and crop whales into equi-sized images and constituted 1 set of image data that was then processed via the EigenFace algorithms so as to extract features for whale classification.

# Eigen-face based Whale Recognition (Sundar Srinivasan)

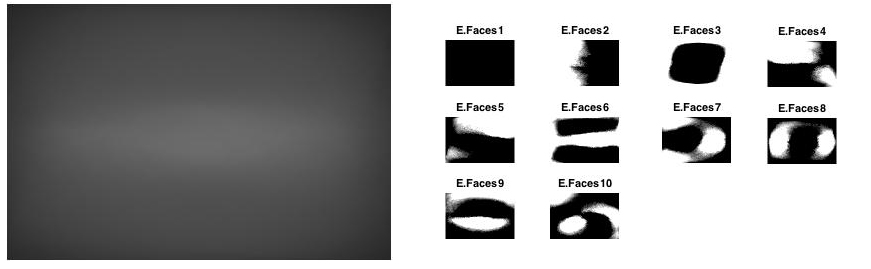
From the “T” training images (each RGB image is MxNx3 matrix of pixels), we determined whale ‘eigenfaces’ using the approach of Turk and Pentland (“Eigenfaces for Recognition”, J Cognitive Neuroscience, 3:71-86, 1991). Briefly, from the 4544 training images (cropped and oriented by Eric, previously), we determine eigenfaces and keep the top M as the eigenface basis that could explain 90% of image space variance.  We then projected each of our training set images (known ids) into the face basis and to obtain a vector of their ‘eigen weights’. These weights served as the set of ‘features’ that were then used to train models and classify images by whale ID.

Then for each new ‘test’ image, the image (transformed into a M\*N\*3 vector for RGB images; into a M\*N vector for Grayscale images) was projected into the whale eigenface basis to obtain the set of weights associated with the test image, and used in concert with the trained model to classify the whale ID associated with the test image.

In our first attempts at using this algorithm to classify images, we used the approach of Euclidean distance between the test eigen weights vs that of the training eigenweights to assign whale IDs to the test image. The idea was to ultimately use these weights to design more sophisticated, ensemble ML models for classification (by Anton).

However, some issues rapidly became apparent including memory limits that required conversion of RGB images to gray scale and resizing images to 25% of original size (which required acceptance of the accompanying loss of image resolution).

We then applied the Eigenface algorithm to the cropped oriented images that Eric had generated at a first pass, and images of the resulting ‘average’ whale (merging of all training images), and the top 10 whale eigenfaces are shown below:



The eigenface basis, and associated weights were then determined for each of 4544 training images. This data set was then partitioned into a 70:30 train-test split, Euclidean distance used to classify (the test images) and the accuracy of classification was determined.

For the 1st instance where whales were oriented and cropped, the classification accuracy of the eigenface based recognition method was extremely poor at 0.22 + 0.15%.

This result may not have been surprising, given that the number of pictures per unique whale was extremely skewed in the training set (ranging from n=1 to n=47 images/unique whale). Additionally, the first pass with orientation of whales and cropping had some misses (apparent on visual inspection, but not otherwise). The other possibilities for this failure included: 1) the noisiness of the images where whale’s images were obtained across a range of lighting conditions, camera angles whale positions (e.g., anterior-posterior to medial-lateral or face up – face down positions), 2) the range of whale ‘pixel’ resolution (due to camera elevation), 2) loss of information in coversion from RGB to grayscale, 3) loss of image resolution due to memory issues, and 4) the inappropriateness of the Eigenface approach for the classification of noisy image data. We attempted to (somewhat) systematically address each of these issues by:

1. Researching alternative means to better orient, crop and possibly identify features that could be used to ID whales (Anton).
2. Grouping whales (exploratory clustering) so that the large training set could be split up into smaller sub-sets based on similarity measures
3. Exploring the appropriateness of the Eigenface approach for whale recognition (Sundar and Eric).

# Exploratory Clustering of Images (Eric Dybdahl)

In part due to the (poor) classification accuracy The 4544 test images are very diverse in brightness and color saturation. In order to classify the images better HSV color distribution statistics for the five distributions were added to the train table. As follows:

hsum, hmean, hvar: is the magnitude, mean and variance for the hue distribution.

vlsum, vlmean, vlvar: is the magnitude, mean and variance for the lower value distribution

vhsum, vhmean, vhvar: is the magnitude, mean and variance for the high value distribution.

slsum, slmean, slvar: is the magnitude, mean and variance for the low saturation distribution

shum, shmean, shvar is the magnitude, mean and variance for the high saturation distribution.

Preformed unsupervised learning kmean clustering on the following columns, to establish a classification of photos: hmeam, hvar, vlmeam, vhmean, slmean, shmean. K of 4 was the most stable and the centroids are as follows:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Size | hmean | hvar | vlmean | vhmean | slmean | shmean |
| Cluster1 | 1501 | 0.5644 | 0.0034 | 0.5090 | 0.9241 | 0.1028 | 0.4368 |
| Cluster2 | 912 | 0.5540 | 0.0105 | 0.4757 | 0.9149 | 0.1138 | 0.2694 |
| Cluster3 | 1280 | 0.5552 | 0.0054 | 0.3538 | 0.9239 | 0.0933 | 0.5015 |
| Cluster4 | 798 | 0.5117 | 0.0185 | 0.2994 | 0.8955 | 0.1038 | 0.2788 |
| NaN | 53 |  |  |  |  |  |  |

Cluster 1 – brightest photographs from vlmean=0.5090 and high color saturation of shmeam=0.4368. Water color is in a consistent blue-green color range; hmean=0.5644 with a narrow variance of hvar=0.0034.



Cluster 2 – still bright but grayer photographs with vlmean=0.4757and low color saturation of shmeam=0.2694. Water color is more variable between a grayer green to blue range; hmean=0.5540 with a larger variance of hvar=0.0105.



Cluster 3 – are darker but still have high color saturation with vlmean=0.3538 and shmean=0.5015.The water Color is in the narrow blue-green range: hmean=0.2994 with a narrow variance of hvar=0.0054.



Cluster 4 – are both bark and gray with vlmean=0.2994 and shmean=0.2788. The water color is mostly gray with hues ranging form green to blue: hmean=0.5117 with a larger variance of hvar=0.0185.



NaN – which could be considered cluster 5 have no distributions for either high value or high saturation. They are either very black or very gray.

