# Part 2: Model construction and evaluation

## Construct analytic model(s) to address the project objective

We approached the problem of whale identification with so called Eigenface algorithm -- developed in 1987 by Sirovich and Kirby and later implemented by Turk and Pentland and widely used for face recognition. At the foundation of the algorithm lays the Primary Component Analysis method.

The analytic model we constructed to address the project objective is a distribution of data-points of the labeled whale images within so called “Eigenface-space” determined by the eigenvectors extracted from the set of labeled images.

The process of determination of the eigenvectors includes generation of the “average” image. On Fig. 1 is shown an example of such average image for a subset of training images.



Figure . The average "face"

Fig. 2 presents the visual representation of the final product of the PCA – top 10 eigenvectors generated for a subset of 929 labeled images (the “Eigenfaces”):



Figure . Top 10 Eigenfaces

Mathematically each of the images is represented by a vector of coordinates of its location in the Eigenface space.

The process of identification of a new image is performed through classification of its representation in the Eigenface space. We’ve used Random Forests and Euclidian distance for classification.

Due to the wide variance in quality of the images to be analyzed, we decided to divide the original image-set into groups based on K-means clustering of the HSV values. Centroids of the resulting clusters are shown on Fig. 3:

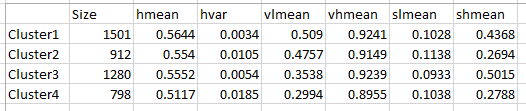


Figure . HSV cenroids of the image-set

Distinction between the clusters can be seen on the line-up of the representatives of each cluster on Figure 4.

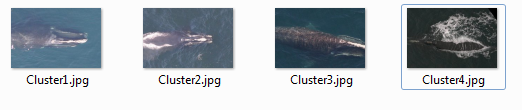


Figure . HSV-based cluster samples

## Evaluate the model outcomes

This is a work in progress.

## Iterate and improve the model when necessary

This is a work in progress.

## Justify the final model and its output

This is a work in progress.