Whale Detection & Identification from Aerial Photography

Aditya Mahajan, Adrien Perkins
{mahajan1, adrienp}@stanford.edu
Department of Aeronautics & Astronautics, Stanford University, California.

1 Introduction

The North Atlantic Right Whale is a critically endangered species of baleen whale, with an estimated fewer than 500 individuals living in our oceans [1, 2]. In the interests of science and conservation, it is necessary to track and monitor these animals. Currently, only a few experienced researchers are able to identify and tag individual whales through the use of custom software and this process is extremely time consuming [1, 3]. This project focuses on detecting and identifying whales from aerial imagery using techniques from machine learning and image processing.

A large dataset (comprised of about 11,000 aerial images) has been made available by Kaggle for this task. Approximately 4,500 of these images are labeled as training data and the remainder make up the test set.

The problem was split into the following two parts:

1. Detecting a whale in the image and determining which pixels are whale and which are ocean. Two machine learning algorithms, one supervised and one unsupervised, were employed for this task.
2. Identifying the individual whale (designated by ID number) in a given image using baseline Scale Invariant Feature Transform (SIFT) feature matching techniques and supervised machine learning.

1.1 Dataset

The dataset provided for training consists of 4,500 images of 447 unique whales. The images have varying resolutions, between 4–17 megapixels. While many of these images are of excellent quality and capture the whale clearly, others pose several challenges from a computer vision point of view. These are illustrated in Figure 1 and include:

- Poor lighting from reflections on the water or lack of contrast between the whale and the water
- Varying orientations of the whales from breaching to diving as well as feeding and spouting
- Occlusions due to sea foam and distortion from the water

An additional drawback of the dataset was the skewness of the labeled data in terms of images per unique whale. Figure 2 highlights the uneven nature of the training data. It can be seen that more than half of the whales had fewer than 10 images each, whereas one specific whale appeared in 45 different images. This was expected to pose a challenge during the identification process.

1.2 Related Work

Previous work in animal tracking underwater relies on many different types of vision-based sensing like multi-channel imagery and infrared cameras [4, 5]. The dataset used in this work consisted only of images in the visual range, so it was inferred that detecting the whale under the water was similar to detecting elements in an image under a shadow. This process is handled very well in the HSV color space along with K-Means clustering [6, 7].

Feature matching using the Scale Invariant Feature Transform (SIFT) has successfully been used for ecological research on marine fauna like manta rays [8] and leatherback turtles [9]. Given that many of the whales have visually distinct markings on their backs, tails, and noses, SIFT feature matching has been considered as a key step in the whale identification process.

2 Technical Approach

The whale detection and identification problem has been divided into two sub-problems. The tasks of determining which pixels in an image belong to

Figure 1: Many images from the dataset suffer from poor lighting, occlusions and distortions. Above are a few examples that pose a challenge for whale detection.

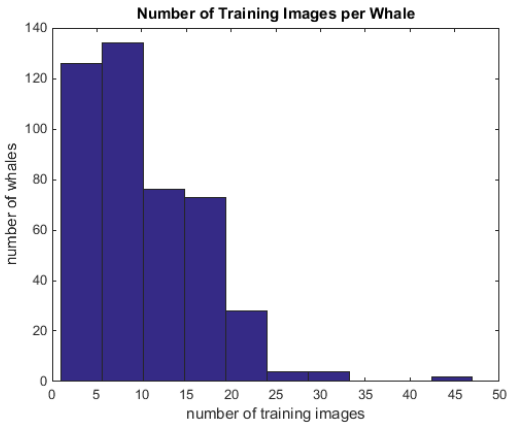

Figure 2: The frequency of the number of training images per whale.

a whale is addressed in Section 2.1, and the task of determining the unique identification number of the detected whale is addressed in Section 2.2.

2.1 Whale Detection

Two means of detecting a whale in an image were considered. The Cascade Object Detector (COD), a supervised learning tool in MATLAB, and K-Means Clustering (KMC), an unsupervised approach for partitioning data into sub-spaces based on the characteristics of each data point.

2.1.1 Cascade Object Detector

The COD is a built-in training image labeling app for MATLAB. Using this app, 500 images were manually annotated with Regions Of interest (ROIs) corresponding to positive and negative instances of a whale sighting, expressing each annotation as a Histogram Of Gradients (HOG) [10]. The HOG is a common descriptor used in computer vision and works by counting gradient orientations in a local neighborhood of pixels. The COD uses the Viola-Jones [11] algorithm, originally developed for face detection, and outputs a set of bounding boxes which cover the ROI of a alleged whale.

2.1.2 K-Means Clustering

KMC was executed for each image, with each pixel represented in 3-dimensional HSV (Hue-Saturation-Value) space. HSV space was chosen over RGB space due to its robustness to different lighting conditions that may be present throughout an image. The difference can be visually seen in Figure 3 and the corresponding KMC results are shown in the same figure.

During KMC, the pixels were placed into one of four clusters (roughly mapping to ocean, whale, reflections and foam). The cluster with the largest number of pixels was assumed to be a binary 'ocean' mask and the remaining pixels were inferred as 'whale'. Due to the possibility of small, disconnected regions (reflection and foam) being classified as whale, the following steps were taken to ensure one contiguous 'whale' object:

- The 'ocean' mask was dilated and eroded to close small holes
- The 'ocean' mask was inverted to obtain a 'whale' mask
- The largest contiguous blob in the 'whale' mask was retained
- The 'whale' was slightly eroded to mitigate feature extraction from surrounding foam

To expedite convergence of KMC, the cluster centroids were initialized as a random sample of the pixels in the image. Intuitively, this increased the possibility of a centroid being initialized near its converged location.

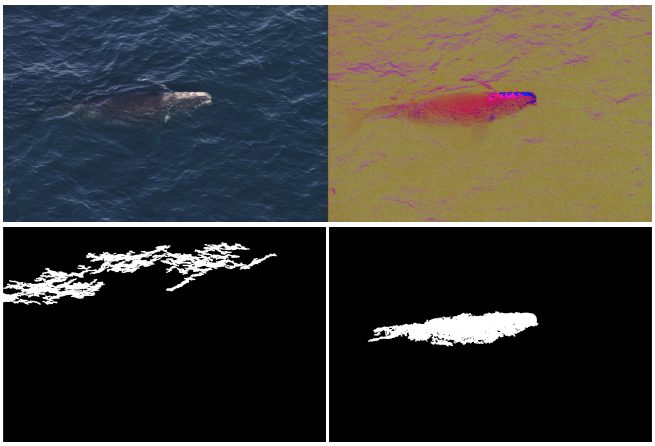

Figure 3: A whale difficult to see (top left) becomes obvious in HSV (top right). KMC performs poorly in RGB space (lower left), compared to HSV space (lower right).

2.2 Whale Identification

Having manually looked at many of the whale images provided, it was noticed that most whales had distinct features on the nose and back and we felt SIFT features may be able to capture that information. Furthermore, SIFT is invariant to rotation and scale, dispelling the need to rotate/scale the whales. Therefore, the first step to identification is the successful extraction of good SIFT features for each of the whale images.

We then looked at two different ways of whale identification, testing these methods using k-fold cross validation with 4 folds. Given that a portion of the whales only have a single training image, the cross validation was modified to guarantee at least one training image per whale.

2.2.1 Feature Extraction & Dimensionality Reduction

Feature descriptors were extracted from the pixel regions, determined in the section above, using SIFT [12]. SIFT is a common feature extraction method and is invariant to changes in lighting and orientation. Furthermore, the descriptors rely on local pixel information, making them robust to minor occlusions.

The result of SIFT is a descriptor for each feature (there can be thousands of features on a whale) in 128-dimensional space. The sheer size of the dataset (11,000 images), combined with the high dimensionality of the feature space, meant that this information had to be condensed for efficient computation.

Principal Component Analysis (PCA) was used to reduce the dimensionality of the data to more manageable 16 dimensional vectors [13]. Note that the objective of PCA was to reduce data dimensionality with minimal information loss and improve feature matching speed. PCA was not used to reduce overfitting.

Each image, n, was represented by a set of m_n features. This meant that the images themselves could be discarded, and instead be represented by a feature matrix, $F_n \in \mathbb{R}^{16 \times m_n}$.

To ensure that the reduced SIFT data performed like the full dimensional SIFT data, descriptor matching was done on a subset of the images with both the 128 dimensional and 16 dimensional data. The resulting matches for the 2 data sizes were very similar, indicating that any information lost during PCA was negligible.

2.2.2 Basic Feature Matching (Baseline)

The baseline method involves simple SIFT descriptor matching. Two descriptors are declared 'matched' if the euclidean distance between them is a threshold factor closer than any other descriptors. Two sets of descriptors (one per image) can be compared to identify those shared across both images.

All of the training images in 3 of the 4 folds were combined to create a set of features to define each of the 447 different whales. E.g. for a given whale, k, its feature set, $S_k = [F_i, F_j]$ where $S_k \in \mathbb{R}^{16x(m_i+m_j)}$ if images i and j contain whale k. Each image in the 'test' fold was compared to each of the 447 different sets of features to get a matching vector $M \in \mathbb{R}^{447}$ which contained the number of unique matching features between the test image and the feature sets. To account for the different number of whale images that made up each of the feature sets, M was normalized by the number of features in the given feature set. Finally, the test image was tagged to the whale that had the highest match value in M.

2.2.3 Support Vector Machine (SVM)

The supervised learning identification method used was an SVM using the libSVM library for MATLAB

[14]. The features used in the SVM were each of the different 16-dimensional SIFT features themselves and the label for each SIFT feature was the ID of the whale that feature came from. This resulted in an SVM with 447 different classes.

Much like the baseline, 3 of the folds would be used to train the SVM and one used for testing. To classify a specific test image, each of the image's 16-dimensional SIFT features were labeled with the SVM and the final identification of the image was that of the most commonly occurring label.

3 Results

This section summarizes the results of the whale detection (Section 3.1) and identification (Section 3.2) algorithms. Training and testing was done with k-fold cross validation with 4 folds.

3.1 Whale Detection

The detection results obtained from COD and KMC are compared in Figure 4. In the upper sample, it can be seen that KMC (right) is able to detect more of the whale than the COD, whereas in the lower sample, KMC avoids selecting ocean pixels due to the 'blob' versus 'bounding box' nature of the ROIs. In an attempt to assess the performance of KMC and the COD, 100 random sample results were examined manually. In this assessment, the KMC performed better than the COD almost 90% of the time. For these reasons, KMC was used to obtain the results in Section 3.2.

3.2 Whale Identification

The performance of the whale detection and extraction of features was crucial to the performance of the whale identification. Unfortunately, while a lot of improvements were made to the detection and feature extraction steps, we were unable to get either of the identification methods to work successfully. However, in this section we describe some of the challenges faced and improvements that were made throughout this process.

3.2.1 SIFT Feature Matching (Baseline)

The use of SIFT feature matching for whale identification posed several different challenges. When using the initial features to do the matching, almost all of the test images matched to the few whales that had the most images in the training set. This was due to the fact that the KMC detection method was not perfect and would include some water or foam from the training images. These features would successfully, albeit incorrectly, match to watery parts of many of the test images.

To minimize water and foam matches, feature sets derived from multiple images of the same whale were pruned. To prune the feature set, only the features within a threshold distance of each other (within 60% of the best match) were retained, and any 'odd' features were discarded. While the pruning did help reduce the number of features for most of the whales, aiding computation time, there was still difficulty with some water features, as that water was common between the images. Despite this, when running SIFT matching, the test images no longer matched to the same whales as before. The test images now had a preference to match to whales that had training sets comprised of only a single image.

This new problem most likely occurred from the dilution of good matches with bad features in the feature sets. For a whale with only a single image, most of the false matches arise from water or foam. While comparing a test image with the correct whale feature set (comprised of multiple images), the good matches on the whale were diluted by the number of bad features there were in the set. Looking at a percentage match, the percentage was reduced due to these bad features. We are still looking into properly tuning the detection and feature extraction steps to produce the best feature sets for this method.

3.2.2 SVM

Training the SVM was incredibly expensive (computationally) and hence time consuming, taking over 20 hours with only a quarter of the training set. With just that quarter of the data, the SVM failed to correctly classify any of the training images. Due to the lengthy computation time and minimal success on the baseline method, we focused our efforts on getting better features to better define each of the whales and testing with the baseline method. It was believed that feature extraction was a priority over implementing the SVM, as the SVM was based on the same features.

Figure 4: Original image (left), and features (red) detected in 'whale' areas as identified by COD (center), and KMC (right).

4 Conclusions

Whale detection was successfully performed on aerial imagery with K-Means Clustering on images in the HSV space. This resulted in superior whale detection in comparison to the Cascade Object Detector. Using SIFT features for identification proved to be extremely challenging and depended strongly on the ability to extract good, reliable features from each image. While many iterations and improvements were made on the SIFT features (like dimensionality reduction using principal component analysis), robust whale identification was unsuccessful using either baseline SIFT feature matching or an SVM with SIFT features as the SVM features.

5 Future Work

The whale detection method used in this paper performed very well. However, there is room for improvement. In order to better extract the whale itself from the image, methods of background removal can be implemented to possibly remove frequency content that corresponds to regular ocean wave patterns. It is understood that the KMC algorithm is prone to problems with local minima (see Figure 5) which can be managed by running KMC several times or attempting to initialize KMC with learned centroid values for whale, ocean, reflections and foam.

For whale identification, the authors would like to suggest many improvements, especially with trying other classification methods, such as Convolutional Neural Networks. Robust whale identification requires impeccable detection performance and even further processing of the images, such as aligning all the whales to be in the same direction.

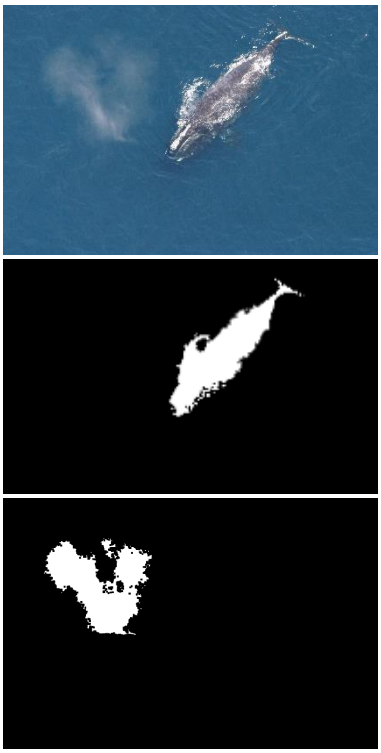

Figure 5: KMC on an image (top) can lead to correct (middle) or incorrect (bottom) whale detection.

References

[1] Kaggle Inc. Right whale recognition, 2015.

[2] NOAA Fisheries. North atlantic right whales (eubalaena glacialis), 2015.

[3] New England Aquarium. Digits: Digital image gathering and information tracking system, 2015.

[4] Yuliya Podobna, Jon Schoonmaker, Cynthia Boucher, and Daniel Oakley. Optical detection of marine mammals. In Weilin (Will) Hou, editor, *SPIE Defense, Security and Sensing*, volume 7317, pages 73170J–73170J–11, 5 2009.

[5] Henry Stark. Considerations in designing a marine-mammalship collision avoidance system based on aerial imagery by an unmanned airborne vehicle. *Optical Engineering*, 42(1):11, 1 2003.

[6] S. Vitabile, G. Pollaccia, G. Pilato, and E. Sorbello. Road signs recognition using a dynamic pixel aggregation technique in the hsv color space. In *Proceedings 11th International Conference on Image Analysis and Processing*, pages 572–577. IEEE Comput. Soc, 2001.

[7] Rita Cucchiara, Costantino Crana, Massimo Piccardi, Andrea Prati, and Stefano Sirotti. Improving shadow suppression in moving object detection with hsv color information. In *ITSC 2001. 2001 IEEE Intelligent Transportation Systems. Proceedings (Cat. No.01TH8585)*, pages 334–339. IEEE, 2001.

[8] Christopher Town, Andrea Marshall, and Nutthaporn Sethasathien. Manta matcher: automated photographic identification of manta rays using keypoint features. *Ecology and Evolution*, 3(7):1902–1914, 7 2013.

[9] P.M. DeZeeuw, E.J. Pauwels, E.B. Ranguelova, D.M. Buonantony, and S.A. Eckert. Computer assisted photo identification of dermochelys coriacea. In *International Conference on Pattern Recognition (ICPR)*, 2010.

[10] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[11] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–511–I–518. IEEE Comput. Soc, 2001.

[12] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 11 2004.

[13] Aly A. Farag and Shireen Elhabian. A tutorial on principal component analysis, 2005.

[14] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.