**Humpback Whale Identification Challenge**

1. **Background and problem statement:**

Whaling, the practice of hunting of whales for products like meat, oil, and blubber, has been a practice since 3000 BC. The emergence of the competitive national whaling industries dominated in the 18th and 19th centuries and the practice of ‘whale harvesting’ in the first half of the 20th century led to the extreme depletion of whale stocks. According to the Canadian Encyclopedia by the 1930s more than 50,000 whales were killed annually. In 1986, the International Whaling Commission (IWC) put an end to commercial whaling. [[1]](#footnote-0)

The recovering whale populations still face difficulties and challenges that directly affect their survival. Whales struggle to adapt to warming oceans and compete with the industrial fishing industry for food. As part of the whale conservation efforts, scientists use photo surveillance systems to monitor ocean activity. The shape of whales’ tails and unique markings from the recorded images help scientists identify the species of whale they are analyzing. Furthermore, this helps them log whale pod dynamics and movements.

For the past 40 years most of this work has been done manually by scientists, a process which is both slow and time-consuming. This has left a plethora of data unutilized. The purpose of this project is to automate the process of whale identification.

**2. Data Description**

1. This training data contains 9850 images of humpback whale flukes. Individual whales have been identified by researchers and given an Id.
2. The challenge is to predict the whale Id of images in the test set. What makes this such a challenge is that there are only a few examples for each of 3,000+ whale Ids.

Data set files description on Kaggle:

1. train.zip - a folder containing the training images
2. train.csv - maps the training Image to the appropriate whale Id. Whales that are not predicted to have a label identified in the training data should be labeled as new\_whale
3. test.zip - a folder containing the test images to predict the whale Id - 15610
4. sample\_submission.csv - a sample submission file in the correct format

**3. Approach**

After doing some research for similar image recognition problems several approaches were identified to solve this problem as given below:

**Image Processing:**

The dataset is a mix of both colored and black and white images. Since the images have very high contrasts of color and quality, the images would be converted to grayscale images for modelling purposes. Images also have varying sizes and thus need to be standardized by reshaping the images to same size. There are also plenty of images with text and white boundaries which would need to be removed since it is irrelevant to our problem at hand. All these tasks can be achieved by using the OpenCV module in python as per our research.

**Neural Networks:**

A variant of neural networks called Convolutional Neural Networks(CNN) is a popular go to model for image recognition. For this competition we would leverage the deep learning libraries and data flow framework in python called Keras and Tensorflow respectively. A relatively new addition to family of neural networks set to replace CNN in the coming future called Capsule Networks would also be used. These neural networks would be trained on preprocessed images and the same working pipeline for preprocessing of images would be applied on the test set while predicting.

**Support Vector Machines (SVM):**

Due to the potential problems with overfitting and biases arising from unbalanced classes in neural networks, SVM would be used in hopes of making a more robust classifier. However, SVM is very sensitive to curse of dimensionality and thus further image preprocessing would be required. HOG features would be calculated which gives unique 32 features of an images which can be used for training the model. An image if standardized to 50 x 50 pixel size would give 2500 columns to train an SVM model which is not practical due to curse of dimensionality.

**Brute Force:**

This approach does not require training of a model at all but algorithms popular in Computer Vision. One such algorithm to generate keypoints in an image is scale-invariant feature transform (SIFT). These keypoints would be calculated for all the training set images. These keypoints can then be correlated with a test image and whichever set of keypoints gives the highest amount of correlation; its label would be the classification of the test image. This approach is very computationally inefficient, but it would be interesting to see how it performs versus model based approaches.

**4. Computational tools and libraries that will be used:**

1. Tensorflow
2. Keras
3. Opencv
4. Numpy
5. Pandas
6. PIL

**5. Team management tools:**

1. Slack - for coordination and communication with-in the team
2. Github - to access and collaborate the project work

---

1. https://en.wikipedia.org/wiki/Whaling#cite\_note-2 [↑](#footnote-ref-0)