A face classifer for North Atlantic Right whales

Matt Smith[1,*], Abhir Bhalerao[2], Ben Graham[3]

[1]Department of Statistics, University of Warwick, Coventry, United Kingdom
[2]Department of Computer Science, University of Warwick, Coventry, United Kingdom
[3]Facebook Artificial Intelligence Research, Paris, France
[*]m.d.smith@warwick.ac.uk

Abstract: Accurate monitoring of individuals in a threatened species is of upmost importance to conservationists and researchers. Human observation is expensive and autonomous ariel photography is becoming an increasingly useful technique regarding animal biometrics [1, 2]. Fewer than 500 North Atlantic right whales are left in the world's oceans. As with many animal biometric inspection processes, tracking and monitoring individuals is an extremely time consuming process. Advances in the implementation and performance of deep learning algorithms have drastically improved performance in object detection and recognition tasks [3]. We employ a wide range of interesting techniques to build a "face-identification" algorithm for ariel photos of 447 unique. We follow a conventional modern face recognition pipeline consisting of the stages: detect, align, represent and classify [4]. We use deep learning algorithms to both detect and classify. A fully convolutional network [5] is employed to semantically segment a given image to detect the location of the whale's head and body, we then use PCA on the resulting image to normalize for the whale's direction. A significant amount of hand labelled masks are needed to generate enough supervised training data to make this work effectively [6]. We tackle this issue by employing semi-supervised learning techniques and histogram matching between images to improve our localization algorithm and find a significant improvement in our results.

1. Introduction

We entered the 2015 Right Whale Recognition online competition issued by Kaggle. Data consists of aeriel images, the vast majority containing a single Right whale. There are M = 447 unique whales, each of which has at least one photo in the training set which contains 4543 labelled images. The test set contains N = 6925 unlabelled images. Evaluation is based on the multi-class logarithmic loss

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{M} y_{ij}\log(p_{ij}), \tag{1}$$

where log is the natural logarithm, y_{ij} is 1 if observation i belongs to whale j and 0 otherwise, and p_{ij} is the predicted probability that observation i belongs to whale j.

The data was collected and labelled over a 10 year period by NOAA (National Oceanic and Atmospehric Administration) scientists via numerous helicopter trips over the northern Atlantic.

We follow the conventional pipeline of alignment and classification and break our task of classifying a given photo, denoted X_i into two main stages, both of which employ using convolutional

neural networks;

- Alignment - We first reduce the dimensionality of X_i and normalize for the distance to and orientation of the whale, generating a headshot of the whale, denoted X_i^h.
- Classification - X_i^h is passed through a classifier which outputs a probability mass function over the 447 whales.

1.1. Related work

This problem is largely analagous to general facial recognition problems, ableit a different species. State of the art results for these problems and general large scale image recognition problems has been obtained by applying deep convolutional neural networks which have seen a huge resurgance in the field since 2012 [7, 8, 9, 10].

This problem also requires an object detection stage or localisation method to extract the key class feature descriptor from the huge dimension of redundant information i.e. the whale's head from the sea. Again deep learning algorithms have been used effectively for object detection by making bounding boxes aroud the region of interest (ROI) [11, 12].

Object localization has also been solved by using convolutional networks to classify each pixel of the input image. This method, known as semantic segmentaion, has gained popularity in many applications, the main advantage being that the bounding contours, as opposed to bounding boxes, can take on complex shapes, don't have to overlap and dramatically reduce the size of the proposed ROI reducing the potential for false positives [13]. Moreover, there is no limit to the number of bounding contours you can have in the image. The main disadvantage being that it requires dense pixel-level labels for training.

The competition's first and second place teams both used a similar localization and identification pipeline [14].

2. Alignment

It is helpful to remove variation in inputs before giving them to a deep learning algorithm and, especially with faces, the success of a learned network is highly dependant on an alignment step [15, 4].

2.0.1. Mask prediction: We randomly choose 550 images from our training set M^{Train} and another random 150 images to generate a test set M^{Test}. Using a graphics editor, for each $X_i \in M^{Train}, M^{Test}$ we create a semantic mask denoted M_i. An example of a pair is shown in figure 1.

(a) (b)

Fig. 1. *Example of* $\{X_i, M_i\}$ *pair; a)* X_i*, b)* M_i*.*

We distinguish between head, body and sea using red, yellow and black respectivley. Having two colors for distinguishing parts of the whale enables us to infer the direcition in which the whale is pointing. We rescale each X_i to dimension $w \times h \times c = 600 \times 900 \times 3$ and each M_i to $w^{'} \times h^{'} \times c^{'} = 19 \times 29 \times 3$. We further process X_i as described later on, creating X_i^1 and feed it into a fully connected convolutional neural network (FCNN) to learn a function $f_1 : \mathbb{R}^{w \times h \times c} \rightarrow \mathbb{R}^{w^{'} \times h^{'} \times c^{'}}$. We are not interested in a huge amount of detail being produced in our predicted $\hat{M}_i$, enough to infer head and body location, hence our choice for relatively small output space.

We describe our neural network architecture as follows;

$$f_1 = \{X_i^1 - \text{down}_0 -, ..., -\text{down}_4 - 3\text{C}_{2D}3/1 - \text{sigmoid } - \hat{M}_i\}, \tag{2}$$

where

$$\text{conv}_\text{n} = \{(48 + 32\text{n})\text{C}_{2D}3/1 - \text{BN} - \text{ReLU}\}, \tag{3}$$

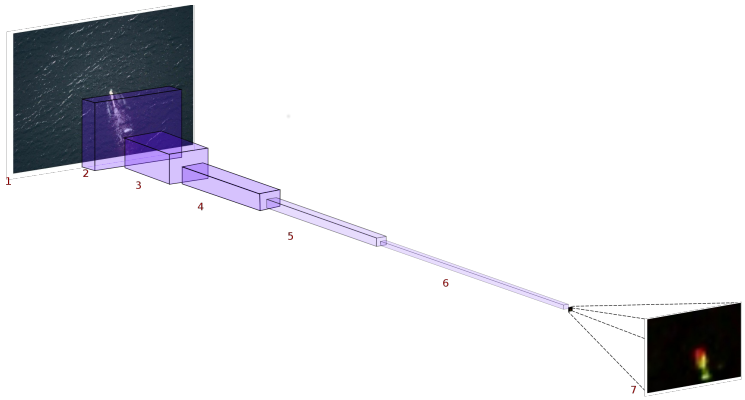
$$\text{down}_\text{n} = \{\text{conv}_\text{n} - \text{MP}3/2\}. \tag{4}$$

The following notation for denoting a network architecture is similar to [9]. $\{a - b - c - d\}$ denotes a network where the intial tensor a is passed through the layers b, the output of which is passed through c finally creating output d. $fC_{iD}k/s$ denotes an i dimensional convolutional layer with kernel size k in each dimension, a stride of s and number of filters f. Similarly MP$k/s$ is a two dimensional max-pooling layer with kernel size k and stride s. Other layer notations; BN = batch normalization, ReLU and Sigmoid are layers of rectified linear activation units and sigmoid activation units respectively. An illustration w of f_1 is given in figure 2.0.1.

We train f_1 for 20 epochs, using the mean squared error (MSE) as our loss function and the Adam optimization algorithm with an initial learning rate of 0.001 and batch size of 5. Together with the MSE we also look at a variant of the Sørensen-Dice coefficient to compare model performance [16], given by

$$\text{QS}(M_i, \hat{M}_i) = \frac{2|M_i \cap \hat{M}_i| + 1}{|M_i| + |\hat{M}_i| + 1}. \tag{5}$$

To artificially increase the size of the dataset we perform random augmentation procedures to each image on the fly; including, flips, rotations, shifts and sheers. However It is well known that a



No.	Shape ($w \times h \times c$)
1	$900 \times 600 \times 3$
2	$450 \times 300 \times 48$
3	$225 \times 150 \times 80$
4	$113 \times 75 \times 112$
5	$57 \times 38 \times 144$
6	$29 \times 19 \times 176$
7	$29 \times 19 \times 3$

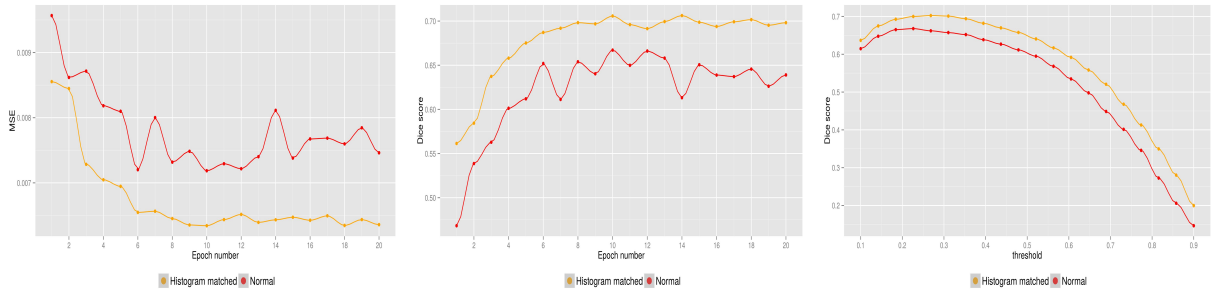
Table 1. *Shape of tensors depicted in figure 2.0.1*

Fig. 2. *Overall architecture of f_1.*

large quantity of varied data is required to train deep networks well, especially to fit complicated functions and generalise well to unseen examples. It is not plausible to learn a huge number of parameters in deep networks reliably in this scenario [17]. Our particular case we hand labelled 700 images, moreover, as can be seen in the particular case of figure 1, they are far from perfect. Large variations in brightness and sea color in X_i ensures even more data is needed to generalize well.

We combated variation in sea color and brightness by using a histogram matching [18] based technique to normalize for such variation. Before each X_i is fed into f_1 we match the cumulative distribution function (CDF) of each channel in its YUV color space to the respective YUV channel of a prespecified baseline target, denoted T. This is illustrated figure 3.

The results of which are given below in figure 4. As we can see, histogram matching has a positive effect on test performance. We believe this method works well in this situation as the input images are largely bimodal with the first, normally smaller mode, the whale, being darker that the sea on most occasions. We therefore do not have to do as much data synthesis such as random illuminations or contrast changes to reduce overfitting.



(a) (b) (c)

Fig. 4. *Test performance of f_1 for histogram matching and non matching (normal) preprocessing; a) MSE per epoch b) Dice score per epoch and c) Dice score over all thresholds after final epoch.*

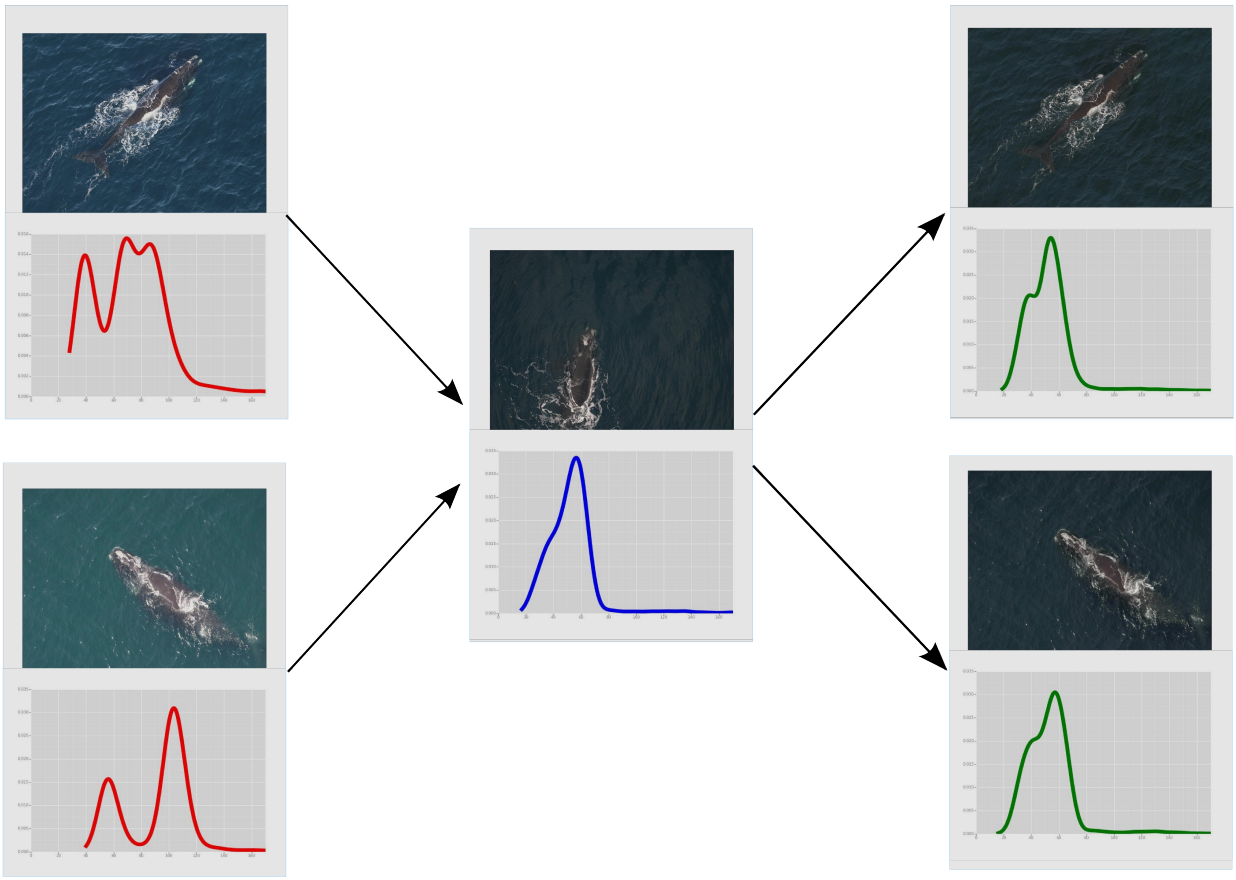


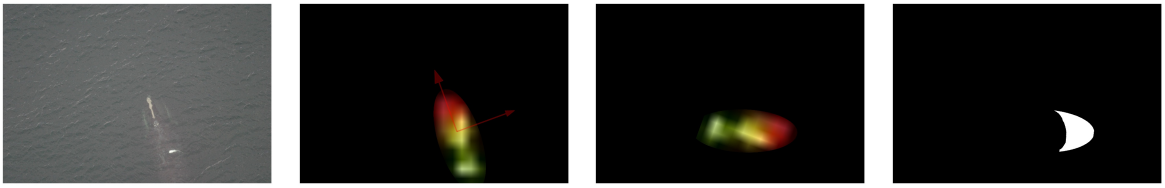
Fig. 3. *Illustration of histogram matching; two examples are histogram matched with our base target image T. Densities below the two respective input images illustrate the prematched density (red) who's CDFs are matched with the CDF of the target density (blue), resulting in the matched output images and their corresponding densities (green).*

2.0.2. Mask post-processing: We find the largest ellipse of the mask in case more than one whale is in the photo, giving us $\hat{M}_i^1$. We then upscale $\hat{M}_i^1$ to the same dimensions as X_i. To infer direction we find the centroids of the red and green channels, denoted m_r and m_g respectivley. This gives us a vector indicating the direction in which the whale is pointing. To provide robustness to direction we calculate the direction of the first principal axis [19], denoted $u_1 \in \mathbb{R}^2$. As u_1 is an eigenvector we flip its direction if the angle between u_1 and $p = m_r - m_g$ is larger than $90°$. An example of this is given in figure 5.

(a) (b) (c) (d)

Fig. 5. *Example of mask post processing stage; a) X_i, b) X_i^1, c) $\hat{M}_i$ and d) $\hat{M}_i^1$ with principal axes illustrated.*

We rotate $\hat{M}_i^1$ and X_i about m_r by $\theta = \text{atan2}(u_{12}, u_{11})$, generating $\hat{M}_i^2$ and X_i^2. As rotation is completed we want to crop around the head of the whale. The rotated mask $\hat{M}_i^2$ is then converted to the HSV color space in order to threshold the hue channel for red and value channel above another tuned threshold as shown in figure 6.



(a) (b) (c) (d)

Fig. 6. *Rotation example in alginment stage; a) X_i, b) $\hat{M}_i^1$, c) $\hat{M}_i^2$ and d) $\hat{M}_i^2$ thresholded for red.*

This gives us our head shot X_i^h as shown in figure 7.

(a) (b) (c)

Fig. 7. *Obtaining the head shot; a) X_i, b) X_i^2 and c) X_i^h.*

3. Classification

3.0.3. Results:

4. Conclusion

5. Figures & Tables

The output for figure is:

Fig. 8. *Insert figure caption here*
a Insert Sub caption here
b Insert Sub caption here

The output for table is:

Table 2 An Example of a Table

One	Two
Three	Four

6. Conclusion

The conclusion text goes here.

7. Acknowledgment

Thanks to Christin Khan and Leah Crowe from NOAA for hand labeling the images. Kaggle for competition.

8. References

[1] L. Koh and S. Wich, "Dawn of drone ecology: low-cost autonomous aerial vehicles for conservation," 2012.

[2] J. Martin, H. H. Edwards, M. A. Burgess, H. F. Percival, D. E. Fagan, B. E. Gardner, J. G. Ortega-Ortiz, P. G. Ifju, B. S. Evers, and T. J. Rambo, "Estimating distribution of hidden objects with drones: From tennis balls to manatees," *PLoS One*, vol. 7, no. 6, p. e38882, 2012.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[6] S. Hong, H. Noh, and B. Han, "Decoupled deep neural network for semi-supervised semantic segmentation," in *Advances in Neural Information Processing Systems*, 2015, pp. 1495–1503.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[9] B. Graham, "Fractional max-pooling," *arXiv preprint arXiv:1412.6071*, 2014.

[10] D. Mishkin and J. Matas, "All you need is a good init," *arXiv preprint arXiv:1511.06422*, 2015.

[11] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[13] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[14] "No free hunch. (2016). noaa right whale recognition, winners' interview: 1st place, deepsense.io. [online]," http://blog.kaggle.com/2016/01/29/noaa-right-whale-recognition-winners-interview-1st-place-deepsense-io/, [Accessed 19 Jan. 2017].

[15] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 9–48.

[16] T. Sørensen, "{A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons}," *Biol. Skr.*, vol. 5, pp. 1–34, 1948.

[17] X. Zhu, "Semi-supervised learning," in *Encyclopedia of machine learning*. Springer, 2011, pp. 892–897.

[18] R. C. Gonzalez and R. E. Woods, "Digital image processing publishing house of electronics industry," *Beijing, China*, p. 262, 2002.

[19] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE transactions on information theory*, vol. 8, no. 2, pp. 179–187, 1962.

9. Appendices

Appendices are allowed but please be aware that these are included in the overall word count.