IDENTIFYING HUMPBACK WHALE FLUKES BY SEQUENCE MATCHING OF TRAILING EDGE CURVATURE

By

Zachary Jablons

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: COMPUTER SCIENCE

Examining Committee:

_____

Dr. Charles Stewart, Thesis Adviser

_____

Dr. Barbara Cutler, Member

_____

Dr. Bülent Yener, Member

Rensselaer Polytechnic Institute
Troy, New York

April 2016
(For Graduation May 2016)

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

This project could not have been completed so quickly and thoroughly without the invaluable help from the rest of the IBEIS team. I would like to thank Jon Crall for helping to write the experimentation framework which has saved a lot of time and effort. I would also like to thank Jason Parham and Hendrik Weidemann for the great discussions and advice, as well as the handy LaTeXtemplate. Additionally, without the development and annotation efforts of Andrew Batbouta, a lot of models could not have been so successfully trained. Importantly, I would like to thank the Wildbook team (Jason Holmberg, Jon van Oast) for providing the dataset and corresponding annotations with which a lot of models were trained and experiments run. Of course, this work could never have been undertaken without the help, advice, and direction of my advisor, Charles Stewart.

ABSTRACT

Humpback whales (Megaptera novaeangliae) are some of the most iconic cetaceans inhabiting our oceans, and have historically been at risk for extinction due to human activity — primarily hunting [6]. While they are currently rated as 'Least Concern' [42] internationally, they are still at risk from various commercial operations, and tracking their population is important. Using photographic identification of the Humpback flukes is a common and reliable method for tracking these populations [4]. These flukes are often patterned and scarred in unique ways, allowing conservationists to identify individuals [25]. However, until recently, most automated identification methods still relied on signficant manual effort to describe and identify the fluke, severely limiting the amount of humpbacks that can be tracked.

This thesis lays out a method that automates the identification of Humpback flukes directly from still images thereof, using the 'trailing edge' of the fluke. Using this method, we achieve a fairly high top-1 ranking accuracy on a large dataset (consisting of about 400 identified individuals). We also show that this method significantly helps the accuracy of a pure appearance based method, Hotspotter [11], giving 93% top-1 accuracy.

To our knowledge, this is the first method that can achieve this level of accuracy on Humpback fluke identification without any manual effort at test-time.

CHAPTER 1
Introduction

1.1 Humpback Whales

Since the international ban on commercial hunting of Humpback whales in 1966, Humpback whales have grown from a population of only 5000 [2] to over 50000 [5]. As the population grows, it becomes more and more important to be able to automatically identify individual Humpback whales in order to accurately monitor their population growth and follow their migration patterns, among other ecological conservation endeavours. One of the most reliable methods for photo-identifying Humpback whales is by taking pictures of their flukes as they breach the water.

1.2 Current Identification Methods

Photo-identification of Humpback whale flukes thas been attempted since the early 90s [33] using computational aid. While early efforts mostly relied on a manual description of the fluke that would then be matched, later efforts have involved matching flukes based on automated analysis of both the patterns on the ventral side of the fluke. It is worth noting however that as shown in [4], the trailing edge changes less with age than surface patterns on the fluke, which means that it can (potentially) be a more reliable identifier over time. That said, trailing edges are hard to photograph well, requiring high resolution imagery and a consistent angle between fluke and photographer.

Each of these annotation methods can be separated into three categories:

- Manually annotated – a human must manually annotate or catalogue features in the fluke
- Semi-automated – a human must guide an algorithm (e.g. by setting control points or highlighting interesting regions) that then automatically identifies

the individual

- Fully-automated – the algorithm can identify individuals from raw images

1.2.1 Based on Trailing Edge

In [33], information about the trailing edge and fluke patterns are manually catalogued and used to match individual whales. This falls under a manual annotation approach. In the I3S contour system [17], the user must input start and end points on the contour, which are then extracted and checked manually, giving a semi-automated system. At the time of writing no published results on this system applied to Humpback whales could be found. Automatically identifying Humpback whales by their entire trailing edge contour is done experimentally in [22], using a technique that is originally designed for Great White Sharks. However the results published on that technique are for a much smaller dataset than the one worked on in this thesis. While trailing edge matching has seen limited use in Humpback whale identification, it is a much more common technique in Sperm whale (P. macrocephalus) identification [21], [3] [51], with varying levels of manual effort.

1.2.2 Based on general Fluke appearance

The primary method for identifying Humpback whale flukes is to use the ventral fluke pattern, as seen in [33], [8], [4], [15], and (in an semi-automated fashion) [27].

1.2.2.1 Hotspotter

Hotspotter [11] is an automated photo-identification algorithm based on SIFT features that has been used in identifying Grevy's Zebras, Plains Zebras, Giraffes, and Elephants [38]. This work is the first to our knowledge of Hotspotter being applied to Humpback whale flukes, and the results are presented in chapter 4.

It should be noted that we used a segmenting convolutional network to aid Hotspotter's predictions, although this is not detailed in this work.

1.3 The Dataset

The main dataset that is used and evaluated in this work is a subset of the dataset collected by the SPLASH project [7]. It consists of about 1400 identified photographs spread over about 860 identified individuals. Of these, only 433 individuals have more than one image associated with them, giving 942 images that can be used in a one-to-one comparison with no distractors.

We also use an external dataset of unidentified Humpback flukes for training some of the models.

1.4 Method Overview

The method for trailing edge identification put forth in this thesis is very nearly fully automated, requiring no human annotation when used (although manual annotation is necessary for training the machine learning models used). On its own, it achieves decent results on a (relatively) large dataset, comparable with the fully automated method used in [22]. Ultimately we find that this method is best used in combination with an automated pattern matching method (e.g. Hotspotter) to provide high accuracy matches. We also explore alternative methods based on more recent advances in deep learning for identification, however it appears that the dataset is too small to properly train these methods.

CHAPTER 2
Background

In this chapter, we provide a series of sections detailing background information on the algorithms from which this method was developed.

2.1 Convolutional Networks

In recent years, convolutional neural networks have provided state of the art results in several challenging computer vision tasks, including general image classification [28], [50], image segmentation [32], [9] and individual identification (specifically for human faces) [13], [45].

The essential idea of a convolutional network is that it we can use the gradient of an error signal to learn hierarchies of convolution kernels separated by nonlinear activation functions. These networks are considered to be a form of neural network where the convolutions provide a meaningful prior to when applied to data with spatial relationships (e.g. image data). Convolutional networks have been around for a long time, but the current incarnation of these networks can be traced back to [30]. More recently however, convolutional networks have grown deeper and are often modeled after the architectural decisions made in [47], [46], and [28]. These decisions are specifically the use of Rectified Linear Units (ReLUs) as activation functions, Dropout [19] after fully connected layers for regularization, and small square kernels with "same" padding alternated with $2\times$ downsampling layers (specifically max pooling layers).

The convolutional networks used in this thesis use the above architectural decisions, and also use batch normalization [23] at every layer.

2.1.1 Facial Keypoint Prediction

Facial keypoints are used in a lot of identification pipelines, as well as in motion capture and expression recognition. There has been recent work in using convolutional networks for facial keypoint prediction [49], [37].

In this work, fluke keypoints are predicted using essentially the same paradigm as the above works but with different underlying convolutional architectures and a slightly different loss function. The essential idea is that the convolutional network predicts points (rather than classifications) in the form of (x, y) coordinates, and is treated as a regression network. This will be explained in more detail in the next chapter.

2.1.2 Fully Convolutional Networks

Classically, convolutional networks reduce a an image to a single (spatially invariant) vector, which is then used for classification (or embedding, regression, etc.) Additionally, these classification networks usually have fully connected (or dense) layers towards the end, which forces the size of the network input (i.e. the image) to be fixed. This ensures that the receptive field of the network (after the convolutional layers) covers the entire image, which is necessary for a lot of applications. However, when dealing with arbitrarily sized images, it is typical to use networks that are 'fully convolutional', in which case the entire network consists of convolution kernels. Convolutional networks that reduce to dense layers can be cast as fully convolutional networks by replacing the dense layers with 1×1 kernels, using the dense units as channels. This technique can also be used for segmentation, as we can simply replace the dense part of the network with convolutional parts that can make class predictions on every pixel of their input. By upsampling and combining different stages of prediction, the authors in [32] produce high quality image segmentations, a technique that we replicate.

In this work, fully convolutional networks are used for predicting the 'trailing edginess' of an image, which allows us to refine the trailing edge contour extraction. This is explained in greater detail in the next chapter.

2.1.3 Embedding Networks

One major difference between the way that individual identification is done with convolutional networks and the standard classification architecture technique is the way that the error to the network is represented. When convolutional networks are used for classification tasks, the standard approach is to use a softmax output

layer and learn with the cross-entropy loss. While an individual identification task could be expressed as a classification task, it becomes a major issue when a new individual is added to the dataset (which for our task should be very common). A better notion for the loss function is to make it teach the network to 'embed' images into some d-dimensional vector (usually constrained to be unit norm). This is generally done by using a loss function that encourages grouping images of the same individual closer than those of different individuals. The most common approach to this is to use the contrastive loss [13] [10], although more recently the triplet loss has risen in popularity [45] [39].

These approaches will be explored in more detail in the next chapter.

2.2 Seam Carving

Seam carving is a technique that tries to resize images without warping or distorting the objects shown in the image [1]. This technique uses a dynamic programming algorithm to find minimal salience paths through an image, where salience is often defined as the gradient. The motivation for this is that these minimal saliency paths are not important to the image, so they can be removed to reduce its size.

While this method is not directly used in this work, the underlying algorithm for trailing edge extraction is essentially a single iteration of the seam carving algorithm, using gradient information.

2.3 Curvature Measures

Contour curvature measures are commonly used to characterize the overall shape of an contour. A lot of work has been done on using curvature information for detection [34], classification [14], and species identification [29]. This curvature information can be broadly broken down into either integral or differential curvature, and is usually computed at multiple scales.

2.3.1 Differential Curvature

Differential curvature can generally be seen as measuring the measures the angle of the tangent normal of the gradient at each point in an image [14]. For

our purposes, we can then take only those points that lie on the contour and use their curvature. While doing this directly can be fast to compute, it tends to be noise sensitive and we found that integral curvature (below) works better for our purposes.

2.3.2 Integral Curvature

Integral curvature works (conceptually) by sliding a circle of some radius r along the contour [40], and measuring how much of the circle is 'inside' the contour. This measurement is usually taken at multiple scales, and has the appealing property of being invariant to rotation and translation (of the entire contour). In this work, we approximate the circular curvature with a square of size r, which appears to perform just as well but can be computed much faster.

This is further explained in the next chapter.

2.4 Dynamic Time Warping

In deciding a sequence comparator, one criterion that is often important is ensuring that small shifts in the sequence do not balloon into large differences. Dynamic Time Warping (DTW) is a sequence comparison method that, roughly, finds the optimal matching between all sets of points in the two given sequences that minimizes the overall distance (for some defined distance function) between the matched points, while keeping the locality of the points intact. This allows for shifts and some warps in the two sequences to be compensated for, and results in a nonlinear mapping of one sequence onto another. The algorithmic complexity of dynamic time warping can be limiting in large datasets, as it is quadratic in both space and time – making a one-to-one comparison a bit daunting.

There are several variants on DTW that give faster speeds [44] [31], however we only use the Sakoe-Chiba bound [43], which both constrains the neighborhood in which points can be matched and gives a complexity of $O(nw)$, where w is a user set parameter.

Sequences of curvature measures have been used with DTW for signature verification [36], however this combination has not been used for matching trailing

edges to our knowledge.

CHAPTER 3
Methods

In this chapter, we detail the finalized algorithm pipeline, as well as some alternative approaches that we found to have limited successs.

3.1 Trailing Edge Extraction

Extracting good, high quality trailing edges images is one of the primary challenges when matching Humpback whales by their trailing edge. In this section, we detail the steps that go into automating the extraction of high quality trailing edges, while trying to minimize human intervention.

One major assumption that we make when extracting these trailing edges is that the Humpback whale fluke is aligned such that its major axis is horizontal. Additionally, we generally assume that all parts of the fluke are present, however theoretically it would be possible to match a partial fluke – although unlikely.

While these assumptions do not make for an incredibly robust system, the nature of the problem (and the dataset that we had at hand) makes these assumptions reasonable.

3.1.1 Basic Contour Extraction Algorithm

The base algorithm that is used for extracting the trailing edge uses the vertical gradient information of the image (denoted as I_y). We extract I_y using a vertically oriented 5×5 Sobel kernel[48].

We then normalize I_y with min-max scaling, giving N_y as

$$N_y = \frac{I_y - min(I_y)}{max(I_y) - min(I_y)} \tag{3.1}$$

With N_y, we then need a starting point and ending point for the algorithm, denoted s and e respectively. For our purposes, we use the left and right tips of the fluke as our start and end. We explain how these points are determined in a later section. We then take the columns corresponding to these points and set every pixel

in $N_y(s_x, \cdot)$ (i.e. the column s_x) and $N_y(e_x, \cdot)$ to ∞, and then set those points in N_y to 0. This forces the path to start and end at these points (as we are finding the minimal path between them).

The minimal path is then found by scanning the columns of N_y from left to right, starting and ending at s and e respectively. For each pixel in a column we set its cost with the update rule

$$C(x, y) = min_{y_c=(y-n)}^{y+n}(C(x-1, y_c) + N_y(x, y)) \tag{3.2}$$

Where n is a neighborhood constraint which we default to 2, meaning that each pixel considers 5 'neighbors' in the previous column.

If y_c is out of bounds, we let $C(x-1, y_c) = \infty$, and if $x - 1 < 0$, we let $C(x-1, \cdot) = 0$.

As C is filled out, we also keep a backtrace matrix B, which keeps track of the index of the minimal candidate chosen in equation 3.2. Once the end column is reached, we work backwards from e to construct the path, adding the coordinate corresponding to $B(x, y)$ at each step.

We can also extend this algorithm by adding extra 'control points' which the path is forced through, using the same methodology as forcing the path through the start and end points. Commonly, we use the bottom of the notch as a control point, although this affects accuracy negatively if it is inaccurate.

While this algorithm has no understanding of Humpback whale flukes, a lot of images that are constrained around the Fluke with oceanic backgrounds (which is a large majority of the dataset at hand) provide high quality trailing edges when put through this algorithm. However, this is (obviously) not a robust algorithm for finding trailing edges, something that we will attempt to fix later on.

3.1.2 Fluke keypoint prediction

One major issue with automating the above trailing edge extraction algorithm is that it requires manual annotation in the selection of the starting and ending points (i.e., the fluke keypoints), as well as any control points (specifically the bottom of the notch). To work around this, we propose a convolutional network that predicts

these tip points as part of the identification pipeline.

The convolutional network does not need full-sized images, so the first step of the keypoint extraction pipeline is to resize the image to 256×256 pixels. This size choice is somewhat arbitrary, but we find that it provides the best performance without using an unnecessary amount of memory. The network then predicts three points (left tip, right tip, and the bottom of the notch) in the range $[0-1]$ for both x and y. This is then rescaled back up to the original image size.

3.1.2.1 Network Design

The overall design of the network follows the pattern of alternating small (3×3) convolutional filters with 2×2 max pooling layers, at each step doubling the number of channels (starting with 8 channels). This is somewhat similar to VGG-16, although with half the trainable layers. After a $32\times$ downsample has been achieved, we attach a decision layer which consists of a dense layer followed by three separate dense layers with separate predictions layers after (one for each point being predicted). While this is not a common approach in keypoint prediction, we found that it gave better performance than having the points predicted as a single vector.

We theorize that this may be because shared units between each of the three predictions leads to stronger correlations between them, reducing overall prediction flexibility.

3.1.2.2 Training Details

Generating the training data for this is straightforward given a set of annotations with the associated points to learn. The dataset that we created for this purpose contains approximately 2700 training images, 900 validation images, and 1200 test images. Despite the small size of the training set, we found that it generalizes pretty well.

First, each image is resized to a fixed width while maintaining the aspect ratio. This is done to somewhat normalize the relative scale of objects in each image on the assumption that they are constrained to contain the fluke. Each image is rescaled to the network size, and then the corresponding targets are rescaled to the range $[0-1]$. The size of the original image is recorded as well. While it would be possible

to treat this as a simple multi-variate regression and use RMSE loss, we achieved better results by averaging the Euclidean distance between the predicted points and true points. We also include a scalar scaling factor α, which scales each point by a proportion of the original image size. Thus, we have the scaled Euclidean loss SE

$$SE(\vec{t}, \vec{p}, \vec{s}, \alpha) = \|(\alpha * \vec{s}) \odot (\vec{t} - \vec{p})\| \tag{3.3}$$

Where

- $\vec{t}$ and $\vec{p}$ are the (x, y) ground truth and predicted values respectively
- $\vec{s}$ is the original image width and height

The networks are trained for 100 epochs with α set to 2e$-$2, the Adam optimizer[26] (with the recommended settings) and $l2$ regularization on the trainable parameters with a decay of 1e$-$4. All of these hyper parameters were tuned using the validation set, although the possible parameter space was not fully explored due to time constraints.

3.1.2.3 Evaluation

On average, the best network achieved a 10 pixel distance error on the validation and testing sets (in the original image scale). While this may seem like a lot, the trailing edge extraction (and subsequent matching performance) was not severely affected when only using the start and end point predictions.

We find that, for the vast majority of images, the network achieves a low pixel distance error, while there are a few that have a much higher error. Qualitative inspection of these images shows that they are of flukes which are not the singular or major object in the image, nor horizontally oriented.

We attempted to use a spatial transformer network[24] to try and handle these cases, but we were unable to get it to perform as well as the standard convolutional network, nor produce sensible transformations.

3.1.3 Trailing Edge Scoring

As mentioned in the beginning of this section, using only the gradient information for extracting the trailing edge works in a lot of cases, but is not a robust method.

If we had a score of each pixel's "trailing edginess" in an image, the trailing edge extractor could make use of this information to make better choices in trailing edge extraction. To do this, we need a prediction of whether or not each pixel belongs to the trailing edge of a fluke, a task that is best suited to a fully convolutional network. In these networks, all convolutions (aside from max pooling layers) are "same" convolutions, which have square, odd filters (usually 3×3) and 1-padding. These "same" convolutions produce a spatial output shape that is the same as the input shape, obviating the need for any interpolation.

The four major variants on trailing edge scoring networks that we evaluated are detailed below. All of these networks function on the same paradigm of taking an arbitrarily sized image and producing an image of the same size but with a class score for each pixel.

The dataset was sourced from existing trailing edges extracted using the gradient based algorithm, however with manual adjustments to fix many of the more common issues we encountered with this algorithm.

To generate the training set, we then extract a 128×128 patch at 128 pixel intervals along the trailing edge (a positive patch), and a corresponding negative patch (with no trailing edge pixels) randomly sampled from the left over space above and below the trailing edge. These patches are then randomly split into training, testing, and validation sets each with 3700, 1200, and 1600 patches respectively.

3.1.3.1 Trailing Edge Scoring Architectures

There are several network architectures that were tried, however here we will report only the major variants. One major consideration that has to be made when selecting a fully convolutional network architecture is its receptive field. If the receptive field is too small, it may not have enough information to accurately determine if a given pixel is part of the trailing edge. However, in order to increase

the receptive field without massively increasing the depth of a network, we must

Simple This network is simply a stack of 6 "same" convolutional layers (of decreasing spatial extent), with no downsampling regions. This has a small receptive field, but can produce detailed predictions. Due to the small receptive field however, at convergence it gives low precision predictions, and seems to (in many cases) produce a lot of the same mistakes that normal gradient based trailing edges do.

Upsample To deal with the small receptive fields, convolutional networks typically downsample at various stages. This downsampling usually takes the form of max pooling (instead of e.g. convolutional kernels with a stride greater than one). The upsample architecture is analogous to the FCN-32s architecture in [32], although we do not go down to a 1x1 spatial extent before upsampling. This produces blocky and nearly unusable trailing edge scores, however they tend to be more connected than the disparate (but fine-grained) predictions made by the simple network.

Jet Following the deep-jet architecture from [32], we combine softmax predictions from various stages throughout the network with predictions from "farther down" the network, cascading until we hit predictions from the last convolutional layer before any downsampling. This gives more fine-grained predictions than the upsample network while giving somewhat better predictions than the simple network.

Residual While the receptive field of the Simple network is small, it can produce very fine trailing edges, which we found to be necessary for good trailing edge extraction. However with such a small receptive field it can be more prone to making mistakes. It is possible to create a very deep network of 'same' convolutions, however training very deep networks like this can run into problems very quickly. Recently, there has been work on making very deep networks feasible by adding residual connections[18], showing that these very deep networks can achieve high accuracy in a challenging benchmark. We replicate this architecture by stacking 64 3×3 'same' convolution kernels, and adding a residual connection every other layer. This performs better than the Simple network, however we do not see a marked

improvement in trailing edge matching as a result.

3.1.3.2 Using the trailing edge scores

Once we have 'trailing edginess' score for each pixel, we need to combine this information with the gradient in a way that causes the trailing edge extraction algorithm to follow those pixels that the scoring map marks as trailing edge. More formally, the trailing edge scoring map gives us an image $T_p \in [0-1]^{w \times h}$ which denotes the network's predicted probability of each pixel being part of the trailing edge. The most simple and obvious way to combine this information with N_y from before is to combine them with some weight β.

$$S_{te} = (1-\beta) * N_y + \beta * (1 - T_p) \tag{3.4}$$

We use $1 - T_p$ in this case because we are minimizing the path through S_{te}. Once done, the trailing edge extraction algorithm procedes as described in the previous section.

For most of the evaluation process, we simply set $\beta = 0.5$.

Another variant on combining T_p and N_y that we tried was to dilate the trailing edge predictions and then forbid the trailing edge from going outside of those predictions. However one of the main issues with this is that if there are any breaks in the prediction (which can happen even with the Upsample architecture), the trailing edge cannot be extracted, leading to broken trailing edges.

3.1.3.3 Training Details

All networks were trained for 100 epochs (or until convergence) with a batch size of 32, with $l2$ regularization using a decay of 1e$-$4. We used the Adam optimizer[26] (with the recommended settings) for calculating weight updates.

One detail that turned out to be important is the class imbalance. The trailing edge pixels (necessarily) make up a small percentage of the total image, meaning that these networks could get a fairly high accuracy (and thus low loss) simply by predicting only background pixels. In order to prevent this, we only sample a negative patch once for every positive patch (as detailed above), and additionally

we weight the loss for the trailing edge pixels 10× higher than the loss for the background pixels.

Due to the nature of the training data, we were concerned that the networks would simply learn to replicate the function that generated the trailing edges, despite human corrections. In order to mitigate this issue, we included some non-spatial data augmentation, namely random Gaussian blur and randomly inverting the pixel intensities. Unfortunately this had limited success.

All of these hyper parameters were tuned using the validation set, although the possible parameter space was not fully explored due to time constraints.

3.1.3.4 Evaluation

Due to the overwhelming class imbalance, the model accuracy is rather meaningless (e.g., the accuracy of an all-background prediction is 99%). Instead, we report the intersection-over-union (IoU) score, which gives a much better idea of model performance. We also report the precision and recall of the model.

	Training			Validation			Testing		
Architecture	Pr.	Re.	IoU	Pr.	Re.	IoU	Pr.	Re.	IoU
Simple	0.59	0.95	0.57	0.59	0.94	0.57	0.60	0.95	0.59
Upsample	0.17	0.88	0.17	0.17	0.86	0.17	0.17	0.87	0.17
Jet	0.62	0.89	0.57	0.62	0.88	0.57	0.63	0.89	0.58
Residual	0.57	0.93	0.54	0.57	0.92	0.54	0.58	0.93	0.56

Table 3.1: Table showing the precision, recall, and IoU of each of the evaluated trailing edge scorers on each section of the trailing edge dataset. For the purposes of this analysis, we use the `argmax` over the classes to determine a positive (i.e. trailing edge) or negative pixel.

Initially, we thought that precision would be the most important metric for evaluating networks (which is reflected in the class weighting). However, we found that even a low precision classifier could give good trailing edges, and that the important measure was how detailed these trailing edges were.

This is likely because small 'blips' of trailing edge would not be chosen by the trailing edge extractor.

3.2 Trailing Edge Matching

Given extracted trailing edges, we define a method for doing a one-to-one comparison between a given query and database trailing edge. The simplest way to do this is to define a (potentially non-metric) distance function between any two trailing edges. Once this distance function is defined, we can identify an individual by its trailing edge (referred to as the query trailing edge) by looking at the identity of the closest trailing edge in the database. As a distance function we propos dynamic time warping over the block curvature measurements, using a weighted Euclidean distance as a local distance function between curvatures.

3.2.1 Curvature Measurement

In order to do this, we must first extract the curvature from the trailing edge. Given the trailing edge as a sequence of coordinates into the original image, we construct a zero-image I_0 of shape similar to the original image. Each pixel corresponding to and below the trailing edge in I_0 is set to 1. Because the trailing edge extractor produces only one coordinate per column, we can do this safely, however this algorithm could be easily adapted to this not being the case. Once this is done, we calculate a summed area table [12] ST from I_0 as follows.

$$ST(x, y) = \sum_{i=0}^{i=y} \sum_{j=0}^{j=x} I_0 \tag{3.5}$$

Conceptually, the next step is to slide a square of shape $s \times s$ centered on each point, and measure the percentage of that square that is within the filled in trailing edge. These values s are the different scales at which we meassure curvature, and are computed as a percentage of the trailing edge width.

To do this, we compute $BC_s(x, y)$ for each (x, y) coordinate in the trailing edge.

$$b(i) = i - \frac{s}{2} \tag{3.6}$$

$$e(i) = i + \frac{s}{2} \tag{3.7}$$

$$BC_s(x,y) = \frac{(ST(b(x),b(y)) + ST(e(x),e(y))) - (ST(b(x),e(y)) + ST(e(x),b(y)))}{s^2} \tag{3.8}$$

The numerator in 3.8 gives the total area within the square that is below the trailing edge, which we then normalize by dividing by the square's area.

The set of scales to choose presents a large parameter space, however we have found that the scales $S = [1\%, 2\%, 3\%, 4\%]$ work well for our purposes. We then treat this curvature measurement as a $\|S\| \times w$ matrix BC, where w is the length of the trailing edge.

3.2.2 Sequence Matching

Given two sequences of curvatures BC_1 and BC_2 (referred to as query and database curvature respectively) we match them using dynamic time warping as follows. First, we create a cost matrix C of size $w_1 \times w_2$. We initialize this cost matrix by setting the first column and row to ∞, and then $C(0,0) = 0$, intuitively forcing the optimal path to match align the beginning of BC_1 with BC_2. Then, for each cell (i,j) in the cost matrix starting with $C(1,1)$, we use the following update rule

$$D_{s_w}(c_1, c_2) = \|s_w \odot (\vec{c_1} - \vec{c_2})\|_2 \tag{3.9}$$

$$C(i,j) = D_{s_w}(BC_1(i,\cdot), BC_2(j,\cdot)) + min(C(i-1,j), C(i,j-1), C(i-1,j-1)) \tag{3.10}$$

Where s_w is a vector giving the weight for each curvature scale, and $\vec{c}$ is a vector of the curvatures at different scales for a point.

Additionally, we impose the Sakoe-Chiba [43] locality constraint T so that

for each element i in BC_1, we only consider the range over elements j in BC_2 of $j \in [min(i - T, 0), max(i + T, w_2)]$.

We set T as a percentage of w_1, which is the "query" trailing edge. For most of these experiments T is set to 10%, which appears to minimize the time taken for each comparison while preserving the overall accuracy of the algorithm.

It's worth noting that while this distance measure is not a metric distance (i.e. it doesn't satisfy the triangle inequality), it is a symmetric distance as $D_{s_w}(\cdot, \cdot)$ is symmetric[35].

3.3 Alternative Approaches

In this section, we list and briefly describe alternative approaches that were tried, although they did not prove accurate enough to make it into the final system.

3.3.1 Aligning Trailing Edges

One obvious pre-processing step that would make sense when comparing trailing edges is to make sure that they are aligned in image space. However, we found that doing so when comparing curvature was often unnecessary (due to the invariances to rotation and translation, scale was taken care of separately), and using the Euclidean distance between points on the aligned trailing edges (i.e. as $D(\cdot, \cdot)$) did not give good results.

There were two approaches to alignment that we evaluated, although neither achieved top-1 accuracies above 20%.

3.3.1.1 Keypoint Alignment

As noted in the section on fluke keypoints, there are three points to predict – left, notch and right. Originally the intention for recording (and predicting) all three, as opposed to just left and right, was to have three corresponding points with which to estimate an affine transformation from database image onto query image. This would be done prior to any computation of trailing edge or curvature.

One major issue with aligning these iamges however is that if a non-affine transformation was required, the trailing edge itself would be warped in such a way

that made matching difficult.

3.3.1.2 Dynamic Time Warping Alignment

Taking after the AI-DTW approach laid out in [41], we ran an iterative alignment process that used the correspondences found by DTW (using either curvature distance or Euclidean distance as criteria). Essentially this method would find alignments using DTW, and then use these alignments to estimate an affine transformation of the database image onto the query image – and then repeat until convergence.

However, we found that this process oftentimes wouldn't converge, and when it did the alignments provided were of worse quality than those found by aligning the three fluke keypoints. Additionally, the extra time taken to carry out this process was impractical.

3.3.2 Histogram Matching

One early curvature comparison method that we evaluated was to use histograms to match instead of a sequence-based method. This is a common approach for comparing curvatures [29], however we found that for our purposes even high resolution histograms did not provide enough detail to match the trailing edges properly, although this could potentially be explored further.

3.3.3 Embedding via Convolutional Networks

We also made an attempt at training convolutional networks to directly embed the images of flukes into a n dimensional vector, much like [45] and [39]. However, most of the previous literature on this technique is applied to larger datasets such as LFW[20], which is significantly larger than the dataset that we had available. A major factor in this is that these larger datasets often have five to ten images per identity (if not more), whereas most of the identities in our dataset had one or two images associated.

Regardless, we attempted the embedding approach (from raw images), however even a severely overfit convolutional network only achieved half the top-1 accuracy on its training set that the main method is able to achieve. We tried both triplet

loss (a modified version of the one detailed in [45]) and contrastive loss [16] to no avail.

We believe that the small amount of images per identity is the main factor for the failure of these methods, and that a larger dataset would be necessary to properly train them.

CHAPTER 4
Results

In this chapter we present the results that this method achieves on the Flukebook dataset. The main results for the optimal method are given briefly, and then we discuss how different variations on the method affect accuracy.

4.1 Main method

4.1.1 Characterization of Success cases

4.1.2 Characterization of Failure cases

4.2 Variations

4.2.1 Keypoint Extraction

4.2.1.1 Keypoint Extractor

In this section, we detail the training results of the fluke keypoint extractor, and note issues with it. We also show the variation in accuracy between keypoint extractors run on different sized inputs.

4.2.1.2 Manual versus Automatic

One interesting note is to see how the final keypoint extractor compares with the manual annotations provided for the dataset. It is worth noting that while some of these annotations made it into the keypoint extractor's training set, the vast majority of its training set comes from an external dataset. Additionally, we can see from its training results that the keypoint extractor generalizes well regardless.

4.2.1.3 Varying training image sizes

Intuitively, the bigger the image the better the prediction can be, but at the expense of requiring more parameters to handle the input. Resizing all images to 256×256 seems to strike a good balance.

4.2.1.4 STN

We also briefly experimented with an Spatial Transformer Network. This was largely motivated by the tendency of the keypoint extractor to do a terrible job of predicting fluke keypoints on flukes that did not 'fill' the image horizontally. Unfortunately, we could not get the STN to converge at a better accuracy than the standard keypoint extractor, even if we held its parameters fixed for a few training epochs. Usually, the STN would produce nonsensical transformations of the image.

4.2.2 Image Preprocessing

4.2.2.1 Cropping and Image Width

Cropping One easy way to normalize the trailing edges somewhat is to make sure that they are all the same length. While, with dynamic time warping, we theoretically can match sequences of similar or different lengths, the distances are distorted by large differences in actual trailing edge length. Since we are only interested in the width of an image (assuming that the trailing edge is roughly horizontal in the image), we can get every trailing edge to have exactly some fixed length w by the following process

- Crop the image horizontally between the left and right columns found by the keypoint extraction process (or manually determined).
- Resize the cropped image to some fixed width w, while preserving the aspect ratio.

In this way, we standardize the trailing edge length so that image scale does not affect detection accuracy too much.

One major caveat with this process is of course that the keypoint extraction can fail if done automatically, however in practice we found that the keypoint extraction network's predictions were usually good enough to work with the above process.

Trailing Edge Length Determining what w should be is not super obvious. One heuristic is to look at the minimal post-crop width, since increasing w past that would introduce interpolation artifacts that could negatively affect matching when

the image needs to be scaled up. Ideally, one would simply look at the mode / average post-crop width, and try to keep the fixed width around there. Regardless, we found that a post-crop width of 500 pixels gave us the best results.

4.2.2.2 Histogram Equalization

When normalizing the gradient (i.e. going from I_y to N_y), it's generally important to make sure that the overall contrast of the image is normalized. This is easy enough to do with histogram equalization. However, we found that using histogram equalization resulted in a significant drop in matching accuracy, due to a drop in trailing edge quality. This happened even with trailing edge scoring networks that are trained on histogram equalized images.

4.2.3 Trailing Edge Extraction

There are a lot of variants on trailing edge scoring networks that were attempted, often with mixed results. One major result that we found was that, when using the averaging method to combine the trailing edge scores with N_y, having a robust trailing edge prediction wasn't as important as having a detailed trailing edge.

4.2.3.1 Trailing Edge Scorer

In this section, we detail the training results for the final chosen trailing edge scorer.

4.2.3.2 Trailing Edge Scorer variations

The variations are detailed in the previous chapter, and here we go into the accuracies and trailing edges that we extracted from each. Overall, we were unable to get a significant improvement from any trailing edge scorer besides the 'simple' all convolutional scorer. We theorize that this is because it could correct some mistakes made by the gradient only method (although it made many of the same mistakes itself) while providing detailed trailing edges. More importantly, it appears that blocky trailing edge scores (such as those produced by any of the networks that are

required to upsample their predictions) result in blocky trailing edges. Understandably, the blockier trailing edges produce terrible accuracies when matching, as many of them are largely the same.

4.2.3.3 Using the notch

One minor issue to note is whether or not we use the notch when extracting the trailing edge. As explained in the previous chapter, we can use the notch as an additional control point, forcing the trailing edge to go through it. However, it appears that if the notch is inaccurate, this can lead to significant reductions in accuracy compared to simply ignoring it. A bad notch annotation is common with network extracted fluke keypoints, and thus in this case it makes sense to ignore it. However, manual annotations benefit (slightly) from using the notch as a control point.

4.2.3.4 Combining N_y and T_y

While the main method for combining the scores of the trailing edges and the normalized image gradient that is used throughout this work is the 'average' method, there were a few other variants that were briefly attempted.

Additionally, the β term does have an effect on accuracy for different networks, which we show here.

4.2.3.5 Number of neighbors in the extraction

The number of neighbors n effectively limits the slope of the trailing edge. We limit it to an odd number for convenience. On the one hand, a lower n can cause the trailing edge to be limited in vertical breadth, but does prevent it from going way off course. Despite this, with trailing edge scoring in place, it might be beneficial to increase n so as to avoid parts of the trailing edge that continually 'max out' the number of neighbors, i.e. parts of the trailing edge that require a higher slope than would be allowed.

4.2.4 Curvature Extraction

Curvature extraction is one of the least parameterized part of the process, however figuring out what the optimal scales to measure is somewhat unintuitive.

4.2.4.1 Different scales

We can look at each scale as measuring the curvature of some percentage of the trailing edge around the given point. While we default to measuring 1 through 4 percent of the trailing edge, we can see that increasing this further yields diminishing returns.

4.2.5 Dynamic Time Warp Matching

We experimented with several different variants on the function $D(\cdot, \cdot)$ that makes the core decision function of DTW, and settled on using curvature. Originally we attempted to use Euclidean distance of the (aligned) trailing edges as well, but the performance was dismal, and so we do not present those results here.

4.2.5.1 Weighting the different scales

We use the curvature distance CD_{cw} as outlined in equation 3.11. While we default to weighting each curvature scale equally (i.e. $cw = [1 \forall s \in scales]$), intuitively different scales might be more important for certain types of trailing edges.

We explroe a few heuristics for generating cw in this section, however a full exploration of the parameter space was not viable given time constraints.

4.2.5.2 Window size

In this section, we vary the window size as a percentage of the (query) trailing edge length. We default this value to 10%, below which we see a reduction in accuracy. While theoretically the boundary can prevent mismatches by only allowing a certain amount of translation in the correspondences generated, realistically we found that increasing the boundary did not affect match accuracy. Thus, the 10% size is chosen to minimize computation time while maintaining the accuracy of the full dynamic time warp.

4.2.5.3 Aggregating over multiple trailing edges per identity

Determining the identity of a given query image given distances to other images in the database is not entirely simple when there are multiple database images for a given individual. Essentially we need to transform these distances from query image to database image into distances from query to individual. To do so we evaluate two options given a group of distances for an individual – either the average distance or the minimum distance.

We find that the minimum distance version provides significantly better accuracy than the average distance. While this is a bit surprising given that most of the individuals in the dataset only have two images associated, there are still a significant amount of individuals for whom there are more than that. That said, the minimal distance version also is in line with the LNBNN decision criterion used by Hotspotter [11], and as such as theoretical backing.

4.3 In Combination with Hotspotter

4.3.1 Failure cases

4.3.2 Characterization of when to use which method

CHAPTER 5
Discussion

5.1 Issues with current method

5.2 Future work

5.3 Conclusion

REFERENCES

[1] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 26(3), July 2007.

[2] C. Baker, A. Perry, J. Bannister, M. Weinrich, R. B. Abernethy, J. Calambokidis, J. Lien, R. Lambertsen, J. U. Ramirez, and O. Vasquez. Abundant mitochondrial dna variation and world-wide population structure in humpback whales. *Proceedings of the National Academy of Sciences*, 90(17):8239–8243, 1993.

[3] B. Beekmans, H. Whitehead, R. Huele, L. Steiner, and A. G. Steenbeek. Comparison of two computer-assisted photo-identification methods applied to sperm whales (physeter macrocephalus). *Aquatic Mammals*, 31(2):243, 2005.

[4] A. L. Blackmer, S. K. Anderson, and M. T. Weinrich. Temporal variability in features used to photo-identify humpback whales (megaptera novaeangliae). *Marine Mammal Science*, 16(2):338–354, 2000.

[5] T. Branch. Humpback whale abundance south of 60 s from three complete circumpolar sets of surveys. *Journal of Cetacean Research and Management (special issue)*, 3:53–69, 2011.

[6] J. Breiwick, E. Mitchell, and R. Reeves. Simulated population trajectories for northwest atlantic humpback whales, 1865-1980. In *Fifth Biennial Conference on the Biology of Marine Mammals, Boston, MA*, 1983.

[7] J. Calambokidis, E. A. Falcone, T. J. Quinn, A. M. Burdin, P. Clapham, J. Ford, C. Gabriele, R. LeDuc, D. Mattila, L. Rojas-Bracho, et al. *SPLASH: Structure of populations, levels of abundance and status of humpback whales in the North Pacific*. Cascadia Research, 2008.

[8] C. A. Carlson, C. A. Mayo, and H. Whitehead. Changes in the ventral fluke pattern of the humpback whale (megaptera novaeangliae), and its effect on matching; evaluation of its significance to photo-identification research. *Rep. int. Whal. Commn (special issue)*, 12:105–11, 1990.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[10] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE, 2005.

[11] J. P. Crall, C. V. Stewart, T. Y. Berger-Wolf, D. I. Rubenstein, and S. R. Sundaresan. HotSpotter - patterned species instance recognition. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 230–237. IEEE.

[12] F. C. Crow. Summed-area tables for texture mapping. *ACM SIGGRAPH computer graphics*, 18(3):207–212, 1984.

[13] H. Fan, Z. Cao, Y. Jiang, Q. Yin, and C. Doudou. Learning deep face representation. *arXiv preprint arXiv:1403.2802*, 2014.

[14] P. Fischer and T. Brox. Image descriptors based on curvature histograms. In *Pattern Recognition*, pages 239–249. Springer, 2014.

[15] N. Friday, T. D. Smith, P. T. Stevick, and J. Allen. Measurement of photographic quality and individual distinctiveness for the photographic identification of humpback whales, megaptera novaeangliae. *Marine Mammal Science*, 16(2):355–374, 2000.

[16] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 1735–1742. IEEE, 2006.

[17] d. J. Hartog and R. Reijns. I3S Contour MANUAL, 2013.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[19] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[20] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[21] R. Huele, H. U. De Haes, J. Ciano, and J. Gordon. Finding similar trailing edges in large collections of photographs of sperm whales. *Journal of Cetacean Research and Management*, 2(3):173–176, 2000.

[22] B. Hughes and T. Burghardt. Automated identification of individual great white sharks from unrestricted fin imagery. In M. W. J. Xianghua Xie and G. K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 92.1–92.14. BMVA Press, September 2015.

[23] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[24] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2008–2016, 2015.

[25] S. K. Katona and J. A. Beard. Population size, migrations and feeding aggregations of the humpback whale (megaptera novaeangliae) in the western north atlantic ocean. *Report of the International Whaling Commission (Special Issue 12)*, pages 295–306, 1990.

[26] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] E. Kniest, D. Burns, and P. Harrison. Fluke matcher: A computer-aided matching system for humpback whale (megaptera novaeangliae) flukes. *Marine Mammal Science*, 26(3):744–756, 2010.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[29] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J. V. Soares. Leafsnap: A computer vision system for automatic plant species identification. In *Computer Vision–ECCV 2012*, pages 502–516. Springer, 2012.

[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[31] D. Lemire. Faster retrieval with a two-pass dynamic-time-warping lower bound. *Pattern recognition*, 42(9):2169–2180, 2009.

[32] 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*, pages 3431–3440, 2015.

[33] S. A. Mizroch, J. A. Beard, and M. Lynde. Computer assisted photo-identification of humpback whales. *Report of the International Whaling Commission*, 12:63–70, 1990.

[34] A. Monroy, A. Eigenstetter, and B. Ommer. Beyond straight linesobject detection using curvature. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 3561–3564. IEEE, 2011.

[35] M. Müller. *Information retrieval for music and motion*, volume 2. Springer, 2007.

[36] M. E. Munich and P. Perona. Continuous dynamic time warping for translation-invariant curve alignment with applications to signature verification. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 108–115. IEEE, 1999.

[37] D. Nouri. Using convolutional neural nets to detect facial keypoints tutorial, 2014.

[38] J. R. Parham. Photographic censusing of zebra and giraffe in the nairobi national park.

[39] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *Proceedings of the British Machine Vision*, 1(3):6, 2015.

[40] H. Pottmann, J. Wallner, Q.-x. Huang, and Y.-l. Yang. Integral invariants for robust geometry processing. 2007.

[41] Y. Qiao and M. Yasuhara. Affine invariant dynamic time warping and its application to online rotated handwriting recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 2, pages 905–908. IEEE, 2006.

[42] S. Reilly, J. Bannister, P. Best, M. Brown, R. Brownell Jr, D. Butterworth, P. Clapham, J. Cooke, G. Donovan, J. Urbán, et al. Megaptera novaeangliae. *IUCN 2010. IUCN Red List of threatened species*, 2008.

[43] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, 1978.

[44] S. Salvador and P. Chan. Fastdtw: Toward accurate dynamic time. *Warping in Linear Time and Space*, 2007.

[45] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[46] 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.

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

[48] I. Sobel and G. Feldman. A 3x3 Isotropic Gradient Operator for Image Processing. Never published but presented at a talk at the Stanford Artificial Project, 1968.

[49] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013.

[50] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[51] H. Whitehead. Computer assisted individual identification of sperm whale flukes. *Reports of the International Whaling Commission*, 12:71–77, 1990.

APPENDIX