RETRACTED CHAPTER: Towards End-to-End DNN-Based Identification of Individual Manta Rays from Sparse Imagery

Tuana Celik[1(✉)], Benjamin Hughes[2,3], and Tilo Burghardt[1]

[1] University of Bristol, Bristol BS8 1UB, UK
tc13007@bristol.ac.uk, tilo@cs.bris.ac.uk

[2] Save Our Seas Foundation, Geneva, Switzerland
ben@saveourseas.com

[3] The Manta Trust, Corscombe, Dorchester DT2 0NT, UK

Abstract. This paper presents an end-to-end deep learning approach for the fine-grained identification of individual manta rays (*Manta alfredi*) based on characteristic ventral coat patterns where training is restricted to sparse photographic sets of <10 ventral images per individual. The dataset is captured by divers in underwater habitats. Its content is challenging due to non-linear deformations (of the rays), perspective pattern distortions, partial occlusions, as well as lighting and noise-related acquisition issues. We show how a combination of data augmentation, encounter fusion, and transfer learning techniques can address the sparsity and noise challenges at hand so that deep learning pipelines can operate effectively in this uncompromising data environment. We demonstrate that using the proposed approach with an adapted InceptionV3 deep neural network (DNN) architecture significantly outperforms tested baselines including the Manta Matcher approach, the so-far best performing traditional, widely used method published for the application at hand.

1 Introduction

Visual detection and subsequent identification of members of a species by recognition of characteristic coat patterns – ideally to the fine-grained granularity of an individual – is a subdiscipline of computational animal biometrics [1]. It is an effective and potentially non-invasive approach to gain knowledge about aspects of a population of interest: be that to estimate presence, abundance, dynamics, or changes in behavior or social networks over time and space [1].

In order to enable modern deep learning approaches to operate successfully in the animal biometrics domain, large datasets that represent the individuals to be identified would appear to be of paramount importance. Yet, there are significant

The correction to this chapter is available at https://doi.org/10.1007/978-3-030-13469-3_113

R. Vera-Rodriguez et al. (Eds.): CIARP 2018, LNCS 11401, pp. 645–653, 2019.
https://doi.org/10.1007/978-3-030-13469-3_75

challenges associated with acquiring high quality visuals at scale, particularly in scenarios where species are rare, move unpredictably across vast areas, or live in habitats that are difficult to monitor (e.g. remote jungle or underwater).

This paper focusses on *Manta alfredi*, a species whose members carry individually characteristic blob patterns on their highly flexible ventral body surface (see Fig. 1). These markings have been exploited in the past, both via manual and semi-automated methodologies [2] using traditional computer vision in order to derive individual animal identities based on photographic evidence.

The objective of this paper is to show that a deep learning approach can be highly effective in our particular problem scenario of individual manta ray identification given sparse ventral pattern imagery. Our approach is depicted in Fig. 2 and combines data augmentation, encounter fusion, and transfer learning techniques to address the sparsity and noise issues at hand – all with the ultimate objective of enabling recent deep learning pipelines to operate successfully in this domain.

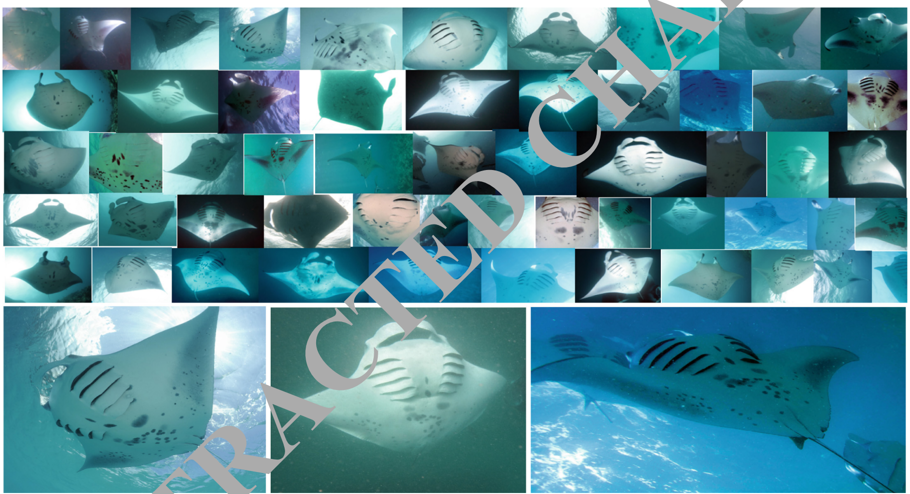

Fig. 1. Ventral Manta Ray Imagery. ***(top)*** Representative samples from the utilized 'Manta2018' data provided by The Manta Trust (see Footnote 2). Note the various non-linear deformations of animals, perspective distortions, partial occlusions, as well as lighting and noise-related challenges. ***(bottom)*** Three sample images showing the same individual under different lighting, pose, and acquisition conditions.

The remainder of the thesis is structured as follows. Section 2 briefly reviews most relevant methodologies and prior work. Section 3 describes the dataset, test architectures, experiments and recorded performance. Section 4 presents results and benchmarks them against those obtained from our re-implementation of the best performing manta identification method published to date [2]. Finally, Sect. 5 provides conclusions and closing remarks.

2 Related Work

For more than a decade now, computational animal biometrics have provided support for non-intrusive, often visual alternatives to traditional invasive tagging and marking methodologies, fueling ecological applications: camera-trapping, visual drone censuses, and colony counts via satellite provide a few commonly used examples [1]. Yet, whilst applicable across a wide range of species and semi-automated application scenarios [1–4], computerized visual identification of individuals widely relied on the use of *hand-crafted features* such as Scale-Invariant Feature Transform (*SIFT*) [5] or related extraction techniques [6].

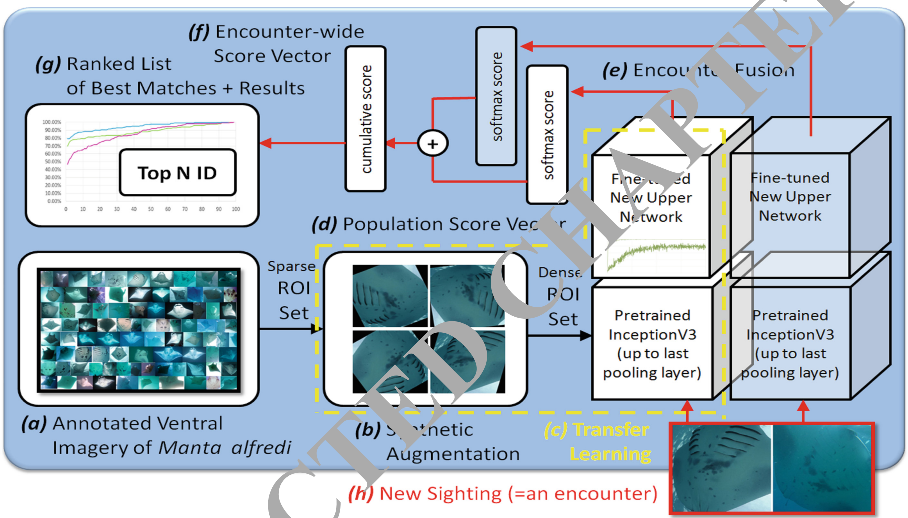

Fig. 2. Overview of Approach. ***(a)*** Field imagery with Region of Interest (ROI) and identity annotations covering a manta ray population of interest is used as system input. ***(b)*** A large pool of visual data is synthesized to enable network training based on domain-specific, geometric data augmentation. ***(c)*** Fine-tuning of a pre-trained InceptionV3-like architecture yields an inference network that can map from entire images to ***(d)*** a score vector over all individuals or, ***(e)*** for optional encounter fusion, two such vectors summed to produce ***(f)*** a score over all individuals produced for an entire encounter set. ***(g)*** A ranked list of best manta matches is then inferred for ***(h)*** new sightings in red. (Color figure online)

In particular, Town et al. in [2] describe a system to identify individual manta rays, one which semi-automatically produces a ranked list of known rays that best match a single provided query image. The system as published requires users to correct for in-plane image rotation and select a rectangular Region of Interest (ROI) aligned with the animal. After noise removal and adaptive contrast equalization, SIFT features are extracted and matched by computing all possible pairings between the feature vectors representing the query image I and every entry J in the feature database. A similarity score between I and J is then computed via all N_{F_i,F_j} matches as:

$$score(I,J) = \frac{\sum_{n=1}^{N_{F_i,F_j}} w_n}{max(|F_i|,|F_j|)} \tag{1}$$

resulting in a score between 0 and 1, where F_i and F_j are the sets of SIFT features of images I and J, respectively, and each matched feature pair is weighted via w_n based on the *significance* and the *strength* of the match as given in [2]. Finally, a similarity score between image I and Manta m is established:

$$Score(I, Manta_m) = mean(score(I, J_m)) \tag{2}$$

where J_m are labeled images that belong to Manta m. For benchmarks, we interpreted [2] to re-implement the pipeline – confirming their results (see Table 1).

Over the past decade or so, limitations of hand-crafted feature approaches have emerged due to inherently suboptimal, *manual* feature designs [5,6]. Representation learning, on the other hand, has established itself as a viable alternative: it utilizes machine learning to evolve features to those best suited to map from inputs to the target domain. Such data-driven end-to-end representation learning, applied via deep neural networks (DNNs), dominates mainstream applications for object detection, classification and identification today [7–13].

In order to apply such deep learning techniques to the task at hand, individual manta ray identification may be understood as a fine-grained classification (FGC) task [14] aiming at differentiating effectively between highly similar classes or objects. In contrast to classic FGC problems such as bird [8] or plant [9] species recognition, we are interested in an intra-species classification of conspecifics here, conceptually in line with recent work for the individual identification of great white sharks [4], gorillas [10] or chimpanzees [3,15]. However, when using deep learning the supervised training of required networks is often crucially dependent on the availability of large, representative, manually annotated training data[1]. If this is not available then an effective application of deep FGC techniques to complex identification tasks is, mainly hampered by overfitting, not straight forward despite the application of regularization, dropout etc.

Yet, large annotated datasets such as ImageNet [17] have led to the training of deep convolutional neural networks (CNNs) such as AlexNet [11], VGG [12] or Inception [13] capable of effectively disambiguating a wide range of visual classes relevant to real imagery. Assuming that visual knowledge encoded in network weights can be 'shared' between related tasks – and visual tasks are indeed related – then starting new optimizations from pre-trained weight settings is potentially beneficial for avoiding narrow generalization. We will explore the use of an InceptionV3-like architecture [13] as basis for late layer fine-tuning (see Sect. 3.2). Note that this network has a reduced footprint on the GPU (i.e. $5M$ compared to the $60M$ of AlexNet [11]) due to extensive kernel factorization.

1 Consider that in [10], for instance, 12,765 images covering 147 individuals are used for training, that is on average 86.8 images per animal. Holstein Friesian cattle identification by Andrew et al. [16] utilizes 46,430 frames describing only 23 individuals.

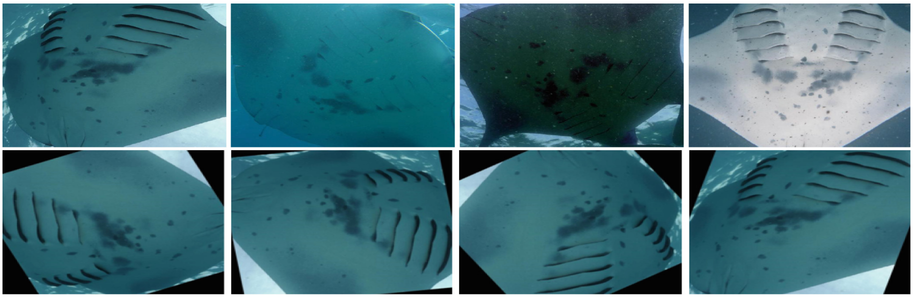

Fig. 3. ROIs and Augmentation. ***(top row)*** Four examples of ROIs of the same individual as used for training re-scaled to 512^2 or 299^2 pixels. ***(bottom row)*** Four representative examples of synthesized training images all from one source image (given at the top left). Shear and rotation produce 60 training images for each input image, overall synthesizing 47,520 training samples from 792 source images. Since ROIs are provided, scale or shift are not augmented.

3 Methodology

3.1 Dataset and Augmentation

Our initial sparse 'Manta2018' dataset of ventral *Manta alfredi* digital photographs is provided by The Manta Trust[2]. Figure 1 depicts a representative subset of the overall 990 class-labeled images with ROIs belonging to 99 individuals – covering exactly 10 images per individual. As exemplified in Fig. 3, provided ROIs contain at most one full single manta instance, potentially less. The data is captured by divers in natural, often murky and poorly lit underwater habitats. Non-linear deformations (of the rays), perspective pattern distortions, partial occlusions, as well as lighting and noise-related acquisition image degradation are prominent in the dataset. All patches given by ROIs are reshaped to fit the network inputs. Each individual's data are split into 8 patches for training and 2 (withheld) for testing. This yields 792 training and 198 testing instances.

Synthetic generation of a 60-times increased training base consists of 50 rotations of patches randomly sampled from a uniform distribution between −180 to 180°, plus a shear transform using a uniform distribution from −30 to 30°, plus 8 cases where we combine a fully random rotation and shear transforms. Together with the original, we thus produce 60 representations of the same image, resulting in each class now having 480 examples in its training pool. Overall, this yields 47,520 training patches – see Fig. 3 (bottom) for samples.

[2] **Acknowledgments:** The dataset has been provided by The Manta Trust, Catemwood House, Corscombe Dorchester, Dorset DT2 0NT, UK. The Manta Trust holds copyrights of all data. Please contact The Manta Trust directly to obtain the dataset.

3.2 Implementation

We compare and experiment with three architectures: (1) the current domain-specific state-of-the-art Manta Matcher pipeline detailed in [2], (2) a custom deep baseline network specified in Fig. 4a, and (3) our InceptionV3-like fine-tuning architecture either used as a single network as detailed Fig. 4b, or as a subnet integrated into an encounter-fusion architecture as explained in Fig. 2.

All deep models were trained on Nvidia P100 GPU nodes with batch sizes of 32 using Adaptive Moment Estimation (Adam) as optimizer over up to 240,000 training steps. Learning rates were experimentally set to 0.0001 for the custom baseline network and to 0.1 for InceptionV3 fine-tuning. We initialize all (non-pre-trained) weights over a random uniform distribution within $(-0.05, 0.05)$ where the custom baseline network is fully trained from scratch. For InceptionV3 fine-tuning, we use pre-trained weights from ImageNet up to the final pooling layer of the network (see Fig. 4b). Transferring layer weights directly, we then train a newly formed fully connected and a final softmax-loss layer with our data. Figure 5 (right) depicts a representative training run with test results in red.

Assuming a user has access to two or more samples of the same manta ray, e.g. acquired during the same dive, we also tested an encounter fusion architecture where we feed all inputs through the fine-tuned subnet in turn, as shown in Fig. 2, before summing output scores over all streams into one output vector.

type	kernel size/stride	filters	activation	input size
conv + BN	3x3	32	*Relu*	512x512x3
MaxPool	3x3/2	–	–	512x512x32
conv + BN	5x5	32	*Relu*	256x256x32
MaxPool	3x3/2	–	–	256x256x32
conv + BN	5x5	64	*Relu*	128x128x32
MaxPool	3x3/2	–	–	128x128x64
conv + BN	3x3	64	*Relu*	64x64x64
MaxPool	3x3/2	–	–	64x64x64
conv + D1 + BN	3x3	64	*Relu*	32x32x64
MaxPool	3x3/2	–	–	32x32x64
conv + BN	3x3	64	*Relu*	16x16x64
MaxPool	3x3/2	–	–	16x16x64
conv + D2	3x3	128	*Relu*	8x8x64
FC1	1x1	8192	*Relu*	8x8x128
FC2	1x1	99	–	8192
softmax − loss	–	99	–	99

(a) Custom deep net

type	patch size/stride or remarks	input size
conv	3x3/2	299x299x3
conv	3x3/1	149x149x32
conv	3x3/1	147x147x32
pool	3x3/2	147x147x64
conv	3x3/1	73x73x64
conv	3x3/2	71x71x80
conv	3x3/1	35x35x192
3xInception		35x35x288
5xInception		17x17x768
2xInception		8x8x1280
pool	8x8	8x8x2048
linear	*logits*	1x1x2048
softmax	*classifier*	1x1x99

(b) InceptionV3-like net

Fig. 4. Deep Net Architectures. The overview provides details on the layer types used, the sizes of kernel and their stride, as well as the layer dimensions.

4 Results

Individual identification results are presented in Table 1. As shown in magenta there, we first confirm that the Manta Matcher approach performs similarly on our dataset as on the one reported in [2] with classification accuracy above 46%.

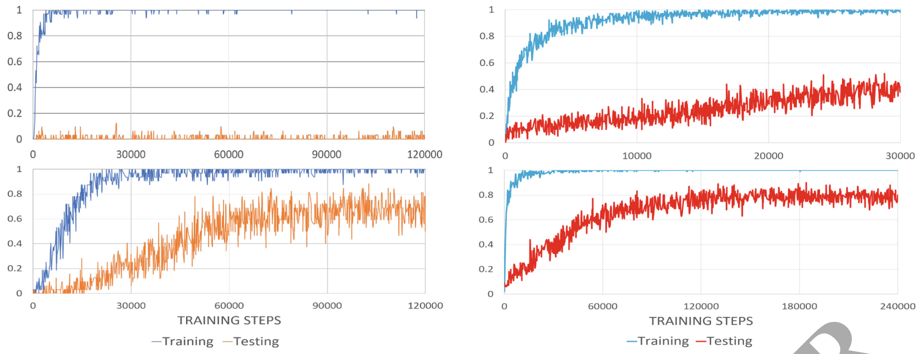

Fig. 5. Accuracy Evolution During Optimization. Graphs depict the development of accuracy (y-axis) along network training steps (x-axis) for our custom model (left) and during InceptionV3 transfer learning (right). ***(top left)*** Custom network optimized without augmentation is unable to generalize training performance (blue) towards testing performance (orange) and overfits the data. ***(bottom left)*** The same network is able to learn more effectively when provided with augmented data. ***(top right)*** Early performance of fine-tuned InceptionV3-like model using the same augmented data, and ***(bottom right)*** long-term learning of this approach. The latter yields competitive benchmarks (also see Table 1). (Color figure online)

Table 1. Top-N accuracy results

Model (and Dataset)	Top-1 accuracy	Top-10 accuracy
Manta Matcher (their 581)	46.82%	65.06%
Manta Matcher (our 198)	46.46%	65.15%
Custom DNN (our 198)	69.69%	79.29%
Fine-tuned InceptionV3 (our 198)	**79.29%**	87.88%
Fine-tuned + Encounter Fusion	78.79%	**91.92%**

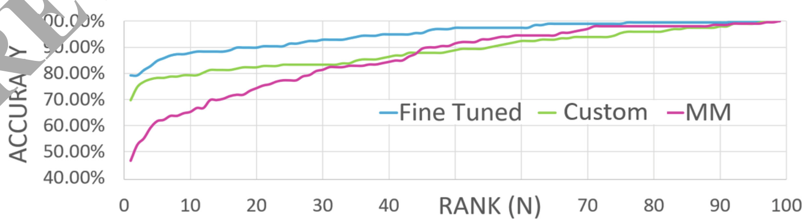

Fig. 6. Top-N accuracy for single image ID on our 198 test samples.

In our case, however, the sparsity of the original training data causes deep learning without augmentation to fail completely w.r.t. generalization, overfitting on the training samples (see Fig. 5, top left). However, augmentation addresses this problem effectively (see Fig. 5, bottom left) yielding a *classification accuracy* just above 69% as shown in ochre in Table 1 and Fig. 6. Our fine-tuned InceptionV3-like model trained over long term (see Fig. 5, bottom right) outperformed both approaches with a classification accuracy above 79% as shown in blue in Table 1 and Fig. 6. Practical applications with a human in the loop can, however, tolerate some ranking error – confirming a match against a dozen or so candidates is practically feasible. Thus, accuracy within the *Top 10 interval predictions* (see Table 1, column three) made by a model is also of interest. Whilst the described encounter fusion gives no gain of the Top-1 accuracy, we observe accuracy improvements in the Top-10 statistics from 87.88% to 91.92%.

5 Conclusion and Future Work

We have shown that, for the problem of photo-based recognition of individual manta rays, a combination of augmentation, transfer learning, and encounter-wide fusion techniques can address sparsity and noise challenges to enable deep learning to operate effectively – potentially assisting field work beyond previous capabilities. We demonstrated that an InceptionV3-like network trained on augmented data and fusing multiple encounter images outperforms the so-far best traditional approach published. Overall, this indicates that deep learning techniques in conjunction with augmentation and regularisation approaches have a role to play in advancing the performance of animal biometrics systems for visual manta ray identification. Future work will target fully automated processing of imagery as well as deep learning extensions that allow for open set identification, that is to avoid retraining of models whenever new individuals are encountered.

References

1. Kühl, H., Burghardt, T.: Animal biometrics: quantifying and detecting phenotypic appearance. Trends Ecol. Evol. **28**, 432–441 (2013)
2. Town, C., Marshall, A., Sethasathien, N.: Manta matcher: automated photographic identification of manta rays using keypoint features. Ecol. Evol. **3**, 1902–1914 (2013)
3. Loos, A., Ernst, A.: An automated chimpanzee identification system using face detection and recognition. EURASIP Image Video Process. **2013**(1), 49 (2013)
4. Hughes, B., Burghardt, T.: Automated visual fin identification of individual great white sharks. IJCV **122**(3), 542–557 (2017)
5. Lowe, D.G.: Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157 (1999)
6. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006). https://doi.org/10.1007/11744023_32

7. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of IEEE CVPR, pp. 779–788 (2016)
8. Branson, S., Van Horn, G., Belongie, S., Perona, P.: Bird species categorization using pose normalized deep convolutional nets. arXiv:1406.2952 (2014)
9. Kumar, N., et al.: Leafsnap: a computer vision system for automatic plant species identification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, pp. 502–516. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33709-3_36
10. Brust, C.-A., et al.: Towards automated visual monitoring of individual gorillas in the wild. In: Proceedings of IEEE CVPR, pp. 2820–2830 (2017)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105. Curran Associates Inc. (2012)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, pp. 1929–1958 (2014)
13. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of IEEE CVPR, pp. 2818–2826 (2016)
14. Freytag, A., Rodner, E., Darrell, T., Denzler, J.: Exemplar-specific patch features for fine-grained recognition. In: Jiang, X., Hornegger, J., Koch, R. (eds.) Pattern Recognition, pp. 144–156. Springer, Cham (2014)
15. Freytag, A., Rodner, E., Simon, M., Loos, A., Kühl, H.S., Denzler, J.: Chimpanzee faces in the wild: log-euclidean CNNs for predicting identities and attributes of primates. In: Rosenhahn, B., Andres, B. (eds.) GCPR 2016. LNCS, vol. 9796, pp. 51–63. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45886-1_5
16. Andrew, W., Greatwood, C., Burghardt, T.: Visual localisation and individual identification of Holstein Friesian cattle via deep learning. In: IEEE International Conference on Computer Vision Workshop, pp. 2850–2859 (2017)
17. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE CVPR, pp. 248–255, June 2009