

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/317558591>

Automatic fish species classification in underwater videos: Exploiting pretrained deep neural network models to compensate for limited labelled data

Article in *ICES Journal of Marine Science* · May 2017

DOI: 10.1093/icesjms/fsx109

CITATIONS

34

READS

2,772

7 authors, including:



Shoaib Ahmed Siddiqui

Deutsches Forschungszentrum für Künstliche Intelligenz

20 PUBLICATIONS 212 CITATIONS

[SEE PROFILE](#)



Faisal Shafait

University of Western Australia

187 PUBLICATIONS 4,305 CITATIONS

[SEE PROFILE](#)



Ajmal Saeed Mian

University of Western Australia

228 PUBLICATIONS 5,397 CITATIONS

[SEE PROFILE](#)



Mark R Shortis

RMIT University

181 PUBLICATIONS 3,323 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



3D Action Recognition from Novel Viewpoints [View project](#)



3D object recognition [View project](#)

Original Article

Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data

Shoaib Ahmed Siddiqui¹, Ahmad Salman^{1,*}, Muhammad Imran Malik¹, Faisal Shafait¹, Ajmal Mian², Mark R. Shortis³, and Euan S. Harvey⁴

¹School of Electrical Engineering and Computer Science, National University of Sciences and Technology (NUST), Sector H-12, Islamabad, 44000, Pakistan

²School of Computer Science and Software Engineering, The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia

³School of Science, RMIT University, GPO Box 2476, Melbourne, VIC 3001, Australia

⁴Department of Environment and Agriculture, Curtin University, Kent Street, Bentley, WA 6102, Australia

*Corresponding author: tel: +92(0)51 9085 2400; fax: +92(0)51 9085 2002; e-mail: ahmad.salman@seecs.edu.pk

Siddiqui, S. A., Salman, A., Malik, M. I., Shafait, F., Mian, A., Shortis, M. R., and Harvey, E. S. 2017. Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data. – ICES Journal of Marine Science, 75: 374–389.

Received 20 February 2017; revised 19 May 2017; accepted 23 May 2017; advance access publication 4 July 2017.

There is a need for automatic systems that can reliably detect, track and classify fish and other marine species in underwater videos without human intervention. Conventional computer vision techniques do not perform well in underwater conditions where the background is complex and the shape and textural features of fish are subtle. Data-driven classification models like neural networks require a huge amount of labelled data, otherwise they tend to over-fit to the training data and fail on unseen test data which is not involved in training. We present a state-of-the-art computer vision method for fine-grained fish species classification based on deep learning techniques. A cross-layer pooling algorithm using a pre-trained Convolutional Neural Network as a generalized feature detector is proposed, thus avoiding the need for a large amount of training data. Classification on test data is performed by a SVM on the features computed through the proposed method, resulting in classification accuracy of 94.3% for fish species from typical underwater video imagery captured off the coast of Western Australia. This research advocates that the development of automated classification systems which can identify fish from underwater video imagery is feasible and a cost-effective alternative to manual identification by humans.

Keywords: deep learning, fish classification, fisheries management, neural networks, stock assessment, underwater video.

Introduction

Data on the relative abundance, length and distribution of fish is important for monitoring the status and health of fish assemblages, and in particular those species targeted by fisheries. Due to global changes in the oceanic environment, fish and fish habitats are under increasing pressures (Bennett *et al.*, 2015; Wernberg *et al.*, 2016). Standardized, cost effective and reliable tools are essential for routine monitoring of fish across multiple depths and habitats (McLaren *et al.*, 2015). Manual methods of fish population estimation and species classification that involve

destructive and time consuming measures, such as the physical capture of samples and underwater visual census by divers, are still a common approach (Cappo *et al.*, 2003a,b; Mallet and Pelletier, 2014). In contrast, underwater video-based fish monitoring approaches are gaining popularity as an effective, portable, non-invasive and non-destructive method of fish population sampling (Shortis *et al.*, 2009; Murphy and Jenkins, 2010; Whitmarsh *et al.*, 2017). Off-the-shelf 'action' cameras and video recorders provide high quality, high-resolution images at affordable cost, leading to their predominance as the preferred sampling

technique (Murphy and Jenkins, 2010). However, manual analysis of video sequences by species experts is a time consuming, and therefore costly process (Greene and Alevizon, 1989). An automatic species recognition system has great potential to improve the efficiency of the analysis, leading to more rapid responses and wider use of video monitoring to assess the status and health of the marine environment. Video-based automatic estimation of fish populations and species recognition is a two-stage process: (i) fish detection in the video frames followed by (ii) species classification. Fish detection is a process of distinguishing fish from non-fish objects, e.g. aquatic plants, coral reefs, kelp, sponges and seabed structures in the video. Fish species classification on the other hand aims to recognize the species of the detected fish from the set of various classes.

In the last two decades, numerous image processing and machine learning algorithms have been proposed for fish species classification. Early methods proposed for this task can perform species classification in controlled environments only, for example using dead fish samples on the fishing vessel or in the laboratory, based on the shape and colour information (Strachan and Kell, 1995). 3D modelling of fish using laser light was proposed by Storbeck and Daan (2001) to measure fish dependent features like length, height and thickness of certain species. Unconstrained underwater fish classification is a more challenging task as the videos acquired in such environments depict high variation of luminosity, turbidity of water and background confusion due to reef structures and moving aquatic plants. The similarity in shape, colour and texture of different fish species (inter-class similarity) and differences among the same fish species (intra-class dissimilarity) due to changes in orientation of freely moving fish pose yet another challenge in accurate species classification. To address this issue, two different classical methods are proposed by Rova *et al.* (2007) and Spampinato *et al.* (2010) to classify fish based on their texture pattern and shape in natural, unconstrained environments. However, those methods only produce good results for fish species that were highly distinguishable due to rich and clearly defined texture patterns.

Recent trends in fish species classification are moving towards the application of machine learning algorithms in conventional computer vision-based approaches. In this regard, early attempts with modest success rates include applying principal component analysis (PCA) (Turk and Pentland, 1991) and linear discriminant analysis (Mika *et al.*, 1999) to extract key features. More recently, sparse representation classification (SRC) in combination with PCA (Hsiao *et al.*, 2014) has been applied to fish classification with success rate of 81.8% on their dataset of 1000 fish images of 25 different species. These approaches are all based on the assumption that visual features of fish are linearly separable from the surrounding underwater variability. Natural variation and the rich background increases the requirement for nonlinearity in the mathematical modelling which results in a compromise on the performance of the algorithm in the classification task. Gaussian mixture modelling and support vector machines (SVM) were employed in Huang *et al.* (2015) to train on fish images. Their algorithm yielded improved results over linear PCA and standalone SVM with 74.8% recognition rate on a dataset of about 24 000 images of 15 different fish species (Duan and Keerthi, 2005; Huang *et al.*, 2015). artificial neural networks (ANNs) were first introduced in the middle of 20th century, but went out of favour due to the high levels of supervised training required and their inability to solve highly complex problems.

ANNs have seen a resurgence as a preferred method for image classification because Krizhevsky *et al.* (2012) were able to achieve a substantial 10% reduction in error rate on the large-scale ImageNet benchmark dataset (Deng *et al.*, 2009) for object recognition by using an ANN variation known as Convolutional Neural Networks (CNNs). The importance of this special type of ANN was quickly recognized, along with a very broad range of applicability in the domain of artificial intelligence. The term Deep Learning suddenly became ubiquitous for neural networks as they learned hierarchical representations of data and the representation improved with the increase in the number of layers. These layers in ANN can be categorized based on their mathematical operation to extract certain features that represent the input image. For example, convolution layers in CNN perform convolution operation to find the correlation among the same class features using tuneable weights. This is followed by a nonlinear activation function. There are several types of nonlinear inducing layers like Rectifying Linear Units, Sigmoid and Hyperbolic Tangent (LeCun *et al.*, 2004; Simonyan and Zisserman, 2014; He *et al.*, 2016). The choice of nonlinear layer is dependent on nonlinearity and complexity of input data for better data-to-feature mapping. Subsampling layers (often called pooling layers) picks the relevant data out of convolution layer based on their significance and discards the rest. The output layer with number of nodes equal to the number of classes to be classified, is usually a fully connected layer that produces predicted labels or scores per node that represents each class. The predicted label is then matched with the ground truth label to calculate accuracy.

In the last three decades, different types of layers have been proposed which can be plugged into a neural network and trained via back-propagation of gradients (Rumelhart *et al.*, 1986; LeCun *et al.*, 1989, 2015). Convolution is a well-known operation in the signal processing and computer vision community. Conventional computer vision techniques make frequent use of the convolutional operation, especially for edge detection and noise reduction. LeCun *et al.* (1989) showed that the filters of the convolution operation that are useful for the task at hand can be automatically learned in a neural network. CNNs and their variants are considered to be state-of-the-art in image classification tasks with promising performance on handwritten numerical digit classification, facial recognition and generic object recognition (Larochelle *et al.*, 2009; Lee *et al.*, 2009; Simonyan and Zisserman, 2014). Each convolutional layer in the CNN is followed by a nonlinearity which allows the network to capture nonlinear dynamics of input data. In our case, water murkiness, abrupt changes in the underwater luminosity and variation in the sea bed are some of the main factors that require complex yet nonlinear mathematical modelling for the problem of automatic classification as their data distribution cannot be modelled by a linear classifier. Therefore, choice of nonlinearity makes a significant impact on the overall performance and is still an active area of research (LeCun *et al.*, 2015). Classification is performed directly using additional fully connected layers at the top of the CNN (LeCun *et al.*, 2004; Chatfield *et al.*, 2014) using features of the last convolution layer or by combination of features from several convolution layers (Ouyang and Wang, 2013). Alternatively, a separate final classifier is also used, with SVM and K-nearest neighbour being the popular choices in the literature. A primary disadvantage of the increased number of layers in neural networks especially CNN is the requirement for larger amounts of training data and hence, computational burden. A CNN approach

adopted by [Salman et al. \(2016\)](#) for fish species classification in an unconstrained underwater environment demonstrates the effectiveness of large datasets. The approach yielded average classification rate of over 90% on the LifeCLEF14 and LifeCLEF15 benchmark fish datasets (<http://imageclef.org/>).

In this article, we have employed deep CNNs pre-trained on large, publicly available image sets, for extraction of features from images of 16 different species of fish from the coastal waters of temperate and subtropical Western Australia, and applied a final classification step utilizing a linear SVM with a one-vs.-all strategy. Training a deep model from scratch requires thousands of labelled image examples, which were not available for the specific dataset used. Therefore, our approach opts for a transfer learning methodology to classify the fish species in the video dataset. To realize transfer learning in our methodology, we use several deep CNN models that are pre-trained on the vast benchmark ImageNet dataset ([Deng et al., 2009](#)) for Large-Scale Visual Recognition Challenge (<http://image-net.org>) to extract feature representations of fish images in our dataset before classification. The transfer learning approach is beneficial when the available training data is not large enough, as in our case, or has few examples of different variabilities. Unavailability of sufficient training datasets in the case of fish classification causes the network to overfit on the training data, resulting in high recognition accuracy on the training data, but very low accuracies on test data. Using a pre-trained network alleviates the requirement of training a deep CNN which requires large amount of training data along with high computational power.

The objective of this research is to determine the accuracy that can be achieved for fine-grained fish species classification using deep learning techniques. A cross-layer pooling algorithm is proposed that uses a pre-trained CNN as a generalized feature detector, thus avoiding the need for a large amount of training data. Classification is performed by a SVM on the features computed through the proposed method of cross-layer pooling, which results in more accurate predictions by the classifier on test images.

Material and methods

Study area for dataset collection

Videos were collected from several baited remote underwater video sampling programs that occurred between Cape Naturaliste and the Houtman Abrolhos islands in the temperate and subtropical coastal waters of Western Australia during 2011–2013. Videos were collected from kelp, seagrass, sand and coral reef habitats between 5 and 50 m of water depth. The sampling location of fish spread across the coast of Western Australia and does not represent the actual frequency distribution of target fish species in that location. The data used in this study are extracted from the videos in a way to achieve the minimum number of samples required to train the classifiers hence, covers only a fraction of actual recorded data.

Camera system description

The video imagery of fish was captured from baited remote underwater stereo-video systems (stereo-BRUVs). These systems are a practical and cost-effective solution for surveying reef fish across a range of depths and habitats ([Harvey et al., 2013](#)). The stereo-BRUVs that this imagery was collected from consisted of two Sony CX12 high-definition video cameras in purpose-built underwater housings. The housings are mounted on a base bar,

0.7 m apart and inwardly converged at 8 degrees, to provide an optimized overlapping field of view from the two cameras. Detailed information on the design and photogrammetric specifics of these systems is presented in [Harvey and Shortis \(1995, 1998\)](#), [Shortis and Harvey \(1998\)](#), and [Harvey et al. \(2010\)](#). Each system was baited with ~1 kg of crushed Australian Pilchards, (*Sardinops sagax*) which was placed in a plastic-coated wire basket and suspended 1.2 m in front of the two cameras ([Hardinge et al., 2013](#)). Pilchards were chosen as bait because they have been commonly used in stereo-BRUVs studies in Western Australia (e.g. [Watson et al., 2005, 2007, 2009; Harvey et al., 2012](#)), and have been shown to be a strong fish attractant ([Dorman et al., 2012](#)). The systems were left to capture video on the seafloor for 60 min ([Watson et al., 2005; Bernard et al., 2014](#)) during daylight hours.

Video analysis

Imagery of the 16 focal species shown in [Table 1](#) was collected during analysis of the stereo-BRUVs video sequences. The focal species were selected because they were either:

- (i) important species for recreational and commercial fishing (e.g. *Choerodon rubescens*, *P. leopardus*, *Lethrinus nebulosus*, *Carangoides fulvoguttatus*, *Scombridae* spp., *Lethrinus* sp, *Pagrus auratus*, and *Lethrinus atkinsoni*),
- (ii) numerically abundant and made a high contribution to the fish assemblages (e.g. *Coris auricularis*),
- (iii) indicator species for ecosystem based fisheries management (e.g. *C. auricularis*, *Scarus ghobban*), or
- (iv) provided challenges for separation of similar species (e.g. *Pentapodus porosus* and *Pentapodus emeryii*) for identification due to contrasting shapes and movement patterns (e.g. *Abudefduf bengalensis* and *Thalassoma lunare*).











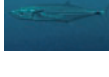

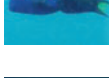


Other species were also included in the video imagery because image data sets generally require an “other” species class containing objects that are not of interest to the task at hand, or are false negatives that are not classified as a species of interest.

Video clips of the 16 species were captured during the routine analysis used to determine the relative abundance of the species. Counts and measurements of the MaxN (maximum number; see [Cappo et al., 2001, 2003a,b, 2006](#)) of fish of any one species identified within the field of view at the same time were made using EventMeasure Stereo software (www.seagis.com.au/event). During the image analysis between 50 and 120 individual 10 second video clips of each of the 16 focal species were captured. From the video clips, still images were extracted, re-sized and cropped to a consistent size of 224 by 224 pixels. CNN architectures work on pre-defined image sizes which were 224 by 224 in our case requiring the original images to be resized for processing. [Figure 1](#) presents samples of the training, validation and test images for the 16 fish species classes and the *other* species class.

Deep CNN architecture



A deep CNN is a mathematical parametric architecture with an input layer, several hidden layers and an output layer ([LeCun et al., 2004](#)). Starting from the input layer, the hidden layers are connected with each other by a set of tuneable weights and each layer represents more complex features of the input image. There are several types of hidden layers in CNN architecture as explained in the Introduction Section.

Table 1. Fish species distribution in the captured data for training validation and test sets along with sample images.

Species	Sample image	Total images	Training set	Validation set	Test set
<i>P. porosus</i>		103	60	12	31
<i>P. emeryii</i>		92	42	15	35
<i>C. rubescens</i>		100	45	22	33
<i>A. bengalensis</i>		100	56	12	32
<i>C. cyanodus</i>		110	59	11	40
<i>P. leopardus</i>		97	43	22	32
<i>L. nebulosus</i>		105	51	11	43
<i>C. fulvoguttatus</i>		108	64	10	34
<i>L. carponotatus</i>		103	63	10	30
<i>S. ghobban</i>		94	50	13	31
<i>Scombridae</i> spp.		89	52	7	30
<i>Lethrinus</i> sp		98	46	15	37
<i>T. lunare</i>		101	58	12	31
<i>C. auricularis</i>		141	91	12	38
<i>P. auratus</i>		101	52	11	38

Continued

Table 1. continued

Species	Sample image	Total images	Training set	Validation set	Test set
<i>L. atkinsoni</i>		105	56	16	33
Other		562	421	26	115
Total		2209	1309	237	663
Total (Without Others)		1647	888	211	548

The mismatch or error between the predicted label from the output layer and actual label is reduced by training the network with respect to the network weights through an iterative procedure such as error back propagation with weight adjustment (Hinton and Salakhutdinov, 2006). An illustrative CNN is shown in Figure 2.

Our pre-trained model on ImageNet dataset is not designed for fish species recognition, but the 1.2 million images provide a very large and diverse number of generic objects. Training on a dataset that contains large variations of objects and backgrounds in the images makes the machine learning algorithms like deep CNNs extract unique information for the objects of interest based on their colour, texture and shape in the images. Therefore, a suitable network learns to capture visual patterns of various objects and if there are an adequate number of examples for each object in the dataset, the generalization capability of the network increases. Such a trained network yields task-specific information on datasets that are not even used for training. Examples of such pre-trained networks include AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan and Zisserman, 2014), and ResNet (He et al., 2016), which the authors have made available freely for the research community to use as generalized feature extractors. We have used the trained CNNs for fish species classification in the transfer learning setup. In our transfer learning setup, deep CNN models used are based on three different architectures; AlexNet (Krizhevsky et al., 2012) is comprised of five convolutional layers and four pooling layers. VGGNet (Simonyan and Zisserman, 2014) is based on a unique architecture in which the size of all convolutional filters is $3 \times 3 \times 3$ Residual Net (He et al., 2016) learns a residual function instead of the whole input to output mapping, which makes the learning easier and hence scalable to very deep models. The network makes use of skip connections along with a separate branch which must learn just a residual function since the original input is propagated to the next layer using the skip connection. These skip connections also alleviate the vanishing gradient problem as the gradient flows directly through the skip connection regardless of the gradient through the branch. Figure 3 conveys the main idea of residual learning. The ResNet-152 model used in our experiments is a Residual Network having 152 layers. The five pooling layers in ResNet splits the architecture into five major modules. Each module is divided into many different operations such that the size of the activations is preserved.

Fine-grained classification method

Traditional classification methods are still prevalent for the task of fish classification despite recent advancements in machine

learning. A prominent example is the work of Huang et al. (2015), who employed shape, texture, and colour features of a fish in combination with a hierarchical classification scheme to identify fish species. These approaches are highly dependent on the feature extraction part which is responsible for the identification of useful features to be fed to the classifier. Most of the development time is devoted to careful feature extraction engineering based on expert domain knowledge. Usually, the classical classification pipeline is composed of pre-processing on images, feature extraction and finally classification (Shortis et al., 2016). This classification course has been completely altered with the recent advancements in the domain of deep learning (LeCun et al., 2015). This usually involves pre-processing on raw images which includes image enhancements, resizing or cropping followed by feature extraction which can be achieved through a trained deep multi-layer neural network to get invariant and abstract representation of input image in feature space. CNNs learn a hierarchical representation of the input in which the initial layers detect very basic patterns like edges and gradients, while layers located on top of the hierarchy learn complex patterns which are useful for the classification task at hand. The concept of hierarchical feature learning is illustrated in Figure 4 where the network was not specifically trained for fish images therefore, patterns found by the network are generic. Recent studies have shown that activations from deep CNNs can be employed as a universal image representation (Razavian et al., 2014). Fine grained classification is difficult because the separation of one class from another is subtle. Relying on the stochastic optimization algorithm to learn the distinct features for each class can potentially lead to poor generalization of the network on unseen examples. Image observations are noisy and the network can overfit the noise instead of discovering the real distribution of different classes. A parts-based pooling method (Zhang and Farrell, 2012) suggests the use of manual parts annotation for fine-grained classification by pooling together only features that belong to a specific part annotation.

This forces the network to focus on the annotated areas in the image for classification. The network is forced to attend to the parts, allowing differentiation between different classes, resulting in much better generalization as compared with using an end-to-end learning approach (Xiao et al., 2015). The final image representation is formed by concatenating different feature vectors obtained after pooling. However, obtaining precise, manual annotation of important image features by a human is expensive. The cross-layer pooling method (Liu et al., 2015) suggests the use of convolutional layer activations at the top of the hierarchy as

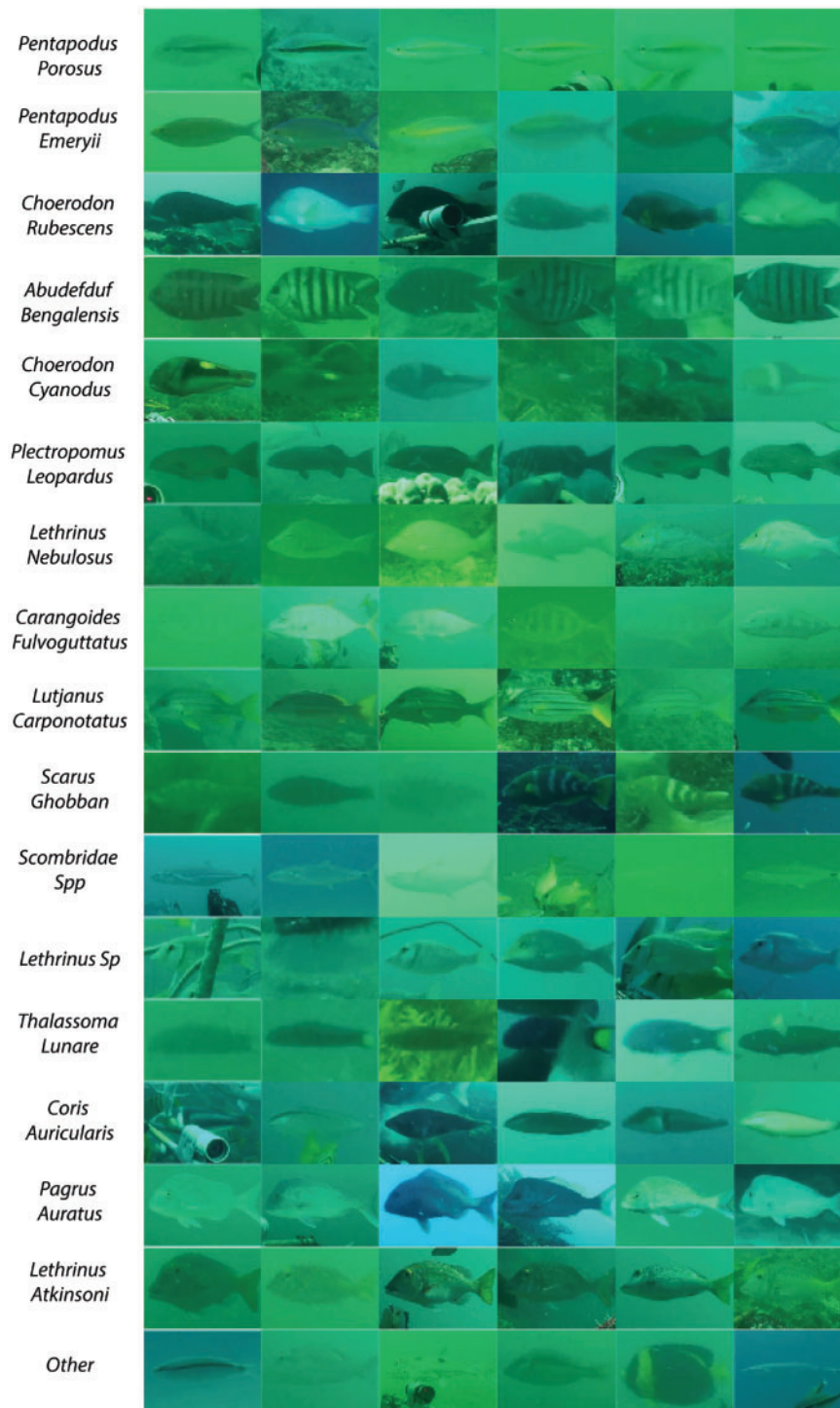


Figure 1. Sample images from the data where each row contains images of a single species. The first two images in each row are samples from the training set, the next two images are samples from the validation set and the last two images are samples from the test set. This diagram also demonstrates the complexity and challenges in the data in the form of murkiness, blurriness, low texture information of fish and luminosity variation.

parts annotation from a network that is pre-trained on a large dataset. The annotation provided by the convolutional layer activations may well be inferior to that provided by a human expert. However, since there are orders of magnitude more annotations due to the large number of feature maps present in the layers on the top of hierarchy, the annotations can give performance

comparable to a careful, manual parts annotation. Cross-layer pooling requires activations from two different layers with the same spatial dimensions, due to point-wise multiplication operation. Lower layer activations serve as local features while higher layer activations serve as parts annotations. Therefore, features in the lower layers are pooled after weighting them with the higher

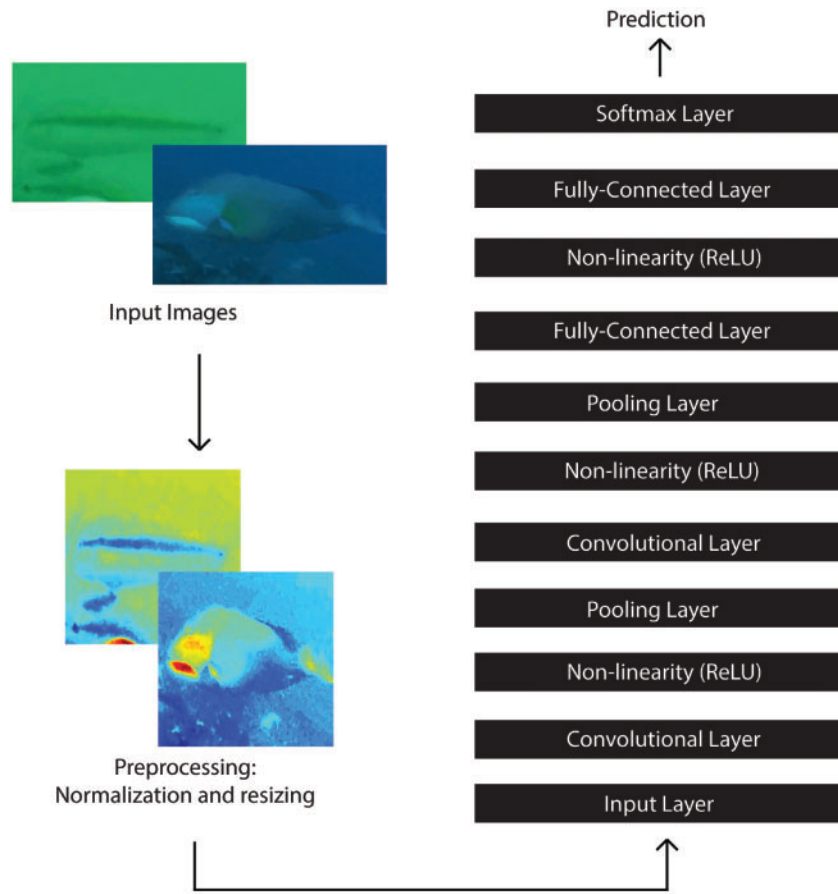


Figure 2. Illustration of a small CNN with alternate arrangement of convolution and pooling layers followed by fully connected layers at the top.

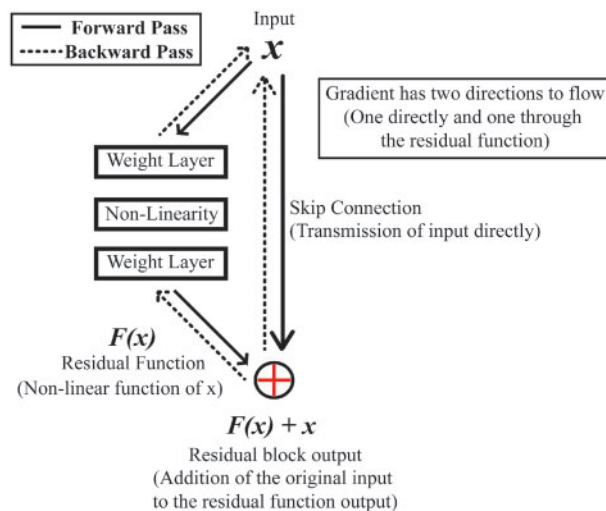


Figure 3. Illustration of residual learning. The input is directly propagated to the output through a skip connection, therefore, the network must learn just a residual function which should be added to the input.

layer activations. We used deep networks that are pre-trained on the ImageNet benchmark dataset for feature extraction. Figure 5 presents an overview of the cross-layer pooling method. The image feature vector is obtained by concatenating pooling results

from all the channels. The size of the feature vector produced for each image after cross-layer pooling is $r \times d \times D$ where d is the number of feature maps in the lower layer L_1 , D is the number of feature maps in the upper layer L_2 , and r is the local region size. Mathematically, the method of cross-layer pooling can be represented as:

$$\mathcal{P} = \left(\Sigma(\mathcal{F}_{L_1} \odot \mathcal{F}_{L_2}^1), \Sigma(\mathcal{F}_{L_1} \odot \mathcal{F}_{L_2}^2), \dots, \Sigma(\mathcal{F}_{L_1} \odot \mathcal{F}_{L_2}^D) \right)$$

where L_1 refers to the lower layer used for extraction of local features, L_2 refers to the upper layer serving as parts annotation, and $\Sigma(\cdot)$ refers to the operation of sum pooling. Element-wise multiplication is performed followed by sum pooling for obtaining the feature vector using a single feature map serving as parts annotation. Features vectors obtained by using D feature maps where D is number of feature maps in L_2 are concatenated together to form final image representation \mathcal{P} . For using larger region sizes, the whole region is multiplied by a single value in the upper layer activations (centre of the region). We experimented with local region sizes of 1×1 and 3×3 for extraction of local features from the lower layer (Figure 5). A dimensionality reduction step was introduced to make the computations tractable. Activations near the image boundary were discarded in case of high dimensional feature maps as the activations were polluted by outliers. We evaluated different pooling strategies and found sum pooling to be

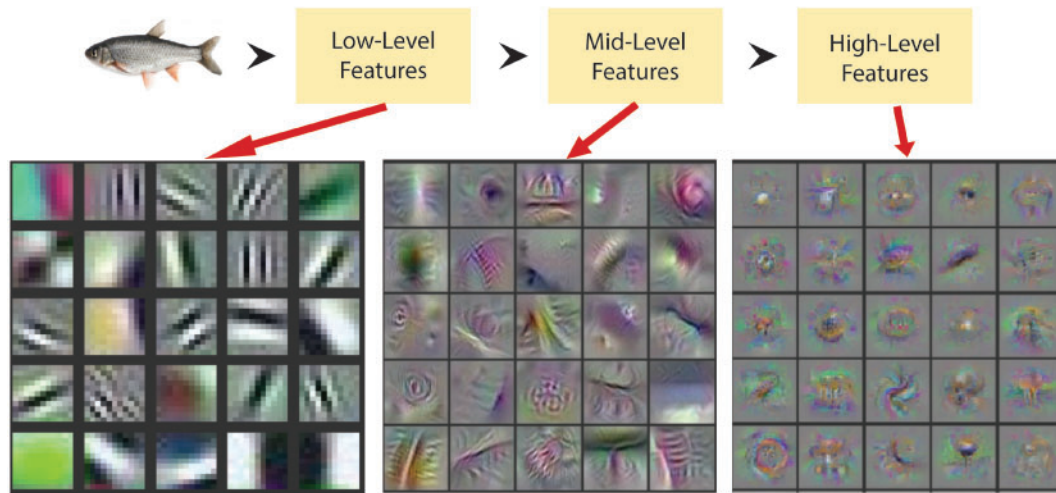


Figure 4. Hierarchical representation learning by a CNN where the initial layer detects simple patterns like edges and gradients while higher layers detect more abstract features (Yosinski *et al.*, 2015).

the most effective technique for fish classification. Power normalization was applied to the obtained feature vector followed by standardizing and transformation into a unit vector. Calculating per channel mean from each image separately for mean normalization instead of using the pre-computed ImageNet means significantly improved the classification performance. The colour distribution between images from ImageNet and the underwater fish images is different, as highlighted in Figure 6. A one-vs.-all linear SVM was trained on the cross-pooled features for classification of the fish species. One-vs.-all is a strategy for training a multi-class configuration in which a separate SVM is trained for each of the classes while treating the rest of the classes as a single negative class. Final classification is made by evaluating the SVMs trained for each class. As the feature extractor is not directly trained on the task at hand, it reduces the probability of network over-fitting on training data, which leads to much better generalization to unseen cases. We experimented with different system configurations for cross-layer pooling. The experiments involved use of the three different pre-trained CNNs on ImageNet dataset. The only constraint on selection of the layers for cross-layer pooling is that they must have the same spatial dimension. The choice of layers was based on empirical analysis, as considering all possible combinations of layers were computationally impractical. We used the fourth and fifth convolutional layer outputs for application of cross-layer pooling from a pre-trained AlexNet as the spatial dimensions were the same. Conv5, comprising a set of three consecutive convolutional operations, is the last convolutional module in VGGNet before the fully connected layers. Accordingly, we used the Conv5_1 and Conv5_3 layers in VGGNet. We chose the fourth module in ResNet-152 model by cross-pooling activations from 15th and 20th layers within the fourth module, annotated as 4b15 and 4b20, respectively. If enough computational resources are available, the best combination of layers for any classification task can be evaluated automatically by performing a grid search on all possible layer combinations and assessing their performance on an independent validation set.

The system was implemented using a commercial software package (MATLAB) with MatConvNet (Vedaldi and Lenc, 2014)

and LibSVM (Chang and Lin, 2011). The implementation was CPU specific. Experiments were conducted on a server with Intel Xeon E5-2630 v3 processor (2.4 GHz) and 64 GB Ram.

Results

A classification accuracy of 94.3% was achieved when classifying only the 16 fish species which we focussed on for this trial. All other evaluations include the *other* species class which contains a number of fish species which are non-relevant to the research goals, but can be encountered while underwater sampling. The results obtained by the different model configurations are summarized in Table 2. The model based on ResNet (He *et al.*, 2016) by cross-pooling activations from layer 4b15 and 4b20, with local region size of 3×3 and training a linear SVM on top of the cross-pooled features, was able to achieve the highest classification accuracy of 89.0% if the *other* species class is included. A change to a local region size of 1×1 produces a slightly inferior result of 86.9% classification accuracy for all classes. In comparison, the AlexNet and VGGNet model configurations produced significantly degraded results. Application of the linear SVM directly to features from the Pool5 layer of the ResNet-152 model resulted in a classification accuracy of 71.49%, highlighting the improved performance produced by cross-layer pooling. We reduced the dimensionality of the 3×3 local features in ResNet-152b model from 9216 ($3 \times 3 \times 1024$) to 512 before cross-layer pooling using PCA. Dimensionality reduction can be useful in order to reduce the size of features from cross-layer pooling, allowing use of larger region sizes as well as more feature maps. Adapting a larger region size even with the use of PCA improved accuracy of the system due to availability of local context. Further analysis of the accuracy of the four CNN models can be conducted using the precision and recall of the classifier. Recall and precision are a more informative way to judge the performance of a classifier, especially if the classes are skewed.

The precision of a classifier is a measure of correctly classified images out of all the images *predicted* to belong to a specific class. Mathematically,

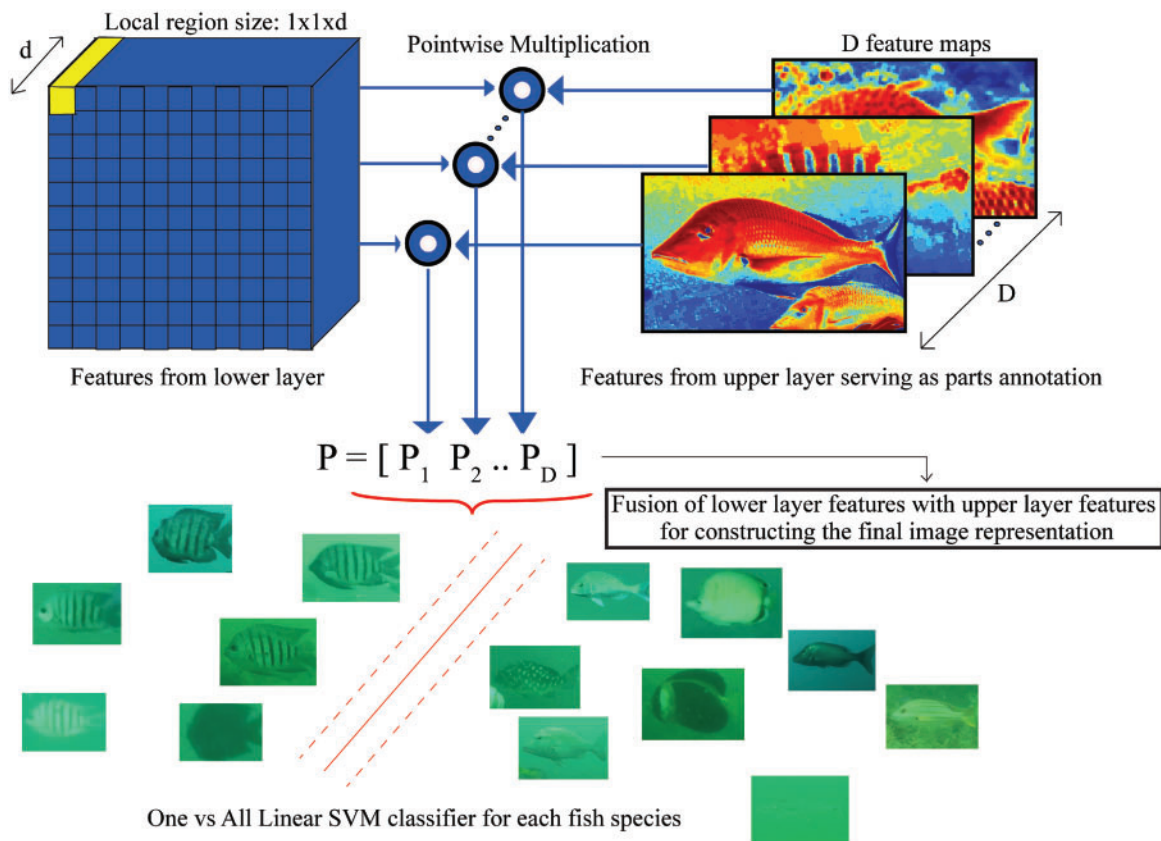


Figure 5. The cross-layer pooling pipeline includes feature extraction from two different layers of a network pre-trained on the ImageNet dataset. Lower layer features serve as local features (top left) while upper layer features serve as parts annotation (top right—colourmap goes from blue to red indicating the magnitude of the features). Features are cross-pooled/fused and passed to a SVM for final classification.

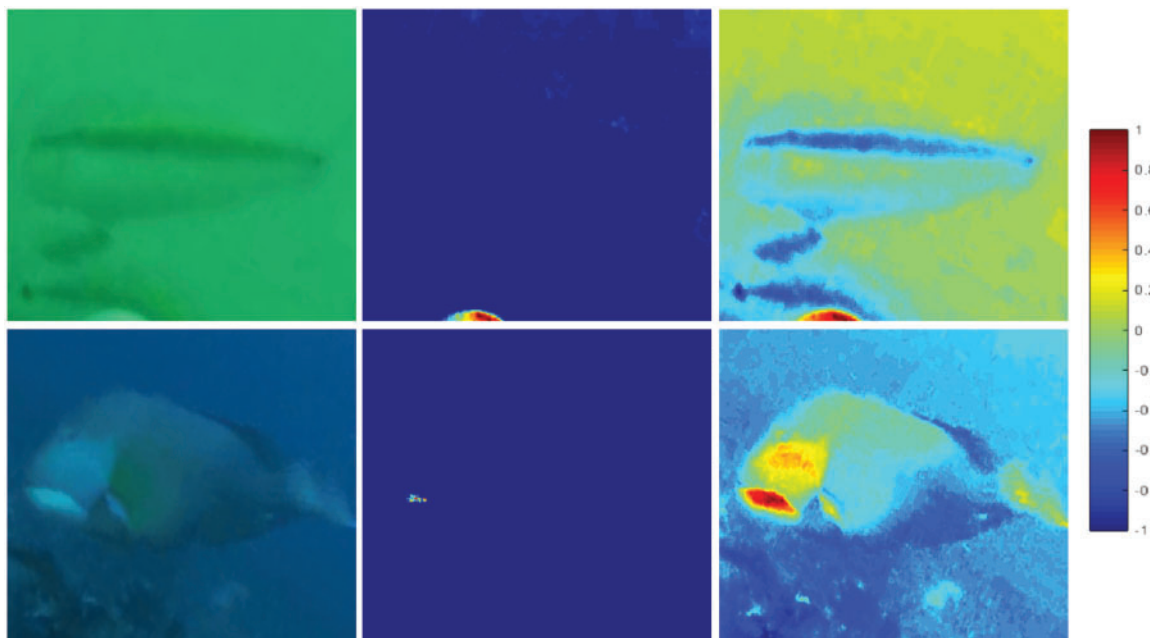


Figure 6. Per channel mean image subtraction from the left image. The middle image is normalized with the pre-computed means from ImageNet dataset. The right image is normalized by computing the per channel mean from the image itself and then converted to grayscale (−1 to 1 range) for visualization using a heat colour map.

Table 2. Results from different model configurations for cross-layer pooling.

Model	Layers used in cross-layer pooling	Accuracy
AlexNet	Convolution layer 4 and 5	65.8%
VGGNet	Convolution layer 1 and 3 of fifth module	78.1%
ResNet-152a	Convolution layer 15 and 20 of fourth module	86.9%
ResNet-152b	Convolution layer 15 and 20 of fourth module	89.0%
ResNet-152c	Classifier trained on MaxPooling layer 5 only	71.5%

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall of a classifier on the other hand is a measure of correctly classified images out of all the images *actually* belonging to a specific class. Mathematically:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The obtained precision and recall graphs for different model configurations are presented in Figure 7.

In general, the precision and recall graphs confirm the classification accuracy levels shown in Table 2, with very few exceptions. In some cases, the ResNet-152 model with a 1×1 local region produces the best result, but these exceptions reflect the small difference in the classification accuracy between the two sizes of local region. Precision is generally higher than recall, indicating that false negatives are more common than false positives. The precision graph shows high values for a number of individual species, indicating that the number of false positives is very low. The lesser precision in

the case of sub-species classes, such as *Scombridae* spp., and the *other* species class indicates higher levels of false positives, as the algorithm incorrectly labelled a larger number of images as belonging to these classes. The sub-species classes and *other* species class are comprised of a number of different species resulting in high intra-class variance. This leads to ambiguity regarding association of an image to a particular class. Unlike the precision graph, the recall graph indicates that, with just one exception, no class achieves a 100% result. For virtually every model and every class, there are examples of images being assigned incorrectly to the *other* class. In other words, in almost every model, a small percentage of fish are not being recognized as belonging to their correct class. The confusion matrix for the ResNet-152b model is presented in Figure 8. Most the false positives and false negatives are caused by resemblance between similar species, or with the *other* species category. For example, inter-class similarities cause false positives and false negatives between *Lethrinus* sp and *L. nebulosus* due to the commonalities between examples from the sub-species class and the specific species class. The confusion matrix confirms that precision should be more favourable than recall because many fish are incorrectly assigned to the *other* species class, as compared with false positives in the specific class. To provide a more detailed analysis of the ResNet-152b model, the misclassified images by the algorithm are presented in Figure 9. Certain fish categories present in the others species class have high resemblance with the rest of the 16 classes. Therefore, the inaccurate assignments in those classes can be clearly observed from the confusion matrix and misclassified images. In situations where the species are either camouflaged or the distinct parts are

unrecognizable, the classifier predicts them to belong to the *other* species class due to higher prior probability of the *other* species class, as compared with the rest of the classes. Classification errors are also introduced due to major occlusions in the image. If we consider the misclassified images for *P. porosus*, *P. emeryii*, *C. rubescens*, and *Lutjanus carponotatus* species, the key areas of patterns and texture distinguishing these species are missing (Table 1). There was only a single misclassification for *A. bengalensis* and *S. ghobban* species due to occlusion with aquatic plants. Background clutter and camouflaged fishes resulted in misclassification of images in *P. leopardus*, *C. fulvoguttatus*, *Scombridae* spp., *C. auricularis*, *P. auratus*, and *L. atkinsoni* species. *Choerodon cyanodus* and *T. lunare* fish species had a small region with a recognizable pattern which allowed them to be distinguished even if the rest of the structure was unrecognizable. Seven images in the *other* species class were labelled as *Scombridae* spp. due to high similarity between *Scombridae* spp. and a specific fish species in the *other* species class.

Discussion

The most important outcome of this research is the classification accuracy achieved. With the *other* species class included, the accuracy of the classification is 89.0%, which is competitive with or exceeds other recently reported results on fish species identification tasks (Hsiao et al., 2014; Huang et al., 2015; Salman et al., 2016). If the *other* species class is excluded, the classification accuracy within the 16 species classes is 94.3%, which is competitive with the identification rates of human experts for similar tasks (Culverhouse et al., 2003).

In Table 3, we compare the performance of our proposed method for species classification with two existing classification methods, using the fish image data set including the *other* species class. The results for the accuracy of the other two classifiers are significantly degraded by the inclusion of the *other* species class, which suggests that these methods are not well suited to fish classification in real-world settings. SRC for fish classification (Hsiao et al., 2014) uses a dictionary learning based approach to find a sparse representation of the input and assumes different fish species and the background to be linearly independent of each other. This linear independence assumption is not true for the captured data, resulting in poor performance for the SRC based classification method. The end-to-end learning framework presented by Salman et al. (2016) was unable to generalize well on the data set captured in the Australian waters, despite being based on CNNs. The major problem associated with the training of any deep network using end-to-end learning is over-fitting on the training set. Despite the use of data-augmentation to increase the size of the captured data by synthetic transformations, the data is nevertheless comprised of a very limited number of images that are not sufficient for training a CNN from scratch (Salman et al., 2016).

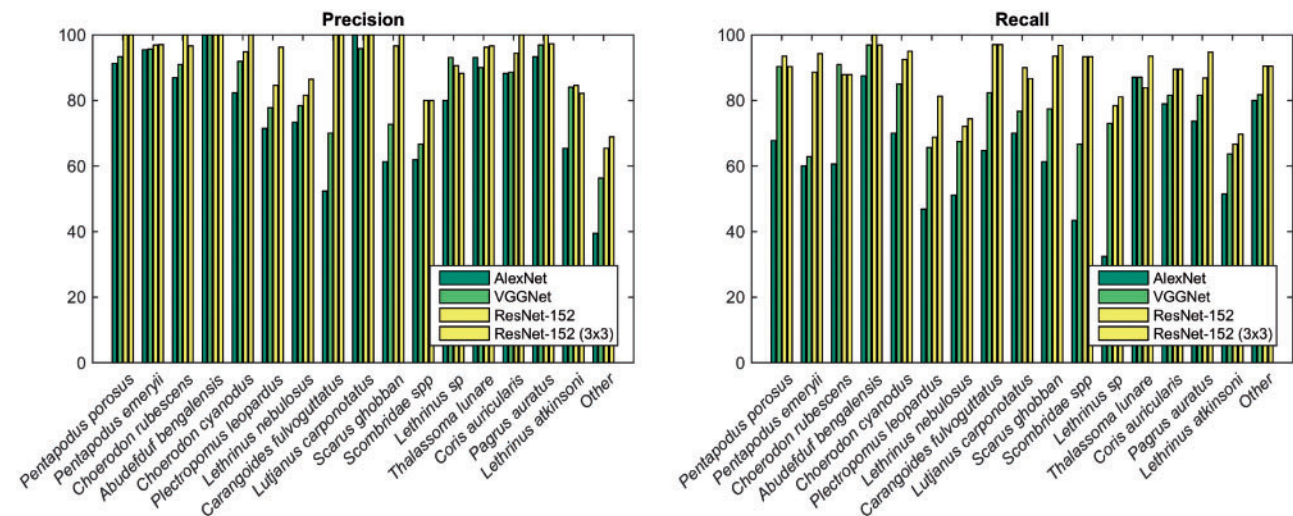


Figure 7. Performance comparison in terms of %precision and %recall between different models for each fish species.

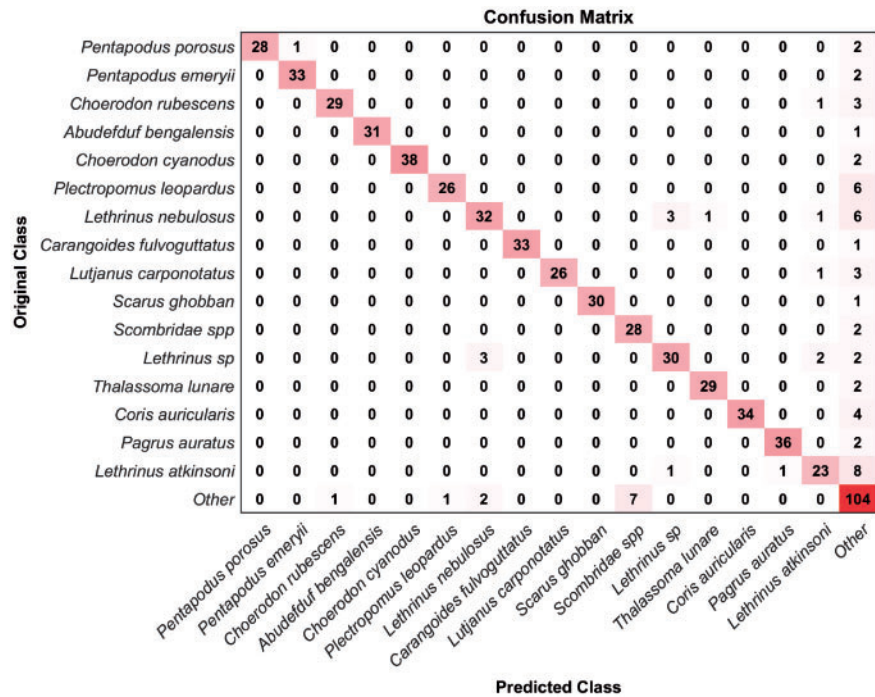


Figure 8. Confusion matrix obtained for the classifier based on ResNet-152 model.

Most of the deep learning based approaches require huge amounts of labelled training data to achieve successful results. To further validate the potential of transfer learning, we also conduct a similar experiment on the benchmark dataset of LifeClef 2015 (<http://www.imageclef.org/lifeclef/2015/fish>) which is extracted from the Fish4knowledge repository (<http://groups.inf.ed.ac.uk/f4k/>). LifeClef15 dataset contain more than 20 000 images of fish divided into 15 classes of species, details of which is given in Salman *et al.* (2016). For each species, this dataset has different number of available images. For this new experiment, the two algorithms in Table 3 i.e. SRC by Hsiao *et al.* (2014) and CNN by Salman *et al.* (2016) were trained end-to-end on randomly selected 5% images of each class of fish species. For our proposed

architecture, we only trained the SVM (for classification) on features extracted by the same pre-trained ResNet-152b with cross-layer pooling on the images of LifeClef15 images. This is the same experimental protocol as was followed with the other dataset. To monitor the training process for optimum accuracy, 5% images per class were reserved for validation set and the rest of the images which compose 90% of the dataset were used for testing. Overall, around 1200 images were used for training. The reason behind using small number of images for training is to emphasize on the better generalization capability of the proposed method under transfer learning protocol that does not require any data for training from the two datasets presented for comparative study. Our proposed method achieved 96.73% accuracy on

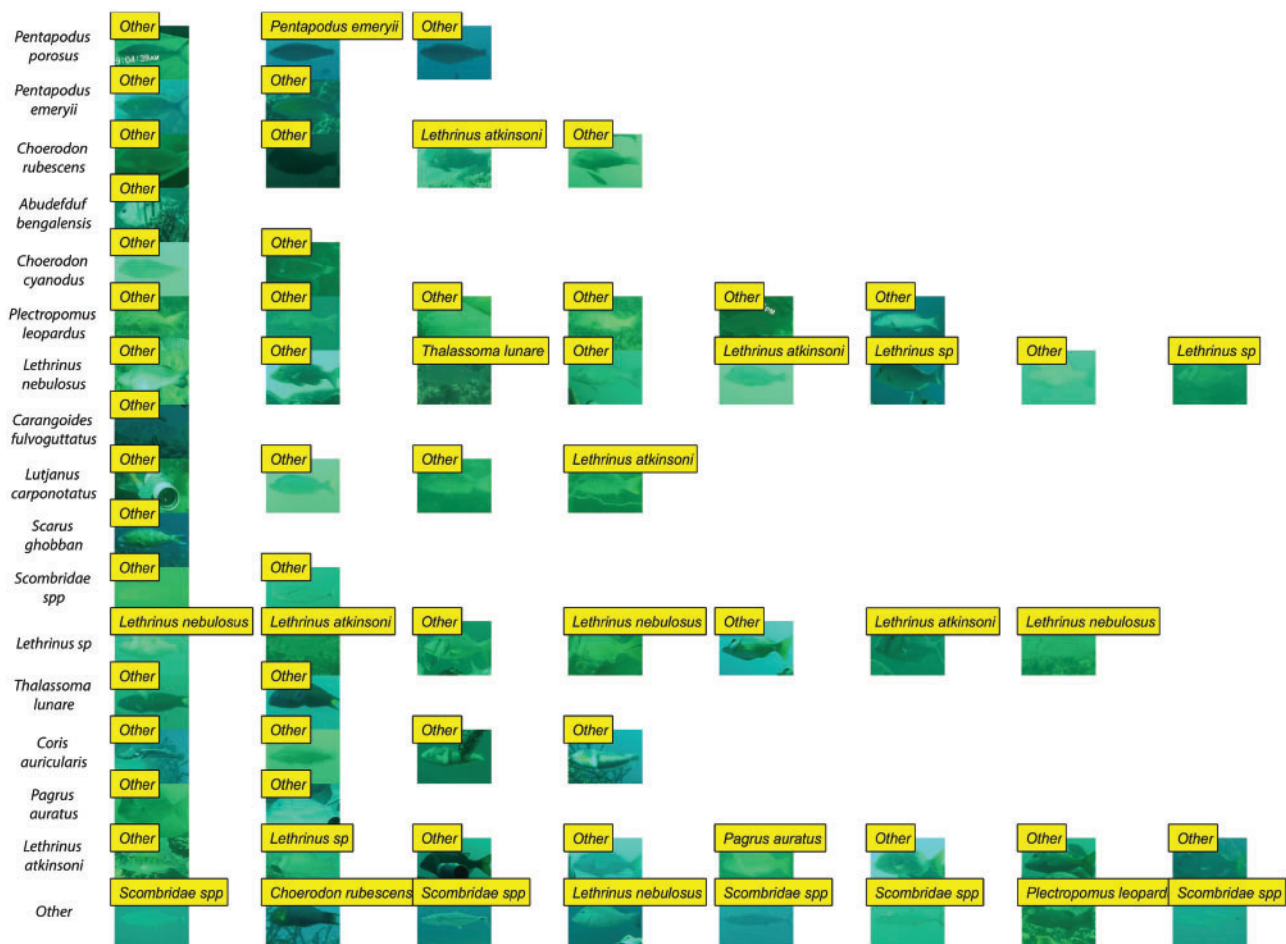


Figure 9. Images with misclassified class tags. The left-most column indicates the actual class.

the test set of LifeClef15 dataset while SRC and CNN by [Salman et al. \(2016\)](#) achieved 65.42 and 87.46% accuracies, respectively. The overall better performance by all three methods in LifeClef15 dataset is due to the less challenging environmental conditions and relatively clear images as compared with the other dataset. If we increase the training dataset size with more than 5000 images, the CNN in [Salman et al. \(2016\)](#) with end-to-end training produces 94.1% accuracy, marginally behind our newly proposed method. Further increase in the training data size may increase the performance due to over-fitting.

The focus of this work is to employ deep learning architecture in the form of a CNN to extract distinguishing features of fish that are uniquely dependent on a specific species. The motivation behind using such an approach is to model highly nonlinear and complex attributes in underwater imagery of fish. These attributes are not modelled effectively by conventional shallow machine learning algorithms and image processing techniques ([Hinton and Salakhutdinov, 2006](#); [Larochelle et al., 2009](#)). Automated surveillance of underwater footage requires fish detection, classification and tracking in unconstrained environments containing variations in lighting, pose, background, and water turbidity. This becomes a fine-grained classification problem due to low inter-class variation and high intra-class variation of the fish species. Fine-grained classification techniques force the network to learn to attend to subtle features that are important for the

classification of species at hand. However, a CNN trained using an end-to-end learning approach on datasets with a large number of classes and a small number of images per class is highly prone to over-fitting. Various approaches to automatic fish species classification are presented in [Rova et al. \(2007\)](#), [Fablet et al. \(2009\)](#), [Blanc et al. \(2014\)](#), and [Hsiao et al. \(2014\)](#), that employ shallow machine learning approaches on different datasets, typically with much less environmental variation. As a comparative study, we used their algorithms on our dataset but failed to achieve significant performance on fish species classification. Hence, we do not report these results here. This strengthens the claim that classification performance is severely compromised on datasets where explicit features of fish, such as texture, colour and shape, are not highly distinguishable due to background confusion, turbidity or poor quality images.

[Qin et al. \(2015\)](#) and [Salman et al. \(2016\)](#) report on fish species classification in unconstrained environments, where CNNs were used to extract fish dependent features from datasets containing inadequate numbers of labelled examples of fish species necessary for training deep networks. Promising results for their datasets were achieved despite the lack of samples, but the approach failed to cope with the more complex datasets where there is difficulty in classifying fish species due to water murkiness, inter-species similarity, intra-species dissimilarity, poor light conditions and changes in orientation of fish (see [Table 2](#)). This gives rise to the

Table 3. Comparative results for fish species classification using our dataset of 16 fish species of interest and an additional class that includes all other species that are not of interest (second column).

Method	Accuracy (our dataset)	Accuracy (LifeClef15)
SRC (Hsiao <i>et al.</i> , 2014)	49.1%	65.42%
CNN (Salman <i>et al.</i> , 2016)	53.5%	87.46%
Proposed method of CNN plus SVM	89.0%	96.73%

Classification accuracy on benchmark dataset of LifeClef15 that is taken from large Fish4Knowledge (<http://groups.inf.ed.ac.uk/f4k/>) repository. Bold entries indicate the best scores on accuracy as compared to all other methods on specific datasets.

necessity of meticulously fine-tuning the generic deep CNN architecture to enhance the effectiveness of the discrimination information in the feature space.

The representation and modelling power of CNN increases with the increase in the number of layers (LeCun *et al.*, 2015) where each successive layer performs a nonlinear computation on its inputs. ResNet-152 has 152 layers which makes it a more powerful model compared with the other two models. Accordingly, the ResNet model achieves the highest species recognition rate of the three models tested. Any pre-trained CNN model can be plugged into the proposed cross-layer pooling framework. We expect better, even more complex models to emerge with time, which can replace the current models to achieve further improvement in the recognition rate. There must be a point of diminishing returns, at which more complexity results in no significant improvement, but at present a larger number of layers does improve the accuracy of classification.

This flexibility can be extended to different fish image data sets. Transfer learning is a strategy for employing pre-trained deep CNNs as generalized feature extractors. The features extracted by the CNN are very generic, allowing the features to be used for classification of any fish species. We demonstrated this generic applicability by testing the proposed method on two dissimilar fish species classification datasets.

The *other* species class has a higher prior probability of selection due to skewed classes. This makes false negatives much more likely, especially in the case of isolated obscuration, resulting in loss of recall for most species and loss of precision for the *other* species class. A network trained to exhibit an explicit attention mechanism can learn to focus on distinct features and ignore misleading parts of the image, including occlusions and background clutter. Attention based models will also allow the network to recognize species based on just a single distinct feature instead of using global image features, which can be misleading, as was the case with *C. cyanodus* and *T. lunare*. Attention based models might reduce the imbalance between precision and recall graphs but complete evasion in the case of skewed classes is still a challenging task.

Since the feature extractor is not directly trained on the data, there is a possibility that it will extract overlapping features for two similar species, reducing the success rate of classification. Distinguishing different species from one another depends on the uniqueness of their features. The less than perfect classification accuracy indicates that there is indeed some overlap between the features extracted by the pre-trained CNN models. Removing the *other* species class which comprises a large number of species will clearly have a positive impact on the recognition performance, as most of the misclassifications are directed towards this class. Notwithstanding these issues, the aims of the species classification task should always be the primary driver for the selection of the number of classes and the inclusion of the *other* species class.

Although *other* species class is virtually indispensable in the unconstrained underwater environment, the minimum number of species classes is a pragmatic decision to ensure that all species of interest can be categorized.

The method of cross-layer pooling adds extra computational complexity making it difficult to adapt to a real-time implementation. A single forward pass of the network followed by dimensionality reduction and cross-layer pooling pushes the processing time of a single image to ~800 ms. Efficient graphics processor unit (GPU)-based implementation of the whole system can reduce the computation time by a significant margin. Some recent fine-grained classification approaches are focused on using a single network with fully connected layers at the final decision stage for classification, substantially reducing the time required to process an image (Lin *et al.*, 2015; Xiao *et al.*, 2015).

Methodologies based on performing just a forward pass on a CNN can allow development of systems with real-time performance. Different attention based models have been recently proposed in the literature for fine-grained classification where the network learns to attend to parts useful for classification, making the predictions more confident and accurate (Lin *et al.*, 2015; Xiao *et al.*, 2015). Fish classification in unconstrained underwater environments is also a fine-grained classification task where the classifier has to ignore the background and clutter, focusing on just the shape and texture of the fish or even a single distinct feature like the tail, ignoring the rest of the features which can be misleading. Spatial Transformer (Jaderberg *et al.*, 2015) is a differentiable component which can be attached to any CNN allowing it to be trained with an end-to-end learning approach. It could prove useful in fish classification, especially in conditions where the input image is not well aligned.

In a real world implementation of a species classifier, fish will be imaged at a wide range of resolutions, will be in a variety of orientations with respect to the image, and can be swimming strongly, along with the well-known problems of partial obscuration and poor image quality due to turbidity. The training set used here comprises individual still images with fish in a near-normal, lateral aspect, whereas in practice, fish may be oriented in different directions and tracked across many individual frames of the video sequence. Whilst the availability of multiple frames affords the ability to majority vote a species classification, the potentially disruptive effects of resolution variations, orientation changes and swimming action will require additional adaptations of the algorithms. For simple measurements of length, tracking across multiple frames has a proven advantage of improved precision of the mean length (Shafait *et al.*, 2017), whereas the changes in body shape and orientation may prove multiple frames to be less of a clear advantage for species recognition. One potential response is for the algorithms to actively select fish only in the near-normal, lateral aspect in order to prevent the disruption to the classification process.

This method is particularly useful in cases where automated classification of fish species is desired. In fact, in this age, rapid marine/fresh water body exploration to monitor trends in species abundance of fish is crucial which is directly affected by higher pace of today's global environmental changes. To achieve this, effective methods in automatic fish species classification are required. The proposed methodology is a step forward in this direction. Our system is based on machine learning and hence needs to be trained (only requires class annotation) for the relevant classes but the number of images required for training is low without compromising on classification accuracy. This has been validated by our experiments. The method can also be coupled with different fish detection (Spampinato *et al.*, 2014) and measurements systems (Shafait *et al.*, 2017) for development of a fully automated system covering a wide range of fisheries related tasks. Optimized implementations of the system can be produced to achieve near real-time performance. Surveys are a particular application which can largely benefit with this automation using the proposed cross-layer pooling method covering both online and offline image classification.

Conclusions

We presented an automatic method that exploits existing pre-trained deep neural network models for fish species classification in videos captured in unconstrained underwater environments. Our results are compiled in a challenging realistic scenario where an additional class, that contains numerous fish species that are not of interest, is included during classification. The major contribution of this work is that it shows how to use CNN models trained for a different classification task, for which sufficient labelled training data is available, for fish species classification where labelled data is scarce. With this strategy, even when our training data was limited, we were able to employ the deepest CNN model so far (with 152 layers) and achieve state-of-the-art results. We proposed a special cross-layer pooling approach that combines features from different layers of the deep network for enhanced discriminative ability. Wide spread use of the proposed and similar automatic techniques will speed up the rate at which marine scientists analyse underwater videos. A bottleneck of the cross-layer pooling method is the extensive computations required which precludes the possibility of real time processing. An essential future research direction is to develop strategies for fine-grained classification by directly feed forwarding the images through a pre-trained network to get the final classification results without the need for an external classifier.

Acknowledgements

The authors acknowledge support from the Australian Research Council Grant LP110201008, which provided the primary funding for this study in addition to funding from a UWA Research Collaboration Award (RCA) grant, Higher Education Commission Pakistan Startup Research Grant 21-949/SRGP/R&D/HEC/2016 and the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research. Ajmal Mian was supported by the Australian Research Council Fellowship DP110102399. The authors also acknowledge Nvidia Corporation, USA for their donation of GPU under their GPU Grant Program. Nvidia GPUs were used to carry out simulations in the work carried out in this article.

References

- Bennett, S., Wernberg, T., Harvey, E. S., Santana-Garcon, J., and Saunders, B. J. 2015. Tropical herbivores provide resilience to a climate-mediated phase shift on temperate reefs. *Ecology Letters*, 18: 714–723.
- Bernard, A. T. F., Götz, A., Parker, D., Heyns, E. R., Halse, S. J., Riddin, N. A. *et al.* 2014. New possibilities for research on reef fish across the continental shelf of South Africa. *South African Journal of Science*, 110: 1–5.
- Blanc, K., Lingrand, D., and Precioso, F. 2014. Fish species recognition from video using SVM classifier. *Proceedings of the 3rd ACM International Workshop on Multimedia Analysis for Ecological Data*, ACM, 1–6 pp., doi: 10.1145/2661821.2661827.
- Cappo, M., Speare, P., Wassenberg, T. J., Harvey, E., Rees, M., Heyward, A., and Pitcher, R. 2001. Use of Baited Remote Underwater Video Stations (BRUVS) to survey demersal fish – how deep and meaningful? *In* Direct Sensing of the Size Frequency and Abundance of Target and Non-Target Fauna in Australian Fisheries, pp. 63–71. Ed. By E. S. Harvey, and M. Cappo. Fisheries Research and Development Corporation. Rottneest Island, Western Australia.
- Cappo, M., Harvey, E., Malcolm, H., and Speare, P. 2003a. Potential of video techniques to monitor diversity, abundance and size of fish in studies of marine protected areas. *In* Aquatic Protected Areas-What works best and how do we know? Ed. By J. P. Beumer, A. Grant, and D. C. Smith. *Proceedings of the World Congress on Aquatic Protected Areas*. Australian Society for Fish Biology, North Beach, Western Australia. pp. 455–464.
- Cappo, M., Harvey, E. S., Malcolm, H., and Speare, P. 2003b. Advantages and applications of novel “video-fishing” techniques to design and monitor Marine Protected Areas. *In* Aquatic Protected Areas - What Works Best and How Do We know? Ed. by J. P. Beumer, A. Grant, and D. C. Smith. *Proceedings of the World Congress on Aquatic Protected Areas*, Cairns, Australia. August 2002. 455–464 pp.
- Cappo, M., Harvey, E., and Shortis, M. 2006. Counting and measuring fish with baited video techniques - an overview. pp 101-114. *In* Cutting-Edge Technologies in Fish and Fisheries Science. Ed. by J. M. Lyle, D. M., Furlani, and C. D. Buxton. Australian Society for Fish Biology Workshop Proceedings, Hobart, Tasmania, August 2006, ASFB. 225 pp.
- Chang, C., and Lin, C. 2011. LIBSVM. *A Library for Support Vector Machines*. *ACM Transactions on Intelligent Systems and Technology*, 2: 1–27.
- Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. 2014. Return of the Devil in the Details: Delving Deep into Convolutional Networks. *Proceedings of the British Machine Vision Conference*, doi: 10.5244/C.28.6.
- Culverhouse, F. P., Williams, R., Reguera, B., Herry, V., Gonzalez-Gil, S. 2003. Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Marine Ecology Progress Series*, 247: 17–25.
- Dorman, S. R., Harvey, E. S., and Newman, S. J. 2012. Bait effects in sampling coral reef fish assemblages with stereo-BRUVs. *PLoS One*, 7: e41538.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
- Duan, K., and Keerthi, S. S. 2005. Which is the best multiclass SVM method? An empirical study. *Lecture Notes in Computer Science*, 3541: 278–285.
- Fablet, R., Lefort, R., Karoui, I., Berger, L., Masse, J., Scalabrin, C., and Boucher, J. M. 2009. Classifying fish schools and estimating their species proportions in fishery-acoustic surveys. *ICES Journal of Marine Science*, 66: 1136–1142.

- Greene, L. E., and Alevizon, W. E. 1989. Comparative accuracies of visual assessment methods for coral reef fishes. *Bulletin of Marine Science*, 44: 899–912.
- Hardinge, J., Harvey, E. S., Saunders, B. J., and Newman, S. J. 2013. A little bait goes a long way: The influence of bait quantity on a temperate fish assemblage sampled using stereo-BRUVs. *Journal of Experimental Marine Biology and Ecology*, 499: 250–260.
- Harvey, E. S., and Shortis, M. R. 1995. A system for stereo-video measurement of sub-tidal organisms. *Marine Technology Society Journal*, 29: 10–22.
- Harvey, E. S., and Shortis, M. R. 1998. Calibration stability of an underwater stereo-video system: Implications for measurement accuracy and precision. *Marine Technology Society Journal*, 32: 3–17.
- Harvey, E. S., Goetze, J., McLaren, B., Langlois, T., and Shortis, M. R. 2010. The influence of range, angle of view, image resolution and image compression on underwater stereo-video measurements: high definition and broadcast resolution video cameras compared. *Marine Technology Society Journal*, 44: 75–85.
- Harvey, E. S., Dorman, S. R., Fitzpatrick, C., Newman, S. J., and McLean, D. L. 2012. Response of diurnal and nocturnal coral reef fish to protection from fishing: an assessment using baited remote underwater video. *Coral Reefs*, 31: 939–950.
- Harvey, E. S., Cappo, M., Kendrick, G. A., and McLean, D. L. 2013. Coastal fish assemblages reflect geological and oceanographic gradients within an Australian zootone. *PLoS One*, 8: e80955.
- He, K., Zhang, X., Ren, S., and Sun, J. 2016. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778, doi: 10.1109/CVPR.2016.90.
- Hinton, G., and Salakhutdinov, R. 2006. Reducing the dimensionality of data with neural networks. *Science*, 313: 504–507.
- Hsiao, Y., Chen, C., Lin, S., and Lin, F. 2014. Real-world underwater fish recognition and identification using sparse representation. *Ecological Informatics*, 23: 13–21.
- Huang, P. X., Boom, B. J., and Fisher, R. B. 2015. Hierarchical classification with reject option for live fish recognition. *Machine Vision and Application*, 26: 89–102.
- Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. 2015. Spatial Transformer Networks. *Proceedings of the Advances in Neural Information Processing Systems*. pp. 2017–2025.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. *Proceedings of the Advances in Neural Information Processing Systems*. pp. 1106–1114.
- Larochelle, H., Bengio, Y., Louradour, J., and Lamblin, P. 2009. Exploring strategies for training deep neural networks. *Journal of Machine Learning Research*, 10: 1–40.
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computing*, 1: 541–551.
- LeCun, Y., Huang, F., and Bottou, L. 2004. Learning methods for generic object recognition with invariance to pose and lighting. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2: 97–104, doi: 10.1109/CVPR.2004.1315150.
- LeCun, Y., Bengio, Y., and Hinton, G. E. 2015. Deep learning. *Nature*, 521: 436–444.
- Lee, H., Grosse, R., Ranganath, R., and Ng, A. 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *Proceedings of the International Conference on Machine Learning*. pp. 609–616, doi: 10.1145/1553374.1553453.
- Lin, T. Y., RoyChowdhury, A., and Maji, S. 2015. Bilinear CNN models for fine-grained visual recognition. *Proceedings of the IEEE International Conference on Computer Vision*. pp. 1449–1457, doi: 10.1109/ICCV.2015.170.
- Liu, L., Shen, C., Hengel, A. 2015. The Treasure beneath Convolutional Layers: Cross-convolutional-layer pooling for Image Classification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 4749–4757, doi: 10.1109/CVPR.2015.7299107.
- Mallet, D., and Pelletier, D. 2014. Underwater video techniques for observing coastal marine biodiversity: a review of sixty years of publications (1952–2012). *Fisheries Research*, 154: 44–62.
- McLaren, B. W., Langlois, T. J., Harvey, E. S., Shortland-Jones, H., and Stevens, R. 2015. A small no-take marine sanctuary provides consistent protection for small-bodied by-catch species, but not for large-bodied, high-risk species. *Journal of Experimental Marine Biology and Ecology*, 471: 153–163.
- Mika, S., Ratsch, G., Weston, J., Scholkopf, B., and Mullers, K. R. 1999. Fisher discriminant analysis with kernels. *Proceedings of the IEEE International Workshop on Neural Networks for Signal Processing*. pp. 41–48, doi: 10.1109/NNSP.1999.788121.
- Murphy, H. M., and Jenkins, G. P. 2010. Observational methods used in marine spatial monitoring of fishes and associated habitats: a review. *Marine and Freshwater Research*, 61: 236–252.
- Ouyang, W., and Wang, X. 2013. Joint deep learning for pedestrian detection. *Proceedings of the IEEE International Conference on Computer Vision*. pp. 2056–2063, doi: 10.1109/ICCV.2013.257.
- Qin, H., Lia, X., Liangb, J., Pengb, Y., and Zhang, C. 2015. DeepFish: Accurate underwater live fish recognition with a deep architecture. *Elsevier Journal of Neurocomputing*, 187: 49–58.
- Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. 2014. CNN features off-the-shelf: an astounding baseline for recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. pp. 512–519, doi: 10.1109/CVPRW.2014.131.
- Rova, A., Mori, G., and Dill, L. M. 2007. One fish, two fish, butterfly, trumpeter: recognizing fish in underwater video. *Proceedings of the IAPR Conference on Machine Vision Applications*. pp. 404–407.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. Learning internal representations by error propagation. *Parallel Distributed Processing: explorations in the Microstructure of Cognition*, 1: 318–362.
- Salman, A., Jalal, A., Shafait, F., Mian, A., Shortis, M., Seager, J., and Harvey, E. 2016. Fish species classification in unconstrained underwater environments based on deep learning. *Limnology and Oceanography: Methods*, 14: 570–585.
- Shafait, F., Harvey, E. S., Shortis, M. R., Mian, A., Ravanbakhsh, M., Seager, J. W., Culverhouse, P., Cline, D., and Edgington, D. 2017. Towards automating underwater measurement of fish length: A comparison of semi-automatic and manual stereo-video measurements. *ICES Journal of Marine Sciences*, doi: 10.1093/icesjms/lsx007.
- Shortis, M., and Harvey, E. S. 1998. Design and calibration of an underwater stereo-video system for the monitoring of marine fauna populations. *International Archives Photogrammetry and Remote Sensing*, 32: 792–799.
- Shortis, M., Harvey, E. S., and Abdo, D. 2009. A review of underwater stereo-image measurement for marine biology. *In Oceanography and Marine Biology: An Annual Review*. Ed. by R. N. Gibson, R. J. A. Atkinson, and J. D. M. Gordon. CRC Press, USA, 47: 257–292.
- Shortis, M., Ravanbakhsh, M., Shafait, F., and Mian, A. 2016. Progress in the automated identification, measurement, and counting of fish in underwater image sequences. *Marine Technology Society Journal*, 50: 4–16.
- Simonyan, K., and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Proceedings of the International Conference on Learning Representations*. arXiv: 1409.1556.
- Spampinato, C., Giordano, D., Salvo, R. D., Chen-Burger, Y. H., Fisher, R. B., and Nadarajan, G. 2010. Automatic fish classification for underwater species behavior understanding. *ACM*

- Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams, Firenze, Italy, 45–50 pp.
- Spampinato, C., Palazzo, S., and Kvasidis, I. 2014. A texton-based kernel density estimation approach for background modeling under extreme conditions. *International Journal of Computer Vision and Image Understanding*, 122: 74–83.
- Storbeck, F., and Daan, B. 2001. Fish species recognition using computer vision and a neural network. *Fisheries Research*, 51: 11–15.
- Strachan, N. J. C., and Kell, L. 1995. “A potential method for the differentiation between haddock fish stocks by computer vision using canonical discriminant analysis”. *ICES Journal of Marine Science*, 52: 145–149.
- Turk, M., and Pentland, A. 1991. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3: 71–86.
- Vedaldi, A., and Lenc, K. 2014. MatConvNet: Convolutional Neural Networks for MATLAB. *Proceeding of the ACM International Conference on Multimedia*. arXiv: 1412.4564.
- Watson, D. L., Harvey, E. S., Anderson, M. J., and Kendrick, G. A. 2005. A comparison of temperate with reef fish assemblages recorded by three underwater stereo video techniques. *Marine Biology*, 148: 415–425.
- Watson, D. L., Harvey, E. S., Kendrick, G. A., Nardi, K., and Anderson, M. J. 2007. Protection from fishing alters the species composition of fish assemblages in a temperate-tropical transition zone. *Marine Biology*, 152: 1197–1206.
- Watson, D. L., Anderson, M. J., Kendrick, G. A., Nardi, K., and Harvey, E. S. 2009. Effects of protection from fishing on the lengths of targeted and non-targeted fish species at the Houtman Abrolhos Islands, Western Australia. *Marine Ecology Progress Series*, 384: 241–249.
- Wernberg, T., Bennett, S., Babcock, R. C., de Bettignies, T., Cure, K., Depczynski, M., Dufois, F. et al. 2016. Climate-driven regime shift of a temperate marine ecosystem. *Science*, 353: 169–172.
- Whitmarsh, S. K., Fairweather, P. G., and Huveneers, C. 2017. What is Big BRUVver up to? Methods and uses of baited underwater video. *Reviews in Fish Biology and Fisheries*, 27: 53–73.
- Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., and Zhang, Z. 2015. The Application of Two-Level Attention Models in Deep Convolutional Neural Network for Fine-Grained Image Classification. *Proceedings in the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 842–850, doi: 10.1109/CVPR.2015.7298685.
- Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. 2015. Understanding neural networks through deep visualization. *Proceedings in the International Conference on Machine Learning Workshop on Deep Learning*.
- Zhang, T. D. N., and Farrell, R. 2012. Pose pooling kernels for sub-category recognition. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3665–3672, doi: 10.1109/CVPR.2012.6248364.

Handling editor: Howard Browman