LIMNOLOGY
and
OCEANOGRAPHY: METHODS

ASLO

Limnol. Oceanogr.: Methods 14, 2016, 570–585
© 2016 Association for the Sciences of Limnology and Oceanography
doi: 10.1002/lom3.10113

Fish species classification in unconstrained underwater environments based on deep learning

Ahmad Salman,*[1] Ahsan Jalal,[1] Faisal Shafait,[1,2] Ajmal Mian,[2] Mark Shortis,[3] James Seager,[4] Euan Harvey[5]

[1]School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad, Pakistan
[2]School of Computer Science and Software Engineering, University of Western Australia, Perth, WA, Australia
[3]School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, NSW, Australia
[4]SeaGIS Pty Ltd, Melbourne, NSW, Australia
[5]School of Science, Department of Environment and Agriculture, Curtin University, Perth, WA, Australia

Abstract

Underwater video and digital still cameras are rapidly being adopted by marine scientists and managers as a tool for non-destructively quantifying and measuring the relative abundance, cover and size of marine fauna and flora. Imagery recorded of fish can be time consuming and costly to process and analyze manually. For this reason, there is great interest in automatic classification, counting, and measurement of fish. Unconstrained underwater scenes are highly variable due to changes in light intensity, changes in fish orientation due to movement, a variety of background habitats which sometimes also move, and most importantly similarity in shape and patterns among fish of different species. This poses a great challenge for image/video processing techniques to accurately differentiate between classes or species of fish to perform automatic classification. We present a machine learning approach, which is suitable for solving this challenge. We demonstrate the use of a convolution neural network model in a hierarchical feature combination setup to learn species-dependent visual features of fish that are unique, yet abstract and robust against environmental and intra-and inter-species variability. This approach avoids the need for explicitly extracting features from raw images of the fish using several fragmented image processing techniques. As a result, we achieve a single and generic trained architecture with favorable performance even for sample images of fish species that have not been used in training. Using the LifeCLEF14 and LifeCLEF15 benchmark fish datasets, we have demonstrated results with a correct classification rate of more than 90%.

Regular sampling of fish populations is important for monitoring the status and trends in the relative abundance, composition, size, and biomass of fish assemblages (Jennings and Kaiser 1998). There is an increasing focus on non-destructive sampling techniques as marine protected areas and areas closed to fishing increase in area, and as these techniques gain popularity as biodiversity management tools (McLaren et al. 2015). Underwater video based monitoring techniques (Harvey and Shortis 1995; Shortis et al. 2009), are being promoted as one of the main tools for non-destructive sampling of fish (Cappo et al. 2003; Mallet and Pelletier 2014). Underwater video systems have been shown to be cost effective, accessible and provide a means of repeatable sampling (Murphy and Jenkins 2010). While manual processing of the resulting imagery decreases the cost effectiveness and availability of numerical data after recording, recent developments in computer vision algorithms leading to automatic species identification can improve the efficiency of image analysis (Shortis et al. 2013).

Automatic counting and recognition of fish can be divided into two components, (1) automatic fish detection in the video sequences and (2) automatic fish species classification in the video frames. Fish detection aims to distinguish fish from "non-fish" objects in the video. Examples of non-fish objects include coral reefs, kelp, sea grass beds and other aquatic plants, sessile invertebrates such as sponges, gorgonians and ascidians, and the physical structure of the seafloor. Fish species classification aims to identify the species of fish out of the pool of various classes or species of interest.

Several image processing and machine learning algorithms have been proposed in the last two decades for these applications. Some early attempts involved classification of dead fish using shape and color dependent features (Strachan and Kell 1995). Another approach, using a laser light source

*Correspondence: ahmad.salman@seecs.edu.pk

Fig. 1. Example of underwater images on Taiwan reef with different background variability (http://groups.inf.ed.ac.uk/f4k/).

to create 3D fish models, was proposed in Storbeck and Daan (2001) to take into account features like height, width, and thickness of the specific species to be recognized. Such systems produced favorable results because they were developed for fish sampling in controlled environments, e.g., fishing vessels or conveyer belts. For real-time underwater fish identification, an effective approach using stereo cameras and controlled lighting conditions was used in Harvey and Shortis (1995). The fish were made to swim through a predefined chamber to capture their images. Unconstrained underwater fish classification involves more complex environments and challenging factors like variation in lighting, turbidity of the water, background confusion due to reef features and underwater plant life, and intra-species variation due to changes in orientation of the freely moving fish. Videos are generally captured using digital cameras and there is no prior assumption about the underwater environment where the cameras are deployed. Due to these confounding factors, underwater fish classification in an unconstrained environment is a real challenge. Two methods for fish classification in the natural environment are presented in, for example, Rova et al. (2007) and Spampinato et al. (2010), based on capturing the texture pattern and shape of fish using image processing. However, fish with only rich and easily distinguishable texture were targeted. Recent trends are moving toward the use of machine learning algorithms for fish classification in video. Such algorithms automatically learn features from labeled training data to differentiate between classes; different fish species in our case. Early machine learning algorithms were based on Principal Component Analysis (PCA) (Turk and Pentland 1991) or Linear Discriminant Analysis (LDA) (Mika et al. 1999). However, these techniques assume that the appearance of each fish species is linearly independent of other species' appearances as well as the background. This assumption does not hold in practice due to the similarities among fish species in both shape and size, and with the ambiguities caused by extremely diverse background consisting of underwater reefs and plant life. Recently, Sparse Representation-based Classification (SRC) has been used together with Eigen-faces (PCA) (Hsiao et al. 2014) for fish classification in the Taiwanese coral reef ecosystem. PCA, LDA, and SRC have been extensively used for other computer vision tasks like generic object recognition and facial recognition from images (Turk and Pentland 1991; Mika et al. 1999; Wright et al. 2009). However, due to their linear nature, these techniques are unable to model the nonlinear differences between the fish species and their complex backgrounds.

Environmental variability in underwater imagery poses a greater challenge towards achieving acceptable performance

15415856, 2016, 9, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lom3.10113 by Universita Di Salerno, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(for example, see Fig. 1). Another approach for unconstrained natural underwater environment uses hierarchical classification trees with Support Vector Machines (SVMs) trained on input image features (Huang et al. 2015). The decision making is based on Gaussian Mixture Modeling (GMM). The reported results are much improved in comparison with PCA and standard SVM classification (Duan and Keerthi 2005; Huang et al. 2015).

In the last few years, deep learning has emerged as a powerful machine learning tool with the ability to overcome the shortcomings of the conventional image classification approaches. Variation in lighting conditions, distortions like poor image quality and noise, changes in orientation and size of the object of interest in the image and variations in the background impose non-linearity in the image data distribution (Bengio 2009). It is difficult for conventional machine learning algorithms to model and adapt to the features of objects of interest in such images. Underwater imagery of fish in their natural habitat includes all of these challenges that must be addressed using a specially designed, non-linear mathematical function to represent the complex features in the data. Multilayer deep neural networks provide such an opportunity to extract unique, invariant and robust fish-dependent features in the presence of the distortions and variability in the images.

We propose to use deep Convolution Neural Networks (CNN) (LeCun et al. 2004) together with classification, based on the standard classifiers like K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) trained on the features extracted by the CNN in supervised deep learning. Fish dependent features learnt in this setup prove to be robust against environmental variability. Inspired by the visual cortex of cats, CNN are marked by their ability to explore spatially correlated sub-regions in natural images for extracting unique and orientation invariant features of objects. CNN has produced promising results in various applications like handwritten digit recognition, facial recognition, and speech recognition (Larochelle et al. 2007; Lee et al. 2009), where a specific architecture was designed for each case. State-of-the-art performance has been achieved using feature extraction through very deep CNN in generic image-based object recognition (Simonyan and Zisserman 2015). They introduced special feature selection layers, called pooling and regularization layers, on top of multiple convolution layers in CNN to sift out distinct features of objects to be recognized. Another pioneering CNN design for object detection in still images was proposed in Ouyang and Wang (2013), where a combination of features selected from each layer of CNN was used to identify the object of interest (pedestrians in their case) in images. We aim to employ a specially designed CNN for the task of fish classification using combination of fish-dependent features identified by each layer of network. Our approach is more suitable for fish recognition in underwater environments as the deep learning CNN technique adapts to the unique challenging situations, in comparison with general object recognition/detection in non-underwater imagery. We present the difference between this approach and the latest CNNs, and the motivation behind our design, in the discussion section.

The remainder of the paper is organized as follows: Materials and procedures section describes the proposed model of the deep learning CNN. Assessment section details the experimental scheme and protocol including the comparative study. Discussion and conclusions section covers detailed discussion about the significance of our results followed by conclusions.

Materials and procedures

In this section, we provide details about the fish dataset used in this article. Then, we describe our CNN architecture designed for extracting fish species-dependent features based on their unique visual characteristics. We also elaborate the motivation and reasoning for designing the proposed CNN and how it is beneficial in extracting information that helps in the classification of fish species.

Fish dataset

The datasets used in this article are taken from LifeCLEF 2014 and LifeCLEF 2015 fish identification tasks (http://www.imageclef.org/). The LifeCLEF 2014 (LCF-14) for fish is a smaller dataset derived from a very large dataset called Fish4Knowledge (http://groups.inf.ed.ac.uk/f4k/, 2015). Fish4Knowledge contains about 700,000 underwater video clips of ten minutes duration each. The videos span a time period of 5 yr of monitoring the marine ecosystem of Taiwan coral reefs, one of the largest fish biodiversity environments in the world with more than 3000 different fish species. The LCF-14 dataset for fish contains about 1000 videos. The labels of approximately 20,000 detected fish in the videos are also provided. A total of 10 different fish species are included in this dataset. LifeCLEF 2015 (LCF-15) is also taken from Fish4Knowledge. LCF-15 consists of 93 underwater videos covering 15 species. There are a total of 9000 annotations provided with the dataset that contain species labels in the videos. In addition, LCF-15 additionally provides about 20,000 sample images with class labels. As compared to LCF-14, LCF-15 provides challenging underwater images and videos marked by noisy and blurry environments and poor lighting conditions. Therefore, using LCF-15 helps us in judging the robustness of fish recognition algorithms in environments with higher variability. Table 1 summarizes the technical details of LCF-14 and LCF-15 datasets. Table 2 provides the categorization of LCF-14 and LCF-15 datasets according to the number of samples for each species.

The robustness of a classification technique is judged by the variability it can handle in the input data. Fish species recognition naturally encounters variability challenges, e.g., quality of the video, water turbidity, algae, background coral

15415856, 2016, 9, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lom3.10113 by Universita Di Salerno, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1. Information about LCF-14 and LCF-15 fish datasets.

Dataset	No. of videos	Format	Resolution	Frames/Sec	No. of labeled images	Species/classes
LCF-14	1000	FLV	640× 480, 320 x 240	24	19,868	10
LCF-15	93	FLV	640 x 480, 320 × 240	24	20,000+	15

Table 2. Species-wise population division in LCF-14 and LCF-15 datasets. Shaded are the common species in both datasets.

LCF-14 species	No. of images
Acanthurus nigrofuscus	3240
Amphiprion clarkii	3863
Chaetodon lunulatus	3411
Chromis margaritifer	3653
Dascyllus reticulatus	3873
Hemigymnus fasciatus	3077
Lutjanus fulvus	866
Myripristis berndti	3390
Neoniphon sammara	2988
Plectroglyphidodon dickii	3036

LCF-15 species	No. of images
Abudefduf vaigiensis	434
Acanthurus nigrofuscus	2770
Amphiprion clarkii	3265
Chaetodon lunulatus	3544
Chaetodon speculum	162
Chaetodon trifascialis	704
Chromis chrysura	3859
Dascyllus aruanus	1749
Dascyllus reticulatus	5327
Hemigymnus melapterus	361
Myripristis kuntee	3231
Neoglyphidodon nigroris	213
Pempheris vanicolensis	906
Plectroglyphidodon dickii	3102
Zebrasoma scopas	343

reef patterns, and light intensity changes. Variation in these parameters will challenge the performance of any classification technique. As shown in Fig. 2, the LCF-14 and LCF-15 data present all of these challenges.

Architecture

Our idea of applying deep CNN has two aims, (1) to find abstract and unique species-dependent fish features implicitly by learning task-specific information, and (2) to apply a suitable classification approach on the learned features. To achieve these goals, we propose a CNN architecture as shown in Fig. 3. Our network is a *K*-layered neural network, i.e., a mathematical parametric model. The first layer is called the input layer and represents the pixels of an image. The k^{th} layer is feature layer where each element of the layer, called a neuron, contributes to the output feature vector. The layers between input and feature layers are hidden layers. Hidden layers are divided into sub-sections denoted as $\Im_k^m$ with $k=1,2,...,K$ being the k^{th} layer and $m=1,2,\ldots M$ is the number of sub-sections in that layer. We stipulate layer $k=0$ as the input layer. The sub-section $\Im_k^m$ is called a feature map. Each layer has different number of feature maps. The number of neurons in each feature map is called the kernel size. In Fig. 3, the weights W_k, $k=1,2,\ldots,K$ are the weight matrices connecting the neurons of feature map m at layer k with the neurons of feature map m at layer $k-1$. However, layer $k=0$ accounts for the pixels of the raw input image. Such an arrangement ensures the detection of object features by the feature maps regardless of their position in the input image or preceding layer (LeCun et al. 2004). The output feature vector $\Im_K^m$ is combined with the selected neurons in the hidden convolution layers $\Im_k^m$, $k=1,2,\ldots,K-1$ to create the final feature representation. The motivation behind combining features from multiple layers is that lower (hidden) layers have more localized information while higher layers have more global information. Thus their combination encodes both local and global information about the fish species. In the supervised learning scenario, conventional CNN based approaches place emphasis on the features represented at the output layer. However, some less dominant local features (for instance, variation in tail shape or main body contour) may be ignored in the higher subsampling layers that select strong feature as a result of max-pooling. Therefore, preserving the information provided by lower level convolution layers is critical in our task.

Algorithm

Suppose input to the architecture, as shown in the Fig. 3, is an image X that is a 2D structure matrix in which each value acts as a pixel. The value of each feature map in the layer $k=1$ is calculated as

$$\Im_k^m=\sigma\left(\left(W_k^{ij}*X\right)+b_k\right) \quad (1)$$

where $(*)$ stands for 2D matrix convolution (LeCun et al. 2004) between weight matrix W_k^{ij} and input image X. The vector b_k is a constant valued bias vector normally used in neural networks to avoid the weight collapse and numerical instability as a result of training (Bengio and LeCun 2007). $\sigma(\cdot)$ is a sigmoid function to introduce a non-linear behavior

15415856, 2016, 9, Downloaded from https://isdpubs.onlinelibrary.wiley.com/doi/10.1002/tom3.10113 by Universita Di Salerno, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Fig. 2. (a) Sample images of various fish species (one per row) in LCF-14 and (b) LCF-15 datasets showing variation in image quality, background, and orientation of fish in each image.

in the network to model the data distribution of the input images which are naturally non-linear (Bengio and LeCun 2007; Bengio 2009). The function is given as $\sigma(z)=\left((1+e^{-z})^{-1}\right)$. As with (1), the value of each feature map in the layer $k=2,3,\ldots,K$ is achieved by

$$\Im_k^m=\sigma\left(\sum_{i=1}^{M_{k-1}}\left(W_k^{ij}*\Im_{k-1}^i\right)+b_k^m\right),\ k=2,3,\ldots K,$$
$$m=1,2,\ldots,M_k,\ i=1,2,\ldots,M_{k-1},\ j=1,2,\ldots,M_k \quad (2)$$

Interpreting (2) and Fig. 3, we can say that each feature map $\Im_k^m$ of layer k is calculated by adding the convolutions of feature maps of layer $k-1$ and weight matrices connecting the feature maps $\Im_k^m$ and $\Im_{k-1}^i$ where $i=1,2,\ldots,M_{k-1}$ accounts for the number of feature maps in the layer $k-1$. W_k^{ij} are the weights connecting M_{k-1} feature maps of layer $k-1$ and M_k feature maps of layer k. Therefore, there will be a total of $(M_k\times M_{k-1})$ weight matrices between the two layers. The main idea behind using the convolution operation is to exploit the correlative behaviour among the structures in an image, in terms of abstract non-linear features. There are additional layers shown in between the convolution layers in Fig. 3, called sub-sampling layers. These layers are not associated with any weight matrix. The purpose of these layers is to select the most dominant outputs and ignore the others. Using this approach, the dimension of the output of any layer can be reduced to enhance the computational efficiency together with sifting the information content from the pool of several neurons. The output of sub-sampling layer in reduced dimension is provided as input to the next convolution layer. In (2), the variable k denotes only the convolution layers and hence a presence of a sub-sampling

15415856, 2016, 9, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lom3.10113 by Universita Di Salerno, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

layer in the equation is ignored for simplicity. As shown in Fig. 3, the feature layer K is subjected to the fully connected neural network (Hinton and Salakhutdinov 2006) to compare the final output with the desired class label vector. A class label vector for N classes is an N-dimensional vector containing 1 at the location corresponding to the correct class label of the input image and all zeros elsewhere. Class label is a number from 1 to N given to each fish species. This is called one-of-n labels and is a common way of defining class labels in supervised learning (Hinton and Salakhutdinov 2006). Therefore,

$$\Im_f = W_f \times \Im_k^m, \;\; k=K, \;\; m=1,2,\ldots,M_K \tag{3}$$

where W_f are the connection weights between output and feature layer. $\Im_k^m$ are the feature maps of the K^{th} layer. The output $\Im_f$ is the result of matrix multiplication between W_f and $\Im_k^m$. The output vector $\Im_f$ is confined to the range [0–1] to be able to compare the result with the class labels. The network is trained using a standard error-backpropagation algorithm (Hinton et al. 2006). The error is defined as the Euclidean distance between the network output y and desired output d

$$E = \|y-d\|^2 \;. \tag{4}$$

The network parameters or weights are trained to minimize the error (4) which will force the CNN to learn the fish species characteristics as a result of supervised constraints. The desired output and consequently the network output y will be different for each fish-species. The symbol $\|\cdot\|$ is the $\mathcal{L}_2$ norm of vectors.

The images of fish in their natural underwater environment may encounter a number of variations. The location of fish as well as background coral reef, sea floor and plants cannot be confined to any fixed location in the image. Moreover, the changing light intensity in consecutive images/frames adds further variation in the videos. Since the CNN architecture is inspired by the biological cortex of the cat's eye (LeCun et al. 2004), each feature map in Fig. 3 (layer-1) is only associated with a small region of the input image containing fish through a set of weights, making it analogous to a receptive field in the eye. This continues for the feature maps of the higher layer, which take the feature maps of preceding layer as the input. Given the fish dataset is large enough to contain a variety of image conditions, this setting ensures the extraction of features that are invariant to the position and the pose of the fish. Each layer extracts unique and invariant features from the layer below it. Furthermore, the supervised learning criterion with class labels, as mentioned in (4), ensures the filtering of non-fish information from the images. Figure 4 illustrates the fish species-dependent feature extraction as the image propagates to the higher layers of CNN. The visible, first convolution layer acts in a similar fashion to an edge detection layer while the higher layer further extracts invariant yet useful features of the fish structure. As the image propagates to the higher layers, the dimension changes according to the designed architecture of CNN. It is evident from Fig. 4 that the information about fish edges diminishes as the image propagates

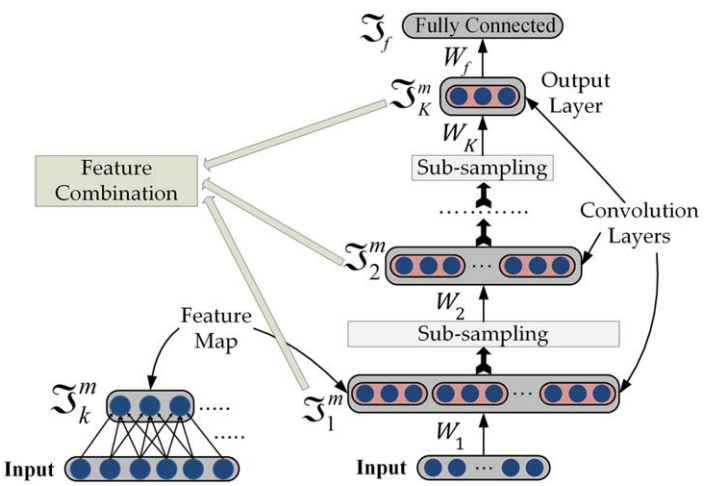

Fig. 3. Proposed deep Convolution Neural Network. On the bottom-left is a feature map whereas on the right is a complete CNN. The input to the network is a two dimensional image and the output is the label vector. The fish species-dependent feature vector is a special combination of the output of convolution layers. Using a large number of labeled training images, the parameters of the hidden layer are optimized so that when an image of a fish is passed through the network, it produces the correct species label.

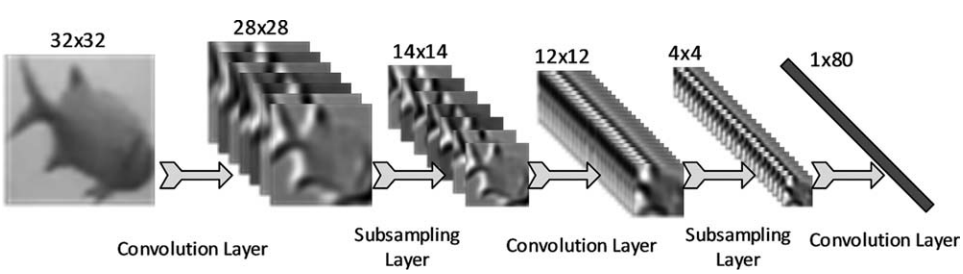

Fig. 4. Image propagation through CNN with each convolution layer as a unique feature extractor. Multiple images in layers are outcome of feature maps in that layer.

15415856, 2016, 9, Downloaded from https://asbpubs.onlinelibrary.wiley.com/doi/10.1002/fom3.10113 by Universita Di Salerno, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3. Data distribution for experimental protocol (no. of images).

Dataset	Original	Generated training set	Generated validation set	Generated testing set	Total generated
LCF-14	20,000	70,000	30,000	6956	106,956
LCF-15	29,000	85,700	32,000	7500	175,200

into the higher layers. Since the edges are not very well defined in terms of pixel strength due to blending with the water color, this information may not appear in the output feature vector. This might create problems in classification, as different species of fish may have a similar tail of fin structure but may exhibit small differences in the main body shape. Therefore, such an information loss may result in errors in classification. To solve this problem, we propose a feature combination approach to combine the useful information extracted by each layer. Given the convolution layer representations of CNN, the final high dimensional feature vector is calculated as

$$\mathcal{F}=\left(\Im_K^m,\ f\left(\Im_{K-1}^m\right),\ldots,f\left(\Im_2^m\right),\ f\left(\Im_1^m\right)\right)\ . \tag{5}$$

which represents the concatination of the output layer feature vector with selected hidden layer features. $f(\cdot)$ is a mathematical function to sift out unique fish species-dependent features learned by the neurons of hidden layers. We use Principal Component Analysis (PCA) to find the orthogonal Eigen vectors representing such features. The dimension of PCA components for all hidden convolution layers were kept the same as that of the output feature vector $\Im_K^m$. These features were then fed to standard classifiers like SVM and KNN to predict the fish species label.

Assessment

In this section, we present several experiments designed to investigate the fish classification accuracy of the proposed CNN neural architecture. We also compare our results with the current state-of-the-art image-based object recognition techniques that have been successfully applied for various computer vision tasks. The experimental settings for training of several algorithms are presented followed by the results and comparisons.

Experimental protocol

To learn fish-specific characteristics, we train the CNN with LCF-14 and LCF-15 fish datasets by organizing them into training, validation and testing sets. The training set is used by the CNN to train the network parameters for optimum performance on the recognition task in multiple iterations. The validation set is used to monitor the performance of the CNN during training. The validation set acts as an intermediate testing of a learning architecture after each training iteration. Once the CNN is trained, the parameters (i.e., network weights), are saved and used to measure performance on the testing set. All three sets (i.e., training, validation, and testing) are disjoint, which means that each set contains unique images of fish that are not used in any of the other sets. The original LCF-14 and LCF-15 datasets provide about 20,000 sample images each. However, to train deep architectures like CNN, it is always beneficial to include more environmental variability so that the architecture learns to suppress such anomalies and extracts class-dependent features in a supervised learning scenario (Raina et al. 2007). To achieve this, we replicate the images in the LCF-14 and LCF-15 datasets with induced image distortions. We use salt and pepper noise to degrade some images, change the light intensity levels, sharpen some images and add blurring to some images through average and Gaussian filtering (Boyle and Thomas 1988; Shapiro and Stockman 2001). With such duplication and degradation, a total of 100,000 images are generated for the LCF-14 dataset for fish including the original 20,000 images. Out of the 100,000 images, 70,000 are used for training and 30,000 are reserved for validation. For the testing set, we use the 6956 images as provided in the dataset, without any modification. Using a similar approach, we generated a total of 175,200 images for the LCF-15 dataset including the original 20,000 images and 9000 annotated images from videos. Out of the 175,200 images, 85,700 are used in training set, 32,000 for the validation set and remaining 7500 images are kept as the testing set. There is no original training and test split provided in LCF-15 dataset. Note that all species are included in training, validation and testing sets for the expanded datasets. Table 3 summarizes the overall dataset distribution.

The effectiveness of a machine learning approach is judged by its ability to correctly classify unknown and previously unseen query images. Unseen means that the particular image was not used at any stage of the training of the machine learning algorithm. This testing protocol is standard in the machine learning literature. However, we made the experimental protocol further challenging by performing cross-dataset classification. In other words, we trained our CNN on the two datasets separately and additionally tested them across the datasets. Thus, we report results for four experimental protocols that are (1) training on LCF-14 and testing on the same using its test set (2) training on LCF-15 and testing on the same using its test set (3) training on

15415856, 2016, 9, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lom3.10113 by Universita Di Salerno, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

LCF-14 and testing on LCF-15 test set (4) training on LCF-15 and testing on the LCF-14 test set. Improved performance on cross-dataset classification validates the effectiveness of fish species features extraction as a result of deep learning in the CNN.

The hyperparameters chosen to train our architecture are as follows: total number of layers are three, i.e., $K=3$. All input images to the CNN are confined to 32×32 pixel resolution given that the datasets contain images of different resolutions. All the images were resized using bilinear transformation (Smith 1981) and converted to greyscale. Color information is not used due to the fact that the colors of the fish are attenuated and are not accurately preserved in the dataset. Thus, we only retain the texture and shape information. We observed that increasing the resolution further did not provide any significant improvement in performance, but increased the computational cost. The first convolution layer, i.e., $k=1$, has eight feature maps with the kernel size of 5 each. The first subsampling layer down samples the output of the first layer by a factor of 2. The convolution layer $k=2$ has 24 feature maps with the kernel size of 3 each. The second sub-sampling layer implements the down sampling by the factor of 3. The last convolution layer $k=3$ has 80 feature maps with kernel size of 4 each. For each input image, the output feature vector has 80 dimensions, which is fed to the fully connected layer with a final output in the form of a class label vector. The CNN trained for LCF-14 has 10 output values for 10 fish classes while the CNN trained for LCF-15 has 15 outputs for 15 fish classes.

The fish species classification task is in fact a fish species identification task in our experiment. In other words, each predicted class of the test image is to be compared with the rest of the classes using the outcome of SVM or KNN classifier and the highest scoring class is selected as the final outcome.

Comparative study

In order to evaluate the effectiveness of our proposed fish species classification approach, we present a comparative study based on various other popular techniques recently used for automatic fish species classification. Support Vector Machines (SVM) based systems are among the state-of-the-art for various applications (Duan and Keerthi 2005; Wang and Casasent 2009; Huang et al. 2015). SVM is basically a binary classifier, i.e., it can discriminate between two classes. However, using one-against-one and one-against-all approaches, it can be used as a multi-class classifier as used in Duan and Keerthi (2005) for fish classification. In addition to SVM, we also present results based on k-nearest neighbor (KNN) classifier, a popular yet simple approach (Cover and Hart 1967; Altman 1992) that is based on exploiting the Euclidean distance among the features of various classes. Classification based on sparse representation of features (SRC) has recently been used for fish species classification in Wright et al. (2009) and Hsiao et al. (2014) with promising results. We have also used SRC in addition to SVM and KNN in our experiments. SRC, SVM, and KNN are trained on raw fish images. Here, we emphasize that the training and test protocols in all these approaches including CNN are kept exactly the same for a fair comparison. The tunable parameters for SVM, KNN, and SRC are chosen on the basis of the training and validation sets (as used for CNN) for their best performance.

For a baseline system, we have also used Principal Component Analysis (PCA) of raw fish images. PCA is generally used for dimensionality reduction of the data. As a result, we choose 10% of the principal components as new features and classify these using the standard SVM and KNN. For score measurement, we have adopted three measures, i.e., Average Count (AC), Average Precision (AP), and Average Recall (AR) (Huang et al. 2015).

$$\mathrm{AC}=\frac{\sum_{j=1}^{c}\mathrm{True\ Positive}_j}{\sum_{j=1}^{c}\left(\mathrm{True\ Positive}_j+\mathrm{False\ Positive}_j\right)} \tag{6}$$

$$\mathrm{AP}=\frac{1}{c}\sum_{j=1}^{c}\left(\frac{\mathrm{True\ Positive}_j}{\mathrm{True\ Positive}_j+\mathrm{False\ Positive}_j}\right) \tag{7}$$

$$\mathrm{AR}=\frac{1}{c}\sum_{j=1}^{c}\left(\frac{\mathrm{True\ Positive}_j}{\mathrm{True\ Positive}_j+\mathrm{False\ Negative}_j}\right) \tag{8}$$

In (6), (7) and (8), c accounts for the total number of classes. The performance comparison between various approaches is shown in Table 4 in terms of AC, AP, and AR. We dub our technique as CNN-KNN and CNN-SVM, i.e., fish species label prediction by KNN and SVM based on the features learned by CNN. Similarly, PCA-KNN and PCA-SVM denote prediction of class label by KNN and SVM classifiers based on the PCA features calculated on raw images of fish. The PCA dimensionality was chosen to be 10 based on the best cross-validation performance across all species of fish.

It is evident from the comparison in Table 4 that CNN-KNN and CNN-SVM perform better when compared to all other techniques if training and testing both are done on LCF-14 fish dataset. CNN-SVM returns the highest classification success rate. PCA-SVM on the other hand shows better performance as compared to KNN, SVM, PCA-KNN, and SRC. Similarly, when training and testing is performed on LCF-15 fish data, CNN-KNN outperforms all other approaches while CNN-KNN performs marginally better than CNN-SVM. These outcomes depict the challenging nature of LCF-15 fish dataset that is marked by higher degradations in terms of light intensity and blurriness together with background confusion with objects of interest, i.e., fish. The CNN trained on LCF-15 fish data is forced to extract fish-dependent features for the challenging environment in the supervised learning scenario. Hence it is capable of suppressing the information unrelated to fish. Although in the same-dataset train-test protocol, CNN based classification outperforms all the others

15415856, 2016, 9, Downloaded from https://isdpubs.onlinelibrary.wiley.com/doi/10.1002/lom3.10113 by Universita Di Salerno, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 4. Performance comparison (percentage values) on same-dataset with the LCF-14- LCF-14 and LCF-15- LCF-15 train-test protocol. Best scores are shown in bold.

	AC		AP		AR	
Method	Train on LCF-14 Test on LCF-14	Train on LCF-15 Test on LCF-15	Train on LCF-14 Test on LCF-14	Train on LCF-15 Test on LCF-15	Train on LCF-14 Test on LCF-14	Train on LCF-15 Test on LCF-15
SVM	83.94	63.41	82.21	63.41	81.12	65.23
KNN	84.56	81.55	83.52	81.50	81.02	83.50
SRC	84.04	26.81	83.75	26.81	80.77	38.02
PCA-SVM	88.54	82.33	86.89	82.74	85.63	82.33
PCA-KNN	86.02	81.37	85.20	81.37	82.71	82.99
CNN-SVM	**96.75**	92.87	**94.47**	91.64	**95.70**	90.97
CNN-KNN	96.23	**93.65**	93.44	**91.99**	95.03	**91.25**

Table 5. Performance comparison (percentage values) of cross-dataset with the LCF-14- LCF-15 and LCF-15- LCF-14 train-test protocol. For cross-dataset experiments, only five fish species common to LCF-14 and LCF-15 were considered. Best scores are shown in bold.

	AC		AP		AR	
Method	Train on LCF-14 Test on LCF-15	Train on LCF-15 Test on LCF-14	Train on LCF-14 Test on LCF-15	Train on LCF-15 Test on LCF-14	Train on LCF-14 Test on LCF-15	Train on LCF-15 Test on LCF-14
SVM	40.80	76.32	40.80	75.12	57.30	90.26
KNN	40.64	82.09	40.64	81.80	60.63	90.01
SRC	44.63	84.02	44.63	88.10	61.35	60.20
PCA-SVM	34.16	80.30	34.16	78.60	54.34	94.29
PCA-KNN	39.88	80.19	39.88	79.66	61.33	90.90
CNN-SVM	**65.36**	**97.41**	**65.36**	**97.18**	74.50	**98.43**
CNN-KNN	63.88	97.22	63.88	96.94	**75.71**	97.99

in general, other algorithms produce reasonable scores except SRC which fails to cope with the challenging variability in LCF-15 dataset. Moreover, it is critical to monitor the robustness of any algorithm and to do critical analysis whether there is any overfitting on a particular dataset by the learning algorithms.

To achieve this in the performance evaluation, two more challenging cases are investigated, i.e., training on LCF-14 fish data and testing on LCF-15 test set and vice versa. The cross-dataset testing evaluates the robustness and generalization ability of the various approaches. It should be noted that the cross-dataset performance can be measured only for the five common classes in LCF-14 and LCF-15 datasets (see Table 2). From Table 5 it is evident that CNN based classifiers, especially CNN-SVM, outperforms all approaches with a large margin in both cross-dataset experiments. It is interesting to notice that all techniques perform better in the case when training is done on the noisy and poorer quality LCF-15 dataset and testing is done on LCF-14. This implies that the machine learning algorithms are able learn to extract useful information regarding fish species in the presence of noise and image distortions and perform well even when those challenges are not present in LCF-14 dataset. On the other hand, when training is done on LCF-14 and testing is performed on LCF-15, the performance is comparatively poor as the variability incorporated in LCF-15 is totally unknown to all classifiers trained either directly on raw images or features extracted by PCA and CNN. In both cross-dataset experiments, SRC produces better scores in terms of AC and AP as compared to SVM, KNN, PCA-SVM, and PCA-KNN but it lags behind in AR. Still, CNN-KNN and CNN-SVM yield the best results among all algorithms, which shows that it is robust enough against overfitting on a specific dataset and learns to extract invariant fish species-dependent features. In contrast, all other algorithms produce worse results while testing on LCF-15 in cross-dataset setup as compared to when they were trained and also tested on LCF-15 (see Table 4).

These experiments strengthen our claim that stand alone shallow architectures like SVM, KNN, and SRC, when trained on either raw images or on features extracted through another shallow mathematical formulation like PCA, fail to accommodate unique, task-specific features in these experiments. On the other hand, our approach based on deep

15415856, 2016, 9, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lom3.10113 by Universita Di Salerno, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

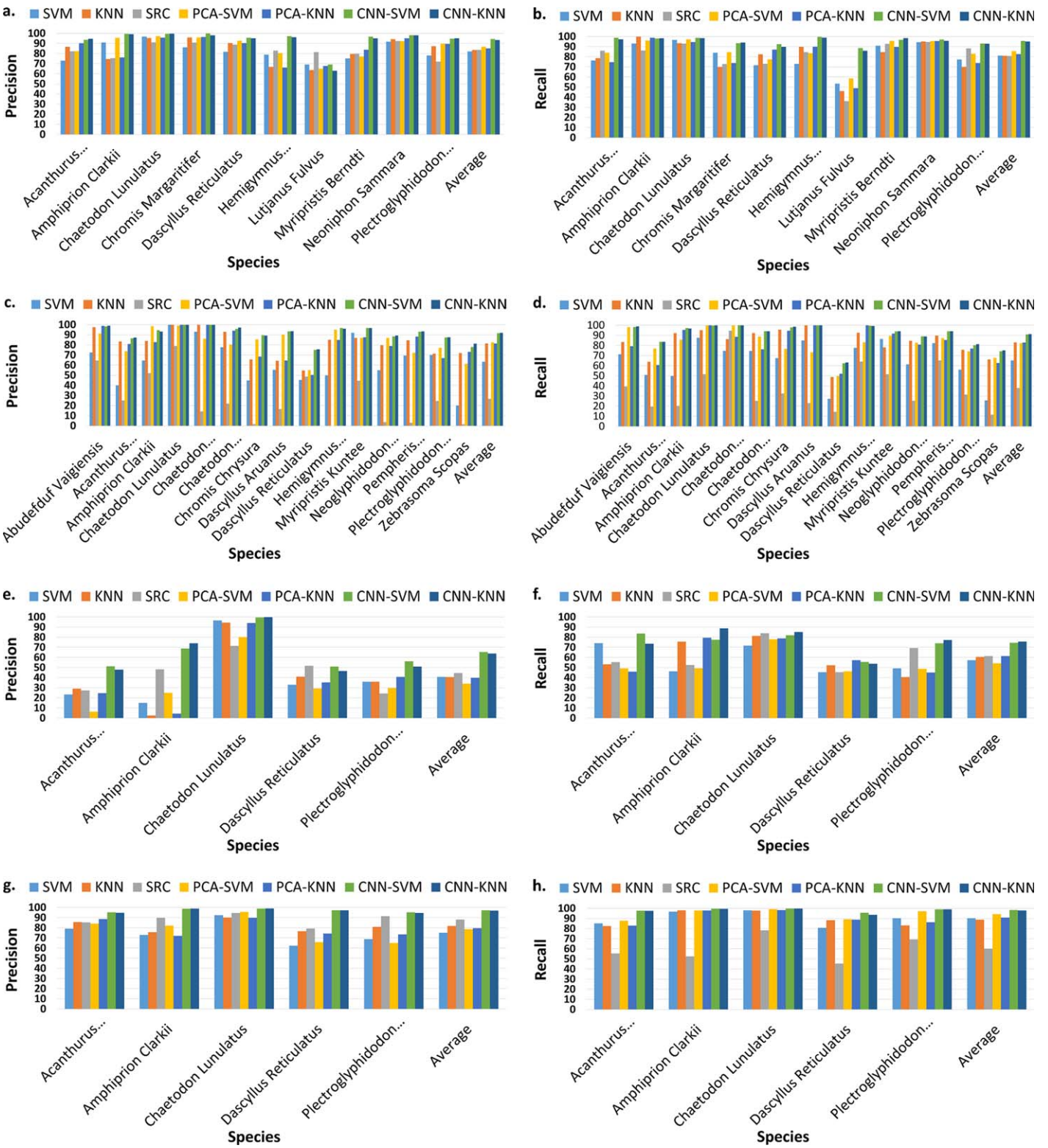

Fig. 5. Performance in terms of %Precision and %Recall for individual fish species. (a, b) Precision and Recall of Same-dataset train-test protocol on LCF-14. (c, d) Precision and Recall of Same-dataset train-test protocol on LCF-15. (e, f) Precision and Recall of Cross-dataset train-test protocol on LCF-14-LCF-15. (g, h) Precision and Recall of Cross-dataset train-test protocol on LCF-15-LCF-14. Cross-dataset graphs (last two rows) are shown for five common species in LCF-14 and LCF-15.

15415856, 2016, 9, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lom3.10113 by Universita Di Salerno, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

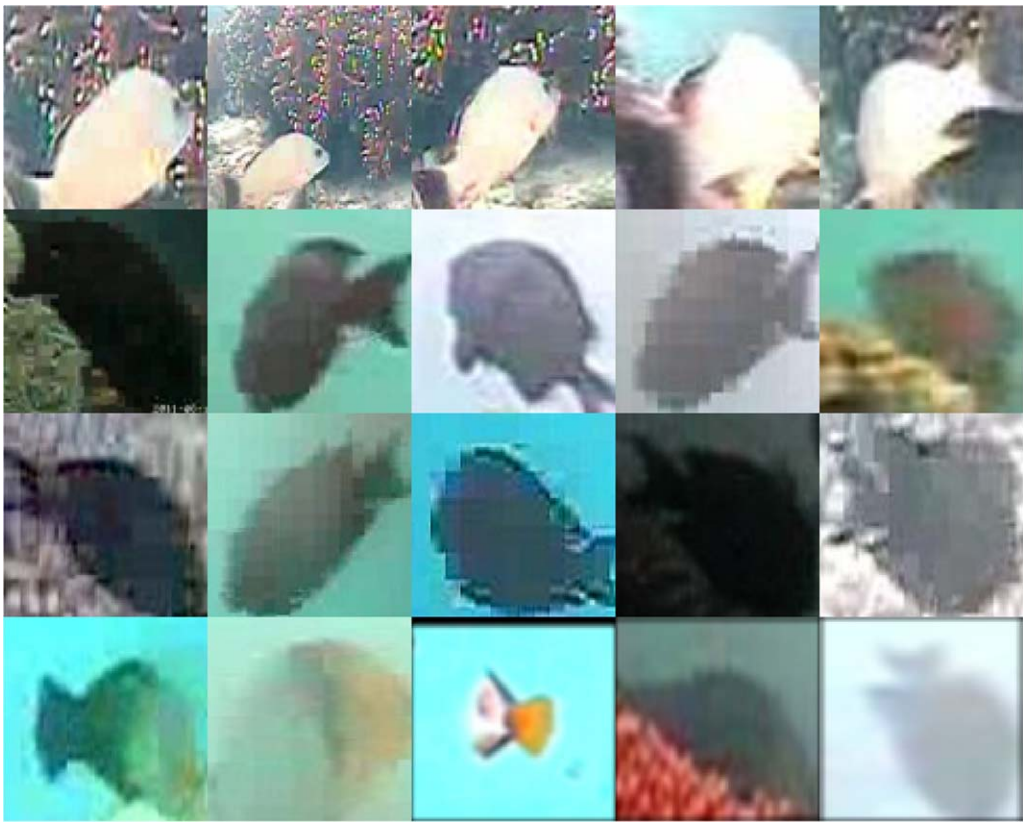

Fig. 6. Examples of fish misclassified by all techniques arranged according to species. (First Row) *L. fulvus* from LCF-14. (Second Row) *A. nigrofuscus* from LCF-14. (Third Row) *A. nigrofuscus* from LCF-15. (Last Row) *H. melapterus* from LCF-15.

learning compensates for the environmental variability and is successful in extracting fish species-specific features. The overall performance of CNN-KNN and CNN-SVM in all four experimental setups remains favorable both in the same and the cross-dataset experimental protocols. Tables 4, 5 tabulate the performance for the average precision, recall and count for all fish species of the LCF-14 and LCF-15 datasets. The test sets used in all experiments are the ones mentioned in Table 3. Figure 5 gives precision and recall of individual fish species for all the seven techniques used in the experiments. The first row of Fig. 5 is the precision (left) and recall (right) with LCF14-LCF14 train-test settings. The second row is the performance with the LCF15-LCF15 train-test setup. The third row is the cross-dataset experiment on five common species with the LCF14-LCF15 train-test scenario. Similarly, the last row is the performance for the LCF15-LCF14 train-test setting. Figure 6 shows the examples of fish species that are misclassified and resulted in the worst performances by all algorithms relative to other species. The first two rows are *Lutjanus fulvus* and *Acanthurus nigrofuscus* of LCF-14 in self and cross-data testing respectively. The last two rows are *A. nigrofuscus* and *Hemigymnus melapterus* of LCF-15 in self and cross-data testing respectively. *L. fulvus* examples are misclassified (first row) apparently due to over exposure by the light source mounted with the camera. *H. melapterus* (fourth row) images are either too dark or extremely blurred, which resulted in misclassification of this species. *A. nigrofuscus* (second and third rows) in both LCF-14 and LCF-15 is not correctly classified due to the same reasons. Therefore, extremely high variability in terms of light and blurriness is responsible for the relatively poor performance. It should be noted that CNN still performs better than all other approaches for all these species as evident in Fig. 5.

In another set of experiments, we test robustness of the algorithms to image blur and noise. Test images of LCF-14 are artificially deteriorated with white Gaussian noise and blurring, a technique proposed in Khan et al. (2015). We choose LCF-14 for this experiment as the images are cleaner compared to those from LCF-15, which already exhibit noise and blurring due to poor quality of images and murkiness of water. As exercised by Khan et al. (2015), we corrupt the test images with 14 different levels of noise and blurring. The algorithms already trained on clean images of LCF-14 are used to recognize fish species in the corrupted LCF-14 images. Figure 7 shows sample images of two fish generated with different levels of blurring and noise. The performance comparison in terms of overall fish species classification accuracies by all machine learning approaches, reported in

15415856, 2016, 9, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lom3.10113 by Universita Di Salerno, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

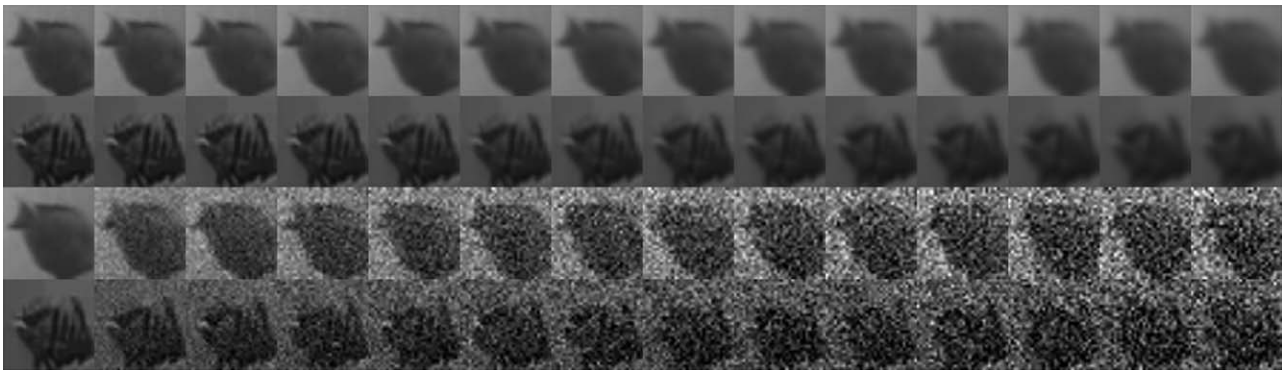

Fig. 7. Sample images of fish with increased levels (from left to right) of Gaussian blurring (first two rows) and Gaussian noise (last two rows).

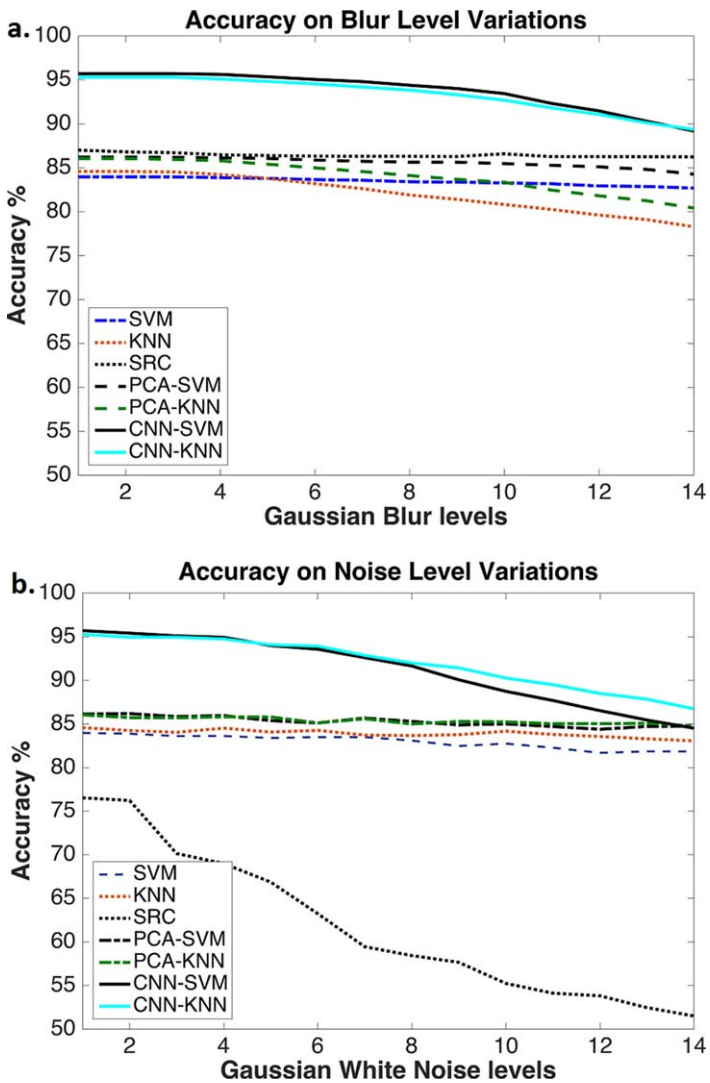

Fig. 8. Fish classification accuracy on various levels of added noise and blurring. Accuracy graphs by various algorithms on LCF-14 when test data is blurred (a) or corrupted with Gaussian white noise (b).

this article, is also given in Figure 8 on the modified LCF-14 test dataset with artificial corruption. It is evident from the results that CNN-KNN and CNN-SVM generate higher accuracies which shows the robustness to severe degradations of the features learned via CNN. KNN is most sensitive to increasing blur levels followed by PCA-KNN. SRC accuracy decays sharply with the increasing levels of noise, although it is second to only CNN based systems in the case of blurring. CNN-SVM lags CNN-KNN with a minute difference while SVM is behind KNN when noise intensity increases. Therefore, it can be concluded that KNN-based systems are more sensitive to blurring and SVM-based systems are more sensitive to noise degradation in general.

Discussion and conclusions

The demand for monitoring and sampling of fish populations in lakes and oceans is inevitable due to its importance in estimating fresh water body and marine conservation status. Before advanced approaches of computer vision in identification and classification using underwater cameras, fish populations were manually sampled and tagged. This practice, which is still popular, demands time and labor costs that are undesirable in this age of rapid marine exploration and real time monitoring. Automatic fish species classification has direct influence on observing and studying underwater ecosystems, which in turn affects our socio-economic activities. Coastal areas provide ideal locations for fish life to flourish as nutrients from deep ocean beds are deposited there as a result of natural oceanic movement (http://www.ecologyandsociety.org). This results in establishment of industry and economic activities related to fisheries in coastal areas. Climate change, water pollution and over fishing are the factors responsible for the observed declines in fish populations of specific species, a problem that needs to be continuously monitored by marine biodiversity conservationists. Monitoring and management are especially critical in coastal areas where human communities get direct benefits from fisheries, but also where the pressure on fish stocks is the greatest. Failure in adopting efficient and cost effective

15415856, 2016, 9, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lom3.10113 by Universita Di Salerno, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ways to sample fish populations may result in extinction of certain fish species and therefore, disruption of the entire marine ecosystem. One example is the severe decline in several salmon species in Northwest Pacific that is contributing to the coast wide closure of fisheries (http://www.ecologyandsociety.org). Hence, a timely warning to regulatory authorities and Government bodies is necessary to implement and impose strict rules for preservation of endangered fish species. Efficient fish identification and classification techniques to keep aquatic life under surveillance can also provide indirect evidence of the degradation of marine ecosystems such as coral reefs and coastal mangroves, both of which are very sensitive to pollution and climate change.

We have proposed a deep architecture in the form of a convolution neural network employed for fish species classification on two benchmark datasets, namely the LCF-14 and LCF-15. Computer vision tasks in general and the fish classification task in particular pose great challenges for the machine learning community to automatically recognize the object of interest from the video sequences in the presence of environmental variability, visual distortions, and image noise. These factors decrease the performance of machine learning approaches, which is directly related to the quality of features used to represent an object in an image (Bengio and LeCun 2007; Bengio 2009). The object of interest in an image in the presence of variability and distortion represents the input space with high non-linearity, making it difficult to model the objects of interest efficiently and effectively in the feature space (Hinton and Salakhutdinov 2006; Larochelle et al. 2009). However, this can be rectified by using non-linear automatic learning systems, like multi-layer neural networks such as CNNs. Each non-linear hidden layer in the network is the input to the next layer making it specifically non-linear and highly complex. The depth of the network is, therefore, related to effectively encoding the non-linearity of the input data. Hence, our network and its parameters are designed to match the non-linearity of the data. The supervised training scenario ensures automatic learning of complex, non-linear and discriminative features (Bengio 2009).

Shallow architectures like SVM, KNN, and SRC do not ensure robust data representation, should the data exhibit degradation and variability. These techniques either fail to perform well or over-fit to a specific environment or dataset. Our results show that when degraded and highly variable data is used in training, conventional shallow machine learning techniques fail to extract useful information from the data and perform poorly when similar variability is encountered in the test data. As shown in our results, KNN, SVM, SRC, PCA-KNN, and PCA-SVM perform relatively poorly in cross-dataset experiments, i.e., when they are trained on LCF-14 and tested on the LCF-15 dataset or vice versa. On the other hand, the results are improved in the same-dataset train-test protocol. This is because of the over-fitting phenomena on the same dataset environment and failing to cope with the variability in test datasets that are unseen during training. CNN-SVM and CNN-KNN on the other hand avoid overfitting to a large extent as the results are much better as compared to the other algorithms in the case when training is done on LCF-14 and testing on LCF-15. In fact, when the CNN is trained on the noisy LCF-15 dataset and tested on the cleaner LCF-14 data, the outcome is even better than the case when both the training and testing is done on LCF-14 dataset. The CNN utilizes the degraded data in training and adapts so as to cancel out the distortions and noise if they appear in the test data, thereby learning to better extract fish species-dependent information. To sum up, too much corruption as in the case with LCF-15 would greatly hamper the learning of useful information in the training of shallow architectures. Deep architectures on the other hand have more capability to filter out distortions. Our experiments with artificially generated test data of LCF-14 with degradation using blurring and noise also supports this idea as the CNN based classifiers produce stable results. These observations are also consistent with the experimentations in Bengio and LeCun (2007).

Our architecture is related to LeNET (LeCun et al. 2004), a pioneering work to implement deep neural network in the form of a convolution neural network. Although CNN has been used in various machine learning tasks including generic object recognition in images, handwriting recognition and speech signal processing (Lee et al. 2009) with favorable results, we have employed CNN using a different strategy. In our case, the network is trained with a fully connected overhead classification layer in a supervised learning style. After training we discard the fully connected layer and use the last convolution layer as output feature vector to represent the input data. In addition, some information from the hidden layers is also utilized and represented in the final feature vectors. This approach turns out to be beneficial in the case of under water fish classification in unconstrained surroundings. Sometimes the shape of fish, especially the edges of the body, fins and tail do not exhibit high contrast due to matching of fish color with background, murkiness of water and low light conditions. Such images, when subjected to training in a strong supervised learning architecture such as the CNN, the less dominant yet important features are ignored in the deeper layers of neural network. Therefore, we have devised a technique to select weak fish species-dependent invariant features in lower hidden layers and append with the highly non-linear and dominant fish features learned in the output layer.

CNN based architectures, with many hidden layers with regularization and maximum valued neuron pooling constraints, have been recently used for generic object recognition from images (Razavian et al. 2014; Simonyan and Zisserman 2015) reporting state-of-the-art results. Similar to the proposed architecture, these networks utilize the output

15415856, 2016, 9, Downloaded from https://isdpubs.onlinelibrary.wiley.com/doi/10.1002/tom3.10113 by Universita Di Salerno, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

layer as a feature vector. Such a setup might be useful when the objects of interest in images are sharp and distinct, and so dominate the input space. However, we also utilize the hidden layer information together with the output layer features to improve the fish classification performance. Another deep network was proposed by Ouyang and Wang (2013) for pedestrian detection. This approach also used the representation of hidden layers in feature extraction, but sets of neurons in hidden layers were trained to detect specific parts of the body and combined to get the overall pedestrian representation. In this case, each hidden layer is associated with a specific part of human body to be detected, which enables their system to detect the general human body shape irrespective of identity. Therefore, the architecture proposed by Ouyang and Wang (2013) cannot be directly used for class-based recognition of fish species. Further, walking or standing pedestrians exhibit very limited variation in poses, which is suitable for their application. In our case, freely swimming fish may appear in any possible orientation, moving sideways, upwards or downwards, so we cannot associate a specific set of neurons in any layer with a specific part of the fish.

To our knowledge, this is the first attempt to utilize a deep learning CNN for the difficult task of underwater fish species classification with state-of-the-art results on the LCF-14 and LCF-15 datasets. With no pre-processing of images and using several complicated image processing techniques before applying some machine learning algorithm for classification, deep learning using CNN has proved to be suitable recipe for creating a single learning module applicable for raw images to extract fish species-dependent features. Based on the results presented here, this technique is both efficient and effective compared to similar work reported in the published literature (Rova et al. 2007; Fablet et al. 2009; Hsiao et al. 2014). Blanc et al. (2014) presents work on LCF-14 dataset, using videos to first detect and then classify the fish species using fish species-dependent features trained using an SVM classifier. The features are explicitly extracted in terms of descriptors invariant to light intensity and color variation. On LCF-14 test data Blanc et al. (2014) reports average precision and recall of more than 55% and 50%, respectively. This relatively lower performance is due to the twofold classification, i.e., detection in videos followed by species recognition. In such cases, detection errors are propagated to the recognition module hence resulting in overall lower recognition scores. In this article, we have reported results on LCF-14 using SVM classifiers trained with raw images and PCA features, demonstrating that detection of fish is not consistent. We perform training and testing of algorithms on still images where the fish have already been detected, as our aim is to demonstrate the effectiveness of robust feature extraction by highly complex and nonlinear CNN models. However, it is clear that good quality features will contribute towards other tasks such as fish detection in videos.

The experiments conducted in this article utilize MATLAB for the algorithm development. Freely available MATLAB toolboxes were used for SVM (https://www.csie.ntu.edu.tw/~cjlin/libsvm/#matlab), KNN and SRC (https://sites.google.com/site/sparsereptool/) and CNN development (http://uk.mathworks.com/matlabcentral/fileexchange/38310-deep-learning-toolbox) implementations. The computation was done on Intel Core i5 2.5 GHz processor with 16GB RAM. The training of CNN took 5–6 h. However, during testing, each fish image takes about 1 ms for classification. No special hardware is required for CNN training or classification. However, the use of GPUs (Graphics Processing Units) can reduce the CNN training time.

To conclude, we have presented and employed CNN deep architecture for the task of fish species classification using two benchmark datasets, LifeCLEF14 and LifeCLEF15. Through same-dataset and cross-dataset training-testing experimental protocols, we have shown that CNN outperforms various other recent approaches employed for fish species classification. Consequently, the performance reported for the classification problem is the best reported so far for the datasets we used.

In future, we aim to further enhance the CNN architecture by designing a better loss function. To demonstrate generalization, further comparative studies with other deep architectures for fish detection and species classification will be performed on several fish image and especially video datasets acquired in the unconstrained underwater environment. It would be interesting to investigate the performance improvement by including the color information in training a deep architecture like CNN.

References

Altman, N. S. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. Am. Stat. **46**: 175–185. doi:10.1080/00031305.1992.10475879

Bengio, Y. 2009. Learning deep architectures for AI. Found. Trends Mach. Learn. **2**: 1–127. doi:10.1561/2200000006

Bengio, Y., and Y. LeCun. 2007. Scaling learning algorithms towards AI. In L. Bottou, O. Chapelle, D. DeCoste, & J. Weston [eds.], Large-scale kernel machines. MIT Press.

Blanc, K., D. Lingrand, and F. Precioso. 2014, November. Fish species recognition from video using SVM classifier. In 3rd ACM International Workshop on Multimedia Analysis for Ecological Data, ACM, 1–6.

Boyle, R., and R. Thomas. 1988. Computer vision: A first course. Blackwell Scientific Publications.

Cappo, M. E., E. Harvey, H. Malcolm, and P. Speare. 2003. Potential of video techniques to monitor diversity, abundance and size of fish in studies of marine protected areas. In J. P. Beumer, A. Grant, & D. C. Smith [eds.], Aquatic protected areas. What works best and how do we know? (Cairns ed.). Queensland: University of Queensland. 1: 455:464.

15415856, 2016, 9, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lom3.10113 by Universita Di Salerno, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Cover, T. M., and P. E. Hart. 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**: 21–27. doi: 10.1109/TIT.1967.1053964

Duan, K., and S. S. Keerthi. 2005. Which is the best multiclass SVM method? Proceedings of the Sixth International Workshop on Multiple Classifier Systems, p. 278–285.

Fablet, R., R. Lefort, I. Karoui, L. Berger, J. Masse, C. Scalabrin, and J. M. Boucher. 2009. Classifying fish schools and estimating their species proportions in fishery-acoustic surveys. ICES J. Mar. Sci. **66**: 1136–1142. doi:10.1093/icesjms/fsp109

Harvey, E., and M. Shortis. 1995. A system for stereo-video measurement of sub-tidal organisms. Mar. Technol. Soc. J. **29**: 10–22.

Hinton, G., and R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. Science **313**: 504–507. doi:10.1126/science.1127647

Hinton, G., S. Osindero, and Y. The. 2006. A fast learning algorithm for deep belief nets. Neural Comput. **18**: 1527–1554. doi:10.1162/neco.2006.18.7.1527

Hsiao, Y., C. Chen, S. Lin, and F. Lin. 2014. Real-world underwater fish recognition and identification using sparse representation. Ecol. Inform. **23**: 13–21. doi: 10.1016/j.ecoinf.2013.10.002

Huang, P. X., B. J. Boom, and R. B. Fisher. 2015. Hierarchical classification with reject option for live fish recognition. Mach. Vision Appl. **26**: 89–102. doi:10.1007/s00138-014-0641-2

Jennings, S., and M. J. Kaiser. 1998. The effects of fishing on marine ecosystems. Adv. Mar. Biol. **34**: 201–352. doi: 10.1016/S0065-2881(08)60212-6

Khan, A. M., R. Mikut, and M. Reischl. 2015. A benchmark dataset to evaluate the illumination robustness of image processing algorithms for object segmentation and classification. PloS ONE **10**: e0131098. doi: 10.137/journal.pone.0131098

Larochelle, H., D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. 2007. An empirical evaluation of deep architectures on problems with many factors of variation. In Proceedings of ICML, p. 473–480.

Larochelle, H., Y. Bengio, J. Louradour, and P. Lamblin. 2009. Exploring strategies for training deep neural networks. J. Mach. Learn. Res. **10**: 1–40. doi:10.1145/1577069.1577070

LeCun, Y., F. Huang, and L. Bottou. 2004. Learning methods for generic object recognition with invariance to pose and lighting. In Proceedings of IEEE CVPR, p. 97–104.

Lee, H., R. Grosse, R. Ranganath, and A. Ng. 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In Proceeding of the ICML, p. 609–616.

Mallet, D., and D. Pelletier. 2014. Underwater video techniques for observing coastal marine biodiversity: A review of sixty years of publications (1952–2012). Fish. Res. **154**: 44–62. doi:10.1016/j.fishres.2014.01.019

McLaren, B. W., T. J. Langlois, E. S. Harvey, H. Shortland-Jones, and R. Stevens. 2015. A small no-take marine sanctuary provides consistent protection for small-bodied by-catch species, but not for large-bodied, high-risk species. J. Exp. Mar. Biol. Ecol. **471**: 153–163. doi:10.1016/j.jembe.2015.06.002

Mika, S., G. Ratsch, J. Weston, B. Scholkopf, and K. R. Mullers. 1999. Fisher discriminant analysis with kernels. In IEEE International Workshop on Neural Networks for Signal Processing, p. 41–48.

Murphy, H. M., and G. P. Jenkins. 2010. Observational methods used in marine spatial monitoring of fishes and associated habitats: A review. Mar. Freshw. Res. **61**: 236–252. doi:10.1071/MF09068

Ouyang, W., and X. Wang. 2013. Joint deep learning for pedestrian detection. In IEEE International Conference on Computer Vision, p. 2056–2063.

Raina, R., A. Battle, H. Lee, B. Packer, and A. Ng. 2007. Self-taught learning: Transfer learning from unlabeled data. In Proceedings of ICML, p. 759–766.

Razavian, A. S., H. Azizpour, J. Sullivan, and S. Carlsson. 2014. CNN features off-the-shelf: An astounding baseline for recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

Rova, A., G. Mori, and L. M. Dill. 2007. One fish, two fish, butterfish, trumpeter: Recognizing fish in underwater video. In IAPR Conference on Machine Vision Applications, p. 404–407.

Shapiro, L. G., and G. C. Stockman. 2001. Computer vision. Prentice Hall.

Shortis, M., E. Harvey, and D. Abdo. 2009. A review of underwater stereo-image measurement for marine biology. *In* R. N. Gibson, R. J. A. Atkinson and J. D. M Gordon [eds.], Oceanography and marine biology: An annual review. CRC Press. 47: 257-292. doi:10.1201/9781420094220.ch6

Shortis, M. R., and others. 2013. A review of techniques for the identification and measurement of fish in underwater stereo-video image sequences. In Proceedings of Videometrics, Range Imaging, and Applications XII, SPIE Vol. 8791, paper 0G. The International Society for Optical Engineering, Bellingham WA, USA.

Simonyan, K., and A. Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In Computer Vision and Pattern Recognition (CVPR), arXiv: 1409.1556v6.

Smith, P. R. 1981. Bilinear interpolation of digital images. Ultramicroscopy **6**: 201–204. doi:10.1016/S0304-3991(81)80199-4

Spampinato, C., and others. 2010. Automatic fish classification for underwater species behavior understanding. ACM Workshop on Analysis And Retrieval of Tracked Events and Motion in Imagery Streams, 45–50. doi:10.1145/1877868.1877881

Storbeck, F., and B. Daan. 2001. Fish species recognition using computer vision and a neural network. Fish. Res. **51**: 11–15. doi:10.1016/S0165-7836(00)00254-X

15415856, 2016, 9, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lom3.10113 by Universita Di Salerno, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Strachan, N. J. C., and L. Kell. 1995. A potential method for the differentiation between haddock fish stocks by computer vision using canonical discriminant analysis. ICES J. Mar. Sci. **52**: 145–149. doi:10.1016/1054-3139(95)80023-9

Turk, M., and A. Pentland. 1991. Eigenfaces for recognition. J. Cogn. Neurosci. **3**: 71–86. doi:10.1162/jocn.1991.3.1.71

Wang, Y., and D. Casasent. 2009. A support vector hierarchical method for multi-class classification and rejection. In Proceedings of IJCNN, p. 3281–3288.

Wright, J., A. Y. Yang, and A. Ganesh. 2009. Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **23**: 210–227. doi:10.1109/TPAMI.2008.79

Acknowledgments

The authors acknowledge support from the Australian Research Council Grant LP110201008, which provided the primary funding for this study in addition to UWA Research Collaboration Award (RCA) grant. Ajmal Mian was supported by the Australian Research Council Fellowship DP110102399.

Submitted 08 February 2016
Revised 18 April 2016
Accepted 30 April 2016

Associate editor: Paul Kemp

15415856, 2016, 9, Downloaded from https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lom3.10113 by Universita Di Salerno, Wiley Online Library on [22/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License