ICES Journal of
Marine Science

ICES CIEM
International Council for the Exploration of the Sea
Conseil International pour l'Exploration de la Mer

ICES Journal of Marine Science (2016), 73(10), 2737–2746. doi:10.1093/icesjms/fsw106

Original Article

Fish identification from videos captured in uncontrolled underwater environments

Faisal Shafait[1,2]*, Ajmal Mian[1], Mark Shortis[3], Bernard Ghanem[4], Phil F. Culverhouse[5], Duane Edgington[6], Danelle Cline[6], Mehdi Ravanbakhsh[3], James Seager[7], and Euan S. Harvey[8]

[1]*School of Computer Science and Software Engineering, The University of Western Australia, Perth, Western Australia, Australia*
[2]*School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad, Pakistan*
[3]*Mathematical and Geospatial Sciences, RMIT University, Melbourne, Victoria, Australia*
[4]*Mathematical and Geospatial Sciences, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia*
[5]*CNRS, University of Plymouth, Plymouth, Devon PL4 8AA, UK*
[6]*Monterey Bay Aquarium Research Institute, Moss Landing, CA, USA*
[7]*SeaGIS Pty Ltd, Melbourne, Australia*
[8]*Department of Environment and Agriculture, Curtin University, Perth, Australia*

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

Shafait, F., Mian, A., Shortis, M., Ghanem, B., Culverhouse, P. F., Edgington, D., Cline, D., Ravanbakhsh, M., Seager, J., and Harvey, E. S. Fish identification from videos captured in uncontrolled underwater environments. – ICES Journal of Marine Science, 73: 2737–2746.

Received 18 August 2015; revised 14 April 2016; accepted 15 April 2016; advance access publication 18 July 2016.

There is an urgent need for the development of sampling techniques which can provide accurate and precise count, size, and biomass data for fish. This information is essential to support the decision-making processes of fisheries and marine conservation managers and scientists. Digital video technology is rapidly improving, and it is now possible to record long periods of high resolution digital imagery cost effectively, making single or stereo-video systems one of the primary sampling tools. However, manual species identification, counting, and measuring of fish in stereo-video images is labour intensive and is the major disincentive against the uptake of this technology. Automating species identification using technologies developed by researchers in computer vision and machine learning would transform marine science. In this article, a new paradigm of image set classification is presented that can be used to achieve improved recognition rates for a number of fish species. State-of-the-art image set construction, modelling, and matching algorithms from computer vision literature are discussed with an analysis of their application for automatic fish species identification. It is demonstrated that these algorithms have the potential of solving the automatic fish species identification problem in underwater videos captured within unconstrained environments.

Keywords: computer vision, fish classification, fish identification, image analysis, image sets, species recognition.

Introduction

Biomass estimation of various fish species is of key importance to marine scientists, environmental conservation agencies, as well as fisheries. Changes in the distribution and relative abundance of fish species in different parts of the oceans can indicate natural or anthropogenic changes in ecological conditions some of which can be managed with appropriate actions (e.g. enforcement of species-specific fishing bans and quotas). Regular surveys are performed in the oceans to estimate the relative biomass and distribution of target or indicator fish species. Two key technologies used for this purpose are acoustic surveying (Fernandes, 2009; Fablet *et al.*, 2009) and video-based monitoring (Harvey and Shortis, 1995; Shortis *et al.*, 2009). Underwater video-based surveying is being more widely deployed due to the low-cost of digital video cameras and possibilities to verify the estimates at a later point in time, based on video recordings (Cappo *et al.*, 2003; Mallet and Pelletier, 2014).

Due to the fundamental importance of fish species recognition, several computer vision-based techniques have been proposed over the last two decades to automatically identify the species of fish in a given image. These techniques can be broadly categorized into three main application areas based on their scope:

- Recognizing dead fish (e.g. on a conveyor belt) under controlled indoor or outdoor conditions
- Recognizing live (swimming) fish under controlled underwater conditions (e.g. during aquaculture transfers)
- Recognizing live (swimming) fish in unconstrained underwater conditions (e.g. free swimming fish in their natural habitats imaged by static cameras)

Earlier work in fish species recognition focused on the recognition of dead fish (Strachan *et al.*, 1990; Strachan and Kell, 1995). The primary features used for discriminating various fish species were shape descriptors and invariant moments. The use of colour features besides shape features was investigated by Strachan (1993). The key application for these methods is sorting of fish in commercial and research fishing vessels. A method to distinguish fish species using a laser light source and a camera to extract features from three-dimensional fish shape (height, width, thickness) was developed by Storbeck and Daan (2001). Recent research in this direction employs custom designed imaging and conveyor belt systems with controlled illumination to achieve over 99% sorting reliability on several fish species of importance (White *et al.*, 2006). The set of image features used for recognizing fish species has also expanded from primarily shape-based features to deformable shape modelling and texture-based features (Larsen *et al.*, 2009).

Pioneering work in the area of deploying stereo camera systems in controlled aquaculture environments was done by Ruff *et al.* (1995) and Harvey and Shortis (1995). The primary focus of these techniques is decision support for farm managers and marine scientists. Fish species identification and length measurements are typically done manually by human operators in the laboratory. Methods for fish sorting and species identification in freshwater fish farms which grow several fish species together in a pond were developed by Zion *et al.* (1999, 2000). To improve the accuracy and speed of the system, they developed a computer vision system that images fish swimming through a narrow channel with their sides to the camera to get a profile view of each fish. Background illumination was used to overcome water opaqueness and to generate high image contrast. Using this system they were able to achieve over 95% fish identification accuracy in real-time (Zion *et al.*, 2007). Another system for species identification and size measurement in a fish ladder, a narrow special passage in dams that makes it possible for fish to bypass the structure of the dam, was developed by Lee *et al.* (2008). They used a controlled illumination setup and employed colour features to extract fish from the imagery. Then, contour matching was employed to do species recognition as well as size measurement.

Automatic methods for fish species identification in unconstrained environments assume that a fish has already been detected in the image and a rough bounding box around the fish is available. Detection of fish can be achieved by using simple frame differencing with the background frame (Shortis *et al.*, 2013), or more sophisticated saliency (Walther *et al.*, 2004) or foreground modelling based methods (Spampinato *et al.*, 2008; Nadarajan *et al.*, 2011). The primary reason for assuming a pre-detected fish is the challenge involved in accurate fish detection in unconstrained conditions. Major problems are posed by the free swimming direction of the fish, which can cause a huge variation in the outline of the fish as projected on the two-dimensional image captured by the camera. Due to the challenging nature of the problem, fish species recognition in natural underwater environments has only been recently addressed.

One of the earliest efforts for free swimming fish species identification in unconstrained settings was made by Rova *et al.* (2007). They present a method for classifying similarly shaped fish based on their texture alone in unconstrained environments. This approach is limited to applications where target fish species exhibit rich texture. In Spampinato *et al.* (2010), a method was presented to combine shape information with texture to classify fish in natural environments. The main challenge in using shape-based features for fish classification in natural underwater environments is the rich texture of seabed and reef scenes that make segmentation of the fish from the background a very challenging problem. Besides, due to the variations in perspective (determined by the pose of the fish with respect to the camera), the outline of the fish in the image exhibits much larger shape variation than the outline for only the profile views of the fish. To reduce dependency on shape features, recent work by Huang *et al.* (2015) uses a rich feature descriptor employing colour, texture, and shape. Furthermore, separate features are extracted from different parts of the fish (snout, tail, body, etc.) to enhance their descriptive as well as discriminative power. Owing to the limitations of feature-based approaches, recently the approach of using complete fish images instead of explicit feature extraction was introduced in Hsiao *et al.* (2014). They used complete fish images in a sparse representation based framework to perform fish classification.

All of the previous studies, with the exception of Huang *et al.* (2015), based their classification decision on a single image. However, in the case of fish identification in the wild, there is almost always a video sequence in which the same fish appears in multiple frames. Owing to the swimming direction and behaviour of the fish, each of those frames does not necessarily carry the same level of complexity for image classification. An illustration of this phenomenon is shown in Figure 1. Therefore, to be able to classify fish in their natural unconstrained environment, it is important to consider multiple recognition candidates for the same fish. Such recognition candidates can be obtained by tracking the detected fish across multiple frames. Hence, basing the final decision on a sequence or set of images has the potential to produce more reliable results as compared to using a single instance of the fish. A simple approach to exploit this redundancy was used by Huang *et al.* (2015), who used voting on classification decisions of individual instances to make the final decision. However, this simple approach of majority voting has several limitations. First, the individual classification decision is still based on a single image. Second, the relationship between different images of the same tracked sequence is not taken into consideration.

In the computer vision literature, a new paradigm of set-based classification has recently emerged for object recognition (Hu *et al.*, 2012). The core idea of image set classification is to represent the object to be recognized as a set of images. It holds more promise for accurate classification because image sets contain more information compared to a single image. Within an image set, individual images either share a common semantic

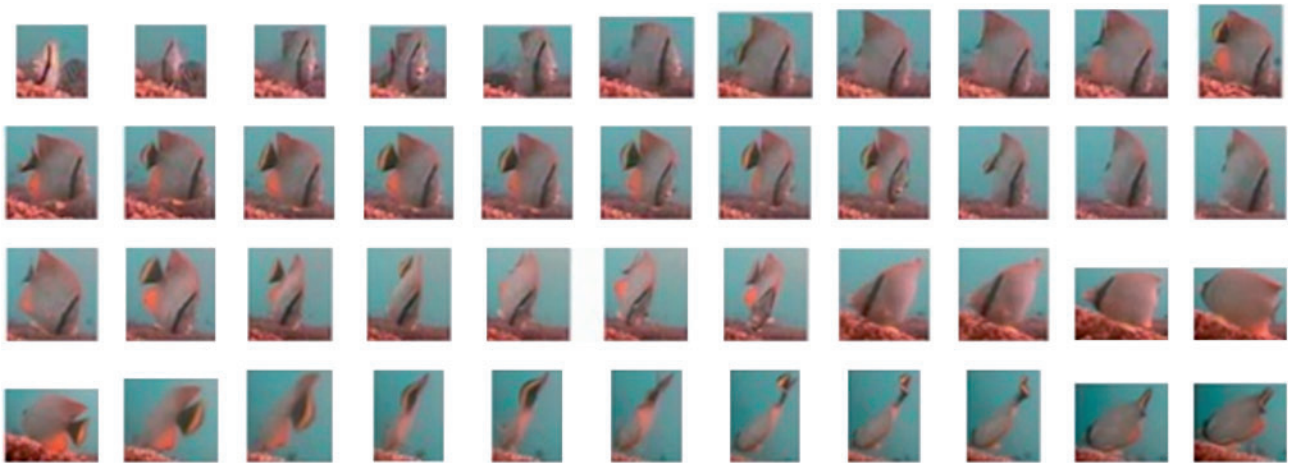

Figure 1. A sample image set showing how individual images in the sequence pose a varying level of complexity for fish species recognition. In several images, the orientation and position of the fish do not allow extraction of reliable shape features.

relationship or complement the appearance variations of the object. Finally, classification is performed for the whole set by using some distance criteria between image sets. Defining novel representations of image sets and useful distance measures to classify an image set are active research topics in the computer vision community. This study demonstrates how the recently emerging techniques of image set-based classification can be used for fish species recognition in unconstrained environments. We propose a framework for this purpose based on state-of-the-art algorithms for tracking (to create an image set from a single fish instance), and image set classification (to identify the species of the fish). The actual tracking and image set classification algorithms used in this work can be replaced by other algorithms without the need for modifying the presented framework. Furthermore, we discuss implications for practical application of the presented method for fish species recognition beyond the species examined in this study.

Material and methods

To evaluate the effectiveness of image set classification for determining fish species, the image collection from ImageCLEF 2014 Fish Task (Spampinato *et al.*, 2014) has been utilized. The dataset contains pre-defined training and test splits for ten fish species. Individual samples in the dataset are obtained from a wide variety of videos containing diverse backgrounds and water conditions. An automatic fish detection algorithm (Spampinato *et al.*, 2008) was used to detect fish in the videos, followed by manual identification of fish species. Sample images of different fish species from the dataset are shown in Figure 2.

Classification using image sets

Image set classification involves the comparison of: (i) a set of images containing a single but unknown species of fish (in the analysis that follows, the test set), with: (ii) multiple sets of images each containing a single known species of fish (the training sets). The goal is to determine the closest match between a training set of images and a test set of images, in order to establish the species of the test set. Importantly, both the test and training sets contain multiple images of fish of a single species, encompassing variation in image characteristics such as pose, lighting, background, etc. This fact makes the technique particularly suitable for identifying fish in unconstrained environments.

We denote an individual set of images, whether test or training, by the notation $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$, where each $\boldsymbol{x}_i$ represents a single image of a fish, and the set contains N images in total, all of a single species. In the case of test data sets the species is unknown; in the case of training data sets it is known. When the species of a specific fish in a given video frame is to be automatically identified, the fish is tracked in the frames immediately before and after the given frame. As a result of tracking, several images of the fish are obtained making one test set. Individual images in the test set represent different appearances of the fish as a result of the swimming behaviour of the fish.

To develop the training sets, multiple images of each target species across different conditions (background, viewing angle, body deformations as a result of swimming motion, lighting, etc.) are required, and the species must be identified. Training data are made of one or more sequences of images generated by tracking each fish species of interest in the video sequences. In practice, a set of videos are manually inspected by an operator, the species noted, and the position (bounding box location, width, height, orientation) of each fish marked. The training data are used by the employed machine learning algorithm (Hu *et al.*, 2012) to build a model of the appearance of these fish species. An illustration of the training and test sets of a particular fish species is shown in Figure 3.

Fish tracking for image set construction

The goal of fish tracking in this study is to generate a set of test image patches (rectangular regions in different frames containing only the fish being tracked) that can be used in image set classification to determine the species of the target fish (refer to Figure 4 for an example). Starting from a known position (bounding box location, width, height, orientation), and possibly motion parameters (speed and acceleration) of a particular fish in a video frame, the tracking method (also known simply as a tracker) will determine the state of the fish in the previous and the next frames. This is done in a probabilistic way using a Bayesian sequential sampling technique called particle filters. We refer the reader to Arulampalam *et al.* (2001) for an overview of particle filters. It is noteworthy to mention that particle filters have been used

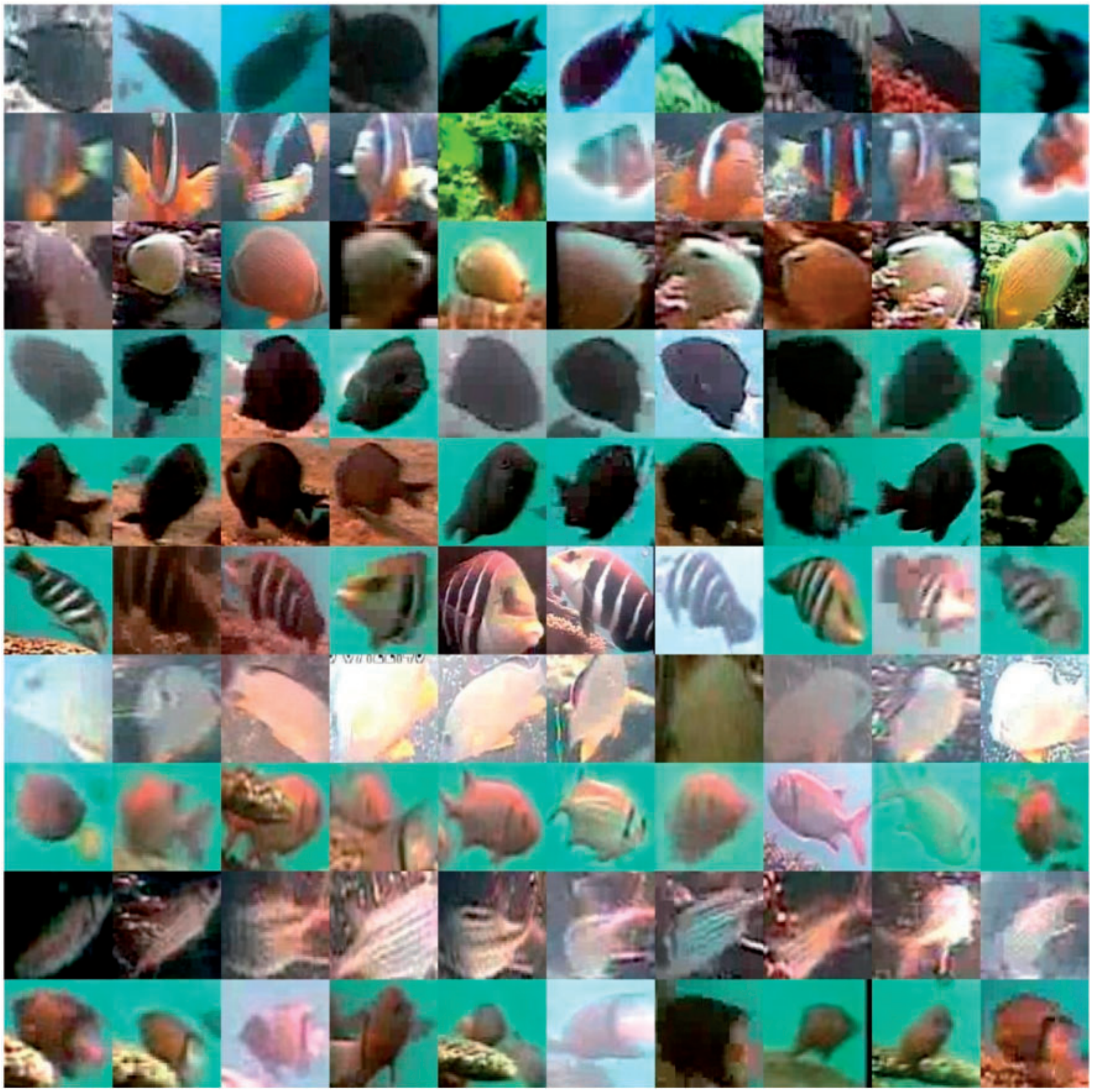

Figure 2. Sample images of various fish species (one per row) in ImageCLEF 2014 dataset.The fish species are (top to bottom): *Acanthurus nigrofuscus, Amphiprion clarkii, Chaetodon lunulatus, Chromis margaritifer, Dascyllus reticulatus, Hemigymnus fasciatus, Lutjanus fulvus, Myripristis berndti, Neoniphon sammara, and Plectroglyphidodon dickii*. Note the variations in images in terms of textured backgrounds, viewing angles of the fish, fish swimming directions, shape deformations, partial occlusions, motion blur, and low image resolution.

extensively for general purpose object tracking in video sequences as illustrated by Smeulders *et al.* (2014). In general, the particle filter tracker aims to maintain an accurate estimate of the posterior distribution of the current state of the tracked fish given all the previous and the current *observations* of the fish, i.e. image patches depicting how the view of the fish is evolving over time. This allows a convenient framework for estimating and propagating the posterior regardless of the underlying distribution through a sequence of prediction and update steps; thus, generalizing the well-known Kalman filter. Details of this framework are presented in Zhang *et al.* (2013, 2015) for further reading.

Image set representation

The first challenge after construction of an image set is how to extract and represent the information from an image set. The image set constructed by tracking the fish across a large number of frames consists of images containing different poses of the fish as well as the changing background. To automatically discover commonalities between these individual images, subspace-based methods are often used. A subspace effectively represents the span of images that can be generated from a set of common *basis images* U_i. For instance, one can consider the *mean image* μ_i,obtained by averaging the pixel grey values at each location across all images in an image set, as a reference. Any given image $\boldsymbol{x_i}$ in the image set can be constructed using the reference image μ_i and a weighted sum of the basis images U_i, that is $\boldsymbol{x}_i = \mu_i + U_i v_i$, where v_i represents the weights. A set of basis images for a particular fish species can be generated by performing a mathematical process called Principal Component Analysis (PCA) on a large set of images depicting fish of that species appearing in different poses, illumination conditions, and backgrounds (Abdi and Williams, 2010). Basis images can be considered to be a set of

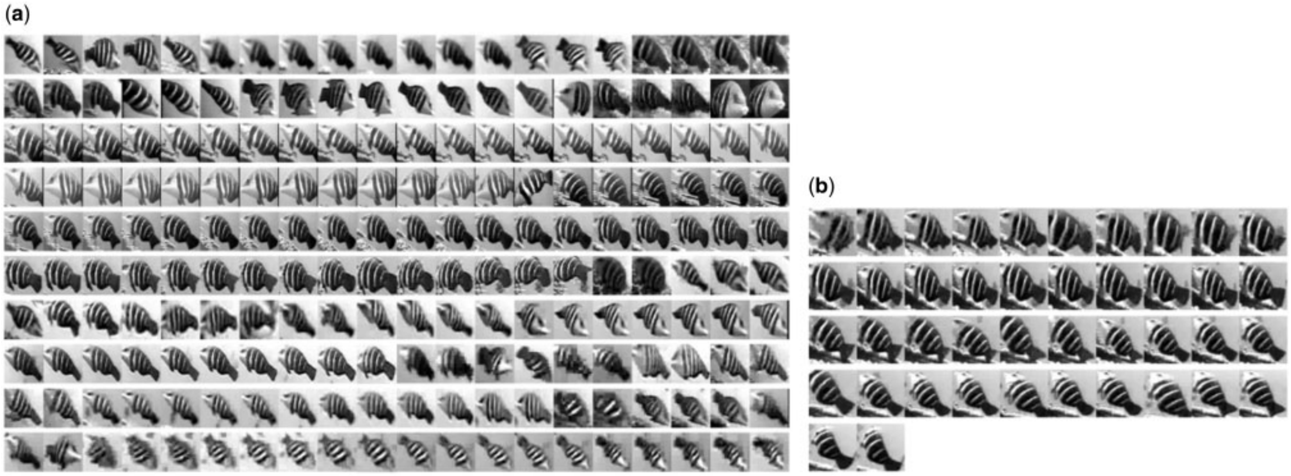

Figure 3. An illustration of training and test image sets for the species *Hemigymnus fasciatus*. Note that the training set (a) contains a large variety of pose and illumination variations. The test set (b) contains the sequence of images obtained by tracking the target fish across multiple frames. All individual images in the training as well as test sets are resized to a fixed dimension (32 X 32 in this case) for a unified image set representation.

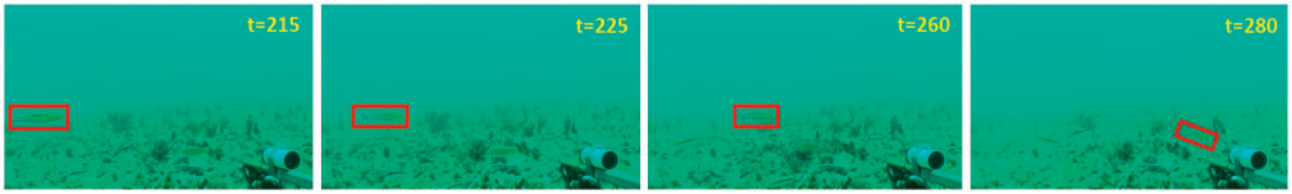

Figure 4. A visual example of a fish being tracked within an uncontrolled environment using the method proposed in Zhang *et al.* (2015). The detected bounding boxes are the current best states of the fish in each of the video frames. These image patches, indicated by the overlaid rectangle in each frame, can subsequently be used to construct the test set for the image set classification method outlined in section Classification using image sets.

ingredients, derived from statistical analysis of many pictures of the object (fish) of interest. Any fish image can be considered to be a weighted combination of these basis images which appear as light and dark areas that are arranged in a specific pattern. An example of the basis images computed using PCA is shown in Figure 5. An image set i is then represented jointly with the set X_i of individual images as well as the basis images $(\mu_i,\ U_i)$ computed for that set.

Similarity computation between image sets

A simple way of computing similarity between two image sets is to use a similarity measure (e.g. correlation between pixel values) between two individual images and then employ it to compute an average similarity between the training and test sets (for instance by comparing each image in one set to each image in the other set). This procedure is not only computationally expensive, but also prone to errors due to noise and presence of outliers in the training or test sets. Such outliers may appear, for example, as the result of a tracking failure. A more elegant way of computing the similarity is to *construct* two synthetic images, one from the training image set and the other from the test image set, such that the synthetic images are as *similar* to each other as possible. An arbitrarily large number of synthetic images can be constructed from a given basis $(\mu_i,\ U_i)$, using the relationship $\boldsymbol{x}_i = \mu_i + U_i v_i$ by varying the weights v_i of the basis. Hence, an optimization algorithm is applied to minimize the difference $D(\boldsymbol{x}_i, \boldsymbol{x}_j)$ between the reconstructed image $\boldsymbol{x}_i$ from the training image set and $\boldsymbol{x}_j$ from the test image set. Numerous image difference measures have been reported in the literature (Mahmood and Khan, 2012) and any of these measures can be used to compute the difference D. Given a test image set, its difference is computed from all training image sets. The test image set is then assigned the label (species name) of the training image set that has the smallest difference from it. This approach is commonly known as one-nearest-neighbour classification (Cover and Hart, 1967) in the machine learning literature.

Since the space of all images that can be created using a given basis is very large, it is possible to construct two images with a very small difference between them even for two image sets of different fish species, which adversely affects the classification performance. Hence, it is desirable during the construction of the images to restrict them to have an appearance visually similar to the samples in the corresponding image set. This objective can be achieved by constructing the synthetic image using a linear combination of the images in the image set, that is $\boldsymbol{x} = \sum_{n=1}^{N} \alpha_n \boldsymbol{x}_n$ or written in the matrix form $\boldsymbol{x} = X\boldsymbol{\alpha}$, where α is the weight vector. Hence, a joint optimization is performed (Hu *et al.*, 2012) to construct the images $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ from the training image set basis (μ_i, U_i) and the test image set basis (μ_j, U_j) respectively; such that the difference $D(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is minimized while keeping the construced images close to their corresponding image set samples X_i and X_j. The outcome of the optimization algorithm is the optimized values of the basis image weights (v_i and v_j) and the image sample weight vectors (α and β) for the training and test image sets, respectively. Hu *et al.* (2012) have demonstrated that improved results are achieved when the weight vectors α and β are *sparse*, that is most of their elements are zero. A sparse weight vector allows the optimization algorithm to choose only a few samples from the corresponding image sets, allowing it to effectively ignore outliers and non-representative samples samples that might occur due to the errors made by the tracking algorithm. Hence a sparsity constraint is also enforced in the optimization algorithm such that the weight vectors (α and β) returned by the algorithm are sparse, which makes the algorithm robust to tracking failures (see Hu *et al.*, 2012 for details). An illustration of the construction of sparsely approximated samples from training and test image sets and the corresponding optimization criteria is shown in Figure 6. Note that the simultaneous optimization of the test image set with each of the training image sets has to be done when classifying the test image set. Hence, the training phase of the presented algorithm only involves collecting the labelled training samples and constructing the image set models for all training image sets. It does not involve any optimization during the training stage, which is an inherent advantage of using a nearest neighbour-based classification approach. On the flip

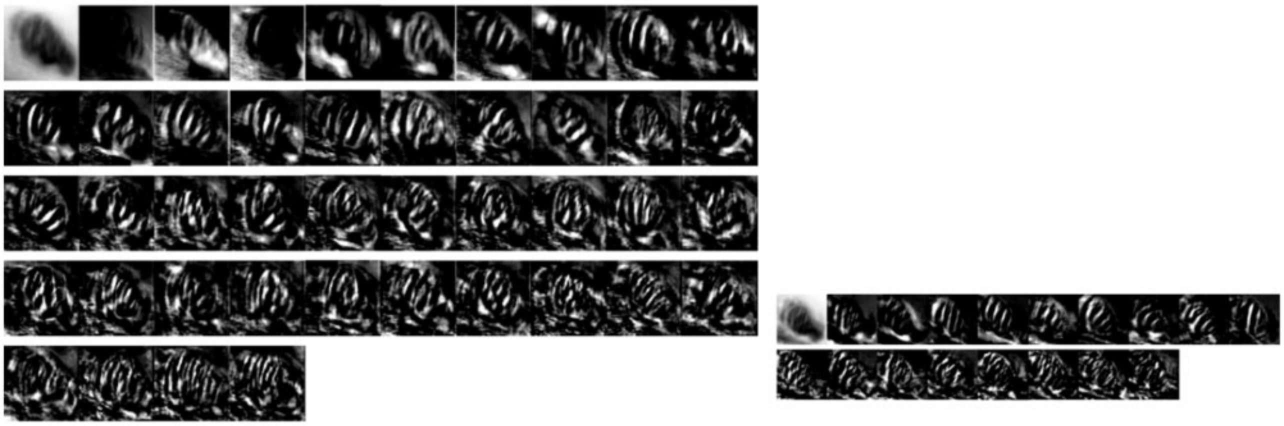

Figure 5. An illustration of *mean* and *basis* images for the training (left) and test (right) image sets shown in Figure 3. The mean image is depicted first, followed by individual basis computed using Principal Component Analysis. Note that the number of basis images is larger for the training set as compared to the test set to cater for the larger diversity of images in the training set.

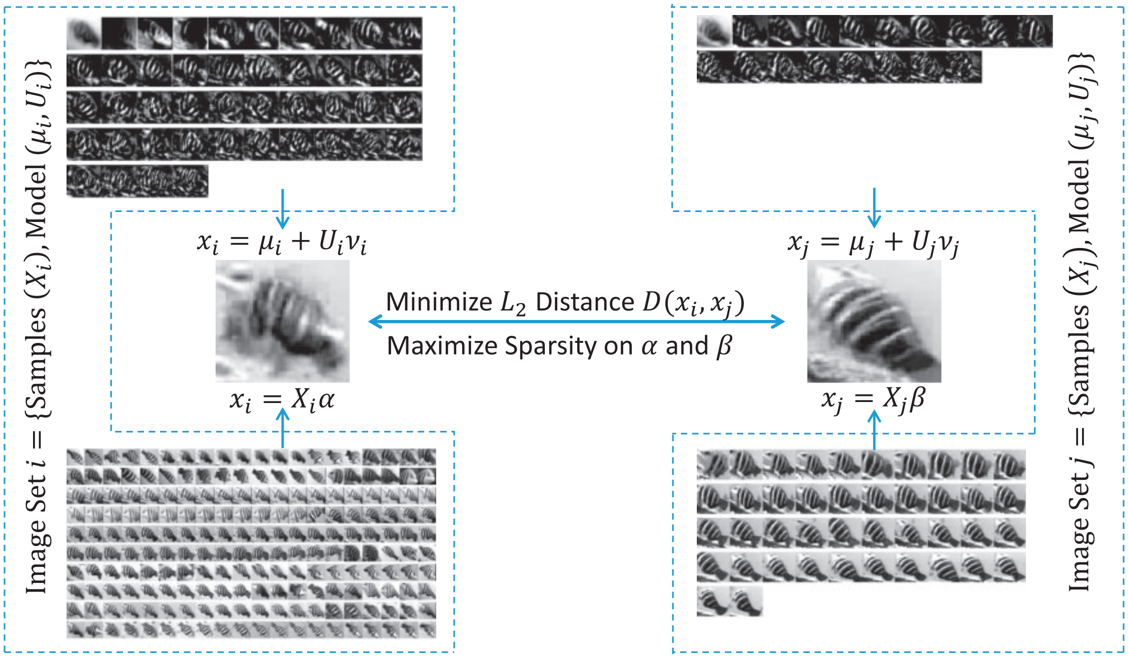
Image Set i = {Samples (X_i), Model (μ_i, U_i)}
$x_i = \mu_i + U_i v_i$
$x_i = X_i \alpha$
Minimize L_2 Distance $D(x_i, x_j)$
Maximize Sparsity on α and β
$x_j = \mu_j + U_j v_j$
$x_j = X_j \beta$
Image Set j = {Samples (X_j), Model (μ_j, U_j)}

Figure 6. An illustration of sparsely approximating new images in the training and test set using the corresponding image set models with the aim of finding the most similar images in the two sets.

side, the testing phase is more computationally involved as the test image set has to be jointly optimized with all training image sets to find its closest matching training set.

Performance evaluation protocol

Both the training and test images in the ImageCLEF 2014 Fish Task data contain image sequences where multiple instances of the same fish are present. Hence, each of these image sequences can be used as a single image set. During testing, one needs to label each tracked sequence independently. Therefore, test image sets were created based on grouping tracked image patches of a single fish into a single image set. A similar approach could also be used to create training image sets. The drawback of this approach is that modeling of image sets using a few samples is not as accurate as modeling the set with a large number of samples. Therefore, all training images of a particular fish species were considered as a single training image set for that species. The advantage of this approach is that the large number of images contain a richer set of variations that are more effectively encoded in the corresponding image set representation.

The accuracy of the presented algorithm is measured by comparing the predicted label of an image set with its true, manually determined label. If both labels match, the image set is considered to be correctly classified. Otherwise, the classification decision is considered incorrect. Classification accuracy is defined as the percentage of the correctly classified image sets with respect to the total number of test image sets.

Table 1. Summary statistics of training and test sets, and the accuracy achieved by the classifier.

Species	Training images	Test images	Test sets	Classification accuracy (%)
Acanthurus nigrofuscus	2511	725	32	90.6
Amphiprion clarkii	2985	878	45	100.0
Chaetodon lunulatus	2494	917	29	100.0
Chromis margaritifer	3282	371	17	100.0
Dascyllus reticulatus	3196	681	14	100.0
Hemigymnus fasciatus	2224	852	47	83.0
Lutjanus fulvus	720	146	07	71.4
Myripristis berndti	2554	840	33	90.9
Neoniphon sammara	2019	969	58	100.0
Plectroglyphidodon dickii	2456	577	16	100.0

The total number of image samples for each fish species in the training and test partitions is listed. In addition, the number of independent test sets, comprised of at least five images of a tracked fish, is listed. The partition of the data into training and test sets has been kept identical as in the original ImageCLEF 2014 Fish Task dataset for easier reproducibility and comparison of results. Note that our method achieves 100% recognition rate for six fish species.

Results

For effective modelling of image sets, test sets having less than five images were ignored in the first experiment. It is a reasonable assumption from a practical perspective that most transits of the fish across the field of view would generate larger sized sets. The results of image set classification on the test sets containing five or more samples are presented in Table 1. For six species, 100% recognition results were obtained. It is important to note that the images were directly used in image set modelling, without extraction of any shape, colour, or texture features. The dataset contains a wide variety of backgrounds, lighting conditions, as well as fish orientations. Despite such large variations, an overall recognition rate of 94.6% was obtained. Table 2 provides a detailed analysis of the failure cases using a confusion matrix of true classes versus predicted classes. It is interesting to note that a number of image set instances of *Hemigymnus fasciatus* were mis-classified as other fish species. A closer inspection of the results revealed that those cases arose when the illumination was quite poor and hence the zebra-striped pattern of the species was not visible. Therefore, the correct match to the training set of *H. fasciatus* could not be established.

Further experiments were performed to study the effect of the number of images in the test set on the recognition performance. To analyse this effect, the experiment was repeated using test set size thresholds in the range from 2 to 10. For each experiment, all test image sets that had a size less than the chosen threshold were discarded and the corresponding recognition accuracywas computed. The cumulative results for all fish species are plotted in Figure 7. The graph shows the accuracy both at the set level and individual image level. The accuracy at the image level is determined by counting all images in the set as correctly labelled if the decision at the set level is right, and as mis-classified otherwise. Classification accuracy is then computed as the percentage of correctly classified images among all test images. Hence in image level classification, mis-classifying a larger sized image set is penalized more than mis-classifying a smaller sized image set. Note that image level accuracy is higher than set level accuracy indicating that most mis-classifications happened for small sized test image sets. In practice, the method will deliver improved classification accuracy for higher frame rate videos as more image patches can be obtained for the same fish as a result of tracking.

Table 2. Confusion matrix indicating the number of correctly classified sets for each fish species and mis-classification errors made by the algorithm.

Species	Acanthurus nigrofuscus	Amphiprion clarkii	Chaetodon lunulatus	Chromis margaritifer	Dascyllus reticulatus	Hemigymnus fasciatus	Lutjanus fulvus	Myripristis berndti	Neoniphon sammara	Plectroglyphidodon dickii
Acanthurus nigrofuscus	**29**			**1**	**2**					
Amphiprion clarkii		**45**								
Chaetodon lunulatus			**29**							
Chromis margaritifer				**17**						
Dascyllus reticulatus					**14**					
Hemigymnus fasciatus	**2**	**1**		**2**	**1**	**39**				**2**
Lutjanus fulvus							**5**		**2**	
Myripristis berndti					**1**			**30**		**2**
Neoniphon sammara									**58**	
Plectroglyphidodon dickii										**16**

The rows represent the human labelled species, whereas the columns indicate the computer labelled species of the fish.

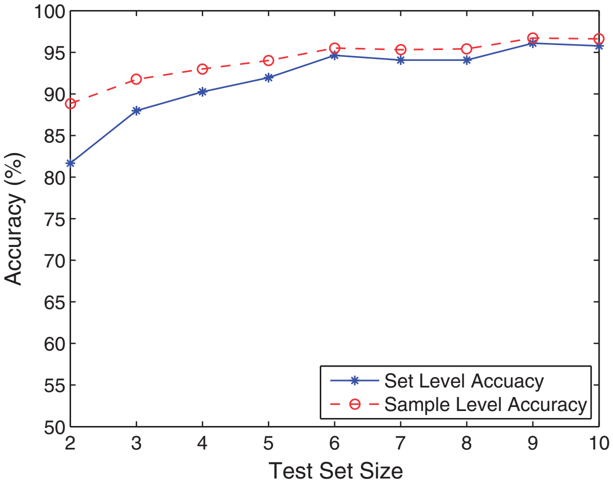

Figure 7. The effect of test set size on recognition rate. Note that the larger the number of images in a given set, the more accurate model can generally be obtained from that set. The results indicate that the recognition accuracy saturates as the number of images in the set increases to six.

Discussion

Underwater video systems have shown to be cost effective, accessible, and provide a means of repeatable sampling of fish populations (Murphy and Jenkins, 2010). While manual processing of the resulting imagery decreases the cost effectiveness and availability of numerical data after recording, use of computer vision algorithms to automate species identification can significantly improve the efficiency of analysis of captured imagery (MacLeod *et al.*, 2010). Most of the existing methods for automated fish species recognition (e.g. Lee *et al.*, 2008; Larsen *et al.*, 2009; Spampinato *et al.*, 2010) rely on accurate delineation of fish boundary to extract shape features. This not only limits the applicability of those methods to cases where the background is easily separable, but also constrains them to identify only profile views of the fish. This study has demonstrated how modelling of multiple fish images directly can yield competitive recognition rates without the need to do explicit feature extraction, which becomes very challenging for free-swimming fish in unconstrained environments. Due to these characteristics, the method has neither of the above-mentioned limitations. The only prior work that used this concept for fish identification is Hsiao *et al.* (2014). A sparse representation-based framework was used and was able to achieve a species recognition rate of 81.8%, which is much lower than the results reported in this paper. Note that the dataset used in this study originates from the Fish4Knowledge project, which was also used by Hsiao *et al.* (2014). However, the sample set is different as the test images used by Hsiao *et al.* are not publicly available. In order to compare the performance of the image set classification technique to that used in Hsiao *et al.* (2014), we implemented their method and trained and tested it on the same data used in this study. An extensive parameter tuning of their method was conducted to optimize its performance on the training partition of ImageCLEF 2014 Fish Task dataset. The optimized algorithm was able to achieve an accuracy of 84.04% on the test set, which is a threefold increase in error rate as compared to the presented method.

Blanc *et al.* (2014) present another study utilizing the ImageCLEF 2014 Fish Task data, where videos have been used to first detect and then classify the fish species using fish species-dependent features trained using a Support Vector Machine classifier. They report an average precision and recall of 55% and 50%, respectively. The relatively lower performance is due to the joint process of fish detection and species recognition. Hence, a direct comparison with these approaches cannot be made. Direct comparison is further complicated because this study uses multiple fish images to make a single classification decision, while prior works based their classification decisions on single images.

The presented method was able to correctly identify fish species for cases where fish are imaged against the sea floor or coral reefs that have a rich texture, making accurate segmentation of fish an extremely challenging task. A major advantage of the presented method is that it is not based on species-dependent feature extraction, and hence can be easily applied to any fish species in general. Moreover, due to image set modelling, fish instances that are not in profile view can also be correctly classified. One disadvantage of the method as compared to the state-of-the-art is that multiple images of the same instance are required to make a decision about the species. However, a fish is usually visible in the field of view of the camera for at least five frames thereby making this issue practically insignificant.

Experiments on the public dataset yielded 100% recognition rates for six fish species, namely *Amphiprion clarkii*, *Chaetodon lunulatus*, *Chromis margaritifer*, *Dascyllus reticulatus*, *Neoniphon sammara*, and *Plectroglyphidodon dickii*. The recognition rates were the lowest for the species *Hemigymnus fasciatus* and *Lutjanus fulvus*. A closer investigation of the failure cases for *Hemigymnus fasciatus* revealed that in most cases, the visbility of the fish was very poor due to low levels of lighting, making the zebra-like stripes barely visible. The poor results for *Lutjanus fulvus* appears to be a combination of the small number of image sets and the relatively featureless appearance of this species.

For the practical application of the presented image set based method, several considerations need to be made. First, the training image set for species of interest has to be collected. The training images should comprehensively capture variation in fish shape due to swimming motion and viewing angle of the camera, as well as image degradations that are expected to be encountered in a particular scenario. A few hundred representative images of each species of interest are usually sufficient to achieve good accuracy. This factor was tested using the ImageCLEF Fish Task data by artificially reducing the size of the training data. A training set built from 200 randomly selected image samples per fish species yielded an accuracy of 88.2%, which is already significantly higher than existing methods.

In set-based classification, the commonly used paradigm is to make individual training sets for each capture instance (Hu *et al.*, 2012). For instance, if for a particular object several videos are available during training, each video constitutes a separate training set for that object; effectively constituting a single sample for nearest neighbour classification. This is in line with the traditional nearest neighbour classification scenario where several training samples are available for each class. However, in our experiments with fish species identification, improved results were achieved by collecting all training samples of a species into a single training image set for that species. This can be attributed to two factors. First, the larger amount of data in the training image set allows the construction of a more powerful and representative model of that species. Since this model is jointly optimized with the test image set, a comprehensive model encompassing a larger image diversity yields improved performance. Second, in several cases, a fish simply swims across the field of view of the camera into a particular direction. Therefore, the variation in the shape of the fish in different images of the image set is minimal, limiting the model variance, and adversely affecting system performance.

To collect and label training images, the use of a fish detection approach (such as Spampinato *et al*., 2008) that can segment out fish from the given imagery is recommended. Once the segmented (but unlabelled) images are obtained, they can be manually grouped according to their species using any file manipulation program (like Microsoft Windows Explorer). Using this approach, we were able to manually classify about 7000 samples in the ImageCLEF test data in <3 h.

A MATLAB implementation using an Intel Core i7 Machine (2.6 GHz, 16GB RAM, 512 GB SSD) required ∼20 min to process all test imagery, resulting in an average classification time of <4 s per test image set. It is interesting to note that the computation time as well as the accuracy decreases as the number of samples in the training set are reduced. To test this a training set was generated using approximately one-tenth of the 20 000 available images from the full ImageCLEF 2014 Fish Task training data. Using a training set built from 200 randomly chosen samples from each of the ten fish species reduced the average computation time from 4 s to 240 ms per test image set. This improvement in computation time comes at the cost of a reduction in accuracy from 95% to 88%. In many time-critical applications this could be a reasonable compromise to make, keeping in view that the achieved accuracy is still improved over that of existing single-image-based methods.

Despite the high accuracy of the presented image set-based approach under challenging conditions, it has certain limitations. One problem inherent with the set-based approach is the dependence of accuracy on image set size—the larger the set size, the higher the accuracy. In applications where continuous monitoring has to be done for a long period of time, video frame rates are usually kept low to cater for the limited storage and transmission capacity. At low frame rates, fish tracking becomes particularly challenging and hence it might become difficult to automatically obtain a reasonably sized image set. Furthermore, the computational complexity of the presented approach is higher than that of most single-image classification methods. This is due to the joint optimization of each test set with the entire training set during classification. However, it should be noted that the method presented is able to classify all images in the set simultaneously. Hence, for an average image set size of ten samples, a single image classification algorithm that is ten times faster than the presented method would still take the same overall computation time. Due to this characteristic, the computational disadvantage gets largely compensated for in the case of high frame rate videos that naturally yield larger image set sizes.

There are several directions in which the presented method can be improved. A particularly promising direction for future research would be to incorporate active learning into the presented framework. Active learning (Settles, 2010) is a special case of semi-supervised machine learning in which a learning algorithm is able to interactively learn from user corrections to obtain higher accuracy at new data points. Note that the algorithm presented is based on nearest neighbour classification which is quite suitable for use in active learning scenarios. Another interesting direction could be to explore generation of image sets using synthetic, computer-generated models (Rabasse *et al*., 2008) of fish movement and image degradations. Based on the synthetic generation of image sets, it might be possible to just use a single image as a seed image and generate the whole training image set synthetically. Yet another possibility to apply the approach effectively on low frame rate videos would be to use unsupervised image clustering techniques to group together fish instances based on image similarity measures. This could compensate for tracking failures in challenging scenarios and create reasonably large-sized image sets to achieve higher accuracy.

Conclusions

This article has presented an image set-based approach for fish species identification in unconstrained environments. The overall classification accuracy for all ten species studied in this work was around 95%, which shows strong potential for application of image set-based fish classification methods in practical applications. Accordingly, the presented method shows huge potential for fish identification from routinely captured video data where fish tracking provides a natural mechanism to construct image sets. Once the image training sets are established, this technique can provide a very high level of automation of species recognition in video sequences acquired to monitor species abundance or biomass. The results of the classification can be used as an effective and accurate tool to provide decision support to fisheries and marine park managers for stock assessment and species conservation.

Acknowledgements

We would like to thank the anonymous reviewers whose detailed and insightful comments have helped us in bringing this manuscript to its current form. In addition, 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 and King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research. A.M. was supported by the Australian Research Council Fellowship DP110102399.

References

Abdi, H., and Williams, M. J. 2010. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2: 433–459.

Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. 2001. A tutorial on particle filters for on-line nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on Signal Processing, 50: 174–188.

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

Cappo, M. E., Harvey, E., Malcolm, H., and Speare, P. 2003. Potential of video techniques to monitor diversity, abundance and size of fish in studies of marine protected areas. Aquatic Protected Areas-What Works Best and How Do We Know, pp. 455–464.

Cover, T. M., and Hart, P. E. 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13: 21–27.

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.

Fernandes, P. G. 2009. Classification trees for species identification of fish-school echotraces. ICES Journal of Marine Science, 66: 1073–1080.

Harvey, E., and Shortis, M. 1995. A system for stereo-video measurement of sub-tidal organisms. Marine Technology Society Journal, 29: 10–22.

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.

Hu, Y., Mian, A. S., and Owens, R. 2012. Face recognition using sparse approximated nearest points between image sets. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34: 1992–2004.

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.

Larsen, R., Olafsdottir, H., and Ersboll, B.2009. Shape and texture based classification of fish species. Scandanavian Conference on Image Analysis, pp. 745–749.

Lee, D. J., Archibald, J. K., Schoenberger, R. B., Dennis, A. W., and Shiozawa, D. K. 2008. Contour matching for fish species recognition and migration monitoring. *In* Applications of Computational Intelligence in Biology: Current Trends and Open Problems. Ed. by T.G. Smolinski, M.G. Milanova and A.E. Hassanien. Springer-Verlag, Germany.

MacLeod, N., Benfield, M., and Culverhouse, P. 2010. Time to automate identification. Nature, 467: 155–156.

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.

Mahmood, A., and Khan, S. 2012. Correlation-coefficient-based fast template matching through partial elimination. IEEE Transactions on Image Processing, 21: 2099–2108.

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.

Nadarajan, G., Chen-Burger, Y., Fisher, R. B., and Spampinato, C. 2011. A flexible system for automated composition of intelligent video analysis. Proceedings Image and Signal Processing and Analysis, pp. 259–264.

Rabasse, C., Guest, M., and Fairhurst, C. 2008. A new method for the synthesis of signature data with natural variability. IEEE Transactions on System, Man, and Cybernetics—Part B, 38: 691–699.

Rova, A., Mori, G., and Dill, L. M. 2007. One fish, two fish, butterfish, trumpeter: recognizing fish in underwater video. IAPR Conference on Machine Vision Applications, IAPR, Japan. pp. 404-407.

Ruff, B. P., Marchant, J. A., and Frost, A. R. 1995. Fish sizing and monitoring using a stereo image analysis system applied to fish farming, Aquaculture. Engineering, 14: 155–173.

Settles, B. 2010. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, USA. pp. 1–67.

Shortis, M., Harvey, E., and Abdo, D. 2009. A review of underwater stereoimage 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.

Shortis, M. R., Ravanbakskh, M., Shafait, F., Harvey, E. S., Mian, A., Seager, J. W., Culverhouse, P. F. et al. 2013. A review of techniques for the identification and measurement of fish in underwater stereo-video image sequences. Videometrics, Range Imaging, and Applications XII, SPIE Vol. 8791, paper 0G. The International Society for Optical Engineering, Bellingham WA, USA.

Smeulders, A. W. M., Chu, D. M., Cucchiara, R., Calderara, S., Dehghan, A., and Shah, M. 2014. Visual tracking: an experimental survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36: 1442–1468.

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, pp. 45-50.

Spampinato, C., Palazzo, S., Boom, B., and Fisher, R.B. 2014. Overview of the LifeCLEF 2014 Fish Task, Proceedings of CLEF2014. Springer-Verlag, Germany. pp. 616-624.

Spampinato, C., Chen-Burger, Y., Nadarajan, G., and Fisher, R. B. 2008. Detecting, tracking and counting fish in low quality unconstrained underwater videos. Proceedings Computer Vision Theory and Applications, pp. 514–519.

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., Nesvadba, P., and Allen, A. R. 1990. Fish species recognition by shape analysis of images. Pattern Recognition, 23: 539–544.

Strachan, N. J. C. 1993. Recognition of fish species by colour and shape. Image and Vision Computing, 11: 2–10.

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.

Walther, D., Edgington, D. R., and Koch, C. 2004. Detection and tracking of objects in underwater video. IEEE Computer Vision and Pattern Recognition, pp. 544–549.

White, D. J., Svellingen, C., and Strachan, N. J. C. 2006. Automated measurement of species and length of fish by computer vision. Fisheries Research, 80: 203–210.

Zhang, T., Ghanem, G., Liu, S., and Ahuja, N. 2013. Robust visual tracking via structured multi-task sparse learning. International Journal of Computer Vision, 101: 367–383.

Zhang, T., Liu, S., Ahuja, N., Yang, M. H., and Ghanem, B. 2015. Robust visual tracking via consistent low-rank sparse learning. International Journal of Computer Vision, 111: 171–190.

Zion, B., Shklyar, A., and Karplus, I. 1999. Sorting fish by computer vision. Computers and Electronics in Agriculture, 23: 175–187.

Zion, B., Shklyar, A., and Karplus, I. 2000. In-vivo fish sorting by computer vision. Aquacultural Engineering, 22: 165–179.

Zion, B., Alchanatis, V., Ostrovsky, V., Barki, A., and Karplus, I. 2007. Real-time underwater sorting of edible fish species. Computers and Electronics in Agriculture, 56: 34–35.

Handling editor: Howard Browman