Marine Technology Society Journal, Volume 50, Number 1, Feb 2016, pp. 4-16

Progress in the automated identification, measurement and counting of fish in underwater image sequences

Mark R. Shortis[1], Mehdi Ravanbakhsh[1],

Faisal Shafait[2], Ajmal Mian[3].

[1] School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, Australia

[2] National University of Sciences and Technology, Islamabad, Pakistan

[3] School of Computer Science and Software Engineering, The University of Western Australia, Crawley, Australia

Corresponding author: Mark Shortis

mark.shortis@rmit.edu.au

Abstract

Underwater video systems are widely used for counting and measuring fish in aquaculture, fisheries and conservation management. To determine population counts, spatial or temporal frequencies, and age or weight distributions, snout to tail fork length measurements are performed in video sequences, most commonly using a point and click process by a human operator. Current research aims to automate the identification, measurement and counting of fish in order to improve the efficiency of population counts or biomass estimates. A fully automated process requires the detection and isolation of candidates for measurement, followed by the snout to tail fork length measurement, species classification, as well as the counting and tracking of fish. This paper reviews the algorithms used for the detection, identification, measurement, counting and tracking of fish in underwater video sequences. The paper analyses the most commonly used approaches, leading to an evaluation of the techniques most likely to be a comprehensive solution to the complete process of candidate detection, species identification, length measurement and population counts for biomass estimation.

Keywords Underwater imaging, video sequences, stereo, candidate detection, fish counts, species classification, biomass estimation

Introduction

The monitoring of fish for stock assessment in aquaculture and commercial fisheries, and in the assessment of the effectiveness of biodiversity management strategies such as Marine Protected Areas, is essential for the economic and environmental management of fish populations. Video based measurement for fishery independent and non-destructive sampling is now widely accepted as an accurate and reliable technique (Murphy and Jenkins, 2010; Shortis et al., 2009). The advantages of

using stereo-video for counting the numbers of fish, measuring their lengths and defining the sample area have been well demonstrated (Harvey et al., 2004; Murphy and Jenkins, 2010). However, the time lag and cost of processing video imagery decreases the cost effectiveness and uptake of this technology. Current research aims to minimise or completely eliminate the involvement of the human operator in the process of length measurement of fish recorded by underwater video surveys. The ultimate goal is to fully automate the identification, measurement and counting of fish, in order to deal with the many thousands of hours of video footage that is routinely captured each year. Advances in automated techniques will substantially decrease the cost of processing, ultimately reduce error rates and make the technology more accessible to a broad spectrum of end users.

Stereo-video systems have the advantages that the measurements are impartial and repeatable (Harvey et al., 2004; Wehkamp and Fischer, 2014), and very reliable levels of accuracy can be achieved for calibrated systems (Boutros et al., 2015; Harvey and Shortis, 1998; Menna et al., 2013). Underwater stereo-video systems have been used in wild fish stock assessment with a variety of cameras and modes of operation (Bouchet and Meeuwig, 2015; Mallet and Pelletier, 2014, McLaren et al., 2015; Watson et al., 2009), in pilot studies to monitor length frequencies of fish in aquaculture cages (Harvey et al., 2003; Phillips et al., 2009) and, more recently, in fish nets during capture (Rosen et al., 2013). Samples taken in aquaculture cages can approach 95% of the population and the measurement technique is non-invasive. Snout to tail fork length and other body spans on the fish are measured from the video recordings and, using a length-weight regression (Pienaar and Thomson, 1969), the weights of the fish are estimated with errors of no more than a few per cent. Commercial systems such as the AKVAsmart, formerly VICASS (Shieh and Petrell, 1998), and the AQ1 AM100 (Phillips et al., 2009) are widely used in aquaculture to determine size distributions based on simple length and span measurements, and thereby deduce total biomass from an estimated number of fish in the cage or tank.

The next significant advances in the technology of video monitoring of fish must be the automation of biomass estimation of species of interest. In aquaculture, single species are the norm in cages or transfers, so the process is limited to identification of candidates, measurement of the lengths and counting the total stock number. In the wild, either in marine protected areas or fisheries, the additional required steps for species classification are the preparation of a training set of species of interest, and the automatic recognition of species so that only fish of interest are included in the counts and biomass estimates. The essential steps in this complete process are depicted in Figure 1.

Once the candidate region is delineated in a pair of stereo images and the object is classified as a species of interest, the body length of the fish can be accurately estimated in 3D (Harvey et al., 2003) using the simple triangulation technique shown in Figure 2. The accuracy of measurement of the 3D locations can be significantly improved using multiple frames from the image sequences in combination with image matching (Gruen and Baltsavias, 1988). Based on the snout to tail fork length, it is then possible to accurately estimate the biomass of the fish using length-weight regression. Single camera systems are generally limited to counts of species because length estimates are too inaccurate to provide useful biomass estimates (Harvey et al., 2002).

Progress toward fulfilling this fundamental series of steps is discussed in the following sections, commencing with a review of published techniques based on single camera and stereo systems. Whilst the essential requirements are to identify, measure and count the fish in the video sequence, the complete extent of the methodology encompasses detection, identification, classification, measurement, tracking and counting of fish.

Marine Technology Society Journal, Volume 50, Number 1, Feb 2016, pp. 4-16

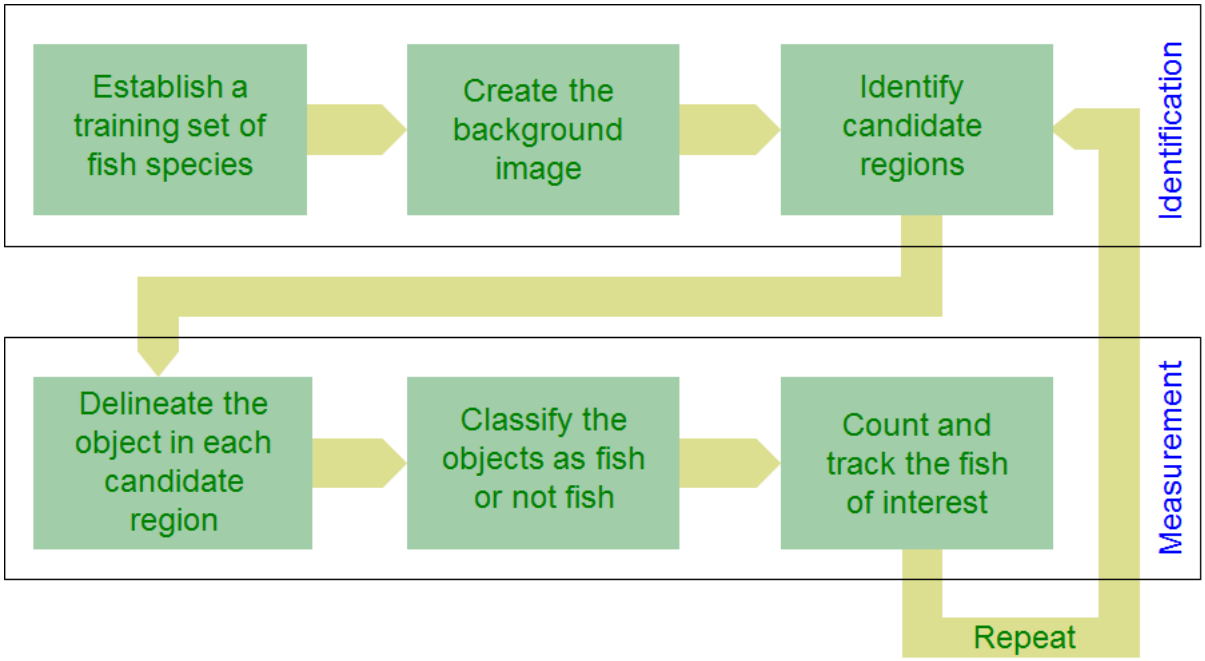

Figure 1. Process required for automated identification, measurement and counting of fish in underwater images.

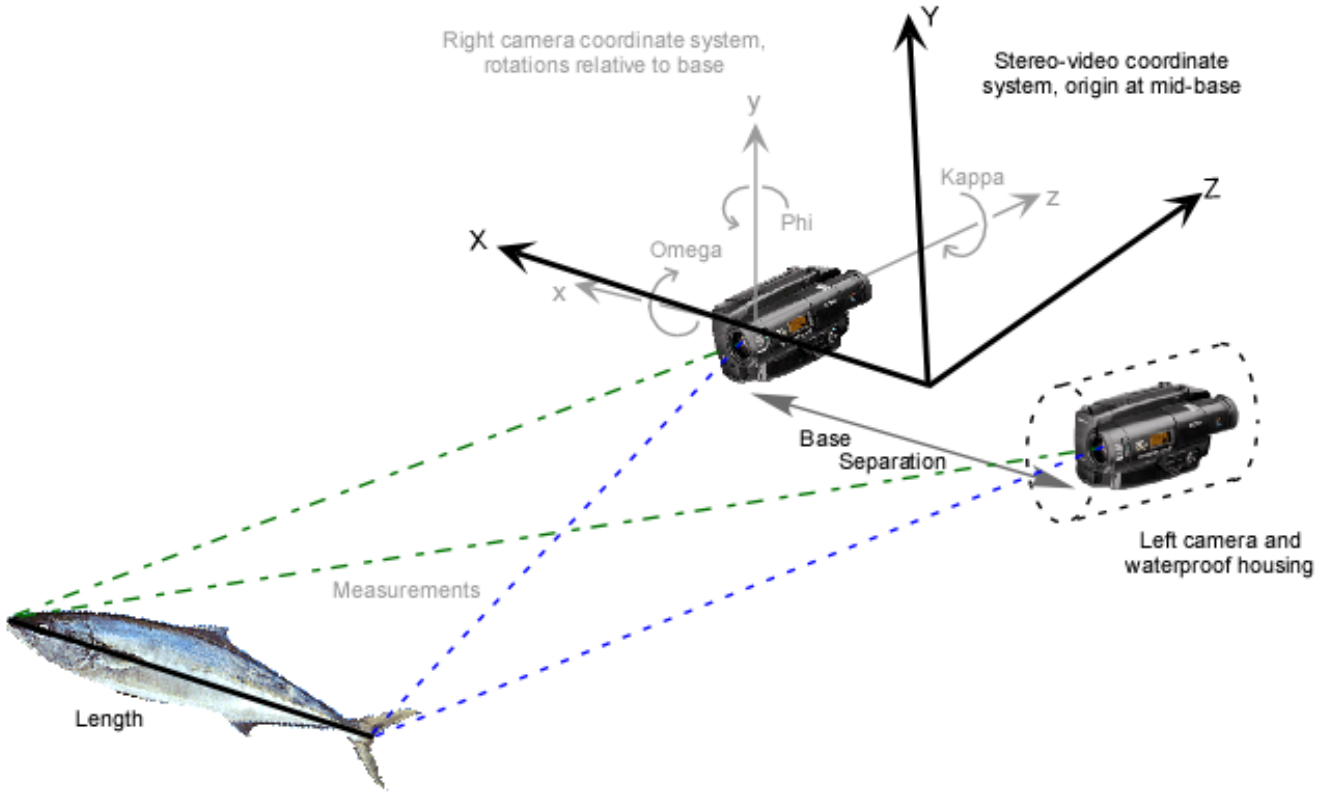

Figure 2. Stereo-video intersection measurement for the snout to tail fork length of a fish.

Status of Identification and Measurement Techniques

Many published papers and reports describe video based systems for automated or semi-automated processes, incorporating some or all of detection, identification, classification, measurement, tracking and counting of fish. Few systems claim to be fully automated, capable of all functions mentioned and can produce high levels of measurement and classification accuracy. The detailed reviews provided in this section are selected primarily on the basis of automation of some or all of the processes within a fully or partially uncontrolled environment.

One of the first successful, operational approaches to the automated measurement of fish in video sequences is described by Tillet et al. (2000). The technique is based on 3D Point Distribution Models (PDM), which are composed of landmark locations on the outline of the fish, in this case Atlantic Salmon held in a small aquaculture tank. The PDM specific to the species is developed from manual stereo-image measurements of a small sample of fish, leading to a mean shape and an estimate of the variations based on Principal Component Analysis (PCA). The PDM is independent of the scale and in-plane orientation, but is limited only to the silhouette of the fish and does not include other landmark features.

The automated measurement process within the stereo-images commences with edge detection based on the Sobel operator. Initial segmentation using a reference image is not required in this case because the tank wall in the background has a relatively uniform intensity. A heuristic threshold is used to eliminate noise and reduce non-fish edges on the basis that higher magnitudes indicate stronger edges. The PDM is then fitted to the edge image based on an energy minimisation function that compares the proximity, the edge gradient strength and the gradient direction of the normal to the PDM. A two-step approach is used that first holds the PDM fixed and searches along the normal to the PDM, the candidate points are then re-projected to detect points on the same edge, and then the PDM is fitted to the best candidate edge points. This process is iterated and the root mean square (RMS) residual compared to a heuristic threshold. Lines et al. (2001) extends this research by applying the technique to low contrast images from sea cages and the introduction of a fish identification step. Candidates are identified based on frame difference images and a binary pattern classifier to locate the characteristic crescent shape of the fish heads.

A success rate of over 90% is reported (Tillet et al., 2000), with failures caused by overlapping fish, large orientation variations and large size variations. The snout to fork lengths derived from the PDM realise a mean error of 5% and a precision of 2.8%, based on comparisons with manual image measurements. Comparisons of the height span of the fish body and accuracy tests against calliper measurements were less favourable. Nevertheless, the fundamental processes of edge detection and then matching to a pre-defined shape are recurrent themes in many approaches to the identification and measurement of fish in video sequences. During the last two decades, there have been many examples of fish detection and identification in fully or partially controlled environments that replicate this basic approach (Larsen et al., 2009; Strachan and Kell, 1995; White et al., 2006).

Spampinato et al. (2008, 2010) reports on a more comprehensive system for detection, tracking, counting and classifying fish that is in continuous operation on a sub-tropical reef in Taiwan waters. Developed by the University of Catania and the University of Edinburgh, the system is based on single video cameras rather than stereo pairs, but is subject to cluttered scenes and poor visibility. Accordingly, the processing steps are significantly more complex than those used in Tillet et al. (2000) and are detailed in point form as follows:

- Two algorithms are combined with an 'and' operation, specifically to reduce false positives in cluttered scenes, to create the background image:
    - A moving average algorithm, based on an adaptive (weighted) update for pixels classified into the background by comparison to a threshold; and
    - An adaptive Gaussian mixture model that classifies background pixels according to likelihood that the pixel is explained by the mixture model;
- Image areas classified as non-background fish candidates are then subject to morphological filters, erosion and closing, dilation and opening, and a median filter to extract the boundary of the fish shape;
- The count of fish in the frame is then determined by a connected component labelling algorithm;
- The fish are then tracked using a combination of two algorithms:
    - a feature vector based on centroid of the image area, the motion vector, the area of the fish and the orientation (angle of the principal axis), comparing changes from frame to frame against heuristic thresholds; and
    - colour matching in HSV space using a comparison of pixel values against the probability that the hue belongs to the histogram of the target object.
- The classification step combines texture and boundary features that are invariant under an affine transformation of the fish, in order to replicate the different views that are present in the video sequences:
    - 70 texture indicators are derived from Gabor filters, the grey level histogram and the grey level co-occurrence matrices and
    - 50 shape features are extracted from a discrete Fourier transform of the boundary and 20 local maxima are extracted from an iteratively smoothed curvature scale space image of the boundary.
- The 120 features are then reduced to a subset of 24 using PCA to create a training set and fish are classified using a discriminant analysis classifier.

The system achieves an average detection rate of 85%, a tracking accuracy of 90%, a counting success rate of 85% (Spampinato et al., 2008) and a classification accuracy of 92% for 10 species (Spampinato et al., 2010). Note that the classification accuracy is only for the detected fish and does not include the 15% of undetected cases.

The approach developed by Spampinato et al. (2008, 2010) also uses edge detection to determine the boundaries of the fish and indirectly uses a pre-defined shape of the fish based on features extracted by the Fourier transform. So whilst the details of the technique are different, the fundamentals echo the approach used by Tillet et al. (2000) and these strategies are consistent in the other recent systems discussed below. In addition, the computer vision techniques of background subtraction to identify foreground objects, and species classification based on the use of image features compared to a training set, are recurrent themes in many systems that deal with uncontrolled underwater environments.

Aguzzi et al. (2009) describe a system for species recognition of deep sea fauna which employs a number of similar elements and fundamental techniques. The system is also a permanent observatory, in this case using time lapse images from a single video camera installed at a depth of 1100m in Sagami Bay in central Japan. Candidate fish and crustaceans are identified using frame subtraction from a reference image derived as a moving average of 100 frames. Regions are refined using grey level and area thresholds to eliminate noise. Objects defined by the regions are then classified against

a manually acquired training set using a combination of 20 Fourier descriptors and the average RGB colour space values. The classifier used is a supervised K-Nearest Neighbour analysis that selects one class or no class for each candidate. Classification accuracies range between 43% and 78%, however the misclassification rate for false positives is less than 1%.

A similar system for the analysis of video sequences captured by ROVs and time lapse images from fixed underwater monitoring stations in Monterey Bay is described in Cline et al. (2010). The Automated Visual Event Detection and Classification System (AVEDac) has been developed at the Monterey Bay Aquarium Research Institute to automatically detect, identify and classify objects of interest that appear in the field of view of the cameras. For the ROV images, a segmentation of objects is based on an adaptive threshold by building a histogram based upon the image, and then the threshold is determined by the value that maximizes the between-class variance of the grey level histogram. For the fixed cameras a graph-cut algorithm is used to detect foreground objects against a running average background image. Several invariant characteristics of the foreground objects, such as average luminance and gradient magnitude, are computed over four different scales. Classification is based on a Gaussian mixture model to classify objects against a training set. Successful classification rates of 95-100% are achieved for ROV video sequences, whilst the rates for the time lapse images are relatively poor and require additional characteristics to be included to improve the classification accuracy.

Wilder et al. (2010) describes the deployment of a stereo camera system at the New York aquarium to automatically detect and classify coral reef fish. The overall success rate of detection and classification of three species of reef fish for the system, when used for a limited sample in this circumstance of partially controlled conditions, is 96%. Background modelling and subtraction is followed by a graph cut algorithm and noise reduction to segment the images. An epipolar search and stereo measurements are then used to determine the pose and size of the fish. The species recognition is then based on two sets of features. The fish shape perimeter, normalised in translation, rotation and scale, is re-sampled to yield a small number of features using PCA. The colour features are extracted from histograms and compared to the histograms from the training sets. Shape and colour features are then processed in a weighted nearest neighbour classification.

Khanfar et al. (2010) and Charalampidis et al. (2012) describe a research project investigating recognition and tracking of fish in video sequences, developed at the University of New Orleans and tested at the Southeast Fisheries Science Centre of the NOAA National Marine Fisheries Service. The initial process applied to the video broadly uses background subtraction, histogram analysis, segmentation, region growing, edge detection and boundary definition using an ‘elastic’ circle (Khanfar et al., 2010). A more detailed description of the steps in the process is as follows:

- create a background image by averaging 200 images without any fish present;
- on a pixel by pixel basis, divide the frame by the background;
- analyse the histogram of the resulting image on the basis that the largest peak is the background and smaller peaks are candidate fish;
- apply the identified thresholds and create a binary image, remove any regions less than a heuristic threshold, crop the original image to the rectangle containing the remaining candidate regions;
- extract the histogram of the cropped image, identify the largest peak as the background and the smaller peaks as candidate fish, set thresholds and extract the regions;
- apply dilation to remove noise, expand and merge adjacent regions, then apply erosion to restore the external boundaries of the regions;

- track, and count, the regions across a series of frames using the location, average intensity, length, width and area of each region;
- isolate each region and apply a Canny edge detector, remove any regions less than a heuristic threshold, identify and connect the closest end-points of the curves if the Euclidean distance is less than a heuristic threshold;
- use a decreasing radius 'elastic' circle to define the outer edges of the region; and
- update the background image based on the non-candidate regions.

The report indicates that the subsequent length measurement process is semi-automated, as an operator must identify the snout and tail fork of each fish on the left image from the stereo-pair, which is then automatically detected on the right image using cross correlation (Thompson, 2013).

The approach to the image processing has been further refined to provide more sophistication and revised to place the emphasis on species recognition (Charalampidis et al, 2012). Some of the basic processing is retained in the four main steps of background subtraction, object detection, object tracking and feature extraction. The amended steps in the process are:

- compute the background image based on the median pixel intensity from multiple frames, subtract the background from the current image, identify non-background pixels using a sample variance from the median, update the background image;
- use morphological image processing (dilation, erosion, median filtering) to remove noise, eliminate small regions and join parts of candidate objects;
- candidate objects are tracked using the corner points of the bounding box and the velocity of the centre point, a Kalman filter is used to predict the location of each region in the next frame;
- Euclidean distances and a measure of the relative change in area are used to match regions from frame to frame, tested against heuristic thresholds;
- If a region matches to the distance but not the area, then it is assumed that a merge or a split of a candidate region has occurred;
- The outline of the fish object is extracted using the elastic circle approach;
- A nearest neighbour classifier (NNC) is used to match feature vectors between candidate fish objects and a training set of three species of fish based on a minimum, weighted Euclidean distance;
- Species specific features are extracted from the candidate region using vertical and horizontal Gabor filters (GF) to identify vertical and horizontal body stripe features associated with each species, if the NNC and GF classifications disagree then the candidate region is labelled as a non-fish;

Once a candidate fish is classified, the tracked region is labelled as the species or as a non-fish, except at the edges of the frame where the candidate fish enters or leaves the scene. Charalampidis et al. (2012) report that, in more than 3,000 frames used as a test sample, all regions were correctly identified. Amongst all of the operational systems documented in the research literature, this system appears to be the most sophisticated and is successful in a completely uncontrolled environment. In harmony with the other systems described, the approach incorporates the fundamental techniques of background image subtraction, edge detection, pre-defined fish shapes and a classification based on feature matching against a training set.

Marine Technology Society Journal, Volume 50, Number 1, Feb 2016, pp. 4-16

Progress on New Methodologies

New approaches to the problems of fish detection, measurement and classification are constantly under development. In this section some recent innovations in some specific tasks are described and placed in context of the overall task of identification, measurement and counting of fish in uncontrolled environments.

Identification and Delineation

Detection of candidate fish comprises two steps: identification of regions and subsequent delineation of the fish outline. The existing work on fish detection from under-water image sequences employs either the differences between successive images (Lines et al. 2001; Spampinato et al. 2008) or histogram-thresholds (Khanfar et al. 2010) to segment a varying number of candidate regions in the frames. Whilst the former approaches appear to be autonomous, the latter method depends on the prior knowledge of the background portion of frames derived from a large number of frames with no fish present.

The identification step is followed by accurate delineation of the fish silhouette. As previously noted, the common approach is the use of edge detectors such as Sobel or Canny. For example, Costa et al. (2006) used edge detection followed by dilation and erosion to remove noise and reduce fragmentation of the fish outlines. Khanfar et al. (2010) used an edge detection algorithm to detect the initial fish outline followed by a shrinking circle to capture the outer boundaries more accurately. These methods, however, can fail due to scene complexity of an uncontrolled marine environment and poor contrast of fish boundaries. A model-based approach to fish detection can offer a solution to this problem.

Active contours, also known as snakes, (Kass et al., 1987) are especially useful for delineating objects like fish bodies that are difficult to model with rigid geometric primitives. Moreover, active contours can be independent of edge gradients with flexibility in initialisation (Chan and Vese, 2001). Figure 3 shows the application of active contours on a low quality, low contrast image.

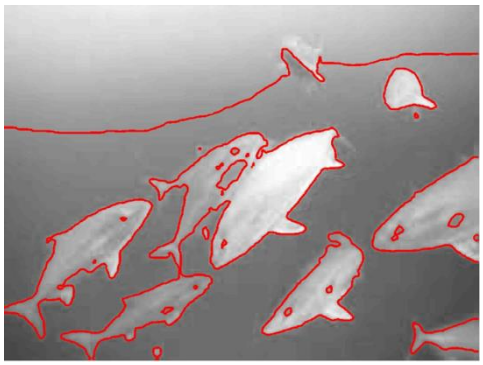

Figure 3. Active contours applied to a low quality, low contrast image.

The area-based active contour model (Chan and Vese, 2001) is based on the techniques of curve evolution and level sets. While parametric active contours cannot automatically handle changes in topology, level sets (Osher and Sethian, 1988) allow for splitting and merging in a natural way and are thus more suited for detection of an unknown number of fish in a video image. Furthermore, the incorporation of high-level prior knowledge about the shape of various fish species within the level set framework can provide an effective solution for images with both poor contrast and occlusions

caused by the uncontrolled and variable nature of marine environment. In the literature, there are some references to successful object extraction using shape information in the presence of image noise, clutter and occlusions (Bailloeul et al. 2005, Cremers et al. 2006, Leventon et al, 2000). The method has been successfully used by Clausen et al. (2007) to segment overlapping fish in an aquaculture environment.

Of various shape prior models, statistical shape models based on Principal Component Analysis (PCA) have been used to establish fish silhouettes (Tillet et al., 2000; Wilder et al., 2010) and have been shown to produce promising results to delineate complex structures (Leventon et al., 2000). PCA enables the representation of global shape variation of the object of interest through a training set of shape templates. While incorporation of global shape information into the Mumford-Shah functional, as reported by Chan and Vese (2001), can detect objects in strongly cluttered scenes, gradient information enables precise capture of local shape variations. Thus, a level sets approach that comprises three separate energy terms, namely shape prior, region- and boundary-based components, can be employed to automate the detection and delineation of fish silhouettes (Ravanbakhsh et al., 2015).

Southern Bluefin Tuna (SBT) are monitored for quota compliance using underwater stereo-video systems during transfers from capture nets to aquaculture cages (Harvey et al., 2003). The SBT must be counted and the snout to tail fork lengths measured to determine the total weight in the transfer. For initialisation, a training set of the silhouettes of 25 SBT samples was generated and the mean shape was used as a shape prior (Ravanbakhsh et al., 2015). The initial shape template was manually digitised and subsequently transformed into the image space through a transformation whose parameters vary within a specified range of values. The initial transformation parameters, however, were determined by estimating the location and orientation of the fish of interest using a Haar classifier (see the next section). A disadvantage of the PCA approach is that areas of high curvature are smoothed through the combination of approximation and averaging.

In Figure 4, some experimental results for SBT segmentation are shown and compare cases in which prior shape information has been incorporated into level sets (Ravanbakhsh et al., 2015). In Figure 4-b, poor contrast and occlusion caused by the adjacent fish produces incorrect results. In Figure and 4-d, incorporation of shape information effectively overcomes these limitations and accurately delineates the foreground SBT despite the uncontrolled background interference from the other SBT.

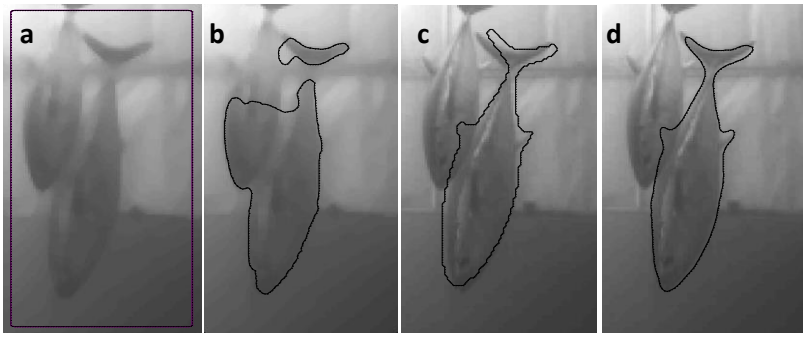

Figure 4. Initial (a) and final (b) unconstrained level set curves, compared to intermediate (c) and final (d) level set curves with shape prior information (Ravanbakhsh et al., 2015).

Measurement

The measurement of the fish snout to tail fork length is very often done manually (Harvey et al. 2003). It has been shown unequivocally that a favourable orientation of the fish to the cameras and

multiple measurements within the sequence of frames does improve the precision of the measurement (Harvey et al., 2002). The shape of fish deforms with swimming action and the Euclidean distance from snout to tail diminishes accordingly. A sinusoidal signal results from multiple measurements taken as the fish swims (Harvey et al., 2003) and the maximum value of the length curve can be used to estimate the true length of the fish. To reduce the effect of swimming motion on length measurement, operators visually inspect different frames in which a particular fish appears. Amongst these frames, measurements are made only on those frames in which the body of the fish appears to be straight.

Manual localization of fish snouts and tails, as well as selection of video frames in which these measurements are made, is labour intensive and an automated solution would greatly improve the efficiency and utility of video monitoring. Once detected, the fish can be tracked using motion models and techniques such as Kalman filters (Charalampidis et al, 2012) that predict the location of the fish in the next frame. Matching within the limited neighbourhood of the predicted region further refines the prediction (Trucco and Plakas, 2006). By iteratively predicting and matching, a swimming fish can be automatically tracked across a sequence of frames in a captured video. Accurate detection followed by tracking of swimming fish establishes correspondences between fish appearing in multiple frames, thereby allowing multiple measurements of the same fish to be taken.

However to acquire accurate measurements, the automatic detection of the fish snout and tail is necessary. Techniques such as bounding box dimensions and region extents (Charalampidis et al, 2012) can be affected by systematic errors in the distance estimation. Edge detection and shape matching (Tillet et al., 2000; Ravanbakhsh et al., 2015) require an initial estimate of the location of the fish in the image and hence *a priori* detection of the snout and tail of the fish is generally a pre-requisite for these methods.

Template matching is one of the primary methods that may be employed to accurately locate the fish snout and tail in the video frames. First, individual templates (rectangular image regions) centred on the snout and tail mid-points are extracted from the sample videos. Then, an efficient template matching strategy, such as that developed by Mahmood and Khan (2012), is employed to locate these templates in the target videos. However, the detection accuracy of template based methods may degrade in the presence of perspective or affine transformations. This requires either multiple templates that capture appearance variations from different viewing angles, or more sophisticated matching techniques that are invariant to these transformations. Mahmood and Khan (2010) proposed a technique to reduce the cardinality of the set of templates required to cover a specified range of transformations. Rova et al. (2007) reported a technique that uses deformable image templates and texture-based classification to identify fish species. However, these types of enhancements may significantly increase the computational complexity of the template matching step.

In addition to the geometric transformations, radiometric variations may also be present between the templates and the target frames. A certain degree of robustness against illumination variations can be achieved by using correlation coefficient as the match quality measure, instead of using Minkowski distance based measures (Kruskal, 1964). Despite the enhancements in template matching methods, they remain error prone due to sudden, spatially non-uniform changes in illumination, variability in background, and the proximity of other similarly-looking fish.

A more effective methodology for locating the snout and tail is to use Haar-like features in a boosted classifier setup. This technique has demonstrated high object detection accuracy, as well as being able to operate in real-time. The method is in wide use for face detection (Viola and Jones, 2001),

even employed by low-cost consumer cameras for real-time face-priority focus. To train the classifier, manually cropped images of the target object (snout or tail) are used so that the classifier can learn which features (among a set of possibly thousands of features) can locate the target with a high level of accuracy. These features, once learned, are then used to construct the object classifier that can locate the presence of the object in cluttered scenes. Due to their high detection speed and ability to perform a scale-space search, Haar classifiers are a promising candidate for locating snout and tail of fish in underwater images, especially in aquaculture setups where a single species of interest is present.

Figure 5 shows an example of a Haar classifier used to identify the snouts and tails of SBT during a transfer between cages. The top row shows the recognition results when a high rejection threshold is used to make conservative decisions about the presence of snout and tail, whereas the bottom row shows how detection results change when the detection threshold is relaxed. It is important to note here that although Haar detector is robust to various image distortions, differences in background image and spatially non-uniform illumination make it difficult to successfully identify all snout and tail candidates. Moreover, Haar detectors are not rotation invariant and would not detect fish that are swimming at an angle that is significantly different to the one at which it was trained. Note that the number of detected snouts and tails in the stereo pair shown in the figure does not match in either case, making the subsequent stage of stereo correspondence quite challenging. Even after relaxing the detection threshold, it is evident that some snouts and some tails are missed due to significant illumination and background variations.

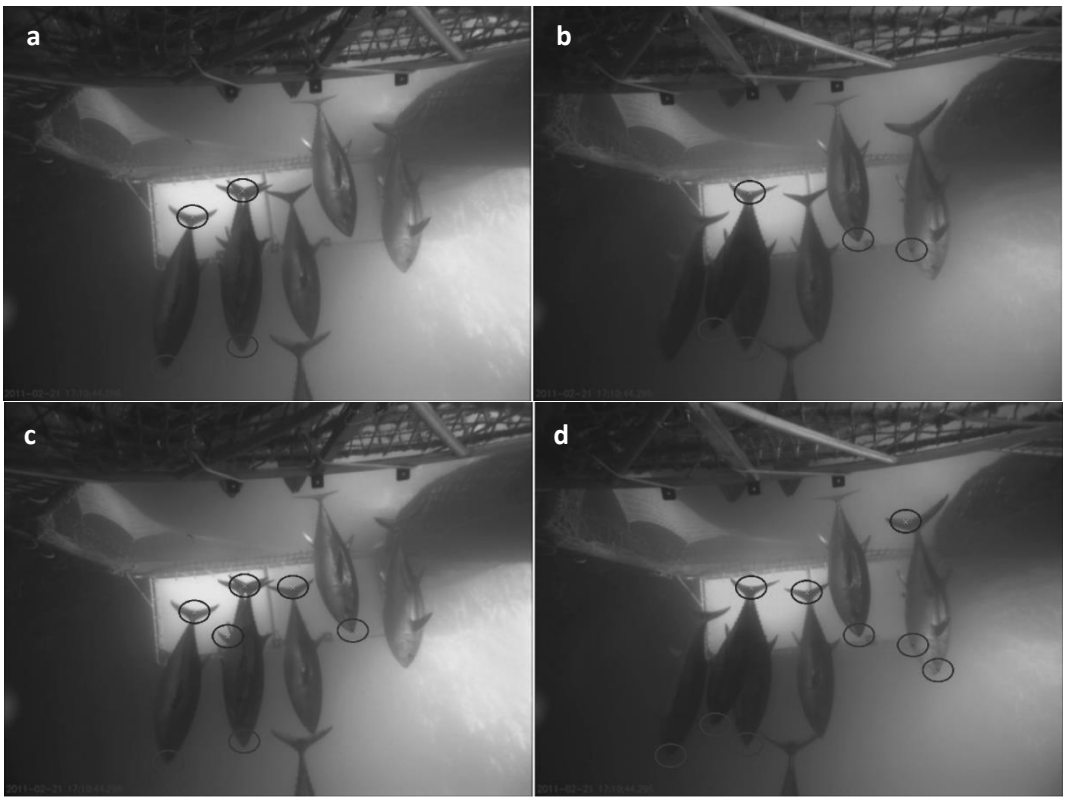

Figure 5. Results of Haar snout and tail detection on a sample stereo image pair. The top row (a and b) shows the results when thresholds for Haar detector are set to a relatively high value to avoid false detections and the bottom row (c and d) shows the results when detection threshold is relaxed. Note

that the apparent vertical orientation of the SBT is caused by the stereo-video camera attachment to the wall of the cage and the water surface is to the right of the images.

The results of independent detection of the snout and tail using Haar detectors can be further improved using relationships between the detected snouts and tails, for instance by constraining the search for tail detection based on the results of snout detection and vice versa. However, the presence of false alarms (e.g. dual snout detection on the right image) and missing both snout and tail of a fish make successful applications of such heuristics difficult and error prone.

To obtain accurate measurements, it is essential to accurately locate the reference points (e.g. tip of the snout and the valley point of the tail). Once the Haar detector gives the location of snout and tail, pin-pointing reference points can be achieved by matching templates of snout and tail having the reference point explicitly marked in the shape prior.

The last step is to establish correspondences between the stereo image pairs such that the snout and tail of each fish is correctly associated with the corresponding image points in the stereo pair. Corresponding points in stereo image pairs lie on epipolar lines (Luhmann et al., 2006). Therefore, the search space for correspondences is effectively one-dimensional. Automatically establishing correspondences in stereo image pairs is a well-studied problem in photogrammetry (Gruen and Baltsavias, 1988; Remondino, 2006) and computer vision, both in a dense and a sparse fashion. Dense methods for correspondence establishment usually rely on block matching to compute the disparity map (Scharstein and Szeliski, 2002). Sparse methods, on the other hand, first identify key points in the images and then capture texture around those locations in a feature descriptor such that correspondences can be established between key-points extracted from the stereo image pairs (Lowe, 2004). Use of a Haar classifier followed by template matching already provides the locations to be matched in the stereo pair. Hence, the search for the corresponding snout and tail reference points in one image can be restricted to the epipolar line corresponding to the snout and tail reference points in the second image. Figure 6 illustrates the use of epipolar geometry to obtain a region of interest in the right stereo frame based on the snout and tail detection in the left stereo frame.

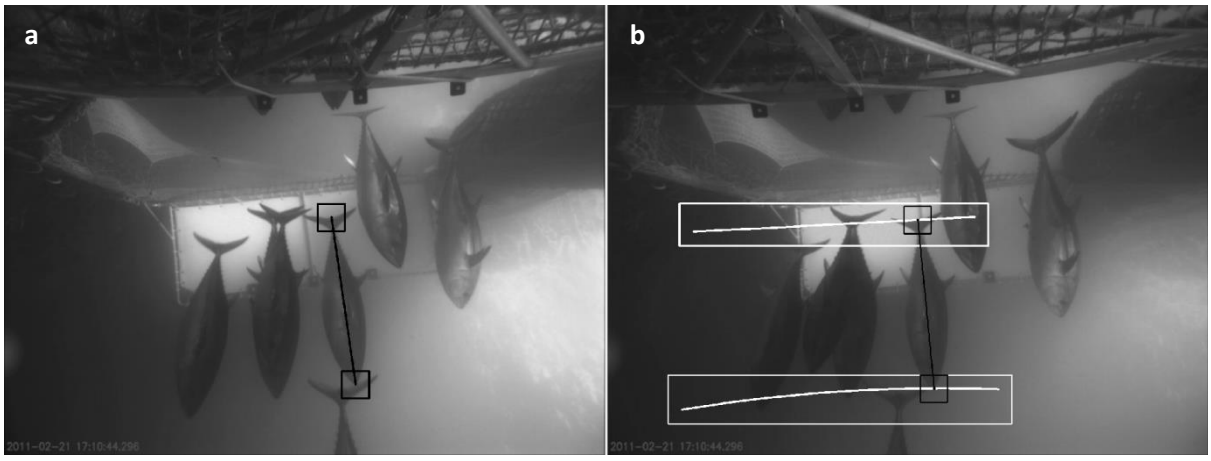

Figure 6 Epipolar lines and regions of interest shown in white in the right image (b) of the stereo-pair correspond to the marked snout and tail locations in the left image (a) of the stereo-pair. Note that due to the substantive radial lens distortions present in these images, the epipolar lines become curved rather than straight, especially at the extremes of the field of view of the camera.

Finally, using epipolar geometry and intersections, 3D locations of the snout and tail of a fish are computed (see Figure 2). The length measurement of the fish resolves to a simple 3D Euclidean distance between the snout and tail. However, due to the swimming motion of fish, measurements taken from a single frame are not reliable. Automated measurements should be made in all frames where the complete fish is visible and the maximum Euclidean distance is adopted as the straight body length (Harvey et al., 2003).

A Synthesis of Methods

The body of knowledge around the effective techniques for the recognition and measurement of fish in image sequences is expanding rapidly. Many of these techniques have their origins in computer vision and have been adapted and customised to increase the effectiveness when applied to cluttered scenes typical of the underwater environment. In a similar fashion to any maturing science, various algorithms and techniques will be developed, applied and analysed. The case studies and recently developed techniques identified in this paper are a small sample of algorithms and techniques that have demonstrated successful outcomes, but have been chosen specifically for the context of the uncontrolled underwater environment.

This analysis presents the opportunity to predict the likely further development and synthesis of the various techniques, based on the demonstrated successes. Selection of individual algorithms and techniques for specific steps in the processing is unlikely to be completely accurate, however some of the steps in the process, such as the use of edge detectors, morphology operators and Kalman filtering, are very well established. Further, the increasing processing power available in the future will encourage the use of majority voting from multiple techniques. The success of this approach has already been demonstrated in some of the case studies.

Accordingly, a synthesis of the current approaches to the recognition and measurement of fish in video sequences, based on the analysis of the successful methodologies, leads to a likely broad scenario of the following steps:

- establish a training set of images of the species of interest;
- capture many images without fish in the scene and compute an average or median image, or a Gaussian model, for the background;
- use differences between the current frame and the background model, or either histogram thresholds or variance from the median, to identify candidate regions;
- apply morphological operators, connected component labelling and geometric tests to remove noise and refine the candidate regions;
- use a Haar or a similar detector, based on the training set, to identify the snout, tail and potentially other distinctive features of the fish, and thereby establish an initial position and orientation of the fish;
- employ edge detection and geometric algorithms or level sets with shape priors to delineate the silhouettes of the fish in each of the candidate regions;
- simultaneously validate the candidate regions as fish or ‘not fish of interest’, and for the fish classify the species using a combination of multiple techniques such as:
    - matching the delineated silhouette of the fish to shape priors,
    - a Nearest Neighbour Classifier to match feature vectors to the training set, and/or
    - Gabor filters to identify body markings on the fish;
    - ANN or SVM classification based on species-dependent visual features
- count the fish based on the validated regions;

- compute the biomass of the fish based on the fish length using an established length-weight regression;
- predict the trajectory using a Kalman filter and track the fish using a combination of feature vectors and colour space matching;
- identify merged and split regions using frame to frame geometric differences for the candidate regions, retaining the non-fish candidate merged regions in the tracking set;
- update the background image or model using the non-fish areas of the image,
- repeat from the difference image step for each frame in the image sequence.

The steps outlined here are presented in a simplistic linear fashion, however in reality the detection, identification, classification and measurement will be an iterative process. As fish and fish candidates are tracked through a sequence, classification as a species of fish at any point in the trajectory will result in re-processing of the images in the sequence to extract a revised estimate of the fish count as well as further refinement of the key dimensions of each fish in the sequence. The optimisation of the length and classification data will provide the maximum confidence in the population count or biomass calculation.

Conclusions

The sustainability of wild fish stocks is of universal concern (Pauly et al., 2002). Commercial fishing commonly targets the large, predatory fish species. History has demonstrated many times that the removal of these top level predators can have unexpected, unpredictable and sometimes catastrophic impacts on marine ecosystems. Fishing has been shown to result in changes in species composition, and in the mean length, biomass and length frequency of target and non-target fishes (Watson et al., 2009). Early detection of the impacts of fishing, especially high catch efforts, is paramount. Accordingly, rapid, accurate and reliable length information is critical for stock assessment.

With declining fish stocks, aquaculture is becoming increasingly important. Like other areas of the world, the Australian aquaculture industry has flourished, and species such as Southern Bluefin Tuna and Atlantic Salmon are being farmed in large numbers (Naylor et al., 2000). The increasing reliance on aquaculture generates the requirement for rapid, continuous monitoring of the growth rates of cage populations in order to maximise the efficiency of feeding regimes.

In all cases, invasive methods of sampling fish to assess biomass and growth rates, such as removal of individuals from the cage or rare species from an ecosystem, have the potential to adversely affect the fish through handling or injury, and consequently may produce skewed data. The measurement of a statistically significant population in a cage, tank, fishery or ecosystem, based on stereo-video measurement, is an accurate, non-contact and non-invasive approach. However, in the absence of advancements in the automation of monitoring, marine scientists will be severely limited in the capacity to rapidly assess the impact of changes in the marine environment. In a scenario of declining biodiversity and reducing catch sizes, the sensitivity, accuracy and especially the speed of the measurement used to assess changes become critical.

To address the need for efficient and accurate assessment of fish populations, this paper has presented a review of the status of automated techniques used for the detection, identification, classification, measurement, tracking and counting of fish in underwater image sequences. The review has identified the common approaches and their shortcomings, leading to an evaluation of two novel techniques that are very likely to contribute to a general solution for the automated, efficient identification and measurement of fish in the wild or in aquaculture facilities. Finally, the paper has

predicted an approach that is likely to provide a baseline system to automate the process, and this will be the subject of further research and development in the future.

Acknowledgements

The authors acknowledge the support of the Australian Research Council through the Linkage Grant LP110201008 'Automation of species recognition and size measurement of fish from underwater stereo-video imagery'. The authors also acknowledge the substantial contribution to the project made by Professor Euan Harvey of Curtin University. Some parts of this paper are based on revised and updated sections of Shortis et al. (2013).

References

Aguzzi, J., Costa, C., Fujiwara, Y., Iwase, R., Ramirez-Llorda, E., Menesatti, P., 2009. A novel morphometry-based protocol of automated video-image analysis for species recognition and activity rhythms monitoring in deep-sea fauna. Sensors, 9(11): 8438-8455.

Bailloeul, T., Prinet, V., Serra, B., Marthon, P., 2005. Spatio-temporal prior shape constraint for level set segmentation. Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR), LNCS (3757), pp. 503-519.

Bouchet, P. J. and Meeuwig, J. J., 2015. Drifting baited stereo-videography: a novel sampling tool for surveying pelagic wildlife in offshore marine reserves. Ecosphere 6(8):137. 29 pages. tp://dx.doi.org/10.1890/ES14-00380.1

Boutros, N., Shortis, M. R. and Harvey, E. S., 2015. A comparison of calibration methods and system configurations of underwater stereo-video systems for applications in marine ecology. Limnology and Oceanography: Methods, 13(5): 224-236.

Chan, T. F., Vese, L. A., 2001. Active contours without edges. IEEE Trans. on Image Processing, 10(2), pp. 266-277.

Charalampidis, D., Gundam, M., Joginipelly, A., Quinteros, M., Ioup, G., Ioup, J., Yoerger, E. J., Thompson, C. H., 2012. Feature analysis for classification of fish in underwater video. Final Report, LA Board of Regents Contract NASA(2011)-STENNIS-02, December 2012. 19 pages.

Clausen, S., Greiner, K., Andersen, O., Lie, K., Schulerud, H., Kavli, T., 2007. Automatic segmentation of overlapping fish using shape priors. 15th Scandinavian Conference on Image Analysis, Aalborg, Denmark, LNCS (4522), pp. 11-20.

Cline, D.E., Edgington, D.R. A Detection, Tracking, and Classification System for Underwater Images. Visual observation and analysis of animal and insect behaviour (VAIB), 20th International Conference on Pattern Recognition (ICPR 2010), Istanbul, Turkey, August 22, 2010.

Costa, C., Loy, A., Cataudella, S., Davis, D., Scardi, M. 2006. Extracting fish size using dual underwater cameras. Aquacultural Engineering, 35(3): 218-27.

Cremers, D., Osher, S., Soatto, S., 2006. Kernel density estimation and intrinsic alignment for shape priors in level set segmentation. International Journal of Computer Vision, 63(3), pp. 335-351.

Gruen, A., Baltsavias, E. P. 1988. Geometrically constrained multiphoto matching. Photogrammetric Engineering and Remote Sensing, 54(5): 633 - 641.

Harvey, E. S. and Shortis, M. R., 1998. Calibration stability of an underwater stereo-video system: Implications for measurement accuracy and precision. Marine Technology Society Journal, 32(2): 3-17.

Harvey, E. S., Shortis, M. R., Stadler, M., Cappo, M., 2002. A comparison of the accuracy of measurements from single and stereo-video systems. Marine Technology Society Journal, 36(2): 38-49.

Harvey, E. S., Cappo, M., Shortis, M. R., Robson, S., Buchanan, J., Speare, P., 2003. The accuracy and precision of underwater measurements of length and maximum body depth of southern bluefin tuna (Thunnus maccoyii) with a stereo-video camera system. Fisheries Research, 63: 315-326.

Harvey, E. S., Fletcher, D., Shortis M. R., Kendrick, G., 2004. A comparison of underwater visual distance estimates made by SCUBA divers and a stereo-video system: Implications for underwater visual census of reef fish abundance. Marine and Freshwater Research, 55(6): 573-580.

Kass, M., Witkin, A., Terzopoulos, D., 1987. Snakes: Active contour models. International Journal of Computer Vision. 1(4): 321-331.

Khanfar, H., Charalampidis, D., Ioup, G., Ioup, J., Thompson, C. H., 2010. Automated recognition and tracking of fish in underwater video. Final Report, LA Board of Regents Contract NASA(2008)-STENNIS-08, 4 June 2010. 40 pages.

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

Leventon, M., Grimson, W., Faugeras, O., 2000. Statistical Shape Influence in Geodesic Active Contours. International Conference of Computer Vision and Pattern Recognition, pp. 316–323.

Lines, J.A., Tillett, R.D., Ross, L.G., Chan, D., Hockaday, S., McFarlane, N.J.B., 2001. An automated image-based system for estimating the mass of free-swimming fish. Journal of Computers and Electronics in Agriculture, 31(2): 151–168.

Lowe, D. G., 2004. Distinctive image features from scale-invariant key points. International Journal of Computer Vision 60(2): 91-110.

Luhmann, T., Robson, S. and Kyle, S., 2006. Close range photogrammetry: Principles, techniques and applications. Caithness, U.K., Whittles.

Kruskal J., B., 1964. Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. Psychometrika 29(1):1-27.

Mahmood, A., Khan, S., 2010. Exploiting transitivity of correlation for fast template matching. IEEE Transactions on Image Processing 19(8): 2190-2200.

Mahmood, A., Khan, S., 2012. Correlation-coefficient based fast template matching through partial elimination. IEEE Transactions on Image Processing 21(4): 2099-2108.

Mallet, D. and D. Pelletier, 2014. Underwater video techniques for observing coastal marine biodiversity: A review of sixty years of publications (1952–2012). Fisheries Research 154: 44-62.

McLaren, B. W., 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. Journal of Experimental Marine Biology and Ecology 471: 153-163.

Menna, F., Nocerino, E., Troisi, S. and Remondino, F., 2013. A photogrammetric approach to survey floating and semi-submerged objects. Videometrics, Range Imaging, and Applications XII, SPIE Vol. 8791: paper 87910H. The International Society for Optical Engineering, Bellingham WA, USA.

Murphy, H. M., G. P. Jenkins. 2010. Observational methods used in marine spatial monitoring of fishes and associated habitats: A review. Marine and Freshwater Research, 61: 236–252.

Naylor, R.L., Goldberg, R.J., Primavera, J.H., Kautsky, N., Beveridge, M.C., Clay, J., Folk, C., Lubchenco, J., Mooney, H. and Troell, M., 2000. Effect of aquaculture on world fish supplies. Nature, 405: 1017–1024.

Osher, S. and Sethian, J.A., 1988. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. Journal of Computational Physics, 79(1): 12-49.

Pauly, D., Christensen, V., Guenette, S., Pitcher, T.J., Sumaila, U.R., Walters, C.J., and D.R. Watson. 2002. Towards sustainability in world fisheries. Nature, 418: 689-695.

Phillips, K., Boero Rodriguez, V., Harvey, E., Ellis, D., Seager, J., Begg, G. and Hender, J., 2009. Assessing the operational feasibility of stereo-video and evaluating monitoring options for the Southern Bluefin Tuna Fishery ranch sector. Fisheries Research and Development Corporation report 2008/44, ISBN 978-1-921192-32-6, 46pp.

Pienaar, L.V. and Thomson, J. A., 1969. Allometric weight-length regression model. Journal of the Fisheries Research Board of Canada, 26:123-131.

Ravanbakhsh, M., Shortis, M. R., Shafait, F., Mian, A., Harvey, E. S. and Seager, J. W., 2015. Automated fish detection in underwater images using shape-based level sets. The Photogrammetric Record, 30(149):46-62.

Remondino, F., 2006. Detectors and descriptors for photogrammetric applications. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 36(3): 49-54.

Rosen, S., Jörgensen, T., Hammersland-White, D., Holst, J. C. (2013). DeepVision: a stereo camera system provides highly accurate counts and lengths of fish passing inside a trawl. Canadian Journal of Fisheries and Aquatic Sciences, 70(10), 1456-1467.

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, Tokyo, Japan, pp 404-407.

Salman, A., Jalal, A., Shafait, F., Mian, A., Shortis, M. R., Seager, J. W. and Harvey, E. S., 2015. Fish species classification in unconstrained underwater environments based on deep learning. Submitted to Limnology and Oceanography: Methods.

Scharstein, D. and Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision 47(1-3): 7-42.

Shafait, F., Ghanem, B.,Mian, A., Harvey, E. S., Shortis, M. R., Ravanbakhsh, M., Seager, J. W., Culverhouse P., Cline, D. and Edgington, D., 2015. Image set based fish identification in unconstrained environments from underwater videos. Submitted to ICES Journal of Marine Sciences.

Shieh, A. C. R., Petrell, R. J. 1998. Measurement of fish size in Atlantic salmon (salmo salar l.) cages using stereographic video techniques. Aquacultural Engineering, 17(1): 29-43.

Shortis, M. R., Harvey, E. S., Abdo, D. A., 2009. A review of underwater stereo-image measurement for marine biology and ecology applications. In Oceanography and Marine Biology: An Annual Review, Volume 47, Gibson, R. N., Atkinson, R. J. A. and Gordon, J. D. M. (Editors). CRC Press, Boca Raton FL, USA. ISBN 978-1-4200-9421-3. 342 pages.

Shortis, M. R., Ravanbakhsh, M., Shafait, F., Harvey, E. S., Mian, A., Seager, J. W., Edgington, D, Cline, D., Culverhouse P., 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.

Spampinato, C., Chen-Burger, Y.-H. Nadarajan, G., Fisher, B., 2008. Detecting, Tracking and Counting Fish in Low Quality Unconstrained Underwater Videos. Proceedings of 3rd Int. Conf. on Computer Vision Theory and Applications (VISAPP), Vol. 2, pp. 514-519.

Spampinato, C., Giordano, D., Di Salvo, R., Chen-Burger, J., Fisher, R., Nadarajan, G., 2010. Automatic fish classification for underwater species behaviour understanding. Proceedings of First ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams. Firenze, Italy, ACM: 45-50.

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(1):145-149.

Thompson, C. H., 2013. Personal communication, February 2013.

Tillett, R., McFarlane, N., Lines, J. 2000. Estimating dimensions of free-swimming fish using 3D point distribution models. Computer Vision and Image Understanding, 79, 123-141.

Trucco, E., Plakas, K., 2006. Video tracking: A concise survey, IEEE Journal of Oceanic Engineering, 31(2): 520–529.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features, IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), Kauai, HI, USA, pp. 511-518.

Watson, D.L., Anderson, M.J., Kendrick, G.A., Nardi. K, Harvey, E.S., 2009. Effects of protection from fishing on the lengths of targeted and non targeted fish species at the Houtman Abrolhos Islands, Western Australia. Marine Ecology Progress Series, 384: 241-249.

Wehkamp, M., and P. Fischer. 2014. A practical guide to the use of consumer-level still cameras for precise stereogrammetric in situ assessments in aquatic environments. Underwater Technology, 32: 111–128. doi:10.3723/ut.32.111

White, D. J., C. Svellingen, C. and Strachan, N.J.C., 2006. Automated measurement of species and length of fish by computer vision Fisheries Research, 80(2):203-210.

Wilder, J., Russell, G. J., Boyle, P., 2010. An automated, real-time identification and monitoring system for coral reef fish. Report of the National Marine Fisheries Service on Automated Image Processing Workshop. NOAA Technical Memorandum NMFS-F/SPO-121, Williams, K., Rooper, C. and Harms, J. (Eds), pp 42-44.