Using Machine Vision to Estimate Fish Length from Images using Regional Convolutional Neural Networks

Monkman, Graham G.; Hyder, Kieran; Kaiser, Michel J.; Vidal, Franck P.

Methods in Ecology and Evolution

DOI:
10.1111/2041-210X.13282

Published: 01/12/2019

Peer reviewed version

Using Machine Vision to Estimate Fish Length from Images using Regional Convolutional Neural Networks

Graham G. Monkman [a,*], Kieran Hyder[d,e], Michel J. Kaiser [c], Franck P. Vidal [b]

[a] School of Ocean Sciences, Bangor University, Menai Bridge, Anglesey LL59 5AB, United Kingdom

[b] School of Computer Science and Electronic Engineering, Bangor University, Dean Street, Bangor LL57 1UT, United Kingdom

[c] The Lyell Centre, Institute of Life and Earth Sciences (ILES), School of Energy, Geoscience, Infrastructure and Society, Heriot-Watt University, Riccarton, Edinburgh EH14 4AS, United Kingdom

[d] Centre for Environment, Fisheries & Aquaculture Science, Pakefield Road, Lowestoft, Suffolk NR33 0HT, United Kingdom

[e] School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk NR4 7TJ, United Kingdom. Tel. +44 (0)1502 524501

* Corresponding author at: School of Ocean Sciences, Bangor University, Menai Bridge, Anglesey LL59 5AB, United Kingdom. Tel.:+ 44 (0)1248 382842.

Email addresses: gmonkman@mistymountains.biz (G.G. ,Monkman); m.kaiser@hw.ac.uk (M.J. Kaiser), kieran.hyder@cefas.co.uk (K. Hyder); f.vidal@bangor.ac.uk (F.P. Vidal)

ORCID

G.G. Monkman http://orcid.org/0000-0002-5645-1834, K. Hyder http://orcid.org/0000-0003-1428-5679, M. J. Kaiser http://orcid.org/0000-0001-8782-3621, F.P. Vidal https://orcid.org/0000-0002-2768-4524

Keywords fiducial marker, photogrammetry, European sea bass, regional convolutional neural network, CNN, videogrammetry

Summary

1. An image can encode date-time, location and camera information as metadata and implicitly encodes species information and data on human activity, e.g. the size distribution of fish removals. Accurate length estimates can be made from images using a fiducial marker however, their manual extraction is time consuming and estimates are inaccurate without control over the imaging system. This article presents a methodology which uses machine vision to estimate the total length (TL) of a fusiform fish (European sea bass).
2. Three regional convolutional neural networks (R-CNN) were trained from public images. Images of European sea bass were captured with a fiducial marker with 3 non-specialist cameras. Images were undistorted using the intrinsic lens properties calculated for the camera in OpenCV, then TL was estimated using machine vision (MV) to detect both marker and subject. MV performance was evaluated for the three R-CNNs under downsampling and rotation of the captured images.
3. Each R-CNN accurately predicted the location of fish in test images (mean intersection over union, 93%) and estimates of TL were accurate, with percent mean bias error (%MBE [95% CIs]) = 2.2% [2.0, 2.4]). Detections were robust to horizontal flipping and downsampling. TL estimates at absolute image rotations $> 20^{\circ}$ became increasingly inaccurate but %MBE [95% CIs] was reduced to -0.1% [-0.2, 0.1] using machine learning to remove outliers and model bias.
4. Machine vision can classify and derive measurements of species from images without specialist equipment. It is anticipated that ecological researchers and managers will make increasing use of MV where image data is collected (e.g. in remote electronic monitoring, virtual observations, wildlife surveys and

morphometrics) and MV will be of particular utility where large volumes of image data are gathered.

1 Introduction

Only a small proportion of the world's marine stocks are sufficiently data rich for formal stock assessments to be performed, hence most marine fisheries are data poor (Costello et al., 2012; Ricard et al., 2012). This is in spite of legislation (e.g. European Commission Decision 2008/56/EC) which requires marine stocks to be exploited sustainably and managed with consideration of their associated ecosystems. The potential for commercial fisheries to negatively impact stocks and ecosystems is accepted, but recreational fishing can also negatively impact fisheries and their associated ecosystem effects (reviews Lewin et al., 2006; Radford et al., 2018). Marine recreational fisheries in particular can lack current and historical data even in developed countries and monitoring of the sector is poor (ICES, 2017; Hyder et al., 2018).

Fisheries assessments have survey phases in which a metrological measurement of the target species occurs (National Research Council, 2006; ICES, 2012). In commercial and recreational fisheries, measurement has traditionally involved observations by researchers, fisheries managers or the fishers themselves. Observer costs are high in commercial monitoring (e.g. Needle et al., 2015) and in the assessment of recreational fisheries (pers. observ. KH). Hence, there has been an increasing interest in remote electronic monitoring (REM) (e.g. White et al., 2006, Chang et al., 2010, Hold et al., 2015, Bartholomew et al., 2018). Videogrammetry and photogrammetry (hereafter, photogrammetry) are becoming commonplace in non-destructive observational marine research (e.g. Dunbrack, 2006, Deakos, 2010).

The use of REM and related approaches is likely to increase as camera technology improves and equipment costs fall (reviews c et al., 2015, Bicknell et al., 2016). Photogrammetry can provide considerable savings when compared to observers (Chang et al., 2010; National

Oceanic and Atmospheric Administration, 2015). Capturing images produces vast volumes of data which is time consuming to process (e.g. Needle et al., 2015, van Helmond et al., 2017). This problem can be alleviated by using motion detection algorithm(s) to extract salient frames from videos (e.g. Weinstein, 2015), but the extracted frames still require manual processing. Object detection with machine vision (MV) could be used to automate the extraction of data from images. Historically, MV has been used to analyse images which have been captured under controlled conditions (e.g. fixed cameras, backgrounds and lighting). This control makes the isolation of the subject from the background (segmentation) much easier, allowing computationally inexpensive techniques to be applied, e.g. using optical flow (Zion et al., 2007; Spampinato et al., 2010; Hsiao et al., 2014) and segmentation by pixel properties (e.g. White et al., 2006, Jeong et al., 2013).

To date, photogrammetry has typically used multi-laser (e.g. Deakos, 2010, Bartholomew et al., 2018) or multi-camera systems (e.g. Dunbrack, 2006, Rosen et al., 2013, Neuswanger et al., 2016), but the equipment is comparatively bulky and expensive. Single camera systems and a fiducial marker (i.e. an object of known scale placed in the camera’s field of view) have been used (Hold et al., 2015; van Helmond et al., 2017) but control of the camera model or the framing of the fiducial marker and subject is usually required (e.g. Rogers, Cambiè, & Kaiser, 2017). Without this control, length estimates are subject to an unknown error because lenses have different optical properties. The additional challenges in extracting quantitative data from images taken by volunteers—or other scenarios where expensive or less portable equipment is unsuitable—may explain the almost complete lack of a suitable solution. Convolutional neural networks (CNN) outperform other methods at object detection and CNN application programming interfaces (API) are now mature enough to be viable for (merely) competent programmers to use regional CNNs (R-CNN) for object detection.

This article explores the feasibility of using MV to automate the identification and size estimation of an important species from images. The objectives are to (i) introduce the software and methods to achieve length estimation with a cheap and portable fiducial marker; (ii) to show that length estimates can be made with no control over the image background, lighting or specialist cameras using a foreground fiducial marker; (iii) provide region of interest (RoI) labelled images of the European sea bass, *Dicentrarchus labrax* (see Appendix S2 Supporting Information); (iv) to compare the speed and performance of three state-of-the-art R-CNN networks.

2 Methods[1]

2.1 Ethics

European sea bass captures were made by recreational fishers and a commercial vessel as part of their day-to-day activity. All reasonable measures were taken to minimise air exposure time to the fish while photographs were taken. Ethical approval was granted by the Animal Welfare and Ethical Review Board of Bangor University, Wales, UK.

2.2 Training and validation image acquisition

Training (n = 734) and validation (n = 184) images were obtained from online public sources. The RoI for each image was drawn tight to the fish body, to the limits of the caudal fin tips and the snout vertex (Fig. 1a). Training and inference were carried out in Tensorflow (Google, 2018) using transfer learning with the following pretrained R-CNNs; (i) ResNet-101 (He et al., 2016), (ii) Single shot MobileNet detector (Howard et al., 2017) and (iii) NASNet (Zoph & Le, 2017), abbrevs. ResNet, MobileNet and NASNet respectively.

[1] Appendix S1 Supporting Information contains additional methodological detail.

2.3 Fiducial marker selection and image acquisition

Three ArUco fiducial markers (Garrido-Jurado et al., 2014) of side lengths 25 mm, 30 mm and 50 mm were mounted on polypropylene sheets (Fig. 1b). Photographs of European sea bass were taken on the shore and afloat, with the informed consent of fishers and with 3 different non-specialist cameras (henceforth *marker images*). Fish were posed to minimise body distortion and occlusion. Fish total length (TL) was measured and recorded. The marker was placed on the fish (Fig. 1c) and then photographed.

2.4 Undistorting marker images

Images from each camera were corrected for radial and tangential distortion with the OpenCV API (OpenCV team, 2018). Lens calibration profiles were created in OpenCV for each camera at each supported field of view and focal length (henceforth *undistorted images*).

2.5 Length estimation

An R-CNN predicts the rectangle which most accurately bounds the subject within the image and then defines the detection as a rectangle with four vertices. Intersection over Union (IoU) measures the accuracy of object localisation by comparing the area of a manually defined ground truth rectangle which bounds the subject with the bounding rectangle predicted by the R-CNN. Each model outputs an objectness score (*score*) which is interpreted as the probability that the proposed region contains the predicted class (Ren et al., 2017).

When estimating TL, the pixel length of the long side of the detection rectangle approximates to the TL (pixels) of the fish. The real-world length per pixel, $\bar{l}$ was estimated from the four sides of the detected ArUco marker according to, $\bar{l} = \frac{1}{n} \cdot \sum_{1}^{n} l/p_i$ where p_i is the i^{th} side length in pixels, and l is the real-world side length (e.g. 50 mm). The accuracy of $\bar{l}$ was validated manually (Linear Regression, $b = 1.003$, $R^2 = 0.999$) using ImageJ (Schneider et al., 2012). Mean absolute error (MAE) and mean bias error (MBE) are reported and are calculated as follows, $MAE = \frac{1}{n} \cdot \sum_{i=1}^{n} |l_i - \widehat{l_i}\,|$ and $MBE = \frac{1}{n} \cdot \sum_{i=1}^{n} l_i - \widehat{l_i}$ where l_i is the i^{th} estimate of TL and $\hat{y}_i$ is the expected (i.e. actual) TL of the i^{th} element. Hence a negative bias represents an underestimate of TL.

2.6 Detection and length estimation with rotation, flipping and downsampling

The accuracy of TL estimates under three translations were checked, these were; (i) image rotation between -30° and 30° in increments of 1°; (ii) horizontal flipping of the image by the x-axis, i.e. the line $x = 0.5 \cdot width$; and (iii) image downsampling by a factor of 1.5, to a minimum image height or width of 50 pixels. TL estimates for rotated images were corrected based on the geometry of the detection box under increasing rotation in relation to the snout and caudal vertices of the subject.

2.7 Removing outliers and modelling bias

NASNet R-CNN detections were split into training and test data. Training data were used to identify biased outliers using an isolation forest (Liu et al., 2008; Pedregosa et al., 2011) with the variables; (i) ratio of height to width of the detection, (ii) objectness score and (iii) % MBE. Outliers were then removed from the training set and a gradient boost regressor (Friedman, 2002; Pedregosa et al., 2011) trained on the predictors (i) and (ii) above. Outliers were removed from the test dataset and the gradient boost regressor model used to correct bias. Further methodological details are given in Appendix S3 Supporting Information.

Several estimates of length measurements are reported and are listed in Table 1. Means followed by square brackets or the ± notation indicate 95% confidence intervals or standard deviation respectively.

3 Results

For every non-transformed European sea bass image, each CNN generated region proposals with objectness scores > 0.5 (with the exception of a single MobileNet score of 0.01). All regional proposals were at least partially coincident with ground truth, with a minimum IoU of 45% (45% IoU detection shown in Fig. 1b). Negative images had no false detections under any network (score mean of 0.005 ±0.008, n = 30, max = 0.04).

Detection performance between networks was practically indistinguishable on untransformed and horizontally flipped images (Table 2), hence detections were invariant to horizontal flipping (IoU mean; horizontal flip, 93.2% [93.0, 93.4]; untransformed, 92.8% [92.5, 93.0]). This equivalence is despite the large differences in mean detection times (Table 2). Nonetheless, when visualised it is apparent that the NASNet network delivered more consistent object detections with no IoU outliers (Fig. 2). All single MobileNet detections had IoUs > 75% however, ResNet had 7 detections < 75% IoU (1.1% of all detections).

3.1 Length estimates

ArUco markers were consistently recognised using the OpenCV API under natural conditions, with the marker successfully localised in 99.3% of untransformed images. Two detection failures occurred because of over-exposure (Fig. 1e). *Corrected MV-TL* estimates had a MBE of 5.9 mm ±20, compared with MBE derived from *corrected manual-TL* estimation of -0.5 mm ±14.8. C*orrected MV-TL* estimates showed consistent variance in bias across *physical TL* (Fig. 3). On excluding TL estimates made under the noisier ResNet and MobileNet networks, MBE for *corrected MV-TL* estimates was increased by 2 mm to 7.9 mm nevertheless, S.D. decreased to 14.7 mm, matching the precision of manual estimates of TL (*corrected manual-TL*).

Corrected manual-TL and *MV-TL* estimation errors tended to be less accurate and precise (mean squared error, MSE) when made on the shore rather than afloat (Fig. 4, MSE; Afloat, 7.9; Shore, 25.9), and there was no apparent systematic bias in length estimation introduced by the camera model when comparing *corrected manual-TL* estimates (which have lower variance than MV-TL length estimates) with platform as a covariate (ANCOVA, $F_{(2,\ 1787)}$, p = 0.15).

Mean %MBE for *corrected manual-TL* estimates were 0.7% ±4.6, 1.1% ±4.0 and 0.7% ±4.1 for the GoPro Hero 5 action camera, Samsung s5690 smartphone and Fujifilm XP30 camera respectively.

The increased %IoU outliers observed during detection with ResNet and—to a lesser degree—the MobileNet single shot detector manifest as the %MBE outliers in Fig. 4. The ResNet detector produced 9 of the top 10 MV associated underestimates (fully corrected percent errors of -16.4% to -38.0%). These errors arose because detections followed the approximate pattern observed in (Fig. 1d), with the ResNet detector occasionally truncating the detection. This behaviour was not observed in the other detectors on untransformed images (i.e. an image which has not been flipped, downsampled or rotated).

3.2 Scale

ArUco marker detection was robust to downsampling to approximately 30% of the original image size (original image size, mean = 1355 by 1029 pixels, or 1.5M pixels^2). ArUco markers were approximately 18 pixels^2 at 30% of original image size and images were approximately 400 by 300 pixels (120k pixels^2). At 30% image size the marker detection rate was 93% however, this dropped to 53% at the next scaling factor of 20% (Table 3). The networks on average, maintained objectiveness scores of ~98% at the 20% scaling factor, where the mean image size was 41.4k pixels^2 (i.e. ~203 pixels^2). At this image size, the average ground truth RoI was 158 by 23 pixels. NASNet produced marginally more accurate TL estimates under downsampling. For each network %MAE increased in increments of between 1% and 2% until the downsampling factor exceeded ~30% (mean ground truth width = 238 pixels), after which %MAE began to increase in larger increments. Each network responded similarly to downsampling (Fig. 5), at 20% image size, %MAE = 9.9% ±7.8 which increased markedly to 15.9% ±8.4 at 13% of the original image size at ~153 pixels^2.

3.3 Rotation

The NASNet and ResNet networks behaved similarly under image rotation (Fig. 6) and detection was robust to small rotations, with over 90% of objectiveness scores greater than 50% at absolute rotation ≤ 20% for the NASNet and ResNet networks. At 20° absolute rotation the MobileNet network had 67% of objectiveness scores below 50%. As the absolute rotation angle increased beyond ~15°, NASNet and ResNet predictions of *corrected MV-TL* exceeded 5% %MBE however, %MBE was 2.5% for the MobileNet network (Fig. 6, absolute rotation = 15°, %MBE; NASNet, -5.0% [-5.3, -4.6]; ResNet, -5.3% [-5.9, -4.7]; MobileNet, 2.7% [2.2, 3.3]). This apparently good performance of the MobileNet CNN masks the greatly decreased confidence in regional proposals under this network (score series, Fig. 6) and a corresponding loss of valid detections.

The geometric rotation correction (variable *rotation corrected MV-TL*) did not consistently decrease bias for all rotations (see Appendix S1 Supporting Information) and bias reduction was only marginally improved for the NASNet and ResNet networks (1.2% and 0.5% respectively) however, bias was increased for the MobileNet network (1.0%). The NASNet and ResNet networks displayed a consistent hyperbolic pattern in TL estimation bias through the rotation range and prediction error was consistent across rotations (Fig. 6).

Combining outlier removal and adjusting *rotation corrected MV-TL* per sample with the trained gradient descent regressor model produced a marked reduction in %MBE across rotations. This correction centred bias at ~0% for absolute rotations ≤ 20° (Fig. 7; Table 4). The overall improvement on applying all corrections to MV estimates following lens correction only are unambiguous, with unadjusted *MV-TL* estimates of %MBE = -11.4% [-11.6, -11.2].

4 Discussion

This study introduced a methodology to estimate fish TL using state-of-the-art open-source R-CNNs and associated software applications (e.g. Abadi et al., 2015, OpenCV, 2018). It was shown that the position of an organism in an image could be accurately predicted without strict control over lighting conditions or subject background. The high degree of accuracy of the predicted RoI (> 90% IoU) enabled the accurate estimation of TL. Estimation was achieved without reliance on specialist cameras, multi-camera systems (e.g. Dunbrack, 2006; Rosen et al., 2013) or paired lasers (e.g. Deakos, 2010, Rogers et al., 2017).

Photographing a well-posed subject with a foreground fiducial marker is faster and more convenient than manually measuring and recording the subject length (pers. observ.). Possessing photographs of subjects provides a persistent record which can be used to derive additional measurements, to cross check data and for validation by third parties. In volunteer based research additional data are typically required (e.g. GPS position, date/time, species) and these data can be automatically captured at image acquisition. The potential for automatic recording of much of the required data—including the onerous task of physically recording a dimension—reduces the recording burden on volunteers which can improve participant retention, the volume of data submissions and data quality as observed in surveys (Galesic, 2006; Hoerger, 2010).

4.1 Networks

Of the three networks, NASNet outperformed the ResNet-101 and MobileNet networks. NASNet was particularly effective at limiting outlier detections. However, the NASNet network had the slowest detection speeds of the three and was the most resource intensive. During learning, NASNet had to be limited to a batch size of 1 to fit within the 6 Gb of memory of the NVIDIA 1060 GTX card (configuration files are available in the Supporting Information). This is unsurprising as the NASNet has many more parameters than ResNet (Zoph & Le, 2017).

Neither ResNet nor NASNet are currently capable of performing real-time detections however, MobileNet can be deployed on mobile devices. The performance of MobileNet in this task was arguably better than ResNet and real time detection would be of particular benefit in volunteer based data collection applications where users could be given immediate feedback on the success or failure of a particular recognition task (Fishbrain, 2018; International Game Fish Association, 2018).

4.2 Length estimation

Fish length measurements (TL, fork length FL and standard length SL) are particularly suited to estimation by R-CNN based networks because the longitudinal dimension of an ideal detection corresponds with the distal extremes of the morphological features which delineate these lengths. In this manuscript, TL was used to demonstrate the methodology, but other measurements (including FL and SL) may be estimated by changing the RoIs defined in the training and test images or using previously determined morphometric relationships (e.g. Needle et al., 2015). To date, rectangular ROIs have no history of providing length data in fisheries assessments because R-CNNs are a recent development in MV. However, our results demonstrate the accuracy which can be achieved where body distortion can be limited. Where curvature cannot be controlled, lengths can be estimated by identifying depth midpoints and calculating the line bisecting these midpoints (Strachan, 1993; White et al., 2006) or line fitting to subject contours (Miranda & Romero, 2017), which requires segmentation of the subject from the background. Tensorflow supports this (He et al., 2017; Google, 2018) but further work would be required to validate.

The fiducial marker deployed was particularly easy to identify in fully automated MV processing pipelines and performed well as evidenced by the low bias and high detection rates. Length was more accurately estimated on afloat platforms than on the shore, because a flat surface was available to measure and photograph the subject. Across both platforms and all camera models there was a small but consistent overestimate of size (mean bias error, 1.6%; 6 mm). Possible explanations include an underestimate of lens-subject distance during camera calibration which did not account for the internal distance between the lens and the glass cover of the cameras, or incorrect estimation of the parameters (e.g. mean profile height) used in the length correction calculation.

Bias magnitude was consistent across the range of fish lengths measured (25 cm to 65 cm) hence a correction could be estimated empirically during training. The model used for rotation correction was successful in eliminating bias (%MBE = -0.1%), which brought the error magnitude in line with methods which control the imaging conditions (Hold *et al*. 2015, *0.6% in lobster*; White *et al*. 2006, 0.3%, in halibut), use paired lasers (Deakos 2010, 0.4% in manta rays) or multiple cameras (Rosen *et al*. 2013 1.0% across 3 fusiform fish species).

Despite bias being largely eliminated, outliers in TL estimates were observed (minimised under NASNet). Without rotation, this error was largely attributable to errors arising from the subject pose in the image. Parallax errors arising through depth differences across the fiducial marker and the subject will be a major source of error which are typically dealt with by excluding images following manual review (e.g. Deakos, 2010, Rogers et al., 2017). Correction for tangential deflection of MV designed fiducial markers is generally supported (Garrido-Jurado et al., 2014), but this is unlikely to be a consistent correction for foreground fiducial markers because the tangential displacement of the marker can differ from that of the subject.

4.3 Transformations

Detections and length estimations were robust to flipping and downsampling. Under decreasing image size the fiducial marker was found to be the limiting factor for the automatic extraction of TL. This is an intrinsic limitation of using a foreground fiducial marker where increasing marker size could obscure salient features. The lowest IoU was observed on the smallest fish sampled, where the marker occluded a comparatively large proportion of the subject (Fig. 1d). The effectiveness of the CNN under substantial downsampling indicates that image sizes can be significantly reduced prior to inference to improve speed and reduce memory requirements.

Length estimates were unbiased and acceptably precise at small degrees of rotation. The bounding box under rotation predicted the x-coordinates of the snout and caudal vertices reasonably well, particularly under the NASNet network (see Supporting Information S4). However, the geometric model (Appendix S1 Supporting Information, 1.4.3) largely failed to improve length estimates under rotation. This failure is attributable to the divergence of the geometric model (detailed in Appendix S1 Additional Methods) from the bounding features of the subject. The CNN detections cannot be represented by the geometry of a rotating rectangle (Appendix S4 Supporting Information). Development of a more accurate geometric correction model would be possible should the use case demand it.

Failure to generalise through all rotations poses a serious limitation in some deployment scenarios. Under volunteer image collection, a significant proportion of subject rotations could exceed the experimental rotation limits. A trivially implemented approach to achieve rotation invariance is the brute force repetition of detection through incremental rotations. The optimal detection among all rotations is then determined by some combination of metrics, e.g. height to width maxima. In this article accurate detections were achieved at absolute rotations to ~15° which suggests that 15° steps could be used to reduce the search space. However, it may be more efficient to train the network on incrementally rotated images. This training is relatively trivial and is supported in most CNN APIs. Nonetheless, data on rotation invariance under rotated training images was not published by Zoph and Le, (2017) and R-CNNs are not intrinsically rotation invariant.

4.4 Applications

A foreground marker is cheap and portable, and volunteers cannot inflate size estimates by moving the marker further away from the subject as possible with a background marker. The methodology applies to many visual markers and to multicamera systems, and to any organism for which morphological estimates are made. Difficulties will arise in unconstrained camera systems where the scale indicator is difficult to distinguish in the image, (e.g. lasers in intense light). None specialist markers can be segmented and length estimated using machine vision, such as a standard ruler (Konovalov et al., 2017). Opportunistic fiducial markers could also be segmented (e.g. human face) and used to produce estimates of fish size from historical images as has been done manually to provide ecological data on some species (McClenachan, 2009; Rizgalla et al., 2017).

Correction for lens distortion is critical for accurate photogrammetry as show in this article, particularly with increased use of robust and waterproof action cameras (Struthers et al., 2015; Schmid et al., 2017) which have significant radial distortion. In small scale projects or where the camera model can be restricted then it may be practical for images to be undistorted on an ad hoc basis. However, to deploy large scale volunteer based metrological data gathering it will be necessary to build a repository of lens correction profiles for each camera model. If a camera supports multiple focal lengths and field of views then each unique combination requires a separate profile. Fortunately cameras typically embed state data (e.g. focal length) and camera model in image metadata which can be used to retrieve the correct profile to remove radial distortion. Profile creation can be embedded in an application and requires the capture of multiple images of a regular pattern (e.g. a chessboard). OpenCV (OpenCV team, 2018) provides the open-source code to undistort images.

This article presents a closed problem with *a priori* knowledge that only a single class would occur in the image, this may not be unusual where interest is in a single species. CNNs are adept at discriminating between object classes (e.g. IMAGENET, 2018) and improved predictive models are frequently released (Google, 2018). The task of generalizing to additional species using R-CNN detectors and the combination of approaches outlined is eminently achievable for many species and CNNs have been used in fine grained species classification (e.g. Sun et al., 2016).

Good results were obtained with fewer than 1000 training images and this may be sufficient for fine grained species classification. CNNs have performed well in classifying images according to bird species with fewer than 100 examples per class (Lin et al., 2015). Nonetheless, data augmentation can be employed to improve the models (Perez & Wang, 2017). Augmentation transforms training images as part of the training pipeline to artificially boost the number of training images. Common transformations include rotation, blurring and elastic transformations, and CNN APIs usually have native support for augmentation. Alternatively augmentation can be managed prior to use in a preferred image processing API (e.g. Jung, 2018). It will be extremely difficult to use MV to discriminate between some species without large numbers of high resolution images. For example, identifying the flatfishes *Pleuronectes platessa*, *Limanda limanda* and *Platichtys flesus* is challenging even for postgraduate marine biologists (pers. observ.).

It will be impossible to obtain perfect object detections and length estimations, particularly in diary like volunteer applications. Pragmatically, users could be prompted to provide "hints" to any application to improve detection. For example, the IGFA fish catch log smartphone application (International Game Fish Association, 2018) prompts users to identify the snout and tail of the fish in an image to improve detection. This process could be used to determine subject rotation. Users could also be prompted to identify species where there may be uncertainty and these images can contribute to the training image set. Another smartphone application has used user contributed images to train a species classifier from submitted images (Fishbrain, 2018). Uncertain classifications and length estimations could be clarified by the general public by crowd sourcing as in other successful citizen science projects (e.g. Joly et al., 2014, Silvertown et al., 2015, Zooniverse, 2017) or by using paid-for crowdsourcing services (e.g. Amazon, 2017).

4.5 Conclusion

Automatically extracting metrological data from images provides opportunities to greatly increase the volume and type of data that can be collected in citizen science programmes, directed surveys, remote electronic monitoring, virtual observers and other applications. Further research is needed to reduce the potential bias and increase precision in extracted data in machine vision (MV) systems to achieve mainstream adoption, but continued technological advances will make automated data processing using machine vision in ecology an increasingly viable option without needing a computer science expert to develop bespoke MV solutions.

5 Funding and Acknowledgements

Graham Monkman was supported by the Fisheries Society of the British Isles under a PhD Studentship. KH was supported by CEFAS Seedcorn (DP227AE).

6 Data Accessibility

Tensorflow configuration files, data and images are available at https://github.com/seabass-detection/seabass-detection. The Tensorflow API is available at https://github.com/tensorflow/models/tree/master/research/object_detection.

7 Author Contribution Statement

GM designed the methodology, collected and analysed all data and authored all software routines for the analysis (excepting 3rd party APIs as noted). FV provided guidance on the methodological approaches. All authors contributed to conception and critically appraised the drafts and gave final approval for publication.

8 References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems [Web Page]. Retrieved 10 March 2018, from https://www.tensorflow.org/

Amazon. (2017). Amazon Mechanical Turk: Artificial Intelligence [Web Page]. Retrieved 2 March 2017, from https://www.mturk.com/mturk/welcome

Bartholomew, D. C., Mangel, C., Alfaro-shigueto, J., Pingo, S., Jimenez, A., Godley, B. J., … Godley, B. J. (2018). Remote electronic monitoring as a potential alternative to on-board observers in small-scale fisheries. *Biological Conservation*, *219*(May 2017), 35–45. doi:10.1016/j.biocon.2018.01.003

Bicknell, A. W. J., Godley, B. J., Sheehan, E. V., Votier, S. C., & Witt, M. J. (2016). Camera technology for monitoring marine biodiversity and human impact. *Frontiers in Ecology and the Environment*, *14*(8), 424–432. doi:10.1002/fee.1322

Chang, S.-K., DiNardo, G., & Lin, T.-T. (2010). Photo-based approach as an alternative method for collection of albacore (*Thunnus alalunga*) length frequency from longline vessels. *Fisheries Research*, *105*(3), 148–155. doi:10.1016/J.FISHRES.2010.03.021

Costello, C., Ovando, D., Hilborn, R., Gaines, S. D., Deschenes, O., & Lester, S. E. (2012). Status and solutions for the world’s unassessed fisheries. *Science*, *338*, 517–520. doi:10.1126/science.1223389

Deakos, M. H. (2010). Paired-laser photogrammetry as a simple and accurate system for measuring the body size of free-ranging manta rays *Manta alfredi*. *Aquatic Biology*, *10*(1), 1–10. doi:10.3354/ab00258

Dunbrack, R. L. (2006). In situ measurement of fish body length using perspective-based remote stereo-video. *Fisheries Research*, *82*(1–3), 327–331. doi:10.1016/J.FISHRES.2006.08.017

Fishbrain. (2018). Fishbrain [Web Page]. Retrieved 19 July 2018, from https://fishbrain.com/mission/

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. doi:10.1016/S0167-9473(01)00065-2

Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics*, *22*(2), 313–328.

Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F. J., & Marín-Jiménez, M. J. (2014). Automatic generation and detection of highly reliable fiducial markers under occlusion.

Pattern Recognition, *47*(6), 2280–2292. doi:10.1016/j.patcog.2014.01.005

Google. (2018). Tensorflow detection model zoo [Web Page]. Retrieved 1 May 2018, from https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/detection_model_zoo.md

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2980–2988). Venice, Italy. doi:10.1109/ICCV.2017.322

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). Retrieved from http://arxiv.org/abs/1512.03385

Hoerger, M. (2010). Participant dropout as a function of survey length in Internet-mediated university studies: Implications for study design and voluntary participation in psychological research. *Cyberpsychology, Behavior, and Social Networking*, *13*(6), 697–700. doi:10.1089/cyber.2009.0445

Hold, N., Murray, L. G., Pantin, J. R., Haig, J. A., Hinz, H., & Kaiser, M. J. (2015). Video capture of crustacean fisheries data as an alternative to on-board observers. *ICES Journal of Marine Science*, *72*(6), 1811–1821. doi:10.1093/icesjms/fsv030

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., … Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *ArXiv Preprint, 1704.04861*. Retrieved from http://arxiv.org/abs/1704.04861

Hsiao, Y. H., Chen, C. C., Lin, S. I., & Lin, F. P. (2014). Real-world underwater fish recognition and identification, using sparse representation. *Ecological Informatics*, *23*, 13–21. doi:10.1016/j.ecoinf.2013.10.002

Hyder, K., Weltersbach, M. S., Armstrong, M., Ferter, K., Townhill, B., Ahvonen, A., … Strehlow, H. V. (2018). Recreational sea fishing in Europe in a global context – participation rates, fishing effort, expenditure, and implications for monitoring and assessment. *Fish and Fisheries*, *19*(2), 225–243. doi:10.1111/faf.12251

ICES. (2012). *Report on the Classification of Stock Assessment Methods developed by*

SISAM. ICES CM 2012/ACOM/SCICOM:01 (Report). Retrieved from http://www.ices.dk/community/Documents/SISAM/Report on the Classification of Stock Assessment Methods developed by SISAM.pdf

ICES. (2017). *Report of the Working Group on Recreational Fisheries Surveys (WGRFS), 6–10 June 2016. ICES CM 2016/SSGIEOM:10* (Report). Nea Peramos, Greece. Retrieved from https://www.ices.dk/sites/pub/Publication Reports/Expert Group Report/SSGIEOM/2016/WGRFS/WGRFS_2016.pdf

IMAGENET. (2018). IMAGENET Large Scale Visual Recognition Challenge (ILSVRC) [Web Page]. Retrieved 6 June 2018, from http://www.image-net.org/challenges/LSVRC/

International Game Fish Association. (2018). IGFA Catch Log [Web Page]. Retrieved 19 July 2018, from http://www.igfacatchlog.org/Default.aspx

Jeong, S. J., Yang, Y. S., Lee, K., Kang, J. G., & Lee, D. G. (2013). Vision-based automatic system for non-contact measurement of morphometric characteristics of flatfish. *Journal of Electrical Engineering and Technology*, *8*(5), 1194–1201. doi:10.5370/JEET.2013.8.5.1194

Joly, A., Goëau, H., Bonnet, P., Bakić, V., Barbe, J., Selmi, S., … Barthélémy, D. (2014). Interactive plant identification based on social image data. *Ecological Informatics*, *23*, 22–34. doi:10.1016/j.ecoinf.2013.07.006

Jung, A. (2018). imgaug: Image augmentation for machine learning experiments. Computer Program. Retrieved from https://github.com/aleju/imgaug

Konovalov, D. A., Domingos, J. A., Bajema, C., White, R. D., & Jerry, D. R. (2017). Ruler detection for automatic scaling of fish images. In *Proceedings of the International Conference on Advances in Image Processing* (pp. 90–95). New York, NY, USA: ACM. doi:10.1145/3133264.3133271

Lewin, W.-C., Arlinghaus, R., & Mehner, T. (2006). Documented and potential biological impacts of recreational fishing: Insights for management and conservation. *Reviews in Fisheries Science*, *14*(4), 305–367. doi:10.1080/10641260600886455

Lin, T., RoyChowdhury, A., & Maji, S. (2015). Bilinear CNN Models for Fine-grained Visual Recognition. In *IEEE International Conference on Computer Vision* (pp. 1–14).

Santiago: IEEE. doi:https://doi.org/10.1109/ICCV.2015.170

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. In *Eighth IEEE International Conference on Data Mining* (pp. 413–422). IEEE Computer Society. doi:http://doi.ieeecomputersociety.org/10.1109/ICDM.2008.17

McClenachan, L. (2009). Historical declines of goliath grouper populations in South Florida, USA. *Endangered Species Research*, *7*(3), 175–181. doi:10.3354/esr00167

Miranda, J. M., & Romero, M. (2017). A prototype to measure rainbow trout's length using image processing. *Aquacultural Engineering*, *76*, 41–49. doi:10.1016/J.AQUAENG.2017.01.003

National Oceanic and Atmospheric Administration. (2015). *A Cost Comparison of At-Sea Observers and Electronic Monitoring for a Hypothetical Midwater Trawl Herring / Mackerel Fishery.* (Report). Retrieved from https://www.greateratlantic.fisheries.noaa.gov/fish/em_cost_assessment_for_gar_herring_150904_v6.pdf

National Research Council. (2006). *Committee on the Review of Recreational Fisheries Survey Methods: Review of recreational fisheries survey methods.* (Report). Washington D.C.: The National Academies Press. doi:/doi.org/10.17226/11616

Needle, C. L., Dinsdale, R., Buch, T. B., Catarino, R. M. D., Drewery, J., & Butler, N. (2015). Scottish science applications of Remote Electronic Monitoring. *ICES Journal of Marine Science*, *72*(4), 1214–1229. doi:10.1093/icesjms/fsu225

Neuswanger, J. R., Wipfli, M. S., & Rosenberger, A. E. (2016). Measuring fish and their physical habitats : Versatile 2-D and 3-D video techniques with user-friendly software. *Canadian Journal of Fisheries and Aquatic Sciences*, *13*(June), 1–48. doi:10.1139/cjfas-2016-0010

OpenCV team. (2018). OpenCV: Camera Calibration and 3D Reconstruction [Web Page]. Retrieved 23 April 2018, from https://docs.opencv.org/master/d9/d0c/group__calib3d.html

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine*

Learning Research, *12*, 2825–2830.

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *ArXiv Preprint*, 8. Retrieved from http://arxiv.org/abs/1712.04621

Radford, Z., Hyder, K., Mugerza, E., Ferter, K., Prellezo, R., Townhill, B., … Weltersbach, M. S. (2018). The impact of marine recreational fishing on key fish stocks in European waters. *PloS One*, *13*(9). doi:https://doi.org/10.1371/journal.pone.0201666

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137–1149. doi:10.1109/TPAMI.2016.2577031

Ricard, D., Minto, C., Jensen, O. P., & Baum, J. K. (2012). Examining the knowledge base and status of commercially exploited marine species with the RAM Legacy Stock Assessment Database. *Fish and Fisheries*, *13*(4), 380–398. doi:10.1111/j.1467-2979.2011.00435.x

Rizgalla, J., Shinn, A. P., Ferguson, H. W., Paladini, G., Jayasuriya, N. S., & Bron, J. E. (2017). A novel use of social media to evaluate the occurrence of skin lesions affecting wild dusky grouper, *Epinephelus marginatus* (Lowe, 1834), in Libyan coastal waters. *Journal of Fish Diseases*, *40*(5), 609–620. doi:10.1111/jfd.12540

Rogers, T. D., Cambiè, G., & Kaiser, M. J. (2017). Determination of size, sex and maturity stage of free swimming catsharks using laser photogrammetry. *Marine Biology*, *164*(11), 1–11. doi:10.1007/s00227-017-3241-7

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. doi:10.1139/cjfas-2013-0124

Schmid, K., Reis-Filho, J. A., Harvey, E. S., & Giarrizzo, T. (2017). Baited remote underwater video as a promising nondestructive tool to assess fish assemblages in clearwater Amazonian rivers: testing the effect of bait and habitat type. *Hydrobiologia, 784*(1), 93–109. doi:10.1007/s10750-016-2860-1

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, *9*(7), 671–5. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22930834

Silvertown, J., Harvey, M., Greenwood, R., Dodd, M., Rosewell, J., Rebelo, T., … McConway, K. (2015). Crowdsourcing the identification of organisms: A case-study of iSpot. *ZooKeys*, (480), 125–146. doi:10.3897/zookeys.480.8803

Spampinato, C., Giordano, D., Salvo, R. Di, Fisher, R. B., & Nadarajan, G. (2010). Automatic Fish Classification for Underwater Species Behavior Understanding Categories and Subject Descriptors. In *Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams* (pp. 45–50). Firenze, Italy. doi:10.1145/1877868.1877881

Strachan, N. J. C. (1993). Length measurement of fish by computer vision. *Computers and Electronics in Agriculture*, *8*(2), 93–104. doi:10.1016/0168-1699(93)90009-P

Struthers, D. P., Danylchuk, A. J., Wilson, A. D. M., & Cooke, S. J. (2015). Action cameras: Bringing aquatic and fisheries research into view. *Fisheries*, *40*(10), 502–512. doi:10.1080/03632415.2015.1082472

Sun, X., Shi, J., Dong, J., & Wang, X. (2016). Fish Recognition from Low-resolution Underwater Images. In *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics* (pp. 471–476). Datong, China. doi:10.1109/CISP-BMEI.2016.7852757

van Helmond, A. T. M., Chen, C., & Poos, J. J. (2017). Using electronic monitoring to record catches of sole (*Solea solea*) in a bottom trawl fishery. *ICES Journal of Marine Science, 74*(5), 1421–1427. doi:10.1093/icesjms/fsw241

Weinstein, B. G. (2015). MotionMeerkat: Integrating motion video detection and ecological monitoring. *Methods in Ecology and Evolution*, *6*(3), 357–362. doi:10.1111/2041-210X.12320

White, D. J., Svellingen, C., & Strachan, N. J. C. (2006). Automated measurement of species and length of fish by computer vision. *Fisheries Research*, *80*(2–3), 203–210. doi:10.1016/j.fishres.2006.04.009

Zion, B., Alchanatis, V., Ostrovsky, V., Barki, A., & Karplus, I. (2007). Real-time underwater sorting of edible fish species. *Computers and Electronics in Agriculture*, *56*(1), 34–45. doi:10.1016/j.compag.2006.12.007

Zooniverse. (2017). Zooniverse: The list of active projects [Web Page]. Retrieved 10 February 2017, from https://www.zooniverse.org/projects?status=live

Zoph, B., & Le, Q. V. (2017). Neural Architecture Search with Reinforcement Learning. In *International Conference on Learning Representations*. Toulon, France. Retrieved from http://arxiv.org/abs/1611.01578

601 # 9 Tables

Table 1. Description of variables used in this article.

Variable	**Derived From**	**Comment**
Physical TL	N/A	The direct measurement of the physical fish with a measure.
Corrected manual-TL	Undistorted image	Manual estimation of the marker and fish length from the undistorted image with ImageJ. Parallax corrections applied (Appendix S1 Supporting Information, 1.4.1 & 1.4.2).
MV-TL	Undistorted image	Machine vision estimates of TL from undistorted images with no other corrections.
Corrected MV-TL	MV-TL	MV TL, corrected for parallax errors (Appendix S1 Supporting Information, 1.4.1 & 1.4.2).
Rotation corrected MV-TL	MV-TL	Corrected MV TL plus a geometric correction based on the height and width of the detected region (Appendix S1 Supporting Information, 1.4.3) to adjust for detections under rotation.
Model corrected MV-TL	MV-TL	Rotation corrected MV TL plus correction with machine learnt models generated from training data to remove outliers and correct bias in test data (Appendix S1 Supporting Information, 1.6). Only test data reported.

602

Table 2. Mean percentage intersection over union (IoU) with standard deviation (S.D.) for NASNet (Zoph & Le, 2017), ResNet-101 (He et al., 2016) and single shot MobileNet detector (Howard et al., 2017). Relative detection time (Rel. Det. Time) compares the relative detection speeds where raw detection speeds were calculated per 1000 pixels^2.

	Untransformed		Flipped		Rel. Det. Time
	Mean IoU	S.D.	Mean IoU	S.D.	
NASNet	93.5	2.5	93.3	2.2	1.00
ResNet	92.5	6.2	93.4	5.1	0.36
MobileNet	92.2	3.5	92.8	3.0	0.10

603

Table 3. ArUco fiducial marker (Garrido-Jurado et al., 2014) detection rates under image scaling (factor = 1.5) with width and height minimum limit of 50 pixels. Marker size is the average side length of the marker in the image. G.T. width is the ground truth horizontal length. Columns are means ±S.D. Obj. score is the mean objectness score across all networks. ND = no detections, px = pixels. % Det. is percentage of markers detected. Scale factor is the proportion by which an image was reduced in size.

Scale factor	N	Width (px)	Height (px)	Marker size (px)	G.T. width (px)	Obj. score	% Det.
1	921	1,355	1,029	63 ±15	874 ±132	1.00 ±0.04	100.0
0.67	921	903	685	42 ±10	536 ±79	1.00 ±0.02	99.3
0.44	921	601	456	28 ±6	357 ±53	1.00 ±0.04	98.7
0.30	921	400	303	18 ±4	238 ±35	0.99 ±0.04	92.8
0.20	921	266	201	13 ±3	158 ±23	0.98 ±0.10	52.8
0.13	921	177	133	10 ±3	105 ±15	0.91 ±0.21	13.0
0.09	921	118	88	7 ±1	70 ±10	0.77 ±0.34	1.3
0.06	918	78	58	ND	47 ±7	0.55 ±0.39	ND
0.04	3	62	50	ND	26 ±0	0.005 ±0.007	ND

Table 4. Mean bias error percentage with 95% confidence intervals (CIs) for fish total length estimates made under NASNet (Zoph & Le, 2017) after corrections for lens distortion only (lens only), parallax and geometric correction (corrected) and application of machine learning to remove outliers and model errors (model corrected). The || notation is the modulus function.

	All rotations		$\lvert\text{Rotation}\rvert \leq 20^\circ$	
	Mean	95% CIs	Mean	95% CIs
Lens only	-11.4	-11.6, -11.2	-9.3	-9.4, -9.1
Corrected	-4.1	-4.3, -3.9	-0.2	-2.2, -1.9
Model Corrected	-0.5	-0.6, -0.3	-0.1	-0.2, 0.1