ICES Journal of Marine Science

Contribution to the Themed Section: 'Applications of machine learning and artificial intelligence in marine science'

Original Article

The Prince William Sound Plankton Camera: a profiling *in situ* observatory of plankton and particulates

R. W. Campbell [1]*, P. L. Roberts[2], and J. Jaffe[3]

[1]*Prince William Sound Science Center, Box 705, Cordova, AK 99574, USA*
[2]*Monterey Bay Aquarium Research Institute, 7700 Sandholdt Road, Moss Landing, CA 95039, USA*
[3]*University of California San Diego, 9500 Gilman Dr, M/C 0238, La Jolla, CA 92093-0238, USA*

Corresponding author: tel: +1 907 424 5800; e-mail: rcampbell@pwssc.org.

A novel plankton imager was developed and deployed aboard a profiling mooring in Prince William Sound in 2016–2018. The imager consisted of a 12-MP camera and a 0.137× telecentric lens, along with darkfield illumination produced by an in-line ring/condenser lens system. Just under 2.5×10^6 images were collected during 3 years of deployments. A subset of almost 2×10^4 images was manually identified into 43 unique classes, and a hybrid convolutional neural network classifier was developed and trained to identify the images. Classification accuracy varied among the different classes, and applying thresholds to the output of the neural network (interpretable as probabilities or classifier confidence), improved classification accuracy in non-ambiguous groups to between 80% and 100%.

Keywords: convolutional neural network, machine vision, Prince William Sound, zooplankton.

Introduction

There is a considerable interest in moving marine resource management away from a single-species approach to a more mechanistic ecosystem approach (e.g. Friedland *et al.*, 2012), but that has proved to be challenging in practice given the complexity and variability of large marine ecosystems. Technologies for measuring physical parameters (temperature and salinity) are mature, and technologies for measuring biogeochemical parameters (e.g. nitrate, phosphate) are also now available and reasonably robust (Johnson *et al.*, 2006). Similarly, the practice of using *in situ* fluorescence as a proxy for primary producer biomass is well established (e.g. Strickland and Parsons, 1972). The assessment of higher trophic levels, particularly fish stocks, is also mature and features an array of well-developed methods (e.g. King, 2007). Zooplankton are the link between primary productivity and fisheries, but zooplankton studies have often been sidelined within

ecosystem studies (Mitra *et al.*, 2014) because they are difficult and expensive to enumerate.

In high latitude ecosystems, secondary producers are mostly small Eumetazoan zooplankton (Longhurst, 2006). Although the dominant large grazers are often crustaceans, there is a diversity of other taxa present: most every phylum within the subkingdom has a member that may be found in the plankton during at least part of their life history. Until recently, the assessment of zooplankton was primarily done by collecting them with nets and examining the resulting samples under a microscope. This method is time consuming and expensive and destroys fragile taxa but is required if species-level taxonomic resolution is desired. There has been much work in recent years on new methods to enumerate zooplankton taxa, both *in situ* and *in manus* (reviewed by Wiebe and Benfield, 2003). One of the more promising methods has proved to be *in situ* imagery, which permits the

discrimination of plankton from abiotic particulates, provides a level of taxonomic resolution, and sizing of imaged plankton and particulates (which is useful because biomass, and many physiological rates scale with size).

A number of *in situ* imagers have been developed, including traditional camera-based systems such as the Video Plankton Recorder (Davis *et al.*, 1992), ZOOVIS (Benfield *et al.*, 2003), and the Scripps Plankton Camera (spc.ucsd.edu). A number of systems have also been developed that employ shadowgraph imagery (Samson *et al.*, 2001; Cowen and Guigand, 2008; Ohman *et al.*, 2019); shadowgraph systems possess a very long depth of field and consequently permit sampling large volumes of water. The disadvantage of shadowgraph systems is that only the silhouette of non-transparent plankton is recovered and only greyscale images may be collected.

Given the high abundance of zooplankton *in situ* (order 10^2–10^6 individuals l^{-1}), most imaging systems collect many more images than may be identified manually and there has also been a parallel effort to develop machine vision techniques to automate the identification of those images (Benfield *et al.*, 2007). Early methods included discriminant analysis (Jeffries *et al.*, 1984), and more recently Support Vector Machines and Artificial Neural Networks (e.g. Culverhouse *et al.*, 1996; Hu and Davis, 2005) and Random Forest (Gorsky *et al.*, 2010) methods have been employed successfully.

With recent advances in computing hardware, most notably the development of cost-effective massively parallel graphics processing unit (GPU) based processors, very deep convolutional neural networks (CNNs) have been developed for solving complex computer vision problems such as image classification (Krizhevsky *et al.*, 2012). CNNs of varying architecture are now commonly employed to address the classification of plankton images from *in situ* imaging systems (e.g. Cui *et al.*, 2018; Luo *et al.*, 2018; Schröder *et al.*, 2018; Bochinski *et al.*, 2019; Cheng *et al.*, 2019). Many studies have focused on smaller phyto- and microzooplankton images based on the publicly available WHOI database (Orenstein *et al.*, 2015; Sosik *et al.*, 2015) and report accuracies in the range of 86–96% (e.g. Lee *et al.*, 2016; Cui *et al.*, 2018; Liu *et al.*, 2018). Among larger zooplankton, Luo *et al.* (2018) used a CNN to identify shadowgraph images to a classification accuracy of order of 90%, if rare difficult-to-classify groups were omitted. Bochinsky *et al.* (2019), using a similar image set, reported accuracies between 69% and 98%. Cheng *et al.* (2019) showed accuracies of between 91% and 98% on a seven-class set of shadowgraph images collected by the ZOOVIS camera (Bi *et al.*, 2013). Transfer learning, the use of pre-trained very deep CNNs has been shown to improve both speed and accuracy when classifying plankton image sets (Lee *et al.*, 2016; Orenstein and Beijbom, 2017; Rodrigues *et al.*, 2018; Schröder *et al.*, 2018).

As a part of the GulfWatch Alaska programme (gulfwatch.com), a long-term monitoring effort in the area impacted by the Exxon Valdez oil spill, a WETlabs Autonomous Moored Profiler (AMP) has been deployed in central Prince William Sound annually since 2013. The AMP site is ∼5 nautical miles southeast of Naked Island, in 200 m water depth. The AMP system is a surface piercing profiler that profiles from a parking depth to surface at a user-specified rate and interval. Once at the surface, the profiler connects to a server computer on land via a cellular data link for data upload and command/control telemetry and then pulls itself back down the line to the park depth with a small onboard winch.

In 2015, an *in situ* zooplankton camera system was developed for the PWS AMP. The camera system was based on the Scripps Plankton Camera, but with larger optics and a higher resolution camera, to sample a larger volume of water to better sample mesozooplankton. The camera system was integrated with the profiler electronics and deployed on the profiler during deployments in 2016–2018. We present here a description of the camera system and a CNN-based classification system that was developed using the images collected during the deployments.

Methods
PWS profiler
The PWS AMP system is based on a WETlabs Thetis profiler, which consists of a positively buoyant frame (∼20 lbs), an electric winch, and a 2.8-mm UHMWPE tether. Starting from a user-specified parking depth, the winch pays out the tether at a specified rate to allow the profiler to ascend. Upon reaching the surface, the profiler enters into a "hold" mode, while an onboard cellular modem connects to the local cellular network. Upon connecting, new profile parameters may be sent to the profiler and a small amount of decimated data from the profile sent out. Following that, or if the profiler is unable to connect to the cellular network before a timeout period (as will occur during heavy weather), it engages the winch and pulls the frame back down to the park depth (Figure 1). The system is powered by a 1.5-kW lithium polymer battery manufactured by Bluefin Robotics for autonomous underwater vehicle use, and with the current configuration it is capable of conducting ∼70 60-m profiles per charge.

The instrument suite on the AMP includes a Seabird model 19 CTD, a WETlabs FLNTU chlorophyll-*a* fluorometer/backscatter turbidometer, a Satlantic SUNA nitrate sensor, and a Seabird SBE43 oxygen sensor. During the 2016–2018 deployments, the profiler was set to conduct twice daily profiles from 60 m depth to the surface. Profiles were usually done within 15 min of the solar minimum and maximum of each day. The ascent rate was set to 30 cm s^{-1}.

PWS Plankton Camera
The optical system of the PWS Plankton Camera (PWSPC) includes a 0.137× 143-mm telecentric lens (Opto Engineering TC2MHR-96) mounted on a 12-MP colour camera (a Point Grey Grasshopper GS3-U3-120S6C-C) inside a large pressure housing with a sapphire glass optical port (Figure 2). Illumination is provided from a second pressure housing on titanium standoffs aimed at the imaging system, with a custom white light emitting diode (LED) array focused through condenser lenses (Edmund Optics 125 mm plano-convex anti-reflective coated lenses) and a white LED ring ahead of the condenser lenses, to produce darkfield illumination of the imaged volume (Figure 2). The LEDs are strobed with a control signal from the camera to synchronize with the frame rate. The imaged volume of the camera is ∼450 ml, and the nominal pixel size is 22.6 μm.

The camera takes 12-bit colour images at a maximum frame rate of 7 frames s^{-1}, which produces more data that can be practically logged to disk (∼500 MB s^{-1}, or ∼5.5 TB for a 1 month deployment of twice daily profiles lasting 3 min). However, mesozooplankton are sparse enough that most of each frame does not contain an image of a particle, it is mostly empty space. The PWSPC thus also incorporates an onboard computer (an Odroid XU4) to segment each image and retain regions of
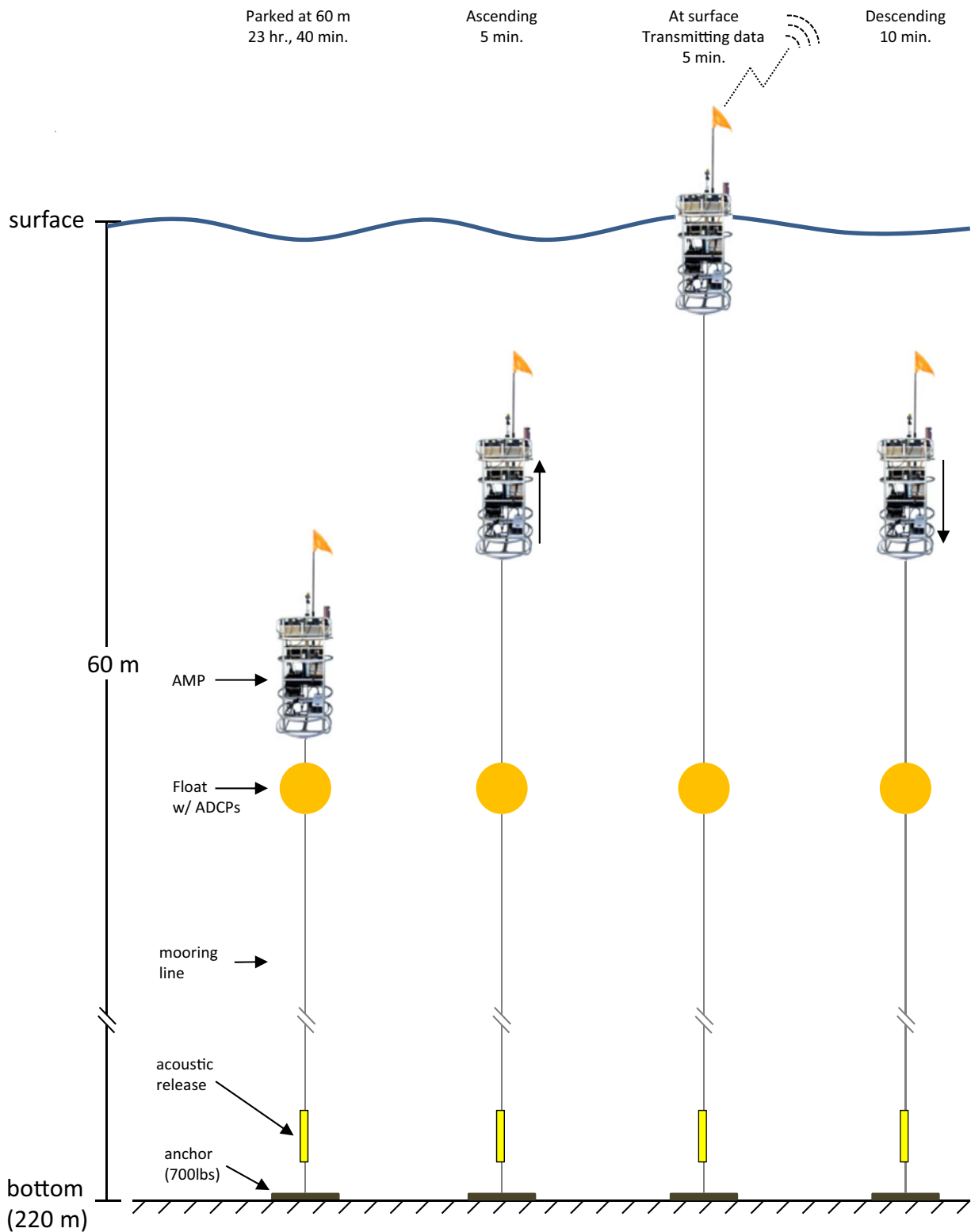
Figure 1. Schematic representation of the PWS profiling mooring and its operation.

interests (ROIs) that contain images of individual plankters. Raw input images were downsampled by a factor of 4 using nearest-neighbour interpolation and then scaled to 8 bits by dividing the pixel values by 256. This approach preserves resolution in one colour channel and avoids the computationally costly debayering operation on the full 12-MP image. ROIs in each frame were detected with the Canny algorithm (Canny, 1986), a multi-step algorithm commonly used to detect edges in
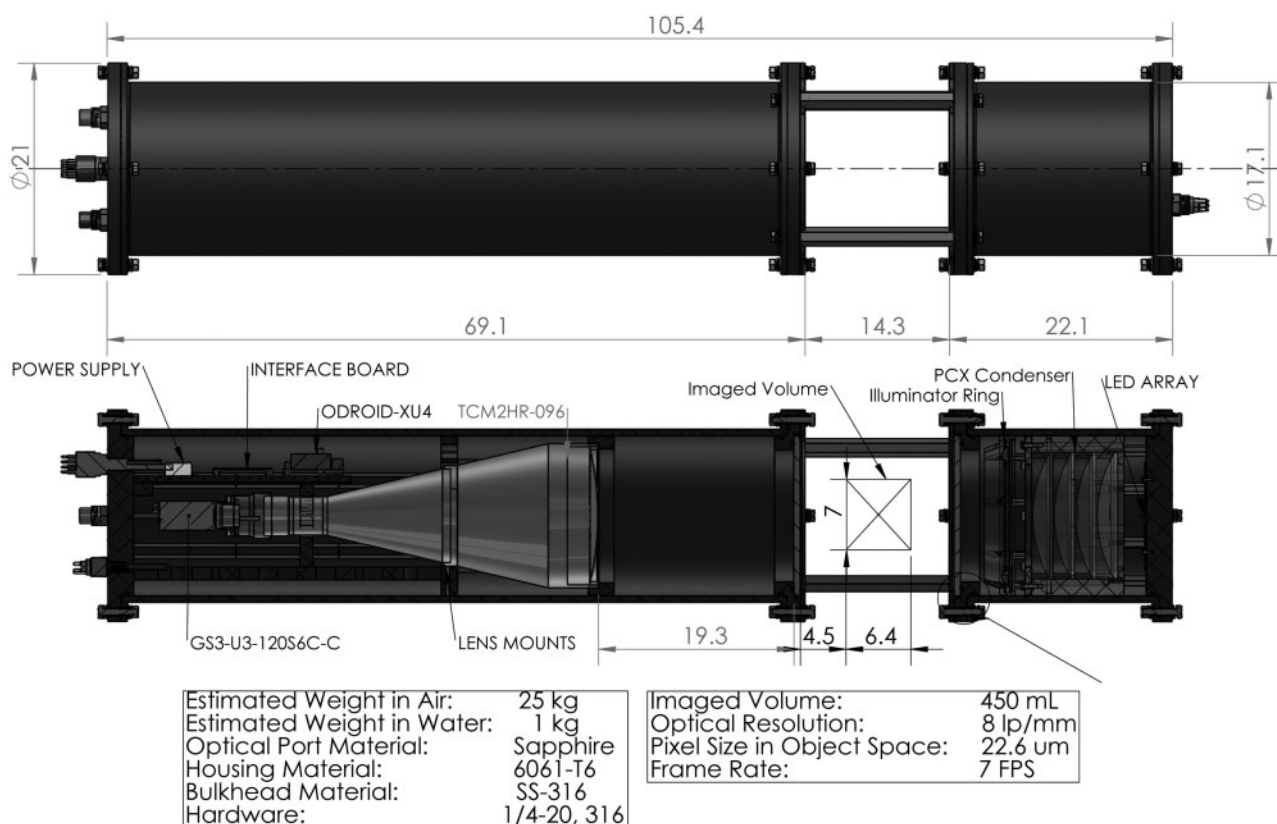
Figure 2. Schematic representation of the PWS Plankton Camera.

images. High and low thresholds were set at 50 and 100, respectively, and kernel size was 3; those thresholds were empirically set and gave good detection of objects in sharp focus or with very high contrast. The edge mask from the Canny operation was then post-processed with a binary morphological closing operation (dilation followed by erosion) with a 5×5 kernel to bridge disjoint edges together. Contours were then detected in the image using the OpenCV findContours function. For each contour with area larger than a threshold of 300 pixels, the contour bounding box was padded by a factor of 50% and upsampled to the scale of the raw image. The coordinates of the padded and upsampled bounding box were then used to extract the ROI from the raw input image and save it to disk in raw 16-bit TIFF format. An upper limit of 50 ROIs per image was imposed by hardware limitations and was not often reached: over the 3 years of deployments, the mean number of ROIs per image was 7.8, and the 50 ROI limit was reached only 0.36% of the time. Images were downloaded from the camera over Gigabit Ethernet during regular service visits to the profiler done every 4–6 weeks.

The PWSPC was integrated with the AMP electronics, and control of the camera system done via an RS232 serial link. Prior to each profile, the AMP control module supplied power to the camera and waited for the onboard computer to boot. After the computer had booted, the AMP sent a string to synchronize the computer clock and an instruction to start logging and then started the profile. As the profile occurred, the PWSPC computer output status messages (time, number of ROIs collected, status messages from the various components) at 1 Hz that were logged by the AMP electronics. ROIs saved to the onboard disk were

Table 1. Specifications of the PWS Plankton Camera.

Exposure time (μs)	10–60
Magnification	0.137×
Field of view (mm)	93 × 70
Pixel size (object space) (μm)	22.6
Optical resolution	8 lp/mm at 30% contrast
Depth of field	64 mm at 8 lp/mm at 20% contrast
Full-resolution imaged volume (ml)	400
Blob detection imaged volume (ml)	>1 000
Frame rate	4 frames/s with ROI processing
Onboard storage (GB)	64
Dimensions (excluding cables)	120 cm L × 18 cm OD
Total system weight (kg)	∼10 (air), ∼2 (seawater)
Power requirements	9–36 V input, 20 W consumption
External communications	RS232, 100 Mbit Ethernet

given timestamped filenames to be used to infer the depth of the profiler at the time each image was taken from the pressure record recorded by the CTD. Following the profile, the AMP controller shut down the computer and removed power to the camera system before returning to the park depth. During profiling the camera and strobes were set to operate at 4 Hz to prevent overlapping images from being taken. The technical specifications of the PWSPC are outlined in Table 1.

Image preprocessing and CNN classifier

Prior to analysis, 16-bit ROIs were debayered to produce a colour image at full camera resolution. These colour images were then

contrast-enhanced by subtracting the minimum and dividing by the maximum of the image. The contrast-enhanced images were then converted to 8 bits by multiplying by 255 and coercing to integer values. The full-resolution colour images were then post-processed using a method similar to the real-time detection method. The images were first converted to greyscale and then filtered with a Sobel edge detector. The edge magnitude image was then thresholded by setting edge magnitudes >2.5 times the median edge magnitude to 255 and others to 0. The edge image was then closed using binary morphological operations, and closed contours are enumerated. The contour with the largest area was then selected as the foreground object. Finally, the contour mask was smoothed with a Gaussian filter and the mask multiplied with the colour image. Each colour channel of the resulting masked image was then deconvolved with the Lucy–Richardson algorithm with seven iterations and a Gaussian point spread function estimate with full width half maximum set to three pixels. The deconvolved colour channels were then combined together to yield the masked, sharpened, colour ROI.

The CNN chosen to classify the PWSPC images was the "Inception v3" model (Szegedy *et al.*, 2015). Inception v3 is a very deep CNN with numerous symmetric and asymmetric neurons that has proved to be adept at image classification problems, and it and its predecessors have consistently ranked highly in the ImageNet Large Scale Recognition Competition (Russakovsky *et al.*, 2015). The ImageNet database for the 2015 competition included ∼1.2 million images in 1000 unique categories; the Inception v3 model had an overall error rate of 5.6%. Because the model is very large, it can be very time intensive to train from scratch, but pre-trained weights (such as from the ImageNet competition) may be used to significantly reduce training time (Tajbakhsh *et al.*, 2016; Orenstein and Beijbom, 2017). To work with the Inception v3 model, the shorter dimension of each image was padded with black values (red-green-blue 0, 0, 0) to make a square image. Both sides of the image were padded to approximately centre the image. Images were rescaled to a dimension of 299 × 299 pixels (the default size of the model).

Resizing the images for input to the CNN necessarily discards size information that is encoded in the image, which can lead to confusion among similar looking but differently sized plankton. For instance, *Pseudocalanus* and *Neocalanus* copepods have a similar appearance but are fairly easily distinguished by size. To reintroduce size information to be used to improve classification, a hybrid architecture was employed, with a second parallel neural network developed to operate on a small set of features extracted from each image, including the major and minor axis lengths and areas in pixels. The first 12 Haralick texture features (Haralick, 1979), a common set of statistics used for image classification (Hu and Davis, 2005), were also included. The features were encoded into a single neuron, batch-normalized, and concatenated with the Inception v3 model prior to the last two layers to produce a hybrid model (Figure 3). The network was implemented in Tensorflow (Abadi *et al.*, 2016) through the Keras front end (Chollet, 2015) in the Python programming language. Training of the network and image classification was done on an NVIDIA Tesla K40 GPU.

CNN training

To produce a training set, images were randomly subsampled from the entire image set. Because the size frequency distribution
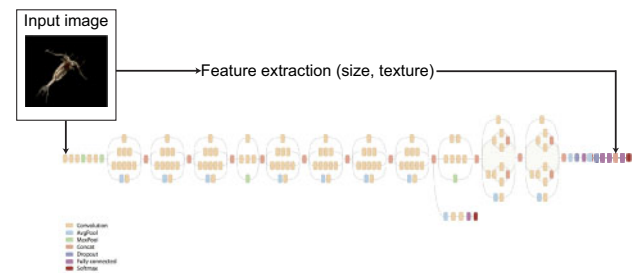

Figure 3. Schematic representation of the Inception v3 CNN and concatenated feature size and texture model.

of the images was roughly lognormal (Figure 4), sampling randomly from the entire set produced batches of images that were mostly smaller particles, which also tended to be of lower resolution and more difficult to identify. Therefore, the images were stratified into four logarithmically scaled size groups based on file size (<1642; >1642, ≤10 000; >10 000, ≤28 183, and >28 183 bytes) before being identified. File size is a useful proxy of image size, and this subsampling scheme allowed more larger images to be classified, which were more likely to be identifiable mesozooplankton. The image set was further stratified by time, such that approximately one-third of the images were taken from each of the 3 years, to provide a subsample representative of all the images.

The amount of training data available is a bottleneck in the training process of CNNs; for complex classification tasks, a large training dataset (10^5 images or more) is desirable. To streamline the identification of the stratified subsets, a custom programme was developed in the Matlab GUIDE framework. The programme consists of a graphical interface that presents an observer with the image displayed at its actual size alongside a larger zoomed version and has a text box into which descriptive text may be entered. Upon entering text and pressing enter, the identification is recorded and the next image in the set presented. Using the GUI, each image could be identified in a few seconds, allowing a large number of images to be identified in a relatively short time.

The training set was produced by an expert zooplankton taxonomist, and each image was identified to the finest taxonomic resolution possible. The training set produced contained 18 868 images within 43 separate classes; some classes were taxon based, while others were based on visual characteristics (Figure 5, Table 2). A number of rare classes (<10 images) were identified during manual classification but were not included for analysis.

The model was initialized with ImageNet weights and was trained using categorical crossentropy as the loss function and the Adam optimizer (Kingma and Ba, 2015); accuracy was the primary metric. The training set was split randomly 90/10 into a training and test set, and 10% of the training set was used for validation purposes during training. Image augmentation (Perez and Wang, 2017) has been shown to improve classification accuracy in classification problems with relatively small amounts of training data and was applied to the images during training. Images were randomly flipped, scaled (±20%), rotated (±90°), or sheared (±8°) as they were input into the model during each training epoch. Network parameters were only retained if they resulted in an increase in validation accuracy.
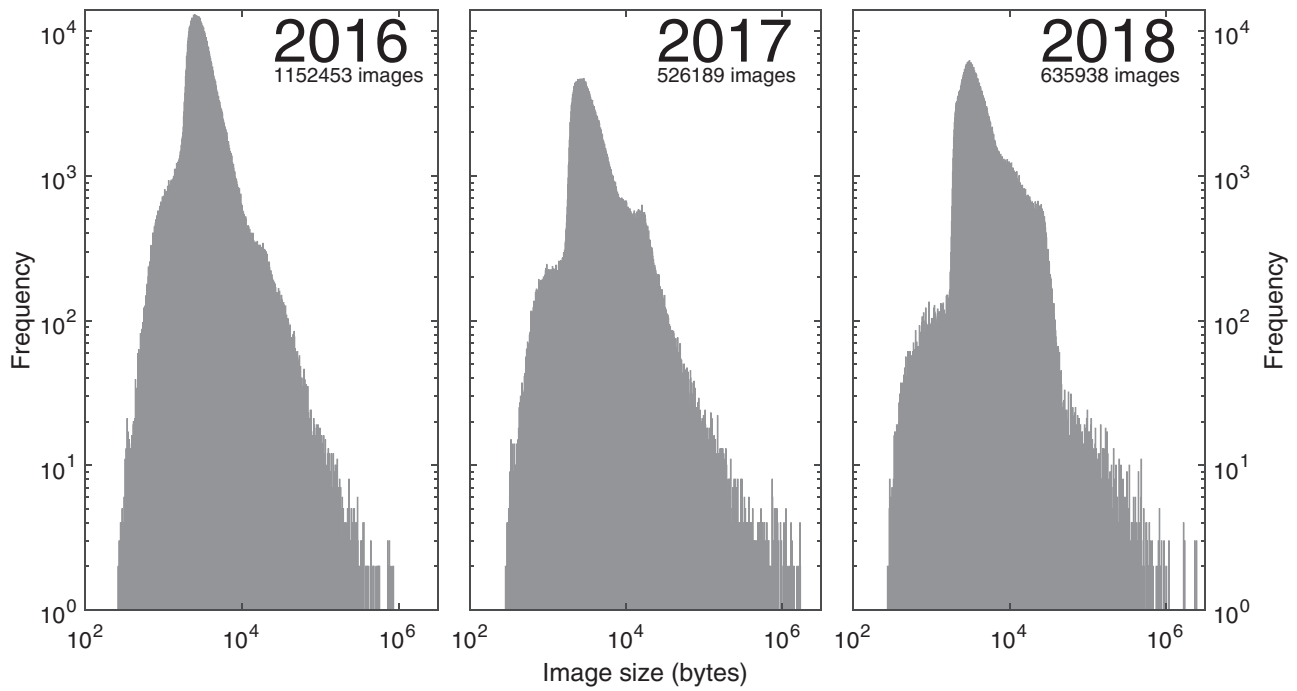
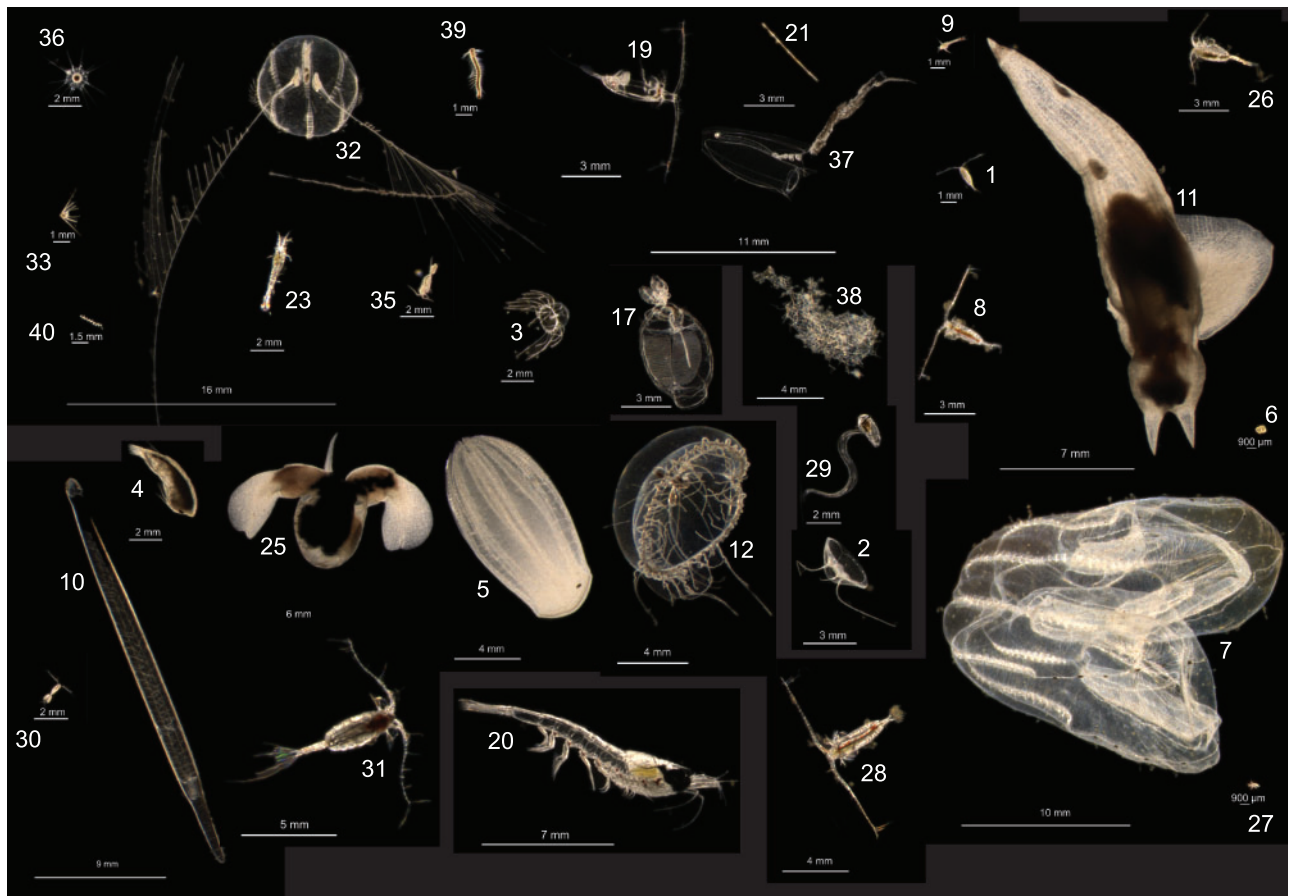**Figure 4.** Size frequency histograms of ROI sizes during the 3 years of deployments.

Figure 5. Examples of non-ambiguous taxa groups among the 43 unique classes identified. Scaling is consistent among different taxa, and the number corresponds to the taxa group number in Table 2.

Table 2. Description of the 43 classes in the training set.

Name	Number	Number of images	Notes
Acartia	1	100	
Aegina	2	147	
Aglantha	3	330	
Amphipoda	4	100	Merged several amphipod species to produce larger group
Beroe	5	100	
Blob	6	2 989	Characteristic opaque large single cell. Possibly Noctiluca
Bolinopsis	7	255	
Calanus	8	613	
Calyptopis	9	198	
Chaetognatha	10	212	
Clione	11	100	
Clytia	12	100	
Cnidaria	13	262	Catchall group of several uncommon species and images not identifiable to species
Cope_lg	14	815	Catchall group of large copepods (approximately Calanus/Metridia sized and larger) not identifiable to species
Cope_sm	15	1 117	Catchall group of large copepods (approximately Pseudocalanus sized and smaller) not identifiable to species
Ctenophora	16	174	Catchall group of Ctenophora not identifiable to lobate groups or Pleurobrachia
Doliolida	17	103	
Dot	18	110	Image artefact: small white dots
Eucalanus	19	177	
Euphausiid	20	97	Juvenile and larger
Filament	21	656	Long thin forms likely diatom chains or large pennate diatoms.
Filaments	22	199	Multiple filaments, often poorly segmented cnidarian tentacles
Furcilia	23	114	Euphausiid furcilia
Larvacea	24	100	Catchall group for non-identifiable and not Oikopleura
Limacina	25	210	
Metridia	26	1 353	
Nauplius	27	256	Nauplii of all types, taxonomically ambiguous
Neocalanus	28	1 574	
Oikopleura	29	207	House usually segmented out
Oithona	30	197	
Paraeuchaeta	31	100	
Pleurobrachia	32	262	
Pluteus	33	308	Primarily echinoderm pluteus larvae
Polychaeta	34	100	Catchall for all polychaetes not identifiable as Spionidae
Pseudocalanus	35	1 004	
Radiolarian	36	251	
Siphonophora	37	204	
Snow	38	172	Amorphous aggregates
Spionidae	39	99	
Spiral	40	177	
Tentacle	41	132	Cnidarian tentacles
Tentacles	42	145	Multiple tentacles in frame.
Unknown	43	2 949	
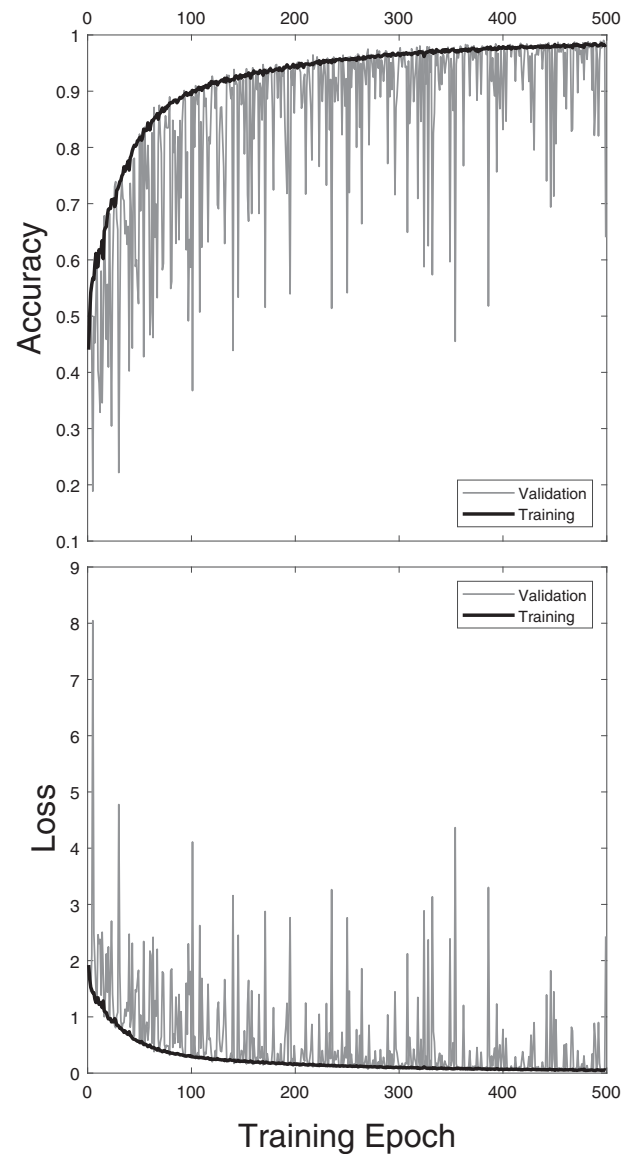
Figure 6. Training and validation accuracy (top panel) and loss (bottom panel) over the 500 training epochs.

Results

A total of 2 424 329 ROIs totalling just over 60 GB were collected during the 2016–2018 deployments (Figure 4).
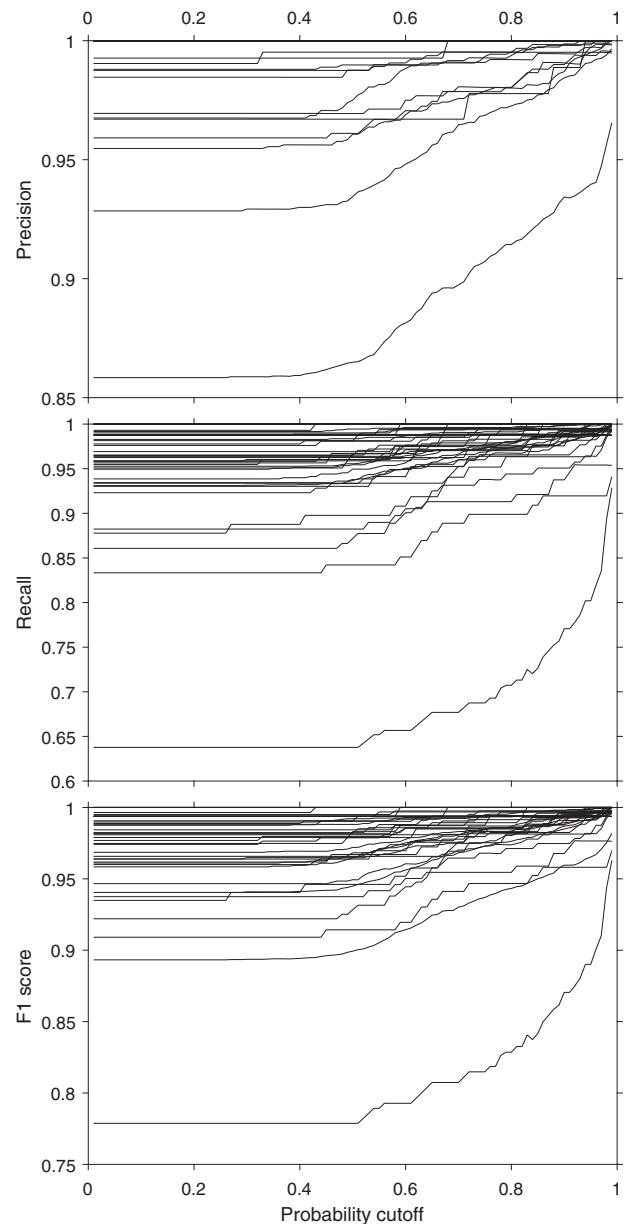
The Inception v3 model was trained on the training set for 500 epochs and took ~275 s per epoch, taking slightly under 40 h. Training accuracy increased to >90% by the 100th epoch, and the rate of increase in training accuracy declined after that, slightly exceeding 98% by the 500th epoch (Figure 6). Validation accuracy and loss was much more variable, presumably due to variability in the image set from epoch to epoch but followed the same trend.

A confusion matrix is a method of representing the accuracy of the classifier (Pearson, 1904; Hu and Davis, 2005; Luo *et al.,* 2018); the confusion matrix of the classifier run on the training data is the theoretical maximum performance that can be expected from the classifier (Figure 7). Furthermore, there are a

Figure 7. Confusion matrix for the classifier applied to the training data only.

number of statistics that maybe be derived to examine the relative success of the classifier. For each class (i), the number of true positives (T), false positives, and false negatives, which may be used to infer the precision, P:

$$P = T_i/(T_i + FP_i).$$

The recall, R:

$$R_i = T_i/(T_i + FN_i),$$

and the F1 score:

$$F1_i = 2P_iR_i/(P_i + R_i).$$

The precision indicates the relative success by the classifier, while recall is a measure of how complete the classifier was; the F1 score is a relative measure of overall accuracy (van Rijsbergen, 1979; Luo *et al.*, 2018).

When used on the training set, the classifier showed comparatively little confusion and generally high scores in the accuracy metrics with precision exceeding recall (Figure 7; Table 3). The "unknown" class had the lowest scores, which is unsurprising since it is by nature a heterogeneous group. A more realistic

Table 3. Accuracy metrics of the training set (left) and test set (right) for each class.

Class name	Training set			Test set		
	Precision	Recall	F1 score	Precision	Recall	F1 score
Acartia	1	1	1	0.7	0.88	0.78
Aegina	1	1	1	0.87	1	0.93
Aglantha	1	0.99	1	0.91	0.83	0.87
Amphipoda	1	1	1	0.95	0.86	0.9
Beroe	1	1	1	1	0.83	0.91
Blob	0.98	0.96	0.97	0.85	0.88	0.86
Bolinopsis	1	1	1	0.88	0.76	0.82
Calanus	0.99	0.98	0.99	0.52	0.53	0.52
Calyptopis	1	1	1	0.78	0.94	0.85
Chaetognatha	1	1	1	0.84	0.82	0.83
Clione	1	1	1	1	1	1
Clytia	1	1	1	0.65	0.62	0.63
Cnidaria	1	0.97	0.99	0.25	0.35	0.29
Cope_lg	1	0.98	0.99	0.47	0.54	0.5
Cope_sm	0.99	0.96	0.98	0.51	0.53	0.52
Ctenophora	1	0.99	1	0.14	0.26	0.19
Doliolida	1	1	1	0.62	0.87	0.72
Dot	0.99	0.73	0.84	0.77	0.53	0.63
Eucalanus	1	1	1	0.92	0.97	0.94
Euphausiid	1	0.96	0.98	1	1	1
Filament	1	0.97	0.98	0.79	0.57	0.66
Filaments	1	0.99	0.99	0.7	0.68	0.69
Furcilia	0.97	0.98	0.97	0.78	0.78	0.78
Larvacea	1	0.99	0.99	0.6	0.6	0.6
Limacina	1	0.98	0.99	0.74	0.74	0.74
Metridia	0.99	1	0.99	0.89	0.84	0.87
Nauplius	1	0.96	0.98	0.88	0.77	0.82
Neocalanus	0.99	1	0.99	0.72	0.81	0.77
Oikopleura	1	0.97	0.99	0.76	0.71	0.74
Oithona	1	0.98	0.99	0.6	0.55	0.57
Paraeuchaeta	1	1	1	0.9	0.9	0.9
Pleurobrachia	1	0.99	1	0.74	0.72	0.73
Pluteus	1	1	1	0.97	0.9	0.93
Polychaeta	1	0.93	0.96	0.8	0.43	0.56
Pseudocalanus	1	1	1	0.7	0.69	0.69
Radiolarian	1	1	1	0.94	0.98	0.96
Siphonophora	1	0.99	0.99	0.71	0.78	0.74
Snow	1	0.95	0.98	0.63	0.79	0.7
Spionidae	1	0.95	0.98	0.85	0.81	0.83
Spiral	1	0.96	0.98	0.92	0.57	0.7
Tentacle	0.99	0.99	0.99	0.89	0.71	0.79
Tentacles	1	1	1	0.86	0.93	0.89
Unknown	0.91	0.98	0.95	0.68	0.7	0.69

Figure 8. Changes in the accuracy metrics (precision, recall, and F1 score) in all of the 43 different classes of the training set as a function of varying the probability cut-off.

assessment of the usefulness of the classifier is to test it on a set of images that it did not see during training (the 10% of images set aside as a "test" set). Deep neural networks tend to overfit to the training set, and image augmentation and aggressive use of dropout layers used in the Inception v3 model are techniques to reduce that (Yamashita et al., 2018). When the classifier was applied to the test set, there were considerably more confusion and lower scores in all of the accuracy metrics (Figure 8; Table 3). If the overall success of the classifier with the different taxa is summarized by sorting by the F1 score (Figure 9), some taxa were resolved quite well, while others were not. Less populated groups

were not less likely to be classified accurately, several of the smaller classes were classified well (many were visually distinctive), and several larger groups (which were more heterogeneous visually) had lower success.

The presence of "unknown" (cannot be identified by a human observer) and novel (not seen before by the network) categories is problematic for CNNs, since their structure assumes a fixed and known set of classes. The softmax function used as the final layer in the Inception model returns scaled outputs that sum to 1 and may be interpreted as probabilities (Bridle, 1990; Goodfellow et al., 2016); the prediction made by the classifier is usually assigned to the category with the highest associated probability.
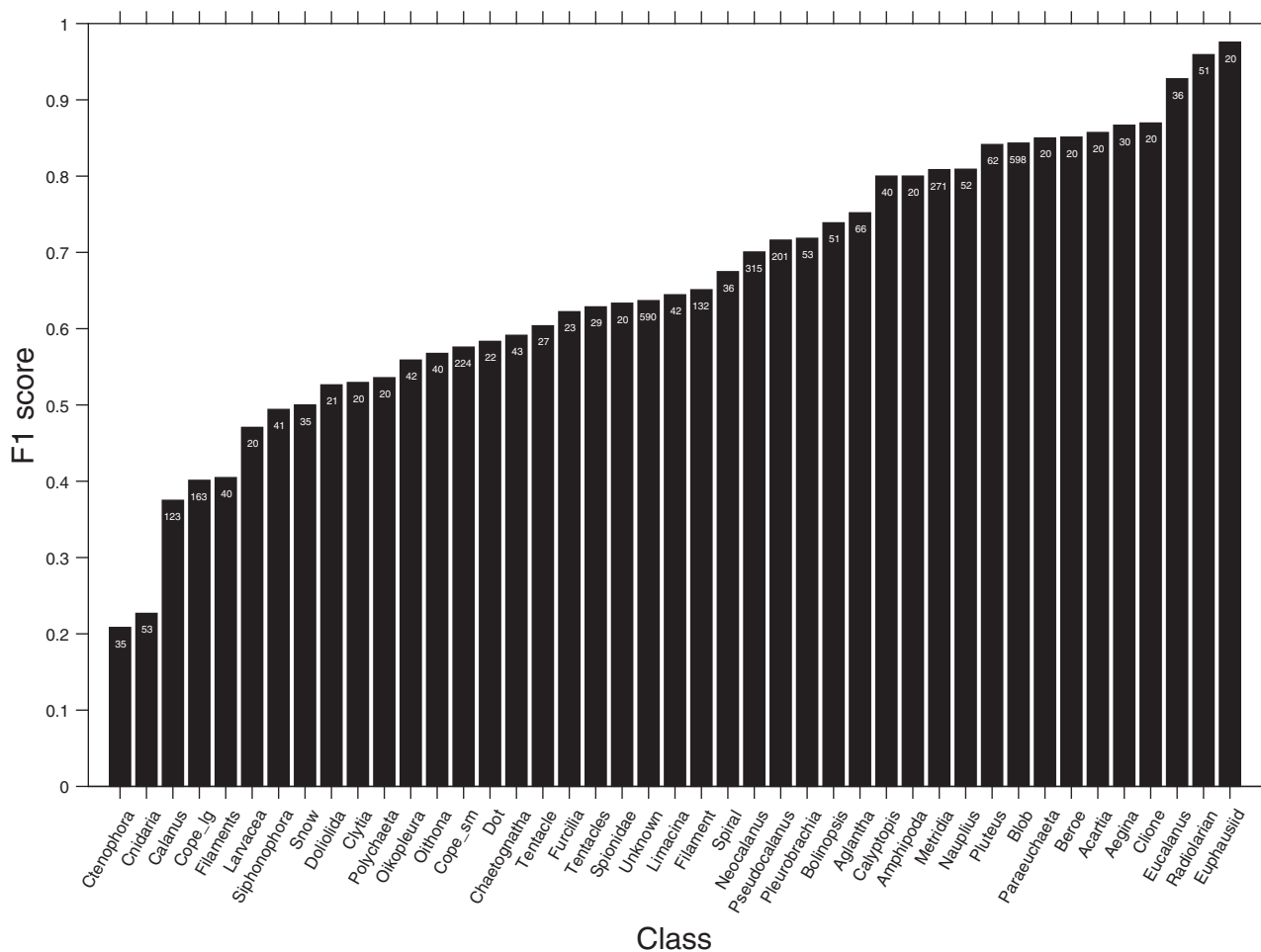
Figure 9. *F*1 scores for each of the 43 different classes in the test set.

The associated probability may also be used as a threshold to reduce the amount of misclassification and new groups that have not been encountered by the model before (Hendrycks and Gimpel, 2018). The technique has been used successfully with plankton images (Faillettaz *et al.*, 2016; Luo *et al.*, 2018).

To examine how a probability threshold might improve classification accuracy, the accuracy statistics were recalculated at varying probability thresholds (i.e. if the prediction for a particular image did not exceed the threshold it was not included in the calculation). Applying this procedure to all classes in the train and test sets (Figures 8 and 10, respectively) produced "trajectories" for each class that generally showed that a more restrictive probability threshold resulted in improvements in classification accuracy. The three taxa with the lowest *F*1 scores in the test set (Figure 9) showed an opposite trend, with a decrease in accuracy metrics at higher probability thresholds. Those classes were among the more ambiguous ones ("Ctenophora", "Cnidaria", "Calanus") that exhibited high confusion with other classes with similar or even overlapping appearance (e.g. "Calanus" and "Cope-lg"; see Figure 11). Because those classes were employed when the human observer had low confidence of the identification, it is perhaps unsurprising that the confidence of the machine classifier remained low as well. A trade-off to this

technique is that as higher probabilities are used, more images are discarded from the analysis (Figure 12). If a 90% threshold is used, the overall error rate drops from ∼30% to ∼10%, but approximately a fifth of the images are discarded. A 95% threshold results in ∼25% of images being discarded. Applying a 90% probability threshold resulted in an increase in most accuracy statistics in most classes (Figure 13, Table 4).

Discussion

The camera system developed here is among the highest resolution *in situ* zooplankton camera systems deployed thus far, with a comparatively large sampled volume as well (Table 5). It is also among the first colour imagers deployed, joining the Video Plankton Recorder (Davis *et al.*, 1992; Lombard *et al.*, 2019) and CPICS (Continuous Particle Imaging and Classification System: Grossmann *et al.*, 2015). Given that the system was designed for battery-limited autonomous vertical profiling (as opposed to long tows), a relatively large sampling volume was desirable, to capture adequate numbers of relatively dilute mesozooplankton (Sheldon and Parsons, 1967) during each profile. Colour information is also useful, because it may be diagnostic of some plankton classes (e.g. red pigments are common in some copepod
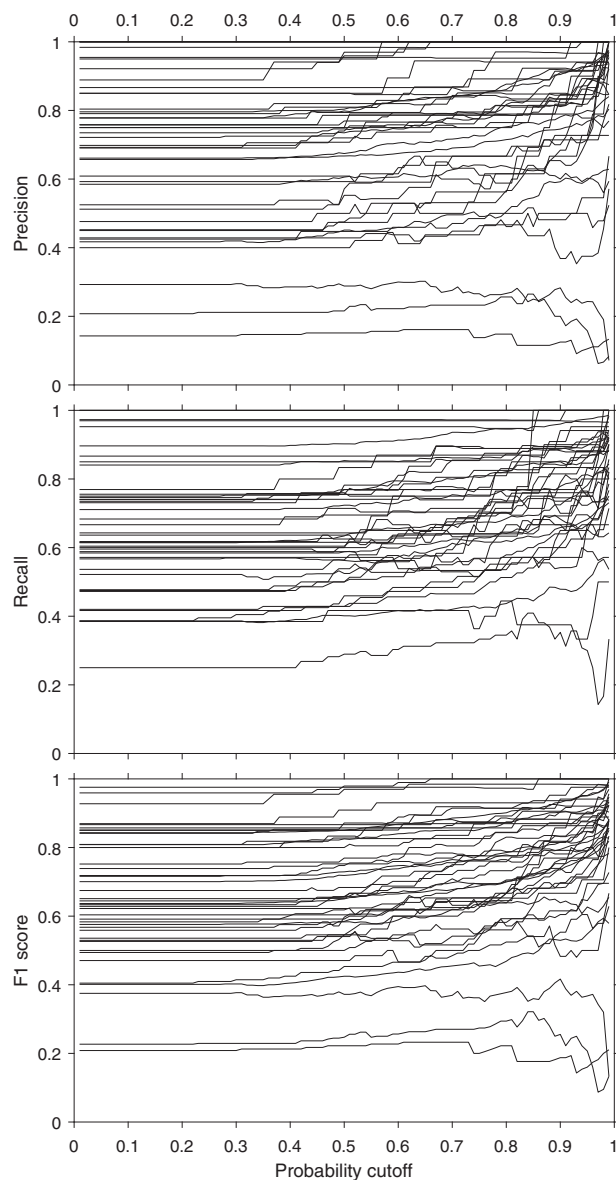
Figure 10. Changes in the accuracy metrics (precision, recall, and *F*1 score) in all of the 43 different classes in the test set as a function of varying the probability cut-off.

well; smaller raw images, when upsized to 299 × 299 pixels for classification, remained less sharp.

Identification of taxa from images is a difficult task, and the error rate of human observers can be significant. In a dino-flagellate classification task, Culverhouse *et al.* (2003) found that expert taxonomists achieved 84–95% accuracy at best, although accuracy dropped considerably among multiple observers (43%). Similar studies in other fields have shown lower ranges in more visually complex situations (69–96%: Austen *et al.*, 2017). The proportion of unknown images can also vary among observers and can depend on the number of classes involved (Cowen *et al.*, 2015). Luo *et al.* (2018) suggest that 90% accuracy be used as a benchmark for automated clas-sification. Those levels of accuracy were possible with the Inception v3 CNN for a number of taxa, particularly if images with lower confidence were not used.

Filtering images by probability, as suggested by Faillettaz *et al.* (2016), improved precision and recall in most taxa by 5–10%. Examination of the confusion matrix post filtering (Figure 13) shows that much of the confusion was between related classes, for instance the large calanoid copepods *Metridia, Calanus, Neocalanus,* and the catchall group Cope_lg and the small cope-pod classes *Oithona, Pseudocalanus,* and the catchall group Cope_sm. There was also confusion among the classes represent-ing gelatinous forms, both cnidarians and ctenophores. The catchall groups Cnidarian and Ctenophora were not well re-solved, while individual taxa within those groups (e.g. the cteno-phore groups *Bolinopsis, Beroe,* and *Pleurobrachia*) were well classified. The catchall groups may have thus likely represented lower quality images (to both human and machine observers) that were more visually heterogeneous and possessed fewer useful features for identification.

Large, very deep CNNs benefit from large training sets (e.g. Cho *et al.*, 2016), and the training set used here is small compared with those used in contemporary machine vision re-search like ImageNet. It is however of similar size to several training sets used in plankton identification studies (order of hundreds to thousands of images per class: Hu and Davis, 2005; Bi *et al.*, 2015; Faillettaz *et al.*, 2016). The roughly lognormal size distribution of plankton populations makes finding less common taxa problematic. The size stratified technique used here attempted to balance the need to obtain examples of as many classes as possible, while not missing out on more rare forms. The classifier developed here discriminated several com-paratively rare (and visually distinctive) taxa with high accuracy. An iterative process where the results of the classifier are checked and added to the training set will aid in producing a larger training set, but that does however leave open the possi-bility of an unknown bias being introduced to the network (i.e. the network probably classifies some images better than other and will bias towards those images). Examination of the unknown class and those removed by probability filtering will also be instructive, though in the case of the latter would involve looking through a very large image set (10^5 images in the case of the PWS image set so far) and would likely need to be subsampled. Training set size will continue to be problematic for plankton studies using imagery, every plankton imager has different optical characteristics, resolution, and lighting, which makes each image set different and not directly compara-ble. Presently, there are several large plankton training image
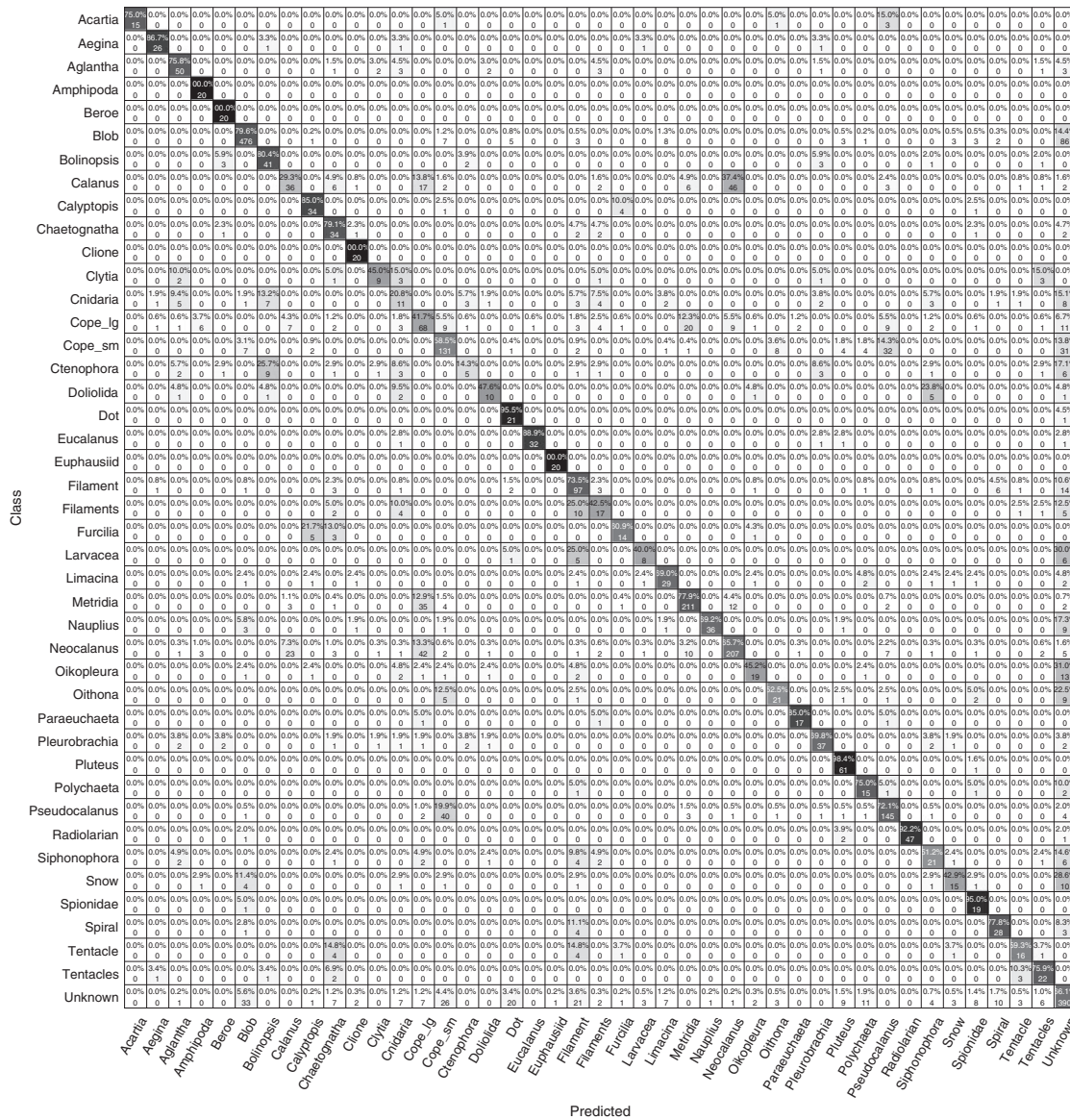
species), and features of the plankters are also discernable in some images (e.g. full guts and lipid sacs in copepods).

The image set collected during the 2016–2018 deployments spanned large phytoplankton to large mesozooplankton and exhibited a diversity of taxa, orientations, and qualities. Although magnification is constant across the depth of field with a telecen-tric lens, they do not have an infinite depth of field and particles on either side of the depth of field will be less sharp than those in the centre. Scattering by small particulates (phytoplankton cells too small to resolve and inorganic particles) may also have re-duced the practical resolution at times. Not all images were sharp enough to detect the features required to identify a plankter to a fine taxonomic level. Larger raw images obviously had more fea-tures (*sensu* Hassaballah and Awad, 2016) that were more useful to a human observer and presumably to a machine observer as

Figure 11. Confusion matrix for the classifier applied to the test set (i.e. images that the classifier did not experience while training).

sets available (Cowen *et al.*, 2015; Orenstein *et al.*, 2015) and, transfer learning, the use of networks pre-trained on other image sets has shown to improve the speed and accuracy of results (Orenstein and Beijbom, 2017; Rodrigues *et al.*, 2018; ICES, 2020).

There is no panacea when approaching the problem of understanding zooplankton dynamics. Zooplankton are dilute, and a large volume of water must be sampled to obtain representative estimates of abundance. Plankton nets sample a large volume of water and allow fine scale taxonomic resolution but are expensive in terms of time and money and damage fragile taxa. Cameras sample a smaller volume of water and provide less taxonomic information but are inexpensive to operate following the initial capital outlay. Obtaining twice daily profiles over several months is simply not tractable with nets (Huntley and Lopez, 1992) but is with a camera. The classifier developed here permits high confidence the discrimination of several species-level and more broadly based groups.
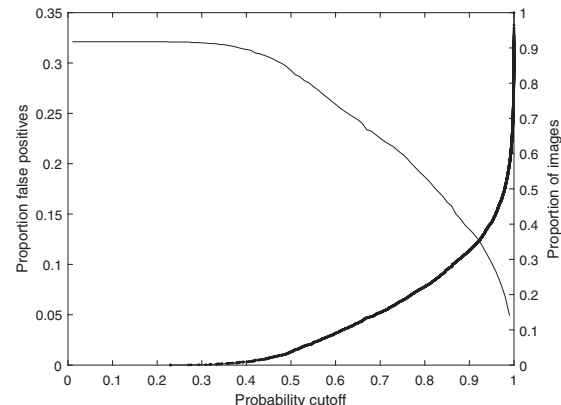

Figure 12. Proportion of false positives (black, left axis) and proportion of images that were rejected (red, right axis) as the probability cut-off was varied in the test set.
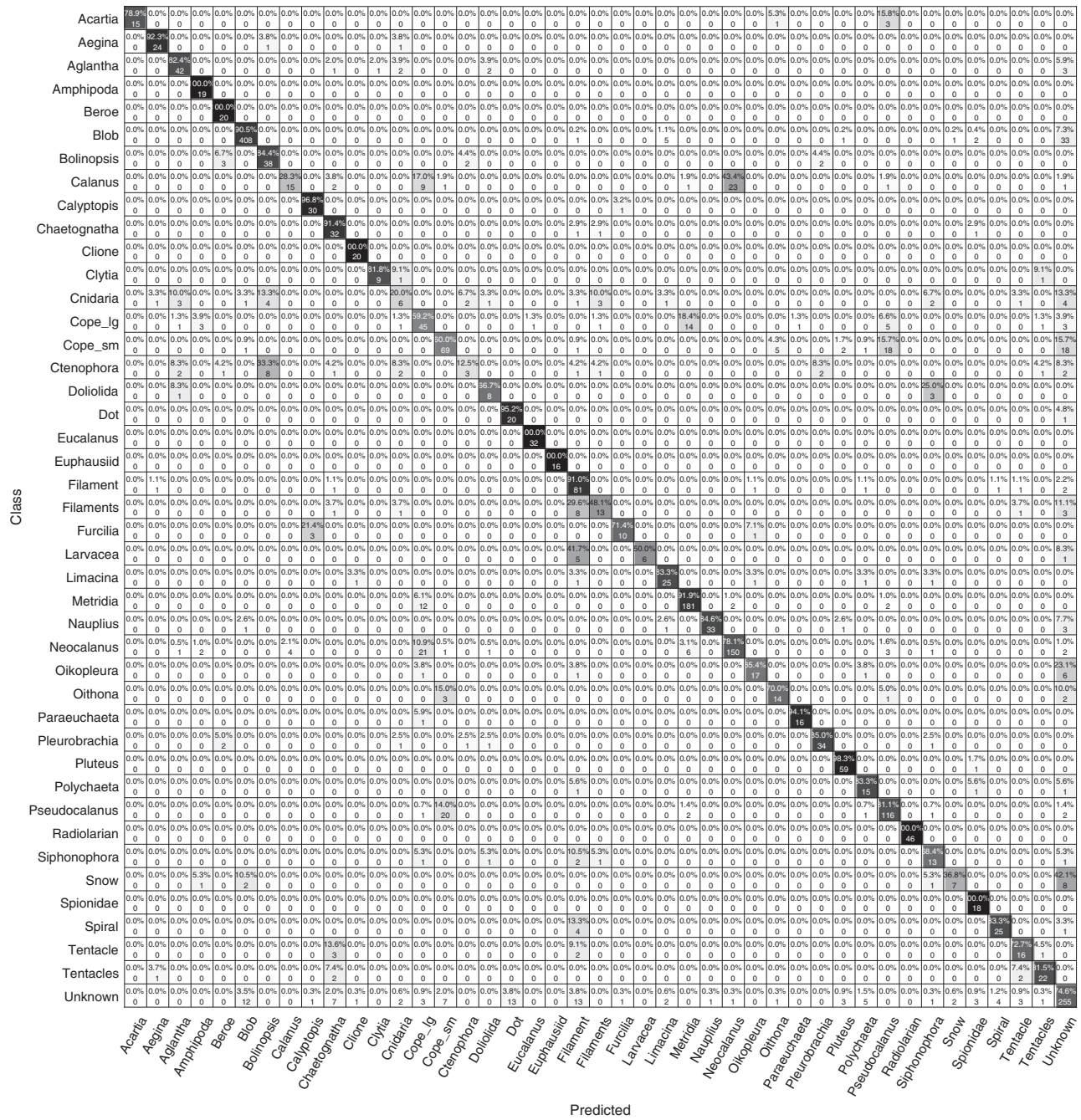
Figure 13. Confusion matrix for the classifier when applied to the test set and using a 90% probability threshold to discard uncertain classifications.

The value and usefulness of automatically classified imagery depends on the questions at hand. Simple information from zooplankton imagery such as abundance and size is easily determined with high confidence. For example, although there was some confusion between copepod species and the generalized copepod group, but if one is primarily interested in the abundance and relative biomass of copepods, that information may be determined with high confidence. Done over several years, estimates of zooplankton biomass could be of value to fisheries and ecosystem managers (e.g. Möllmann *et al.*, 2014). If there is interest in a single species, then more work may be required with the classified images to assure confidence but more inferential questions may be addressed.

Funding

Table 4. Accuracy metrics for each class in the test set, when a 90% probability threshold was applied.

Class name	Precision	Recall	F1 score
Acartia	0.82	1	0.9
Aegina	0.92	1	0.96
Aglantha	0.93	0.86	0.9
Amphipoda	1	0.86	0.93
Beroe	1	0.83	0.91
Blob	0.91	0.92	0.92
Bolinopsis	0.91	0.78	0.84
Calanus	0.59	0.64	0.61
Calyptopis	0.93	0.97	0.95
Chaetognatha	0.92	0.9	0.91
Clione	1	1	1
Clytia	0.68	0.76	0.72
Cnidaria	0.26	0.41	0.32
Cope_lg	0.53	0.62	0.57
Cope_sm	0.58	0.63	0.6
Ctenophora	0.2	0.36	0.26
Doliolida	0.79	0.92	0.85
Dot	0.87	0.65	0.74
Eucalanus	1	0.97	0.98
Euphausiid	1	1	1
Filament	0.87	0.69	0.77
Filaments	0.71	0.73	0.72
Furcilia	0.92	0.86	0.89
Larvacea	0.86	0.75	0.8
Limacina	0.84	0.86	0.85
Metridia	0.95	0.87	0.91
Nauplius	0.96	0.92	0.94
Neocalanus	0.79	0.85	0.81
Oikopleura	0.88	0.75	0.81
Oithona	0.69	0.62	0.65
Paraeuchaeta	0.95	0.9	0.92
Pleurobrachia	0.83	0.81	0.82
Pluteus	0.98	0.95	0.97
Polychaeta	0.93	0.61	0.74
Pseudocalanus	0.8	0.78	0.79
Radiolarian	0.96	0.98	0.97
Siphonophora	0.84	0.87	0.85
Snow	0.68	0.83	0.75
Spionidae	0.88	0.88	0.88
Spiral	0.97	0.71	0.82
Tentacle	0.96	0.81	0.88
Tentacles	0.89	1	0.94
Unknown	0.77	0.79	0.78

Table 5. Comparison of the imaging specifications of published plankton imagers designed for zooplankton.

System	Imager resolution	Pixel resolution	Sampled volume	Illumination	References
CPICS	1 360 × 1 024	30 μm to 20 mm	1 ml	Darkfield	Grossmann et al. (2015)
VPR	Varies	30 μm to 5 cm	1.25–380 ml	Darkfield	Davis et al. (1992), Lombard et al. (2019)
ZOOVIS	2 448 × 2 050	10 μm	240 ml	Shadowgraph	Bi et al. (2013, 2015)
UVP	1 280 × 1 024	174 μm	1 020 ml	Light sheet	Picheral et al. (2010)
ISIIS	2 048 × 17 frames per second (line scan)	68 μm (in vertical)	169 l s^{-1}	Shadowgraph	Cowen and Guigand (2008)
Zoocam	1 280 × 960	40 μm	250 ml	Shadowgraph	Ohman et al. (2019)
PWSPC	4 240 × 2 824	22.6 μm	450 ml	Darkfield	This project

authors are their own and do not necessarily reflect the views or position of the Trustee Council.

Acknowledgements

References

Abadi, A., Agarwal, P., Barham, E., Brevdo, Z., Chen, C., Citro, G. S., Corrado, A. *et al*et al. 2016. Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467, 2016.

Austen, G. E., Bindemann, M., Griffiths, R. A., and Roberts, D. L. 2017. Species identification by conservation practitioners using online images: accuracy and agreement between experts. PeerJ, doi: 10.7717/peerj.4157.

Benfield, M., Grosjean, P., Culverhouse, P., Irigolen, X., Sieracki, M., Lopez-Urrutia, A., Dam, H. *et al*. 2007. RAPID: research on automated plankton identification. Oceanography, 20: 172–187.

Benfield, M. C., Schwehm, C. J., Fredericks, R. G., Squyres, G., Keenan, S. F., and Trevorrow, M. V. 2003. Measurements of zooplankton distributions with a high-resolution digital camera system. Handbook of Scaling Methods in Aquatic Ecology: Measurement, Analysis, Simulation, pp. 17–30. Ed. by L. Seuront and P. G. Strutton. CRC Press, Boca Raton.

Bi, H., Cook, S., Yu, H., Benfield, M. C., and Houde, E. D. 2013. Deployment of an imaging system to investigate fine-scale spatial distribution of early life stages of the ctenophore *Mnemiopsis leidyi* in Chesapeake Bay. Journal of Plankton Research, 35: 270–280.

Bi, H., Guo, Z., Benfield, M. C., Fan, C., Ford, M., Shahrestani, S., and Sieracki, J. M. 2015. A semi-automated image analysis procedure for in situ plankton imaging systems. PLoS One, 10: e0127121.

Bridle, J. S. 1990. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. *In* Neurocomputing: Algorithms, Architectures and Applications (1989). NATO ASI Series (Series F: Computer and Systems Sciences), 68, pp. 227–236. Ed. by F. F. Soulié and J. Hérault. Springer, Heidelberg.

Bochinski, E., Bacha, G., Eiselein, V., Walles, T. J. W., Nejstgaard, J. C., and Sikora, T. 2019. Deep Active Learning for In Situ Plankton Classification. ICPR 2018 Workshops, LNCS 11188. doi: 10.1007/978-3-030-05792-3_1.

Canny, J. 1986. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 8: 679–698.

Cheng, K., Cheng, X., Wang, Y., Bi, H., and Benfield, M. C. 2019. Enhanced convolutional neural network for plankton identification and enumeration. PLoS One, 14: e0219570.

Cho, J., Lee, K., Shin, E., Choy, G., and Do, S. 2016. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? arXiv:1511.06348v2.

Chollet, F. 2015. Keras. https://keras.io/.

Cowen, R. K., and Guigand, C. M. 2008. In Situ Ichthyoplankton Imaging System (ISIIS): system design and preliminary results. Limnology and Oceanography Methods, 6: 126–132.

Cowen, R. K., Sponaugle, S., Robinson, K., and Luo, J. (2015). PlanktonSet 1.0: plankton imagery data collected from F.G. Walton Smith in Straits of Florida from 2014-06-03 to 2014-06-06 and used in the 2015 National Data Science Bowl (NCEI Accession 0127422).

Culverhouse, P. F., Simpson, R. G., Ellis, R., Lindley, J. A., Williams, R., Parisini, T., Reguera, B. *et al*. 1996. Automatic categorization of 23 species of dinoflagellate by artificial neural network. Marine Ecology Progress Series, 139: 281–287.

Culverhouse, P. F., Williams, R., Reguera, B., Herry, V., and González-Gil, S. 2003. Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. Marine Ecology Progress Series, 247: 17–25.

Cui, J., Wei, B., Wang, C., Yu, Z., Zheng, H., Zhenk, B., and Yang, H. 2018. Texture and shape information fusion of convolutional neural network for plankton image classification. OCEANS-MTS/IEEE Kobe 10.1109/OCEANSKOBE.2018.8559156

Davis, C. S., Gallager, S. M., Berman, M. S., Haury, L. R., and Strickler, J. R. 1992. The video plankton recorder (VPR): design and initial results. Archiv für Hydrobiologie Beiheft Ergebnisse der Limnologie, 36: 67–81.

Faillettaz, R., Picheral, M., Luo, J. Y., Guigand, C., Cowen, R. K., and Irisson, J. O. 2016. Imperfect automatic image classification successfully describes plankton distribution patterns. Methods in Oceanography, 15–16: 60–77.

Friedland, D. D., Stock, C., Drinkwater, K. F., Link, J. S., Leaf, R. T., Shank, B. V., Rose, J. M. *et al*. 2012. Pathways between primary production and fisheries yields of large marine ecosystems. PLoS One, 7: e28945.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. Deep Learning. MIT Press, Cambridge.

Gorsky, G., Ohman, M. D., Picheral, M., Gasparini, S., Stemmann, L., Romagnan, J.-B., Cawood, A. *et al*. 2010. Digital zooplankton image analysis using the ZooScan integrated system. Journal of Plankton Research, 32: 285–303.

Grossmann, M. M., Gallager, S. M., and Mitarai, S. 2015. Continuous monitoring of near-bottom mesoplankton communities in the East China Sea during a series of typhoons. Journal of Oceanography, 71: 115–124.

Haralick, R. M. 1979. Statistical and structural approaches to texture. Proceedings of IEEE, 67: 786–804.

Hassaballah, M., and Awad, A. I. (2016) Detection and description of image features: an introduction. *In* Image Feature Detectors and Descriptors. Studies in Computational Intelligence, 630, pp. 1–10. Ed. by A. Awad and M. Hassaballah Springer, Cham.

Hendrycks, D., and Gimpel, K. 2018. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv: 1610.02136v3.

Hu, Q., and Davis, C. 2005. Automatic plankton image recognition with co-occurrence matrices and support vector machine. Marine Ecology Progress Series, 295: 21–31.

Huntley, M. E., and Lopez, M. D. 1992. Temperature-dependent production of marine copepods: a global synthesis. The American Naturalist, 140: 201–242.

Jeffries, H. P., Berman, M. S., Poularikas, A. D., Katsinis, C., Melas, I., Sherman, K., and Bivins, L. 1984. Automated sizing, counting and identification of zooplankton by pattern recognition. Marine Biology, 78: 329–334.

Johnson, K. S., Coletti, L. J., and Chavez, F. P. 2006. In situ ultraviolet spectrophotometry for high resolution and long-term monitoring of nitrate, bromide, and bisulfide in the ocean. Deep Sea Research Part I, 53: 561–573.

King, M. 2007. Fisheries Biology, Assessment and Management, 2nd edn. Wiley-Blackwell, Oxford. 400 pp.

Kingma, D. P., and Ba, J. 2015. Adam: a method for stochastic optimization. arXiv:1412.6980 [cs.LG].

Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. *In* Advances in neural information processing systems, pp 1097–1105.

Lee, H., Park, M., and Kim, J. 2016. Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. IEEE ICIP. doi: 10.1109/ICIP.2016.7533053.

Liu, J., Du, A., Wang, C., Yu, Z., Zheng, H., Zheng, B., and Zhang, H. 2018. Deep pyramidal residual networks for plankton image classification. MTS/IEE OCEANS. doi: 10.1109/OCEANSKOBE.2018.8559106

Lombard, F., Boss, E., Waite, A. M., Vogt, M., Uitz, J., Stemmann, L., Sosik, H. M. *et al.* 2019. Globally consistent quantitative observations of planktonic ecosystems. Frontiers in Marine Science. doi: 10.3389/fmars.2019.00196

Longhurst, A. 2006. Ecological Geography of the Sea, 2nd edn. Academic Press, Cambridge. 560 pp.

Luo, J. Y., Irisson, J.-O., Graham, B., Guigand, C., Sarafraz, A., Mader, C., and Cowen, R. K. 2018. Automated plankton image analysis using convolutional neural networks. Limnology Oceanography Methods, 16: 814–827.

Mitra, A., Castellani, C., Gentleman, W. C., Jónasdóttir, S. H., Flynn, K. J., Bode, A., Halsband, C. *et al.* 2014. Bridging the gap between marine biogeochemical and fisheries sciences; configuring the zooplankton link. Progress in Oceanography, 129: 176–199.

Möllmann, C., Lindegren, M., Blenckner, T., Bergström, L., Casini, M., Diekmann, R., Flinkman, J. *et al.* 2014. Implementing ecosystem-based fisheries management: from single-species to integrated ecosystem assessment and advice for Baltic Sea fish stocks. ICES Journal of Marine Science, 71: 1187–1197.

Ohman, M. D., Davis, R. E., Sherman, J. T., Grindley, K. R., Whitmore, B. M., Nickels, C. F., and Ellen, J. S. 2019. Zooglider: an autonomous vehicle for optical and acoustic sensing of zooplankton. Limnology Oceanography Methods, 17: 686.

Orenstein, E. C., Beijbom, O., Peacock, E. E., and Sosik, H. M. 2015. WHOI-plankton-a large scale fine grained visual recognition benchmark dataset for plankton classification. CoRR 2015; abs/1510.00745.

Orenstein, E. C., and Beijbom, O. 2017. Transfer learning and deep feature extraction for planktonic image data sets. *In* 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, 2017, pp. 1082–1088. doi: 10.1109/WACV.2017.125.

Pearson, K. 1904. Mathematical Contributions to the Theory of Evolution on the Theory of Contingency and Its Relation to Association and Normal Correlation. Dulau and Co., London. 34 pp.

Perez, L., and Wang, J. 2017. The effectiveness of data augmentation in image classification using deep learning. arXiv:1712.04621.

Picheral, M., Guidi, L., Stemmann, L., Karl, D. M., Iddaoud, G., and Gorsky, G. 2010. Limnology Oceanography Methods, 8: 462–473.

Rodriques, F. C. M., Hirata, N. S. T., Abello, A. A., De La Cruz, L. T., Lopes, R. M., and Hirata, R. Jr 2018. Evaluation of Transfer Learning Scenarios in Plankton Image Classification. VISIGRAPP, doi: 10.5220/0006626703590366.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., *et al.* 2015. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115: 211–252.

Samson, S., Hopkins, T., Remsen, A., Langebrake, L., Sutton, T., and Patten, J. 2001. A system for high resolution zooplankton imaging. IEEE Journal of Oceanic Engineering, 26: 671–676.

Schröder, S.-M., Kiko, R., Irisson, J.-O., and Koch, R. 2018. Low-shot. *In* Pattern Recognition, pp. 391–404. 40th German Conference, 2018 Proceedings. Ed. by T. Brox, A. Bruhn. and M. Fritz. GCPR 2018 Stuttgart, Germany, October 9–12.

Sheldon, R. W., and Parsons, T. R. 1967. A continuous size spectrum for particulate matter in the sea. Journal of Fisheries Research Board of Canada, 24: 909–915.

Sosik, H. M., Peacock, E. E., and Brownlee, E. F. 2015. Annotated Plankton Images Data Set for Developing and Evaluating Classification Methods. https://hdl.handle.net/10.1575/1912/7341.

Strickland, J. D. H., and Parsons, T. R. 1972. A practical handbook of seawater analysis. Fisheries Research Board of Canada, 167: 310.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. 2015. Rethinking the inception architecture for computer vision, CVPR, vol. abs/1512.00567, arxiv.org/abs/1512.00567.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. 2016. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Transactions on Medical Imaging, 35: 1299–1312.

van Rijsbergen, C. J. 1979. Information Retrieval, 2nd edn. Butterworth, London.

Wiebe, P. H., and Benfield, M. C. 2003. From the Hensen net toward four-dimensional biological oceanography. Progress in Oceanography, 56: 7–136.

Yamashita, R., Nishio, M., Do, R. K. G., and Togashi, K. 2018. Convolutional neural networks: an overview and application in radiology. Insights into Imaging, 9: 611–629.

Handling editor: Cigdem Beyan