Vol. 615: 15–30, 2019
https://doi.org/10.3354/meps12925

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published April 18

Automated identification of benthic epifauna with computer vision

Nils Piechaud[1,*], Christopher Hunt[2], Phil F. Culverhouse[3], Nicola L. Foster[1], Kerry L. Howell[1]

[1]School of Biological and Marine Sciences, University of Plymouth, Plymouth PL4 8AA, UK
[2]Controlled Frenzy Ltd., THINQTANQ, Fairbairn House, Higher Lane, Plymouth PL1 2AN, UK
[3]School of Computing, Electronics and Mathematics, University of Plymouth, Plymouth PL4 8AA, UK

ABSTRACT: Benthic ecosystems are chronically undersampled, particularly in environments >50 m depth. Yet a rising level of anthropogenic threats makes data collection ever more urgent. Currently, modern underwater sampling tools, particularly autonomous underwater vehicles (AUVs), are able to collect vast image data, but cannot bypass the bottleneck formed by manual image annotation. Computer vision (CV) offers a faster, more consistent, cost effective and sharable alternative to manual annotation. We used TensorFlow to evaluate the performance of the Inception V3 model with different numbers of training images, as well as assessing how many different classes (taxa) it could distinguish. Classifiers (models) were trained with increasing amounts of data (20 to 1000 images of each taxa) and increasing numbers of taxa (7 to 52). Maximum performance (0.78 sensitivity, 0.75 precision) was achieved using the maximum number of training images but little was gained in performance beyond 200 training images. Performance was also highest with the least classes in training. None of the classifiers had average performances high enough to be a suitable alternative to manual annotation. However, some classifiers performed well for individual taxa (0.95 sensitivity, 0.94 precision). Our results suggest this technology is currently best applied to specific taxa that can be reliably identified and where 200 training images offers a good compromise between performance and annotation effort. This demonstrates that CV could be routinely employed as a tool to study benthic ecology by non-specialists, which could lead to a major increase in data availability for conservation research and biodiversity management.

KEY WORDS: Benthic ecology · Computer vision · Automated image analysis · Automated species identification

1. INTRODUCTION

Marine ecosystems cover the majority of Earth's surface but benthic ecologists and biodiversity mangers have long been confronted with a shortage of data (Jongman 2013, Borja et al. 2016) regarding its composition and functioning. With increasing anthropogenic pressure, management measures urgently need to be implemented (Van Dover et al. 2014, Danovaro et al. 2017). These conservation measures must be based on a solid understanding of taxonomic diversity and ecological dynamics of the habitats considered (Hernandez et al. 2006). In many cases, that knowledge is lacking and specialists agree that data collection must be increased to tackle the challenge (Costello et al. 2010, Borja et al. 2016). The amount of data currently available on benthic ecosystems is always limited by how many samples can be collected, stored, and processed at a time. Since the 19th century, various technological innovations have attempted to bypass this bottleneck.

[*]Corresponding author: nils.piechaud@plymouth.ac.uk

Benthic ecosystems are traditionally sampled by trawls, cores and other physical means. These physical samples are costly to collect and process and logistically challenging to store (Clark et al. 2016). While physical samples remain the mainstay of benthic surveys, use of underwater imaging technologies is becoming increasingly popular among marine ecologists (Solan et al. 2003, Bicknell et al. 2016, Brandt et al. 2016, Romero-Ramirez et al. 2016). These technologies offer a less invasive, more cost-effective method of survey, and storage space for image data is virtually unlimited (Mallet & Pelletier 2014). Underwater imaging is now regularly utilised alongside other sampling tools to provide a comprehensive view of the marine environment.

Modern underwater sampling vehicles, and particularly autonomous underwater vehicles (AUVs), have great potential in providing the step-change in the rate of data gathering that is needed to support sustainable marine environmental management. They are capable of collecting large numbers of images of the seabed in a single deployment (Lucieer & Forrest 2016, Williams et al. 2016). For example, a 22 h AUV dive can deliver more than 150 000 images of the seafloor along with other types of environmental data (Wynn et al. 2012). Comparatively, trawls and remotely operated vehicles (ROVs) cover less ground per dive, and the ship and its crew are unable to operate any other benthic equipment while they are deployed (Brandt et al. 2016, Clark et al. 2016).

To translate the information contained in images into semantic data that can then be used in statistical analysis, a step of manual analysis (or annotation) is conducted by trained scientists. Human observers, even those that are highly trained, do not achieve 100% correct classification rates and are highly inconsistent across time and across annotators (Culverhouse et al. 2003, 2014, Beijbom et al. 2015, Durden et al. 2016). Besides, manual image annotation results are subject to observer bias, meaning that interpretations vary depending on the annotator's experience and mood, which can change across the analysis process (tiredness, boredom or stress, etc.) (Culverhouse et al. 2003, Durden et al. 2016). The results (format, taxonomic resolution and nomenclature) of these analyses also tend to differ from one institution, project or individual annotator to another. This lack of standardisation makes merging and comparing data sets difficult (Bullimore et al. 2013, Althaus et al. 2015, McClain & Rex 2015), and the data quality is not always consistent. More importantly, manual analysis is a time-consuming process, which forms the current bottleneck in image-based marine ecological sampling (Edgington et al. 2006, Beijbom et al. 2015, Schoening et al. 2017). The growing trend towards use of AUVs for seafloor biological surveys will only increase the scientific challenge.

Artificial intelligence (AI) and computer vision (CV) provide potential means by which to both accelerate and standardise the interpretation of image data (Culverhouse et al. 2003, MacLeod et al. 2010, Beijbom et al. 2012, Favret & Sieracki 2016). Although using AI for biological research has a long history (Rohlf & Sokal 1967, Jeffries et al. 1984, Gaston & O'Neill 2004), it has always been challenging to implement for non-specialists and requires skills and materials that most biologists do not have access to (Gaston & O'Neill 2004, Rampasek & Goldenberg 2016). CV has been successfully applied to benthic species identification by a growing number of studies (Edgington et al. 2006, Beijbom et al. 2015, Marburg & Bigham 2016, Manderson et al. 2017, Marini et al. 2018b, Norouzzadeh et al. 2018, Schneider et al. 2018), but has yet to be made into an easy-to-use tool that any biologist in the field can implement as an alternative to manual image annotation and integrate with previous analysis. Multiple potential commercial applications, the availability of new tools such as open software, as well as the improvement of hardware capacity are driving new developments in AI (e.g. neural networks and deep learning). This is likely to change how CV can be employed in the field of scientific research (Rampasek & Goldenberg 2016, Weinstein 2018). In parallel, new image analysis and data science software allow easier and more efficient integration of various tools into the research process, from data collection to final scientific or public outreach material (Gomes-Pereira et al. 2016). These new technologies are potentially enabling full automation of the annotation process and could revolutionise ecological research (Weinstein 2018).

While the principle of automated classification (automated assignation of pre-established classes to objects on images) has been validated, few practical examples exist of AI-based methods used to identify benthic organisms from images acquired by AUVs. Consequently, implementing an automated species classifier is a potentially time-consuming investment for an uncertain return. Relying on proven manual methods remains the safe option for researchers. Practical guidance is needed to help ecologists decide whether adopting AI and CV is feasible and would fit their data set and scientific objectives. To make that decision, benthic ecologists need to know (1) what level of accuracy and uncertainty can be expected from CV annotation and whether it matches or approximates the accuracy of human annotators; (2) how much material is needed to train a classifier, and whether a limited amount obtained from a single study is sufficient; and (3) how to assess their own data set to decide whether use of CV is appropriate.

In this study, we investigated these issues by using an open access algorithm to build a convolutional neural network (CNN) to identify benthic organisms in seafloor images, obtained from a single deployment of the UK's Autosub6000 AUV. Technically speaking, we sought to train an automated classifier that is able to determine which taxa an animal on an image most likely belongs to using a list of predefined taxa (or classes). Specifically, we asked the following: (1) What impact does the number of images, on which the classifier is trained, have on its performance? (2) What impact does the number of classes, on which the classifier is trained, have on its performance? In addition, we provide a case study in the application of CV to an unbalanced ecological data set.

2. MATERIALS AND METHODS

2.1. Study area and data collection

All images used in this study were collected by the UK's national AUV Autosub6000 in May 2016 as part of the Natural Environment Research Council (NERC)-funded DeepLinks (JC136) research cruise. The images were taken as part of an 1880 m long transect at Stn 26 of that cruise at 1200 m depth on the northeast side of Rockall Bank, NE Atlantic. This region was selected for the study due to the flat topography and low likelihood of disturbance, making it ideal for AUV deployment. The AUV was equipped with a downward facing Grasshopper2 GS2-GE-50S5C camera (Point Grey Research). The AUV was flown at 1.1 m s^{-1} speed, at 3 ± 0.1 m off the bottom and took images every 1 s, resulting in near overlapping image coverage. The surface area of each image was between 1 and 2.5 m^2, and the resolution was 2448 × 2048 pixels.

In total, 1165 raw photos of the seabed were manually annotated by a single observer with the BIIGLE 2.0 software (Langenkämper et al. 2017) using a regional catalogue of operational taxonomical units (OTUs) developed by Howell & Davies (2016). Within the BIIGLE 2.0 software, location (*x* and *y* coordinates in pixels within the photo for point annotations, or *x*, *y* and radius for individuals marked using a circle) and identity of individual OTUs annotated within each image were recorded and stored.

For each OTU, all individual annotations were visually inspected using the 'Largo' evaluation tool in BIIGLE 2.0, to maximise consistency in identification and reduce error. Later, an assessment of 75% (around 28 000) of the annotations in the final data set used in the model found 41 identification errors. By that assessment, we concluded that the accuracy of identification was above 99%.

2.2. Image data

Manual image annotation resulted in a data set consisting of 41 208 individuals belonging to 148 OTUs. Each individual was then cropped from the raw image, together with its assigned OTU label, using a custom Python (www.python.org) script. For each annotation, a square of 240 pixels or more, positioned manually on *x* and *y* coordinates on the centre of the animal, was fitted and cropped out. For organisms larger than 40 pixels, the size of the square was manually set to encompass the whole individual. These cropped image slices and associated OTU labels (to become classes in the model training design) formed the input used in the CNN.

2.3. TensorFlow and transfer learning

CNNs are a particular architecture of neural networks, more specifically, deep neural networks, particularly suited to image analysis (Krizhevsky et al. 2012, LeCun et al. 2015). A CNN has the capacity to detect and match patterns in images, thereby 'learning' what features are relevant to differentiate objects and, subsequently, classify them accordingly. Rather than train our own neural network, we used transfer learning (Pan & Yang 2010) to retrain the Inception V3 model (Szegedy et al. 2016), a CNN built in the freely available library: TensorFlow (TF) (Abadi et al. 2016).

TF is a C++ based library but has a Python application programming interface (API) that makes it easier to train, tune and deploy neural networks. Transfer learning is a method allowing a CNN built on a large data set to be repurposed into a classifier capable of distinguishing between classes on which it was not initially trained. The strength of this method is that the data set on which it is transferred does not need to be as large as it should be to train a CNN from the beginning. Here, we were able to train a classifier with 10s to 100s of images $class^{-1}$ (in our case, OTUs) instead of millions.

2.4. Classifier training and testing

A random 75–25% split was applied to every OTU in order to separate images used for training the classifier from those used for testing. The training and test data sets for all OTUs were then combined into single 'training' and 'test' data sets.

The OTUs the classifier was trained to identify are referred to as classes, and only those OTUs for which there were a sufficient number of image slices (individual observations) available were selected for use in training. The minimum number of images needed for training was set to 20. This means that for an OTU to be included in the study at least 27 image slices were needed: 20 for training and 7 for testing. Of the 148 OTUs observed, 52 were above that threshold. The remaining 96 OTUs represented 3.19% of the total number of individual annotations and were removed from the data set.

The classifier was trained on the training data set and then predictions were made on the test data set. For each cropped image slice in the test data set, TF gave a score for each of the possible OTU classes for which it had been trained. The scores range from 0 to 1 (the sum of scores for all classes being 1) and represent the model's confidence that the slice belongs to the corresponding class. The final prediction was the OTU class that received the highest score. The prediction was then compared to the manually assigned OTU class.

To measure the effect of the number of training images (or limit, see below) on the accuracy and confidence of the predictions, the training data set was filtered so each OTU class was represented by 20, 50, 100, 200, 500 and 1000 images (Table 1). A classifier was then trained on each of these 6 pools of images and tested using the test data set. Only 7 OTUs were observed frequently enough to be used with these 6 limits (Fig. 1).

The combination of groups and limits is referred to as a treatment, and designation of each treatment follows the nomenclature in Table 1 (e.g. A1000 is group A, limit 1000). Each treatment was repeated 10 times with different random splits between testing and training data for cross-validation.

To measure the effect of the number of OTU classes used to train the CNN on its capacity to correctly classify the test data set, we used 3 training data sets, each with different numbers of classes (referred to as groups) (Table 1). The number of classes is defined by the number of available images per OTU so classifiers can be trained on a set number of images for every class while retaining enough images for testing. Group A contained 7 classes for which more than 1000 images was available; group B contained 27 classes for which more than 100 images were available; and group C contained 52 classes for which more than 20 images were available. Within each group, classifiers were trained with all 6 pools of images (Table 1).

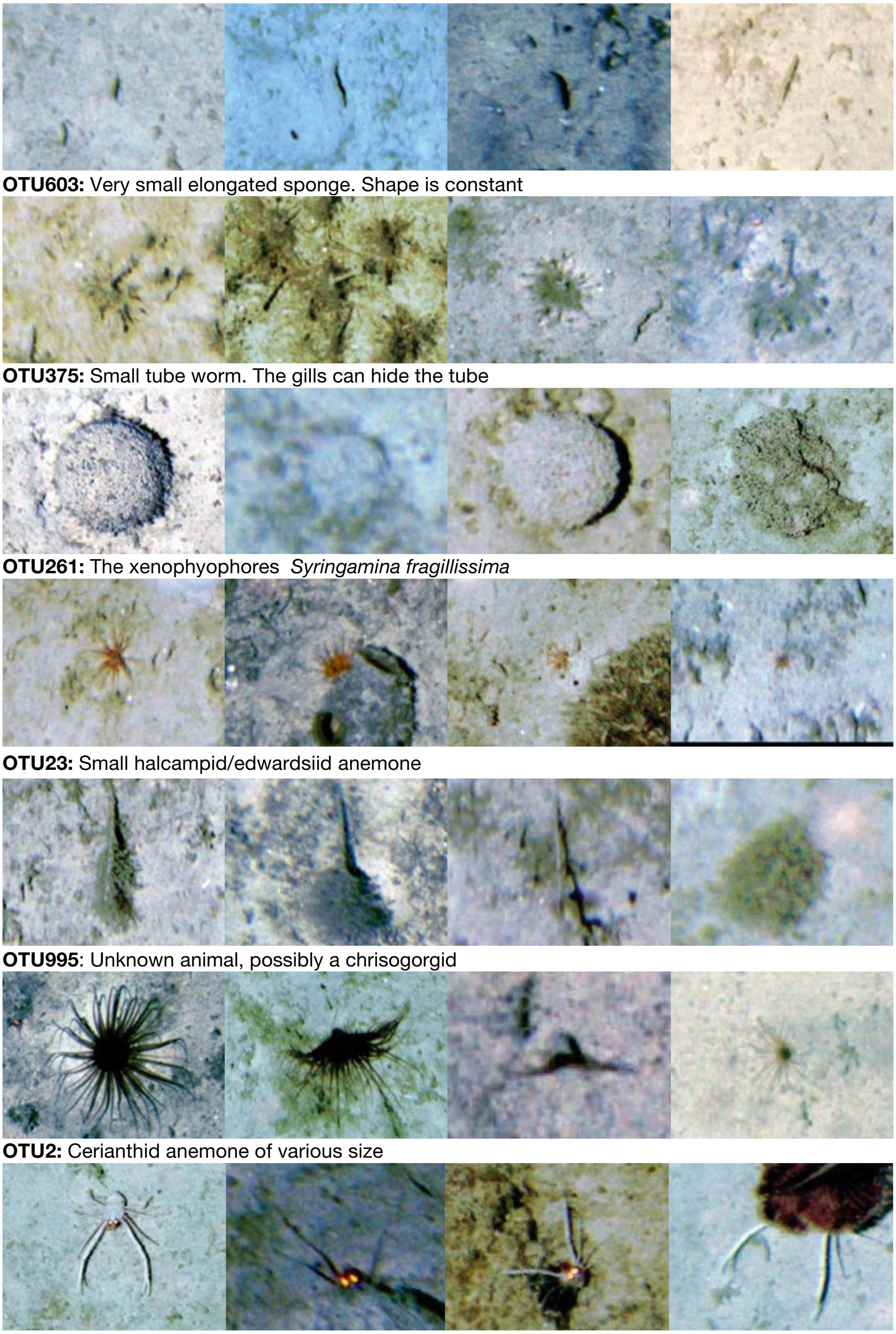
Note that when the limit is above the available number of images, the classes with less images were trained with the maximum number available regardless of the limit. This results in class imbalance in the model training for some treatments in group C with more than 20 images and in B with more than 100 images (balanced treatments are listed in Table 1). To assess the effect of the number of OTU classes used to train the CNN on its capacity to correctly classify the test data set, only balanced designs were used.

In total, 180 (3 × 6 × 10) classifiers were trained and tested. All CNNs were trained in the Google Cloud Machine Learning (ML) (https://cloud.google.com/) remote computing facility.

To be applied to a 'real-life' ecological study, the classifiers have to maximise performances while minimising the initial effort needed to build the training data set. To assess appropriate use of CV on a 'real-life' data set, we considered all possible combinations of numbers of training image and numbers of OTU classes in an unbalanced design. Average performances and individual OTU performances were assessed.

Table 1. Nomenclature of classifier names and characteristics. The different classifier names are a combination of group name and image numbers per operational taxonomical unit (OTU) in training. Groups are defined by the number of different OTUs (or classes) in the training set. In the different groups, the OTUs used are those for which the minimum number of images indicated are available. Within each group, treatments refer to the number of images of each class in training. The same treatments (20, 50, 100, 200, 500 and 1000 images per OTU in training) were applied to each group but only the classifiers names in **bold** *italics* are balanced (equal number of images for every class). In unbalanced designs, the maximum number of available images is used and is therefore different for each OTU

Group	No. of classes	Min. no. of images (for OTU to be in group)	Classifier names
A	7	1000	***A20, A50, A100, A200, A500, A1000***
B	27	100	***B20, B50, B100***, B200, B500, B1000
C	52	20	***C20***, C50, C100, C200, C500, C1000



OTU603: Very small elongated sponge. Shape is constant

OTU375: Small tube worm. The gills can hide the tube

OTU261: The xenophyophores *Syringamina fragillissima*

OTU23: Small halcampid/edwardsiid anemone

OTU995: Unknown animal, possibly a chrisogorgid

OTU2: Cerianthid anemone of various size

OTU339: The squat lobster *Munidasarsi/tenuimana*

Fig. 1. Example images and description of operational taxonomic units (OTUs) abundant enough to be in group A. Scale varies. OTUs are ordered by abundance in the original data set

2.5. Analysis and performance evaluation

Considering each class, the observation can be a presence (the OTU is present on the image) or an absence (the OTU is not on the image and another OTU is). The different possible outcomes or predictions of the classifier are detailed in Table 2. The respective number of each outcome type (the confusion matrix) was used to calculate performance metrics.

Classification accuracy is the percentage of predictions that are correct (prediction matches observation); it is often used to evaluate performances in ML studies. This measure ignores the differences between classes, thus we used 2 model evaluation metrics (sensitivity and precision) which rely on a confusion matrix (Manel et al. 2001) explained in Table 2. Sensitivity, also referred to as the 'true positives rate' or 'recall', varies between 0 and 1. It quantifies the proportion of individuals of a given OTU in the testing set that are correctly identified. A value of 1 means that all individuals of a given OTU are identified as such:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Precision, or 'positive predictive value', also varies between 0 and 1. It quantifies the proportion of true positives among the individuals identified as a given OTU. A value of 1 means all the individuals identified as a given OTU class are indeed that OTU:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

The average and standard deviation for all metrics were calculated for each class within each treatment and then averaged over other grouping factors. This gave an estimation of the overall performance of the classifiers. The performances of the classifiers for each individual class were also carefully analysed.

Differences in metrics were statistically tested with a permutation-based analysis of variance in the 'lmPerm' package in R (Wheeler & Torchiano 2010). We report p-values classified with 5 levels of significance: >0.05 or non-significant, <0.05, <0.01, <0.001 and <0.0001. Relationships between the number of images and performance were extrapolated with a neural network regression in the 'nnet' package in R (Ripley et al. 2016) projected over 1000 to 10 000 images. All data analyses were carried out in R (R Core Team 2014) using the 'tidyverse' package (Wickham 2017).

3. RESULTS

The results are presented in 3 sections. First, questions related to the impact of the number of training images are addressed, then the effect of the number of classes in the training set is assessed, and finally the results relevant to choosing the best method in our case study are presented.

3.1. Impact of the number of training images on performance

Average performance, measured as both sensitivity and precision, increased with an increasing number of training images used (Fig. 2). For sensitivity, there was an average increase from 0.64 to 0.78 when moving from 20 to 1000 images, respectively. This was mirrored by increases in precision from 0.63 to 0.75 when moving from 20 to 1000 images, respectively. Non-linear extrapolations of average sensitivity and precision showed that performances reached with 1000 training images may be close to an asymptote and performances obtained with additional training material probably plateau below 0.78 for sensitivity and 0.75 for precision (Fig. 2). This suggests that the model is unable to achieve perfect performance regardless of how many additional images are used in training.

The number of training images has a clear positive effect on performance. For almost all pairs of

Table 2. Possible outcomes of the classifiers, indicating how the classifier predictions compare to the manual annotation (i.e. the labels) and if it identifies the operational taxonomical unit (OTU) present on an image correctly

Outcome	Description
True positives	Label is OTU and class predicted is OTU ▶ Classifier correctly identified the OTU
True negatives	Label is not OTU and class predicted is not OTU ▶ Classifier correctly recognized the OTU is not in the image
False negatives	Label is OTU but class predicted is not OTU ▶ Classifier misidentified the OTU
False positives	Label is not OTU but class predicted is OTU ▶ Classifier misidentified another OTU

models compared (Fig. S1 in the Supplement at www.int-res.com/articles/suppl/m615p015_supp.pdf), performance values were statistically significantly different ($p < 0.05$) and very often, significance was very high ($p < 0.0001$). There were a few exceptions, like between the A20 and A50 classifiers where the p-value was greater than 0.05 for sensitivity and between 0.01 and 0.05 for precision, or the B1000 classifier, for which there was no significant difference in sensitivity between this classifier and the B500 and B200 classifiers. However, measured differences in performance between sequential models became vanishingly small at higher numbers of training images, such that the difference between A200 and A1000 classifiers was 0.04 for sensitivity and 0.05 for precision. This suggests little to no improvement is gained in model performance by using more than 200 training images.

There were strong between-OTU differences in classifier performance (Fig. 3). All classifiers had high sensitivity for OTU261 and OTU339 and even the A20 classifier (0.88 and 0.77, respectively). For OTU2 and OTU23, classifiers had more variability and lower sensitivity regardless of the number of training images used.

The OTUs for which precision was highest were not necessarily those for which sensitivity was highest. The highest precision observed was for OTU261 but the second highest precision observed was for OTU603, which had a lower sensitivity. For some classes (OTU261 or OTU339), precision was lower with 50 training images compared to 20 training images.

3.2. Impact of the number of classes on classifier performance

Classifiers trained with 7 classes (group A) had significantly better sensitivity (Fig. S1) and precision than equivalent classifiers trained on more classes but the same number of images (Fig. 4). Variability in performance was also lower for classifiers trained with fewer classes. Average sensitivity decreased from 0.71 to 0.38, and average precision decreased from 0.69 to 0.32, when moving from 7 to 27 classes. This suggests a negative effect of the number of classes on performance; however, on average, there was only a minor drop in performance (0.018 in sensitivity and 0.035 in precision) between classifiers trained on 27 and 52 classes. Interestingly, B100 and C100 both had sensitivity of 0.38 (no statistical difference) and C20 had higher (+0.02) sensitivity than B20.

OTUs that performed well in one group tended to perform well in other groups. OTU261 and OTU339 were in the top 10 for each group although their performances were lower in groups B and C.

3.3. Application of CV to an unbalanced ecological data set

When considering all treatments in an unbalanced design (Fig. 4), the average sensitivity per treatment ranged from 0.32 to 0.78. The highest sensitivity was achieved by the A1000 classifier (7 classes, with 1000 training images in each class) while the lowest was

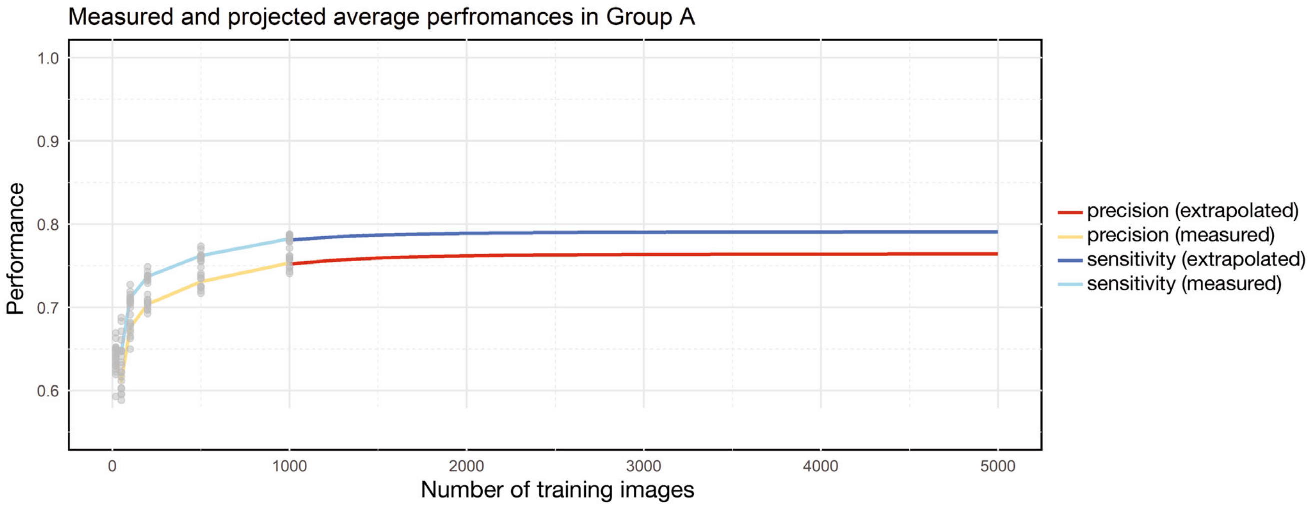


Fig. 2. Classifier performances (sensitivity and precision) per number of training images measured (20–1000) and extrapolated (1000–5000). Grey dots: averaged values across all operational taxonomic units (OTUs) for each classifier

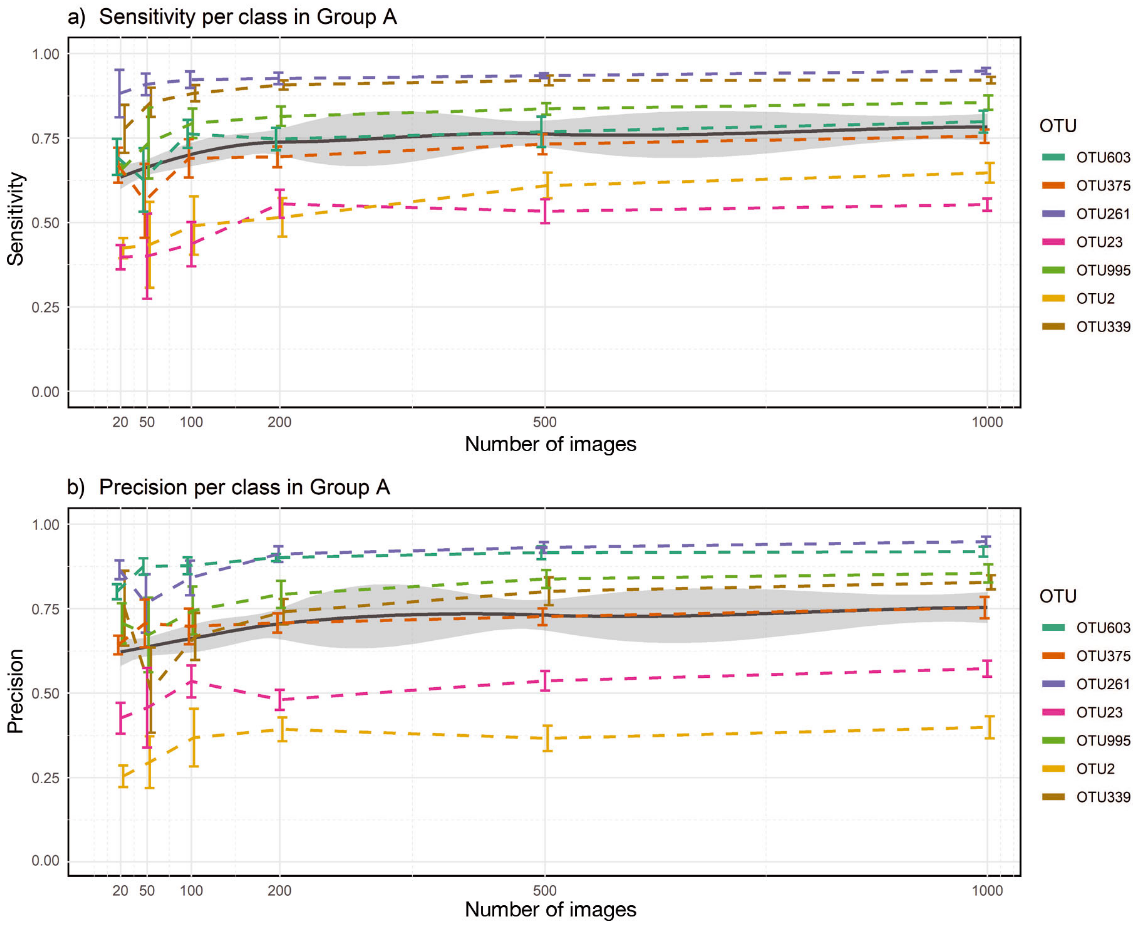


Fig. 3. (a) Evolution of sensitivity in group A classifier trained with an increasing number of images. (b) Differences in precision in group A classifier trained with an increasing number of images. Black line: 'loess' smoothed curve of the average of all the classes; grey area: a *t*-based approximation of the standard error. Error bars represent standard deviation. OTU: operational taxonomic unit

achieved by the B20 and C20 classifiers (27 and 52 classes, respectively, and 20 images in each class). A1000 also had the highest precision (0.75), with the lowest precision observed in the C20 classifier (0.20). Sensitivity of the C1000 classifier (where class imbalance was highest) was lower than in the C100 and C200 classifiers but precision simply increased with the number of training images, although this could be an artefact driven by the improvement of precision on the most abundant classes.

When considering individual OTUs, performance was unacceptably low for most, but not all, as some had sensitivity and precision greater than 0.85. Based on average sensitivity across all treatments, the top 10 and bottom 10 OTU classes were identified. The top 10 classes were large organisms with consistent or distinctive shape, colour and patterning. They were not necessarily the most abundant classes, as 6 of them were only present in group C, for which there were less than 100 training images, and only 2 in A, for which there were at least 1000 training images. Of these OTUs, the 2 present in group A had better average precision than any other OTU class in the top 10. The OTU classes with the worst performances were generally those for which there were fewer training images (group C). They also tended to be smaller organisms, had colours similar to the background and had variable shapes and sizes.

In this data set, CV could be applied to OTU261 and OTU339. These OTUs were both very abundant in the study area, justifying automated annotation, and they both had very high performances, making their identification by the classifier reliable (Fig. 5).

The performance of CV for OTU261 and OTU339 was maximised in the A1000 classifier with only 7 classes and 1000 training images. The A200 classifier also achieved performances close to A1000, despite being trained on 5 times less images. For OTU261, even the A20 and A50 classifiers achieved sensitivity and precision greater than 0.86, and differences

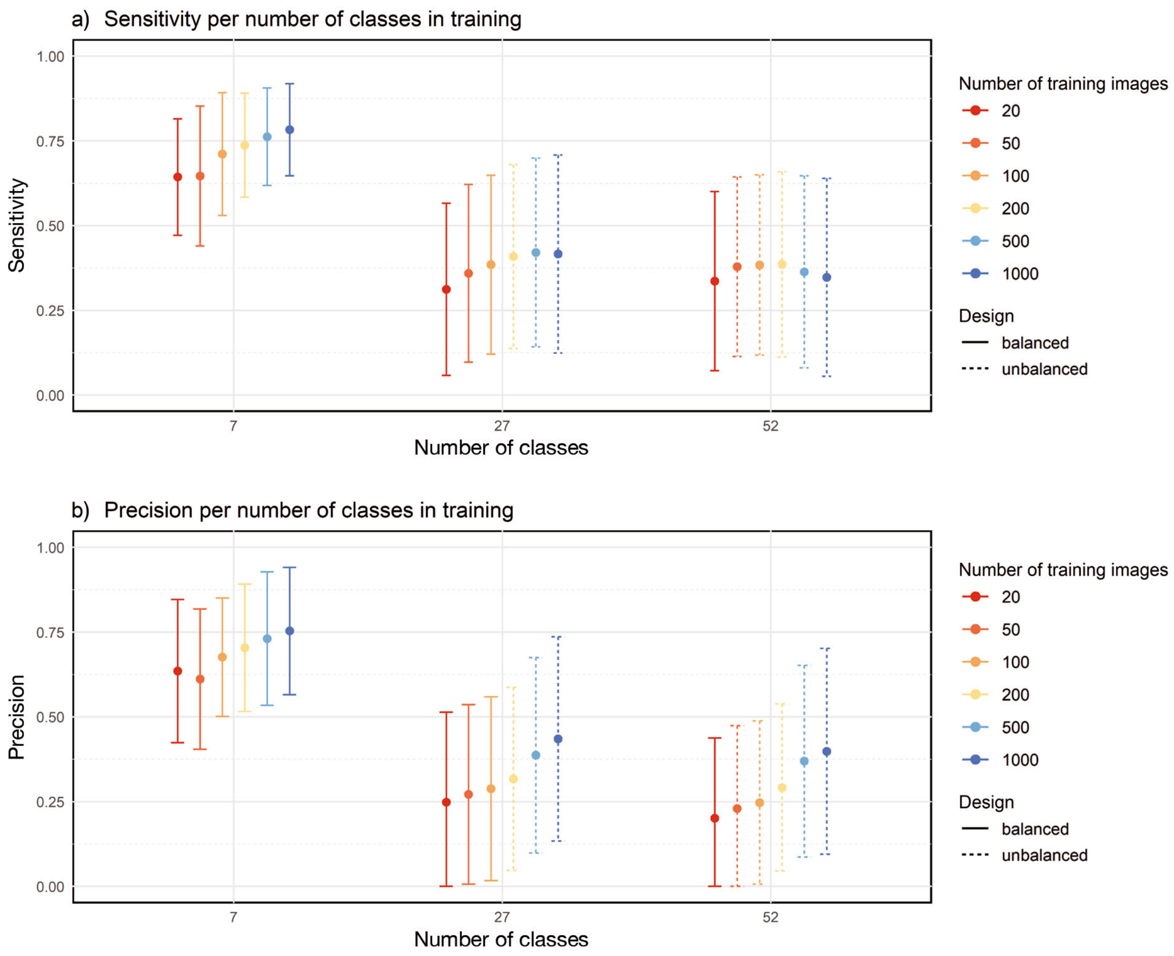


Fig. 4. Differences in (a) sensitivity and (b) precision in classifiers trained with different number of classes and images (7 for group A, 27 for group B, 52 for group C). Unbalanced designs (not all classes have the same number of training images) are displayed with dashed lines. Error bars: standard deviation of the 10 random splits

between the A20, A50 and A100 classifiers were not statistically significant (Fig. 5).

Sensitivity in the C1000 classifier was 0.92 and 0.89 for OTU261 and OTU339, respectively, which is significantly lower than the A1000 classifier ($p < 0.0001$ for both; Figs. S2 & S3) but only a marginal difference (0.03 each). For OTU261, the C200 classifier achieved lower sensitivity than the A200 but they had equal precision. For OTU339, precision was also the same in the A200 classifier and all classifiers in group C (Fig. S4). Note that for both OTUs, precision of all treatments in group C were either not significant or barely significantly different ($p > 0.01$). Thus, classifiers in group C (with 52 classes) achieved performances almost as good as classifiers in group A when training on 200 or less images.

Group B classifiers tended to show slightly lower sensitivity than group A classifiers and slightly lower precision than group C classifiers, although often not with significant differences.

4. DISCUSSION

In this study, our purpose was to test the capacity of a transferred CNN classifier (partially trained on a different data set) to identify benthic organisms and, by extension, to test if this methodology can be successfully applied in ecology by non-specialists with a relatively small data set, open-source software and libraries, as well as a short investment in time after manual image annotation.

4.1. Overall performances

Our classifiers achieved a maximum average performance of 78% (0.78) in sensitivity and 75% (0.75) in precision. In other studies, performances achieved through manual annotation range from 50 to 95% for benthic fauna (Beijbom et al. 2015, Durden et al. 2016) and 84 to 94% accuracy for plankton (Culver-

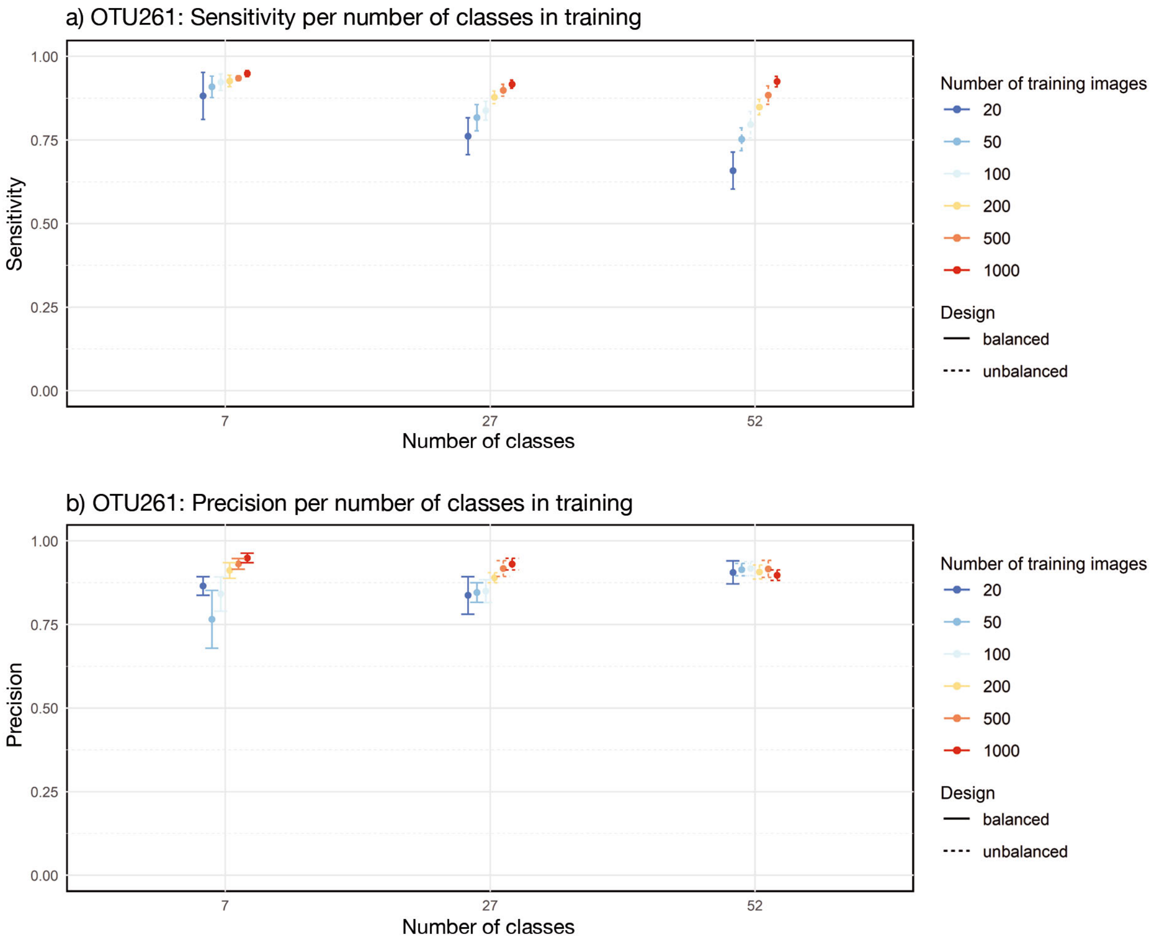


Fig. 5. Differences in (a) sensitivity and (b) precision for operational taxonomic unit (OTU) 261 in classifier trained with different number of classes and images (7 for group A, 27 for group B, 52 for group C). Unbalanced designs (not equal number of images in every class) are displayed with dashed lines. Error bars: standard deviation of the 10 random splits

house et al. 2003). There is no consensus on what an acceptable error rate in the ecological literature is but, to be competitive with experts, automated identification performances should be towards the higher end of those achieved manually. In this regard, Culverhouse et al. (2014) reported an anecdotal value of 90% correct classification cited by experts. Previous studies on marine ecosystems sampled via images that have attempted to automatically classify multiple benthic megafaunal taxa with various methods sometimes achieve performances comparable to those of experts. Beijbom et al. (2012) found that different coral species in shallow reefs were correctly identified 97% of the time. Schoening et al. (2012) found an average sensitivity of 87% and precision of 67% when classifying deep benthic megafauna in the Arctic. Marburg & Bigham (2016) found 89% accuracy when classifying benthic mobile megafauna off the Oregon coast. When considering other faunal groups, CV can achieve even higher performances; for example, Siddiqui et al. (2018) automatically identified various fish species correctly 96.7% of the time on average.

Even at their best performances, our classifiers would misclassify more than 1 out of 5 observations if they were used to make novel predictions. This is not good enough to be considered a suitable replacement for manual annotation. To be the tool benthic ecologists need, average performances need to be increased by at least 10 or 15%.

4.2. Impact of the number of images in training on performances

In our study, average performance measured as both sensitivity and precision increased with the number of images used in training. Performances obtained with 1000 training images were significantly better than those obtained with fewer images, for example, those obtained with 200 images (5 times less). This difference, however was marginal. Extrapolation of the data suggests that performances may never greatly exceed those obtained with 1000 training images regardless of how many images are used.

It has been generally demonstrated that more data is preferable when modelling (Junqué de Fortuny et al. 2013) and training classifiers (Lu & Weng 2007, Maxwell et al. 2018). Unsurprisingly then, our results suggest that the number of training images has a clear positive effect on performance, particularly on sensitivity. Sun et al. (2017) tested their generalist object classifiers with 10, 30 and 100 million images and observed a clear increase in performance. Siddiqui et al. (2018) also found that increasing the size of a data set by 25% (20 000 to 25 000 images) resulted in a 6.6% increase in performance of the same CNN.

More data, however, is not a simple solution to low performance as the relationship between the amount of training data and performance is not linear. Sun et al. (2017) reported a logarithmic relationship between the size of the training set and performance. These authors gained less than 20% increase in performance by adding 90 million images to their training set. This logarithmic relationship has also been reported by Favret & Sieracki (2016) in their fly species classifiers. These authors noted a diminishing return of adding more training data and observed little gain when doubling their training size from 50 to 100 images. Cho et al. (preprint https://arxiv.org/abs/1511.06348), who classified computed tomography images of 6 human body parts, found the same logarithmic relationship and, although it was 95.7% with 200 training images, their desired 99.5% accuracy target was only reached with 4092 images. Thus, there is an optimal size to every data set beyond which more training data results in very little gain. This point can be determined by the goal of the study and what is considered acceptable performance. With our methodology, this point occurred at 200 images for the data set we used, and represents a reasonable amount of manual work for ecologists aiming to build the data set to train a CNN.

4.3. Impact of the number of OTU classes in training on performances

We observed that classifiers with a small (7) number of classes had better performances than those trained with 27 or 52 classes. The difference in performance between the latter 2 was marginal, although significant.

The number of classes in machine learning studies is usually driven by the data set and the research question rather than maximising performance by limiting the number of classes. Thus, few studies have assessed the effect of that number on their performance. In the 24 CV-based animal identification studies cited by Favret & Sieracki (2016) and Weinstein (2018), no significant correlation existed between the number of classes used in each classifier and their respective performances. In their large data set experiment, Sun et al. (2017) also found no difference when training with 1000 or 18 000 classes. But in contrast, Favret & Sieracki (2016) observed a counterintuitive increase in performance as more insect species were included into their training set. They hypothesised that, although a higher number of possible outcomes could increase confusion, the higher number of comparison points helped determine the important features of each category. Further tests are needed to disentangle the effect of the number of classes in training or the relative difference in morphology of these classes on performance. In general, practical applications of CV in ecology would benefit from more information on this effect.

4.4. Potential application of CV to a real ecological data set

To deploy classifiers such as these in a 'real-life' ecological study, reasonable performances must be achieved while retaining time and cost effectiveness of building the training set. In our study, no classifier achieved average performance above 78% (0.78), which would mean 1 misidentification out of 5 predictions, at best. We also observed high interclass variability as some OTUs were consistently well identified while others were, on the contrary, always misclassified. Even if the measured average performances were considered acceptable, it would introduce completely false appreciation of the distribution of some OTUs and local diversity.

This variability in both expert and machine classification performance between classes or taxa has been observed by other authors (Beijbom et al. 2015, Cho et al. preprint https://arxiv.org/abs/1511.06348). Experts in Durden et al. (2016) had various annotation successes for different taxa and Schoening et al. (2012) found that human observers and their semi-automated classifier had variable success at detecting and identifying different taxa but agreed on which one had the best performance. It is therefore sensible to consider the predictions of each OTU class separately and only rely on those for which the classifier achieves good performances.

Good performance obtained by our classifier with some specific OTU classes is encouraging and automated annotations could be an appropriate method to study these OTUs. The top 10 best and worst OTUs ranked by sensitivity shows that the classifiers are better at identifying large sized organisms exhibiting a low intra-class morphological variability. The majority of the top 10 OTUs were rare (e.g. less than 100 training images). If CV were applied to these rare taxa, there would be a proportionally higher impact of any misidentification or false positives (predictions of presence that are in fact another OTU) on the results. Yet given their relatively low number of occurrences (10s to a few 100s), a manual verification step (or semi-automated identification), as performed by Schoening et al. (2012) and suggested by Marburg & Bigham (2016), would be easy to perform for a reasonable time investment and to ensure the reliability of the predictions. For example, the Largo tool in BIIGLE, used in this study to validate the training set, makes a visual check of a large number of annotations much faster than going over the raw images again. It would be an efficient way to validate the CNN's predictions and make results usable. On the other hand, OTU261 and OTU339, both among the top 10 OTU classes, were very abundant in the study area (above 1200 ind.). In a larger data set, manual validation of identifications of these OTUs would be impractical and, to some extent, cancel the gains in speed and objectivity of CV. Ideally, their identification should be fully automated if the classifier is to be deployed in these conditions.

CNNs are considered as 'black boxes' whose internal prediction and decision process are difficult to visualize and understand (Samek et al. preprint https://arxiv.org/abs/1708.08296), yet we can speculate on the reasons why some organisms are better identified than others. OTU261 is very constant in shape and colour and has a distinctive pattern on its outside: this homogeneity probably makes it easily identifiable. OTU339 can be in different pose or orientation within an image but has a number of distinguishing features, such as its reflective eyes, and its long, often spread-out limbs. These features are not found in other OTUs, probably making confusions rarer.

OTU2 and OTU23 are both anemones. OTU2 is a cerianthid (a tube anemone) of various size and orientation and OTU23 is a Halcampidae/Edwardsiidea-like anemone of very small size. They are similar in shape and size, hence distinguishing them is difficult even for human annotators. This could explain the lower performances of the classifier on these OTUs. The fact that, during annotation, the smaller OTUs were localized with point coordinates and an arbitrary radius was used for slicing is a potential source of bias. The 240 pixel square used by default leaves a large surface of the image as background. This feature, common to several small OTUs, including the small anemones, could be a cause for higher rates of confusion in the predictions. With OTU261 and OTU339, high sensitivity (up to 0.95 and 0.92, respectively) and high precision (up to 0.95 and 0.82, respectively) were achieved by the classifiers, meaning they were usually correctly identified and false positives (another OTU wrongly identified as one of them) were relatively rare. These performances are equivalent to those of human experts working on a very similar ecosystem (Durden et al. 2016) without the inconsistency over time by individual observers reported by these authors. Moreover, based on the speed of manual annotation in this study, we estimate that building the training set, validating it, training the classifier and testing it could be achieved in a matter of days rather than months. Therefore, these classifiers can be applied to the remaining unannotated images in our data set and provide useful presence records of these specific OTUs. This would be a valuable contribution to this study of deep-sea ecosystems.

Classifier A1000 had the best performance of all classifiers and would detect almost all individuals of OTU261 and OTU339, but it needed a large training set, while the A200 classifier had very similar performances but needed 5 times less training material and is therefore more cost-effective. These group A classifiers however, risk producing a high number of false positives if they encounter too many individuals of an OTU they have not been trained on. Thus, it is only applicable if diversity at the study site is low or is predominantly represented by a small number of OTUs. These classifiers would not be suitable to survey very diverse ecosystems, like coral reefs.

In the long term, classifiers able to identify as many OTUs as possible, even semi-automatically, are undoubtedly more desirable, even if they perform slightly less well. In our study, the group C classifiers had marginally lower performances than group A classifiers, particularly if training with 200 images, but both sensitivity and precision were above 0.9 for OTU261, which is still comparable to manual annotation. Thus, although this design is still valid for identifying specific OTUs, it has the advantage, as it is trained on 52 classes, of being able to automatically identify more OTUs. Even if some of these identifications need to be manually validated, it is more representative of real field studies where many OTUs could be encountered.

Based on our observations on classifier performances, we recommend the following approach to the use of CV in small-scale benthic ecological studies: (1) build a general classifier to identify OTUs that achieve good performance and quantify the error rate associated with each. This can be an unbalanced design with many OTUs, like group C in the current study. A large number of classes potentially allows more OTUs to be tested. The number of training images should preferably be above 200 plus a 20% or more surplus so the classifier can be tested with independent data. (2) Only use the presence prediction of those OTUs that have good performances and regard any other predictions as unknown or an absence of those. (3) Consider all remaining OTUs as 'unidentified' and leave for manual identification or for later, more efficient, automated classifiers. Alternatively, a one-versus-all classifier could be trained and deployed for each of the target OTUs (Siddiqui et al. 2018), but this approach would become logistically challenging for a large number of target OTUs.

Even if the presence records of some OTUs are not sufficient to understand the composition and dynamics of an ecosystem, it will still contribute and more importantly, take on some of the annotation time, leaving experts free to perform other tasks while providing useful insights in ecology. In the specific case of this study, the automated identification of OTU261 and OTU339 would be useful for deep-sea ecologists, especially if it only requires a few days of work. Indeed, very little is known about the fine-scale distribution of these OTUs. *Syringammina fragillissima* (OTU261) is considered habitat forming, enhances local metazoan abundance (Gooday 1984, Levin et al. 1986, Levin & Thomas 1988) and aggregations of this species are a Vulnerable Marine Ecosystem under United Nations General Assembly Resolution 61/105 (UNGA 2003). The squat-lobsters *Munida sarsi* or *M. tenuimana* may play an important role in the benthic community as predators or scavengers (Hudson & Wigham 2003) and are suited to examining ecological patterns (Rowden et al. 2010). Extracting the location of these 2 taxa from a vast data set would be a valuable way to study or map their extent and distribution as other studies have done with other faunal groups at fine (Milligan et al. 2016) and broad scales (Rex & Etter 2010, Wei et al. 2010). Besides, this would complement the studies carried out by trawling, which can underestimate diversity of benthic crustaceans (Cartes & Sarda 1992, Ayma et al. 2016) and destroy xenophyophores (Roberts et al. 2000). Furthermore, appropriate data to study rhythmic diel and seasonal movements or behavioural changes of megabenthos, including squat-lobsters (Aguzzi & Company 2010, Aguzzi et al. 2013), is currently lacking. By providing more data on abundance and distribution of *Munida* sp., this method could greatly aid this field of research. Also, assuming that benthic decapods can easily be counted with CV, and abundance differences reliably measured, the stock assessment of *Nephrops norvegicus*, an important and carefully monitored commercial species (ICES 2010, Sardà & Aguzzi 2012) could be achieved at greater speed, cost-efficiency and more objectively than by trawl.

This study only deals with the identification of animals and not with their detection within the images, which was performed manually in BIIGLE before cropping images around each individual. Detection is an essential step in automated image analysis and many solutions have been explored (Cheng & Han 2016, Hollis et al. 2016, Sorensen et al. 2017). A step for object detection needs to be added to the protocol described here to completely automate the process. This study also did not deal with the behaviour of the classifiers when presented with novel OTUs. This situation is unavoidable in real-life ecological data sets, and although methods exist for novelty detection (Pimentel et al. 2014), this remains to be integrated into our methodology.

5. CONCLUSIONS

Our results demonstrate that CV-based image annotation cannot entirely replace manual annotation of benthic images at present, but that usable results can be obtained for specific taxa with open-source software, very little tuning and optimisation of the model itself and a relatively small training data set (200 images). These results can inform the distribution of these specific taxa in a more robust way than currently possible. In general, monitoring the abundance of a single taxon for novel research or in routine stock assessment could greatly benefit from this method. It offers greater speed, cost-efficiency, objectivity and consistency than trawl surveys or manual image analysis.

This does not immediately solve the many challenges of marine ecology but could initiate momentum and catalyse further development of CV-based methods in this area as these tools are becoming more accessible to non-specialists. The development of fully automated image annotation, or pragmatic combinations of manual and automated annotation protocols (Matabos et al. 2017), is likely to continue across different platforms capable of gathering large image data sets (Marini et al. 2018a,b). Indeed, there is still much room left for improving classifier performance with better image pre-processing prior to the training or better tuning of the model, and more research could lead to game-changing methodological development. In the age of big data and global open research, the participation of many different actors of research contributing data (Hampton et al. 2013, Hussey et al. 2015), computing power, and above all, taxonomic and informatics expertise (Weinstein 2018) could be synthesised in the development of CV tools able to take on some of the workload of human researchers and increase the pace at which the oceans are explored and sampled and, ultimately, how they are preserved.

Acknowledgements. The authors acknowledge the officers and crew of RV 'James Cook' as well as the Autosub support team who assisted in the collection of the data. The Deep-Links project is funded by NERC NE/K011855/1. N.P.'s PhD, C.H.'s time and the processing of the data in Google cloud were funded by University of Plymouth. We also thank Kirsty Morris, Erik Simon Lledo and Marcus Shirley for their advice on processing the images prior to analysis.

LITERATURE CITED

Abadi M, Barham P, Chen J, Chen Z and others (2016) tensorflow: a system for large-scale machine learning. In 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), p 265–283

Aguzzi J, Company JB (2010) Chronobiology of deep-water decapod crustaceans on continental margins. Adv Mar Biol 58:155–225

Aguzzi J, Costa C, Ketmaier V, Angelini C, Antonucci F, Menesatti P, Company JB (2013) Light-dependent genetic and phenotypic differences in the squat lobster *Munida tenuimana* (Crustacea: Decapoda) along deep continental margins. Prog Oceanogr 118:199–209

Althaus F, Hill N, Ferrari R, Edwards L and others (2015) A standardised vocabulary for identifying benthic biota and substrata from underwater imagery: the CATAMI classification scheme. PLOS ONE 10:e0141039

UNGA (United Nations General Assembly) (2003) Oceans and the law of the sea. Report of the Secretary General. UNGA A/58/65

Ayma A, Aguzzi J, Canals M, Lastras G, Bahamon N, Mecho A, Company JB (2016) Comparison between ROV video and Agassiz trawl methods for sampling deep water fauna of submarine canyons in the northwestern Mediterranean Sea with observations on behavioural reactions of target species. Deep-Sea Res I 114:149–159

Beijbom O, Edmunds PJ, Kline DI, Mitchell BG, Kriegman D (2012) Automated annotation of coral reef survey images. In: Proc 25th IEEE Conf Computer Vision and Pattern Recognition (CVPR), 16–21 June 2012, Providence, RI. IEEE, Piscataway, NJ, p 1170–1177

Beijbom O, Edmunds PJ, Roelfsema C, Smith J and others (2015) Towards automated annotation of benthic survey images: variability of human experts and operational modes of automation. PLOS ONE 10:e0130312

Bicknell AW, Godley BJ, Sheehan EV, Votier SC, Witt MJ (2016) Camera technology for monitoring marine biodiversity and human impact. Front Ecol Environ 14: 424–432

Borja A, Elliott M, Snelgrove PVR, Austen MC and others (2016) Bridging the gap between policy and science in assessing the health status of marine ecosystems. Front Mar Sci 3:175

Brandt A, Gutt J, Hildebrandt M, Pawlowski J, Schwendner J, Soltwedel T, Thomsen L (2016) Cutting the umbilical: new technological perspectives in benthic deep-sea research. J Mar Sci Eng 4:36

Bullimore RD, Foster NL, Howell KL (2013) Coral-characterized benthic assemblages of the deep Northeast Atlantic: defining 'coral gardens' to support future habitat mapping efforts. ICES J Mar Sci 70:511–522

Cartes JE, Sarda F (1992) Abundance and diversity of decapod crustaceans in the deep-Catalan Sea (Western Mediterranean). J Nat Hist 26:1305–1323

Cheng G, Han J (2016) A survey on object detection in optical remote sensing images. ISPRS J Photogramm Remote Sens 117:11–28

Clark MR, Consalvey M, Rowden AA (eds) (2016) Biological sampling in the deep sea. John Wiley & Sons, Hoboken, NJ

Costello MJ, Coll M, Danovaro R, Halpin P, Ojaveer H, Miloslavich P (2010) A census of marine biodiversity knowledge, resources, and future challenges. PLOS ONE 5:e12110

Culverhouse PF, Williams R, Reguera B, Herry V, González-Gil S (2003) Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. Mar Ecol Prog Ser 247:17–25

Culverhouse PF, Macleod N, Williams R, Benfield MC, Lopes RM, Picheral M (2014) An empirical assessment of the consistency of taxonomic identifications. Mar Biol Res 10:73–84

Danovaro R, Aguzzi J, Fanelli E, Billett D and others (2017) An ecosystem-based deep-ocean strategy. Science 355: 452–454

Durden JM, Bett BJ, Schoening T, Morris KJ, Nattkemper TW, Ruhl HA (2016) Comparison of image annotation data generated by multiple investigators for benthic ecology. Mar Ecol Prog Ser 552:61–70

Edgington DR, Cline DE, Davis D, Kerkez I, Mariette J (2006) Detecting, tracking and classifying animals in underwater video. In: Proc OCEANS 2006, 18–21 September 2006, Boston, MA, p 1–5

Favret C, Sieracki JM (2016) Machine vision automated species identification scaled towards production levels. Syst Entomol 41:133–143

Gaston KJ, O'Neill MA (2004) Automated species identification: Why not? Philos Trans R Soc Lond B Biol Sci 359: 655–667

Gomes-Pereira JN, Auger V, Beisiegel K, Benjamin R and others (2016) Current and future trends in marine image annotation software. Prog Oceanogr 149:106–120

Gooday AJ (1984) Records of seep-sea rhizopod tests inhabited by metazoans in the North-east Atlantic. Sarsia 69:45–53

Hampton SE, Strasser CA, Tewksbury JJ, Gram WK and others (2013) Big data and the future of ecology. Front Ecol Environ 11:156–162

Hernandez PA, Graham CH, Master LL, Albert DL (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29:773–785

Hollis DJ, Edgington D, Cline D (2016) Automated detection of deep-sea animals. http://digitalcommons.calpoly.edu/star/370

Howell KL, Davies JS (2016) Deep-sea species image catalogue, on-line version 2. https://deepseacruorg/2016/12/16/deep-sea-species-image-catalogue

Hudson IR, Wigham BD (2003) *In situ* observations of predatory feeding behaviour of the galatheid squat lobster *Munida sarsi* using a remotely operated vehicle. J Mar Biol Assoc UK 83:463–464

Hussey NE, Kessel ST, Aarestrup K, Cooke SJ and others (2015) Aquatic animal telemetry: a panoramic window into the underwater world. Science 348:1255642

ICES (2010) Report of the working group on the assessment of southern shelf stocks of hake, monk and megrim. ICES CM 2006/ACFM:01

Jeffries H, Berman M, Poularikas A, Katsinis C, Melas I, Sherman K, Bivins L (1984) Automated sizing, counting and identification of zooplankton by pattern recognition. Mar Biol 78:329–334

Jongman RHG (2013) Biodiversity observation from local to global. Ecol Indic 33:1–4

Junqué de Fortuny E, Martens D, Provost F (2013) Predictive modeling with big data: Is bigger really better? Big Data 1:215–226

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105

Langenkämper D, Zurowietz M, Schoening T, Nattkemper TW (2017) BIIGLE 2.0 - browsing and annotating large marine image collections. Front Mar Sci 4:83

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436

Levin LA, Thomas CL (1988) The ecology of xenophyophores (Protista) on eastern Pacific seamounts. Deep-Sea Res A 35:2003–2027

Levin LA, DeMaster DJ, McCann LD, Thomas CL (1986) Effect of giant protozoans (class: Xenophyophorea) on deep-seamount benthos. Mar Ecol Prog Ser 29:99–104

Lu D, Weng Q (2007) A survey of image classification methods and techniques for improving classification performance. Int J Remote Sens 28:823–870

Lucieer VL, Forrest AL (2016) Emerging mapping techniques for autonomous underwater vehicles (AUVs). In: Finkl CW, Makowski C (eds) Seafloor mapping along continental shelves: research and techniques for visualizing benthic environments. Springer International Publishing, Cham, p 53–68

MacLeod N, Benfield M, Culverhouse P (2010) Time to automate identification. Nature 467:154–155

Mallet D, Pelletier D (2014) Underwater video techniques for observing coastal marine biodiversity: a review of sixty years of publications (1952–2012). Fish Res 154:44–62

Manderson T, Li J, Dudek N, Meger D, Dudek G (2017) Robotic coral reef health assessment using automated image analysis. J Field Robot 34:170–187

Manel S, Williams HC, Ormerod SJ (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. J Appl Ecol 38:921–931

Marburg A, Bigham K (2016) Deep learning for benthic fauna identification. In: Proc OCEANS 2016, 19–23 September 2016, Monterey, CA, p 1–5

Marini S, Corgnati L, Mantovani C, Bastianini M and others (2018a) Automated estimate of fish abundance through the autonomous imaging device GUARD1. Measurement 126:72–75

Marini S, Fanelli E, Sbragaglia V, Azzurro E, Del Rio Fernandez J, Aguzzi J (2018b) Tracking fish abundance by underwater image recognition. Sci Rep 8:13748

Matabos M, Hoeberechts M, Doya C, Aguzzi J and others (2017) Expert, crowd, students or algorithm: Who holds the key to deep sea imagery 'big data' processing? Methods Ecol Evol 8:996–1004

Maxwell AE, Warner TA, Fang F (2018) Implementation of machine-learning classification in remote sensing: an applied review. Int J Remote Sens 39:2784–2817

McClain CR, Rex MA (2015) Toward a conceptual understanding of β-diversity in the deep-sea benthos. Annu Rev Ecol Evol Syst 46:623–642

Milligan RJ, Morris KJ, Bett BJ, Durden JM and others (2016) High resolution study of the spatial distributions of abyssal fishes by autonomous underwater vehicle. Sci Rep 6:26095

Norouzzadeh MS, Nguyen A, Kosmala M, Swanson A, Palmer MS, Packer C, Clune J (2018) Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proc Natl Acad Sci USA 115:E5716–E5725

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22:1345–1359

Pimentel MAF, Clifton DA, Clifton L, Tarassenko L (2014) A review of novelty detection. Signal Process 99:215–249

Rampasek L, Goldenberg A (2016) TensorFlow: Biology's gateway to deep learning? Cell Syst 2:12–14

Rex MA, Etter RJ (2010) Deep-sea biodiversity: pattern and scale. Harvard University Press, Cambridge, MA

Ripley B, Venables W, Ripley MB (2016) Package 'nnet'. R package version 7-3

Roberts JM, Harvey SM, Lamont PA, Gage JD, Humphery JD (2000) Seabed photography, environmental assessment and evidence for deep-water trawling on the continental margin west of the Hebrides. Hydrobiologia 441: 173–183

Rohlf FJ, Sokal RR (1967) Taxonomic structure from randomly and systematically scanned biological images. Syst Zool 16:246–260

Romero-Ramirez A, Grémare A, Bernard G, Pascal L, Maire O, Duchêne JC (2016) Development and validation of a video analysis software for marine benthic applications. J Mar Syst 162:4–17

Rowden AA, Schnabel KE, Schlacher TA, Macpherson E, Ahyong ST, de Forges BR (2010) Squat lobster assemblages on seamounts differ from some, but not all, deep-sea habitats of comparable depth. Mar Ecol 31:63–83

Sardà F, Aguzzi J (2012) A review of burrow counting as an alternative to other typical methods of assessment of Norway lobster populations. Rev Fish Biol Fish 22: 409–422

Schneider S, Taylor GW, Kremer S (2018) Deep learning object detection methods for ecological camera trap data.

15th Conference on Computer and Robot Vision (CRV), IEEE, p 321–328

Schoening T, Bergmann M, Ontrup J, Taylor J and others (2012) Semi-automated image analysis for the assessment of megafaunal densities at the Arctic deep-sea observatory HAUSGARTEN. PLOS ONE 7:e38179

Schoening T, Durden J, Preuss I, Albu AB and others (2017) Report on the marine imaging workshop 2017. Res Ideas Outcomes 3:e13820

Siddiqui SA, Salman A, Malik MI, Shafait F, Mian A, Shortis MR, Harvey ES (2018) Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data. ICES J Mar Sci 75:374–389

Solan M, Germano JD, Rhoads DC, Smith C and others (2003) Towards a greater understanding of pattern, scale and process in marine benthic systems: a picture is worth a thousand worms. J Exp Mar Biol Ecol 285-286:313–338

Sorensen S, Treible W, Hsu L, Wang X, Mahoney AR, Zitterbart DP Kambhamettu C (2017) Deep learning for polar bear detection. In: Sharma P, Bianchi F (eds) Proc Scandinavian Conf Image Analysis. Springer, Cham, p 457–467

Sun C, Shrivastava A, Singh S Gupta A (2017) Revisiting unreasonable effectiveness of data in deep learning era. In: Proc 2017 IEEE Int Conf Computer Vision (ICCV), 22–29 October 2017, Venice. IEEE, Piscataway, NJ, p 843–852

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proc 29th IEEE Conf Computer Vision and Pattern Recognition (CVPR), 26 June–1 July 2016, Las Vegas, NV. IEEE, Piscataway, NJ, p 2818–2826

R Core Team (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Van Dover C, Aronson J, Pendleton L, Smith S and others (2014) Ecological restoration in the deep sea: Desiderata. Mar Policy 44:98–106

Wei CL, Rowe GT, Escobar-Briones E, Boetius A and others (2010) Global patterns and predictions of seafloor biomass using random forests. PLOS ONE 5:e15323

Weinstein BG (2018) A computer vision for animal ecology. J Anim Ecol 87:533–545

Wheeler B, Torchiano M (2010) lmPerm: permutation tests for linear models. R package version 1

Wickham H (2017) Tidyverse: easily install and load 'tidyverse' packages. R package version 1

Williams SB, Pizarro O, Steinberg DM, Friedman A, Bryson M (2016) Reflections on a decade of autonomous underwater vehicles operations for marine survey at the Australian Centre for Field Robotics. Annu Rev Contr 42: 158–165

Wynn R, Bett B, Evans A, Griffiths G and others (2012) Investigating the feasibility of utilizing AUV and glider technology for mapping and monitoring of the UK MPA network. National Oceanography Centre, Southampton

Editorial responsibility: Lisandro Benedetti-Cecchi, Pisa, Italy

Submitted: October 15, 2018; Accepted: March 6, 2019
Proofs received from author(s): April 15, 2019