JOURNAL OF PLANKTON RESEARCH | VOLUME 32 | NUMBER 3 | PAGES 285–303 | 2010

Digital zooplankton image analysis using the ZooScan integrated system

GABY GORSKY[1,2]*, MARK D. OHMAN[3], MARC PICHERAL[1,2], STÉPHANE GASPARINI[1,2], LARS STEMMANN[1,2], JEAN-BAPTISTE ROMAGNAN[3], ALISON CAWOOD[3], STÉPHANE PESANT[4], CARMEN GARCÍA-COMAS[1,2,5] AND FRANCK PREJGER[1,2]

[1]UPMC UNIVERSITY OF PARIS 06, UMR 7093, LOV, OBSERVATOIRE OCÉANOGRAPHIQUE F-06234, VILLEFRANCHE/MER, FRANCE, [2]CNRS, UMR 7093, LOV, OBSERVATOIRE OCÉANOGRAPHIQUE, F-06234 VILLEFRANCHE/MER, FRANCE, [3]CALIFORNIA CURRENT ECOSYSTEM LTER SITE, SCRIPPS INSTITUTION OF OCEANOGRAPHY, LA JOLLA, CA 92093-0218, USA, [4]MARUM, INSTITUTE FOR MARINE ENVIRONMENTAL SCIENCES, UNIVERSITY BREMEN, LEOBENER STRASSE, POP 330 440, 28359 BREMEN, GERMANY AND [5]STAZIONE ZOOLOGICA ANTON DOHRN, VILLA COMUNALE, 80121 NAPOLI, ITALY

*CORRESPONDING AUTHOR: gorsky@obs-vlfr.fr

Received July 14, 2009; accepted in principle October 26, 2009; accepted for publication November 14, 2009

Corresponding editor: Roger Harris

ZooScan with ZooProcess and Plankton Identifier (PkID) software is an integrated analysis system for acquisition and classification of digital zooplankton images from preserved zooplankton samples. Zooplankton samples are digitized by the ZooScan and processed by ZooProcess and PkID in order to detect, enumerate, measure and classify the digitized objects. Here we present a semi-automatic approach that entails automated classification of images followed by manual validation, which allows rapid and accurate classification of zooplankton and abiotic objects. We demonstrate this approach with a biweekly zooplankton time series from the Bay of Villefranche-sur-mer, France. The classification approach proposed here provides a practical compromise between a fully automatic method with varying degrees of bias and a manual but accurate classification of zooplankton. We also evaluate the appropriate number of images to include in digital learning sets and compare the accuracy of six classification algorithms. We evaluate the accuracy of the ZooScan for automated measurements of body size and present relationships between machine measures of size and C and N content of selected zooplankton taxa. We demonstrate that the ZooScan system can produce useful measures of zooplankton abundance, biomass and size spectra, for a variety of ecological studies.

INTRODUCTION

Historically, zooplankton have been sampled primarily by surveys that use nets, pumps or water bottles to collect specimens for quantifying distributional patterns. While such surveys provide invaluable information on species and life stages, their temporal and spatial resolution is usually limited, owing to the time and resources required for sample analysis by trained microscopists. This limited resolution of zooplankton data sets reduces our ability to understand processes controlling pelagic ecosystem dynamics on multiple time and space scales.

Recent advances in image processing and pattern recognition of plankton have made it possible to automatically or semi-automatically identify and quantify the composition of plankton assemblages at a relatively coarse taxonomic level (Benfield *et al.*, 2007). The importance of this approach was recognized by the Scientific Committee on Oceanic Research (SCOR), who created an international working group to evaluate the state of Automatic Visual Plankton Identification (http://www.scor-wg130.net). The hope is that the advent of digital imaging technology, combined with better algorithms for machine learning and increased computer capacity, will facilitate much

doi:10.1093/plankt/fbp124, available online at www.plankt.oxfordjournals.org.

more rapid means for characterizing plankton distributions assessed from a variety of different sampling methods.

Early attempts to use optical bench-top methods for treatment of plankton samples were undertaken by Ortner *et al.* (Ortner *et al.*, 1979) who used silhouette photography to record the contents of a plankton sample. Silhouette imaging of plankton samples on photographic film or video imaging and a limited digitization of plankton samples followed by automatic identification was further developed in the 1980s (Jeffries *et al.*, 1980, 1984; Rolke and Lenz, 1984; Gorsky *et al.*, 1989; Berman, 1990). A variety of bench top methods is now under development (e.g. Benfield *et al.*, 2007). In addition to developments in the ocean sciences, automated image analysis is commonly applied in other fields of biology and medical sciences. Within the geosciences, machine learning is often applied to quantify the morphology of fossils (Kennett, 1968; Bollmann *et al.*, 2004).

A wide range of image analysis and treatment software exists from these various fields. Most can be adapted for enumeration and measurement of particles, but zooplankton pattern recognition is a much more challenging goal. Most zooplankton taxa display high shape variability. Other difficulties include the diversity of body orientations relative to the imaging plane, differences in extension of appendages, damaged individuals and variable quantities of amorphous organic aggregates that must be distinguished by automated recognition methods. With these challenges, it is not surprising that recent papers show relatively low automated zooplankton classification efficiency (Bell and Hopcroft, 2008; Irigoien *et al.*, 2009).

Progress in scanner technology has made it feasible to digitize good quality images of large numbers of plankton individuals simultaneously. The hardware presented here is not the only system based on scanner technology that can be used for zooplankton image treatment (e.g. Wiebe *et al.*, 2004; Bell and Hopcroft, 2008; Irigoien *et al.*, 2009). We have in the past used such systems by adapting commercial scanners (Grosjean *et al.*, 2004). However, a series of problems led us to build an industrialized, rugged, water-resistant ZooScan suitable for organisms ranging in size from 200 μm to several centimeters, together with dedicated imaging software we call ZooProcess and Plankton Identifier (PkID). ZooScans can be calibrated so that different ZooScan units produce normalized images of identical optical characteristics that can be intercompared among laboratories, facilitating cooperative sample analysis. Such a network of calibrated ZooScan instruments currently exists in the Mediterranean region in the framework of the CIESM Zooplankton Indicators program: (http://www.ciesm.org/marine/programs/zooplankton.htm).

In this paper, we first describe the overall approach used, including ZooScan hardware together with ZooProcess and PkID software. We discuss building and validating training sets, the selection of classification algorithms and the accuracy of body size and biomass estimations that can be derived from the ZooScan system. We propose standards for long-term archiving and sharing of raw and processed images and output files. We demonstrate a semiautomatic classification approach based on human validation of automated zooplankton image analysis that provides highly reliable results that are appropriate for quantitative ecological studies. Second, we illustrate the procedures for sample and data analysis through specific application of the ZooScan system to an annual time series of zooplankton samples from the Bay of Villefranche-sur-mer.

METHOD

Sequential steps for sample preparation and scanning with ZooScan hardware, together with image processing with ZooProcess and PkID software, are explained below. Appendix 1 lists a glossary of terms related to image analysis.

Building learning sets

In experiments to determine the optimal number of objects to sort into each category when constructing a learning set (see Appendix 1), we selected eight categories of organisms scanned from Villefranche-sur-mer, each with more than 950 vignettes. We randomly extracted subsets of 10, 20, 30, ... 900 vignettes from each of the eight categories. These subsets were considered independent learning sets and we ran a classifier on each to assess the recall (percent true positives) and contamination (percent false positives) as a function of size of the learning set.

Morphometric measurements and biomass

ZooScan analyses provide sensitive measures of body size, which can be converted to size spectra. To calculate the biovolume of an object from its cross-sectional area, it is necessary to know the geometric shape of the object, the ratio of its major and minor axes and its orientation relative to the illumination system. Copepods can be represented as ellipsoids (Herman, 1992). The ZooScan provides estimates of body length (here major axis of the best fitting ellipse) and width (here minor axis). We evaluated the accuracy of

ZooScan measurements of body length and cross-sectional area. Automated measurements of preserved zooplankton as recorded by ImageJ in ZooProcess were compared with manual measurements of several zooplankton taxa (appendicularians, chaetognaths, copepods, euphausiids, ostracods and thecosome pteropods) collected in the California Current on CalCOFI (California Cooperative Oceanic Fisheries Investigations) cruises. Specimens were collected along CalCOFI line 80 between February 2006 and August 2008 and preserved in formaldehyde buffered with sodium tetraborate. Manual measurements were made using a calibrated on-screen measuring tool, and compared with machine-measured feret diameter, major elliptical axis, minor elliptical axis and equivalent circular diameter (ECD) of the same individuals, identified manually. ECD was determined from the variable "area excluded," which excludes clear regions in the interior of an organism from the cross-sectional area of the organism. Manual length measurements of curved organisms (e.g. chaetognaths and appendicularians) were made by summing a series of line segments along the central axis of the organism.

For C and N relationships, live zooplankton (copepods, chaetognaths and euphausiids) from the California Current were anaesthetized with carbonated water (diluted 1:4 with seawater), scanned, manually identified and individually measured. Multiple species were included in each higher taxon analyzed, in order to obtain group-specific relationships. The cross-sectional area of chaetognaths and euphausiids was measured manually on-screen using multiple rectangles drawn within the outline of each organism. The area of copepods was measured using two ellipses, one defining the prosome and the other the urosome. These manual measurements were compared with machine measurements of the same individuals. Organisms were dried overnight at 60°C and the carbon and nitrogen content of the individual organisms determined at the analytical facility of the Scripps Institution of Oceanography using an elemental analyzer (Costech Analytical Technologies model 4010) calibrated with acetanilide.

Case study: Bay of Villefranche-sur-mer

To illustrate application of the ZooScan system (www.zooscan.com), we analyzed a series of samples describing annual variation of zooplankton from Pt. B (43° 41.10′N, 7° 18.94′E) in the Bay of Villefranche-sur-mer, France. Zooplankton were sampled with a 57 cm diameter WP2 net with a mesh size of 200 μm retrieved vertically at 1 m s^{-1} from a depth of 75 m to the surface, and fixed in 4% v/v formaldehyde buffered with sodium tetraborate. Thirty vertical hauls were made between 22 August 2007 and 8 October 2008. These samples were scanned by ZooScan in two size fractions, <1 mm and >1 mm, leading to a set of 60 scans. For classification, we began with a learning set of 13 categories (10 zooplankton + 3 non-zooplankton) that we had created previously. This can be downloaded at: (http://www.obs-vlfr.fr/LOV/ZooPart/ZooScan/Training_Set_Villefranche/esmeraldo_learning_set.zip).

Description of the ZooScan system

System overview

Hardware. The ZooScan (http://www.zooscan.com) is composed of two main waterproof elements that allow safe processing of liquid samples. The hinged base contains a high resolution imaging device and a drainage channel that is used for sample recovery (Fig. 1). The top cover generates even illumination and houses an optical density (OD) reference cell. Although the ZooScan permits scanning at higher resolution than 2400 dpi, the optical pathway through two successive interfaces (air to water and water to glass) presently limits the working resolution to this value. With a pixel resolution of 10.6 μm, the ZooScan is well suited for organisms larger than 200 μm.

The imaging area of the ZooScan is defined by the choice of one of two transparent frames (11 × 24 cm or 15 × 24 cm) inserted inside the scanning cell. Both frames have a 5 mm step; water is added above this step to avoid forming a meniscus on the periphery of the image. Both frames permit the acquisition and processing of scans as a single image, avoiding biases that may occur when an image is divided into multiple cells.

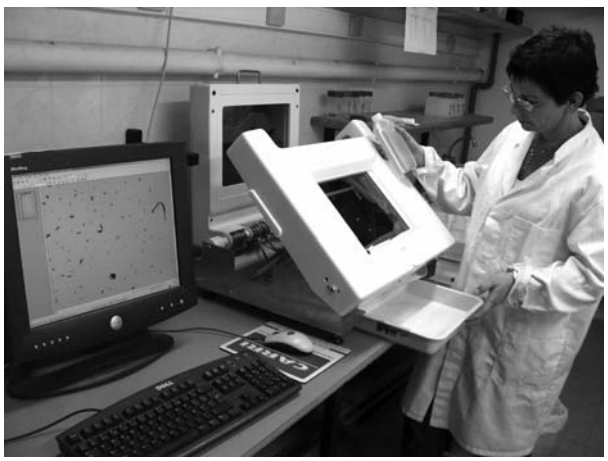


Fig. 1. Sample recovery from the ZooScan, illustrating the top cover, hinged base and sample recovery tray.

Software. The sequence followed in scanning and analysis of zooplankton samples is shown schematically in Fig. 2. The initial steps are completed using ZooProcess software: (i) scan and process a blank background image, (ii) scan the sample to acquire a high quality raw image, linked to associated metadata, (iii) normalize the raw image and convert to full grey scale range, (iv) process images by subtracting the background and removing frame edges, (v) extract and measure individual objects. Subsequent analysis steps are done with ZooProcess in combination with PkID: (vi) create a learning set comprised of representative images from each category of organisms or objects that will be identified, (vii) build a classifier to optimize the capability to accurately recognize the desired categories, and create a confusion matrix (CM) to verify the classifier, (ix) apply the classifier to the suite of unidentified objects and (x) manually inspect the classified objects and move any misidentified objects to the appropriate category. These steps are explained more fully below.

Sample treatment. The aliquot volume of a plankton sample to be analyzed is determined by the abundance and size distribution of the organisms. It is important to minimize coincidence of overlapping animals on the optical surface. At present, a maximum of approximately 1000–1500 objects is scanned in the larger frame, although this value can be exceeded. Because the abundance of organisms usually decreases with increasing body size, it is preferable to scan two (or more) separate size fractions of each sample. One fraction contains larger individuals that are less abundant, obtained from a larger sample aliquot, and the other includes the more numerous smaller individuals, from a smaller aliquot. A mesh size of 1000 μm is efficient for separating large and small size fractions of mesozooplankton.

Only immobile organisms (i.e. preserved or anaesthetized) can be scanned, because they must remain still for ~150 s. Prior to sample processing, the fixative is removed and replaced with either filtered sea water or tap water. Water should be at room temperature in order

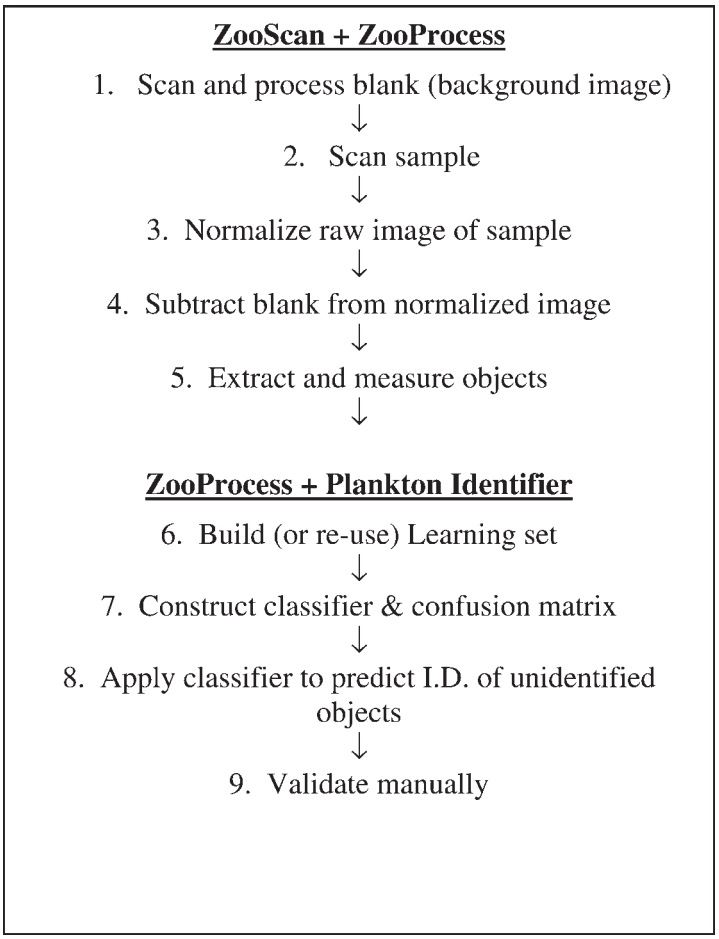


Fig. 2. Schematic illustration of the primary steps in the scanning and analysis of zooplankton samples with the ZooScan/ZooProcess/Plankton Identifier system.

to avoid air bubble formation. We do not stain samples, in order to maintain them unaltered for future comparative studies. Although ZooProcess software provides a tool to separate overlapping organisms once the sample has been scanned, it is important to physically separate touching organisms in the scanning frame and separate them from the frame edges prior to digitizing the sample. Manual separation takes ~10 min per sample.

Detailed description

Zooprocess. ZooProcess software is based on the ImageJ macro language (Abramoff *et al.*, 2004; Rasband, 2005). In addition to guiding the primary steps of scanning, normalization and object detection, ZooProcess provides tools for quality control and is linked to PkID software. Results presented here are based on ZooProcess default parameters. ZooProcess records true raw 16 bit grey images from the ZooScan charged couple device and creates blank images to be subtracted from normalized images.

Grey level normalization. Full grey level normalization of scanned images allows the exchange of images, training sets or data between different ZooScan units. Normalization is done on both the sample image and the background blank image, which is subtracted later. Grey level and size are among the most important variables used in automated plankton recognition (e.g. Hu and Davis, 2005).

The 16 bit raw image is converted to 8 bit source image after determination of both the white point [Wp, equation (1)] and the black point [Bp, equation (2)] from the median grey level (Mg). The OD range that can be resolved by the ZooScan is above 1.8. The sharpness of the background allows setting the white point close to the median grey level independent of the number and size of the organisms in the image.

$$\mathrm{Wp} = \mathrm{Mg} \bullet 1.15 \tag{1}$$

$$\mathrm{Bp} = \frac{\mathrm{Mg}}{1.15 \bullet \log(\mathrm{OD})} \tag{2}$$

ZooProcess provides a tool to check the efficiency of the procedure by scanning standard reference disks (diameter 5.6 mm) with ODs of 0.3 and 0.9. The average grey level values are 150 and 73 ($\pm$10%) for the two disks, respectively, for all ZooScans tested to date after full processing of the images using the default parameters (see Appendix 2). The normalization parameters and the median grey level of both the raw 16 bit image and the final 8 bit image are archived in the image log file.

Image processing. ZooProcess provides two methods for removing a heterogeneous background. A daily scan of the cell filled with filtered water is recommended, because the background image provides a blank and also records instrument stability over time. A background scan is faster to process and requires less computer memory than the second option, the rolling ball method (Sternberg, 1983), which requires no blank image to be scanned. A lower setting of the rolling ball diameter parameter will clean the background, but may create artifacts in zones of uneven contrast on the bodies of larger organisms. Apart from artifacts, measurements made on the same image processed using the two background subtraction methods differ less than 1%, thus the rolling ball can be used as an alternative when no blank image is available.

ZooProcess next measures the grey level in the OD reference disk area. This measurement is compared with the theoretical value calibrated in the factory. ZooProcess then detects the limits of the transparent frame and discards the irrelevant parts of the image. Objects touching the sides are automatically removed from the data set. This image is used for object detection and the extraction of measurement variables from each detected object.

Extraction of vignettes and attributes. The final image is segmented at a default level of 243, thus keeping 243 grey levels for characterizing organisms. Objects having an ECD $>$0.3 mm (default) are detected and processed.

More than 40 attributes (variables) are extracted from every object (Santos Filho *et al.*, submitted for publication). All metadata (Appendix 3), log file and the variables measured (Appendix 4) are stored in a text file called a PID file (Plankton Identifier file). All parameters used during the imaging process are recorded in the log file. Some examples of extracted vignettes and some measurements are illustrated in Appendix 5. After extraction of the vignette (Region of Interest, or ROI), the values of each measurement variable are associated with that vignette.

Additional quality control. Segmented black and white images and the objects' outlines are recorded for quality control. The segmented image is checked for the background subtraction and correct object contours. A 2D "dot cloud" graph allows selection and visualization of vignettes by clicking on a dot or on a selected zone, and the sample image can be visualized with the recognized outlines superimposed.

Plankton Identifier (PkID). PkID permits automatic and semi-automatic classification of plankton images. For the ZooScan application, PkID is interfaced with ZooProcess,

but it can also be used standalone. It has been developed in DELPHI (Borland), because the source code can be compiled. For supervised learning, PkID works in conjunction with Tanagra (Rakotomalala, 2005; http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html), also developed in DELPHI. Source code for PkID can be obtained on request for customization (see Gasparini, 2007).

Three successive steps are followed in applying PkID: (i) "Learning" creates training files that link measurements from groups of similar objects (vignettes); (ii) "Data analysis" permits construction and optimization of classifiers; (iii) "Show results" displays final identifications and statistical reports (see: http://www.obs-vlfr.fr/~gaspari/Plankton_Identifier/userguide.html).

Learning file creation. In the "learning" section, objects are grouped into categories by simple visual drag and drop of object vignettes.

Number of categories. The number of categories of objects selected is a trade-off between the number of retained categories and level of acceptable error (Culverhouse *et al.*, 2003; Fernandes *et al.*, 2009). Typically numerous categories are created, each containing objects of similar visual appearance. Then, from results of classifier performance evaluation, categories showing high cross-contamination are merged. A new performance evaluation is then conducted and the process is applied iteratively until acceptable levels of error rates are reached. For the case of semi-automated classification presented in this paper, a minimal learning set composed of a few dominant zooplankton categories may be sufficient.

Data analysis: algorithm selection. Several supervised learning algorithms are available in the "data analysis" section of version 1.2.6 of PkID (Table I).

Selection of measurement variables. It has been shown that irrelevant variables strongly affect performance and accuracy of supervised learning methods (Guyon and Elisseeff, 2003). In PkID, the user can include different combinations of variables and immediately test their suitability for a particular classification task using cross validation.

Data analysis and performance evaluations. Evaluation of classifier performance requires the examination of a CM, which is a contingency table crossing true (manually validated) and predicted (assigned by the classifier) identification of objects. Object counts on the matrix diagonal represent correct identifications and the sum of counts off the diagonal divided by the total number of objects gives the overall error rate (accuracy) of the classifier. Correct interpretation of the CM requires the examination of each category separately, including the rate of true positives (number of objects correctly predicted/total number actual objects) as well as false positives (number of objects falsely assigned to a category/total number of predicted objects).

There are three ways to build a CM, all available in PkID. The first, re-substitution CM, involves validation of the classification procedure on the same data set used to compute the classification functions. Re-substitution CMs systematically underestimate error rates and even give no errors when algorithms such as Random Forest are used. The second entails random partitioning of the initial data set into n equal fractions, $n - 1$ fractions being used to compute the classification model and one to validate it; this process is repeated m-times to fill up the CM. This procedure, called cross-validation, requires more computation time than the re-substitution CM, but usually gives better error evaluations. However, since data used still originate from the same data set, error levels usually remain underestimated. The third method (Dundar *et al.*, 2004) uses two equivalent and independent learning files describing the same categories with different objects. One is used to build the model and the other to validate it. This procedure, called test, gives good error estimation but requires twice the effort of learning file creation. Moreover, it cannot easily be applied during the learning file optimization procedure unless two learning file optimizations are conducted in parallel.

Table I: The different classifiers in Plankton Identifier (Gasparini, 2007) analyzed in the present study

Name	Short description	Reference
5-NN	k-nearest neighbor using heterogeneous value difference metric	Peters *et al.* (2002)
S-SVC linear	Support Vector Machine from LIBSVM library, using linear functions	Chang and Lin (2001)
S-SVC RBF	Support Vector Machine from LIBSVM library, using radial basis activation functions	Chang and Lin (2001)
Random Forest	Bagging, decision tree algorithm	Breiman (2001)
C 4.5	Decision tree algorithm	Quinlan (1993)
Multilayer Perceptron	Multilayer Perceptron neural network	Simpson *et al.* (1992)

Data management. Here we recommend appropriate practices for archiving ZooScan data and metadata.

ZooScan data include: (i) raw images of zooplankton samples or sub-samples, (ii) raw background images from the system's hardware, (iii) digital images of individual objects (i.e. vignettes), (iv) measurements made by ZooProcess software on individual objects, (v) classification results determined automatically or semi-automatically using PkID and (vi) computed abundances, biovolumes and biomass. ZooScan metadata include: (i) information about sampling and measured variables, (ii) image scan and grey level normalization, (iii) algorithm selection and measured variables, (iv) learning sets and (v) confusion matrices. One of the best practices in data management is to keep data and metadata together and, as much as possible, in the same file. While the latter two types of metadata generated by the ZooScan system come as complementary files, the first three types are included in either the PID files or the log files, along with the data.

Safeguarding ZooScan data and metadata requires that these be published in digital libraries such as National and/or World Data Centres (NODCs and/or WDCs) that have the capacity to archive and distribute images and their associated metadata. NODCs such as US-NODC in the USA, SISMER in France and BODC in the UK are designated by the International Oceanographic Data Exchange programme (IODE) of UNESCO Intergovernmental Oceanographic Commission (IOC), while World Data Centers (WDCs) such as WDC-MARE in Europe, WDC-Oceanography in the USA, Russia, China and Japan are designated by the International Council for Science (ICSU). Part of the data from the annual time series of zooplankton from the Bay of Villefranche-sur-mer, which is presented in the Results section, have been safeguarded at the WDC-MARE and available online by the PANGAEA information system (doi:10.1594/PANGAEA.724540). Access to raw images, log files and PID files is password protected, whereas low resolution images and key variables such as abundances and biovolumes of copepods and total plankton are publically available. With respect to ZooScan data, it is essential that different instruments are inter-calibrated and that software configurations are known.

RESULTS

We first present our results illustrating general characteristics of the ZooScan/ZooProcess/PkID system, including construction of learning sets, selection of classifier algorithms and validation of morphometric and biomass measurements. Then we present a brief case study from the Bay of Villefranche, in order to illustrate the sequential processes involved in sample and data analysis.

Learning set creation

After scanning, normalization, background subtraction and extraction of vignettes, the first step is to create a preliminary learning set or to use an existing learning set to classify ("predict") a small number of dominant groups. Our experiments to determine the optimal number of objects to sort into each category for construction of a learning set showed that sufficiently high recall (true positives) and low contamination (false positives) are achieved when approximately 200–300 objects are sorted per category (Fig. 3), with relatively small additional gains beyond this number. Therefore, we recommend sorting 200–300 vignettes per category of object to be identified.

Choice of the classifier and number of predicted categories

We compared the performance of six classifiers (see Table I) for numbers of categories of objects ranging from 35 to 5 categories (Fig. 4). The categories were balanced in number of vignettes (300 vignettes from each). Each time the number of categories was reduced, a new group of 300 vignettes from that newly combined category was selected for the learning set. The results show, first, that the Random Forest algorithm consistently had the highest recall and nearly always the lowest contamination, regardless of the number of categories predicted (Fig. 4). Support Vector Machine using linear functions had the second best performance. All further analyses were carried out with the Random Forest algorithm. The results also demonstrated that machine classifications were improved when a smaller number of categories were predicted (Fig. 4).

Morphometric and biomass measurements

Comparisons of ZooProcess automatic measurements of digitized zooplankton images with manual measurements of the same images revealed linear relationships between ZooProcess feret diameter and manually measured total length (Fig. 5). There was a greater scatter in the case of appendicularians than in other taxa and a slope <1.0, because the appendicularian tails were often curved, affecting the automated measurements of feret diameter but not manual measurements. Other automated measurements, including major and minor elliptical axes, were also correlated with manual

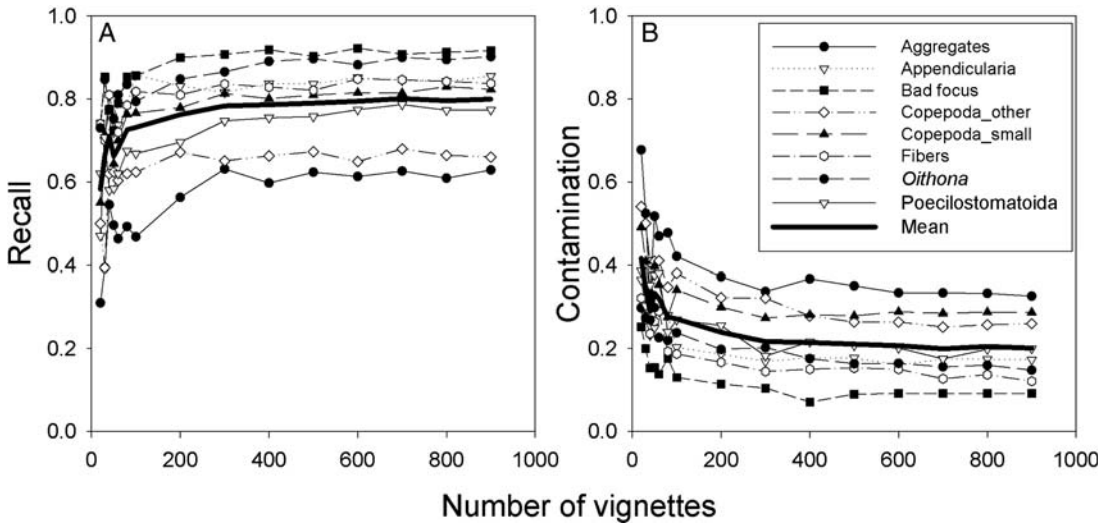


Fig. 3. Dependence of (**A**) recall (true positives) and (**B**) contamination (false positives) rate on the number of vignettes sorted for a learning set. Curves are illustrated for eight categories of organisms or objects, and the overall mean.

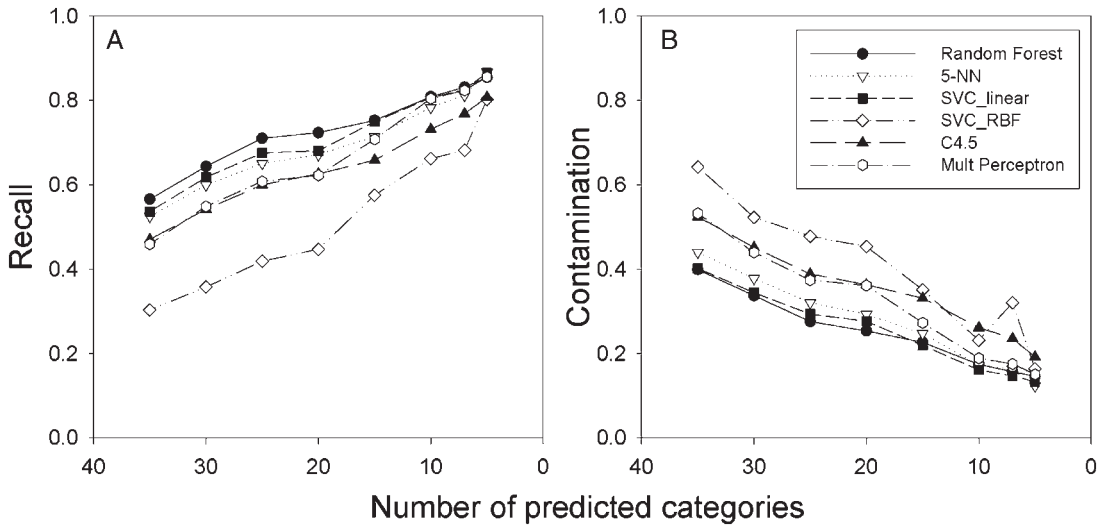


Fig. 4. Dependence of (**A**) recall (true positives) and (**B**) contamination (false positives) rate on the number of categories predicted by the classifier, using different classifier algorithms (see Table I).

length measurements (data not shown), although feret diameter typically showed the best relationship.

Comparison of automated measurements of surface area (as area excluded) with manual measurements of the same individuals was carried out for three taxa (copepods, euphausiids and chaetognaths: Fig. 6). In all cases, there was a linear relationship between manual and automated measurements. The automated measurements were somewhat higher for copepods and euphausiids, but lower for chaetognaths. These results suggest that automated measurements are consistent and reproducible, although their values may differ somewhat from manually determined values.

The relationships between C and N content and automated measurements of linear or areal dimensions were well described by power curves (Fig. 7). Much of the scatter in the relationships shown in Fig. 7 is attributable to the mixture of different species included in these analyses.

The exponents for C and N were similar to each other, implying relatively constant C:N ratios. In the case of both copepods and chaetognaths, the exponents relating C or N content to linear dimensions (feret diameter) were close to

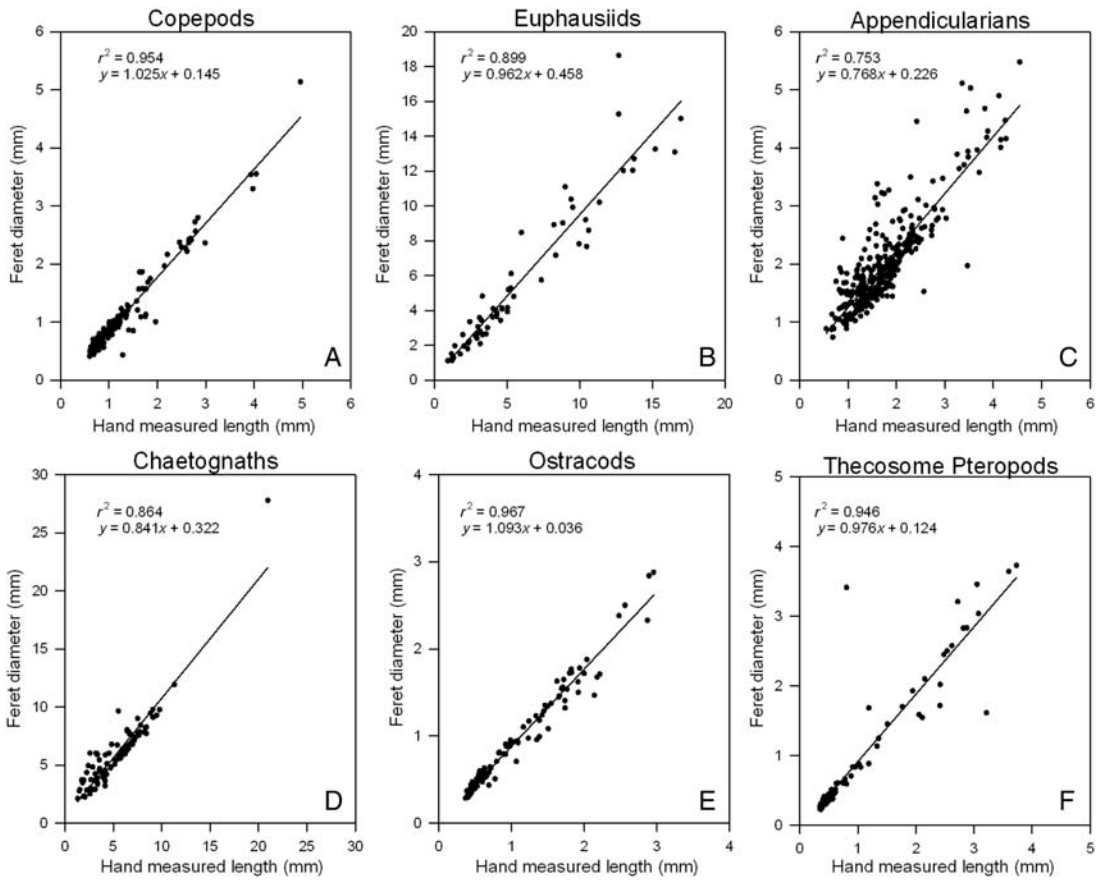


Fig. 5. Relationship between automated measurements of feret diameter and manual measurements of total length, for (**A**) copepods, (**B**) euphausiids, (**C**) appendicularians (tail length), (**D**) chaetognaths, (**E**) ostracods and (**F**) thecosome pteropods from the California Current.

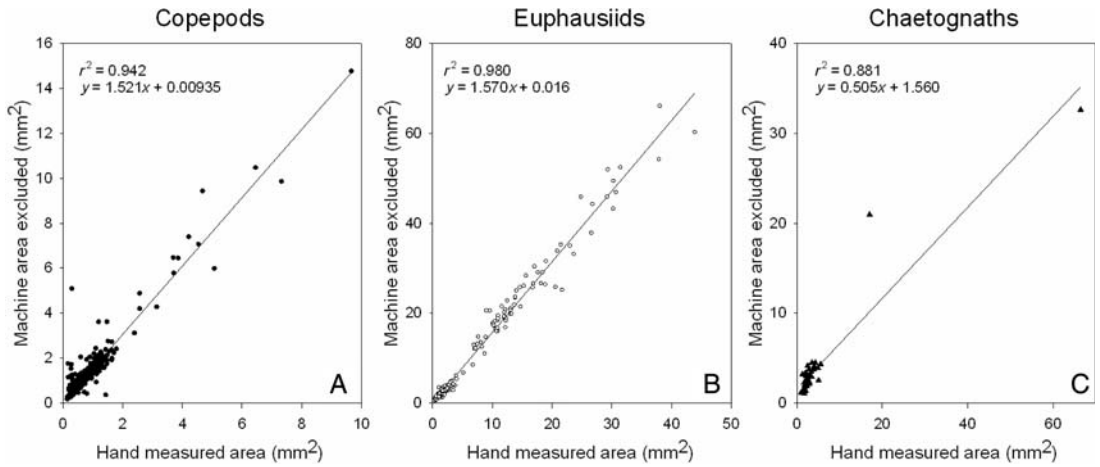


Fig. 6. Relationship between automated measurements of area excluded and manual measurements of projected area for (**A**) copepods, (**B**) euphausiids, (**C**) chaetognaths from the California Current.

3 and the exponents in relation to areal measurements (area excluded) were close to 2. However, for euphausiids, the exponents were close to 2 and 1, respectively. These differences are consistent with the changing body shapes with ontogeny of euphausiids, as the cephalothorax width and depth of euphausiids tends to increase in proportion

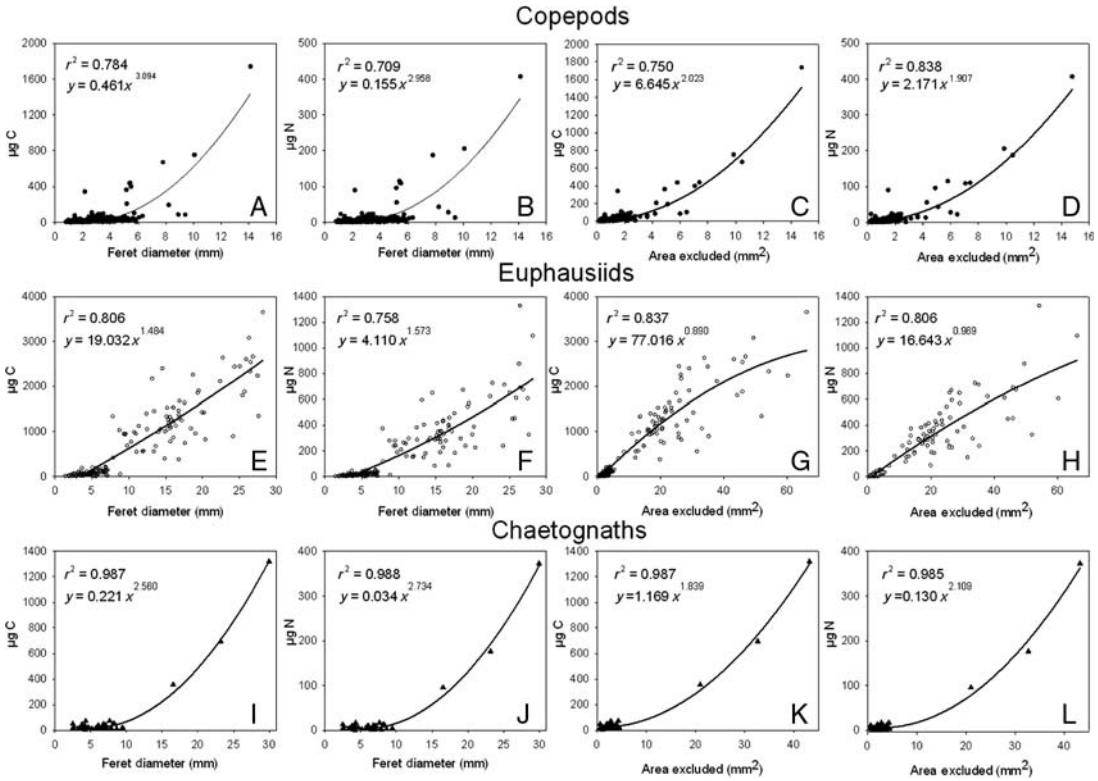


Fig. 7. Relationship between carbon and nitrogen content and automated measurements of body size for copepods (**A**–**D**), euphausiids (**E**–**H**) and chaetognaths (**I**–**L**) from the California Current. Carbon content plotted vs. linear dimensions (feret diameter; A, E, I) or areal dimensions (area excluded; C, G, K). Nitrogen content plotted vs. linear dimensions (feret diameter; B, F, J) or areal dimensions (area excluded; D, H, L).

to total body length through their development. In the case of copepods and euphausiids, area excluded was a slightly better predictor of C or N content than feret diameter, although for all three taxa the results indicate that automated measurements of either linear or areal dimensions of vignettes can be related in a useful manner to the biomass of these organisms.

Results from the Bay of Villefranche-sur-mer

Learning set optimization and application

To create our initial learning set for the Villefranche case study, we utilized a pre-existing learning set (see Methods) to predict 5000 objects from the Villefranche time series. We then manually validated the prediction into 30 categories (which took ca. 4 h). We included categories for "bad focus" objects, artifacts, bubbles and fibers. To improve the classifier, we then randomly selected a fixed number of vignettes drawn from each of these 30 categories from the Villefranche time series and created a new learning set. This second learning set was tested by cross validation in PkID using the Random Forest algorithm. Categories containing only a few objects with low accuracy of detection were not retained (they were left to contaminate the prediction). The accuracy of the prediction for this second learning set was much better than for the first iteration, and the subsequent manual validation was done faster.

As samples were analyzed from different seasons in the Villefranche time series, newly encountered taxonomic categories were added into the learning set when they became sufficiently numerous, provided that confusion with other dominant categories remained low. This occurred, for example, with cladocerans that bloomed only in autumn and were nearly absent during other time periods. Sometimes categories with relatively high contamination were maintained in the learning set because of their ecological value. For example, the *Limacina* category showed a 34.5% error rate (Table II). Nevertheless, it was maintained as a separate group because subsequent manual validation was rapid and the seasonal development of this taxonomic group was important. After the prediction results did not improve significantly with additional iterations, we considered the learning set satisfactory. It contained 14 zooplankton and 6 other categories (Table II) and was applied to the rest of the samples.

Table II: Confusion matrix for the 20 categories in the learning set used for machine classification of the 2007–2008 time series by the Random Forest algorithm

	Percentage in learning set	Aggregates	Aggregates_dark	Appendicularia	Bad focus	Bubbles	Chaetognatha	Cladocera	Copepoda_other	*Oithona*	Copepoda_small	Decapoda_large	Egg-like	Fibers	Medusae	Nectophores	Thaliacea	*Limacina*	Scratch	Pteropoda_other	Radiolaria	Total	Recall	1-Precision
Aggregates	7.2	**683**	27	80	67	0	0	77	99	53	134	8	16	26	9	37	1	1	12	0	25	1355	0.50	0.42
Aggregates_dark	5.7	3	**671**	0	0	39	0	18	38	0	4	0	27	0	0	0	0	193	0	8	43	1064	0.63	0.40
Appendicularia	7.7	85	0	**1243**	0	0	22	0	2	2	0	5	0	73	0	18	1	0	0	2	2	1455	0.85	0.20
Bad focus	7.5	44	0	5	**1230**	0	0	5	27	2	12	0	1	5	0	47	0	0	20	0	12	1410	0.87	0.10
Bubbles	4.5	2	30	0	0	**786**	0	2	0	0	0	0	7	0	0	0	0	10	0	0	3	840	0.94	0.06
Chaetognatha	2.3	0	0	51	0	0	**356**	0	0	0	0	1	0	20	0	0	0	0	0	0	0	428	0.83	0.11
Cladocera	6.2	27	19	2	1	0	0	**1028**	22	0	2	0	5	0	0	1	0	1	0	0	62	1170	0.88	0.16
Copepoda_other	11.3	74	32	4	8	0	0	14	**1608**	119	237	24	1	0	0	4	0	0	0	6	2	2133	0.75	0.24
Oithona	7.5	24	0	28	0	0	0	0	61	**1256**	38	0	0	9	0	0	0	0	0	4	0	1420	0.88	0.17
Copepoda_small	7.8	130	3	0	6	0	0	3	139	43	**1141**	0	0	0	0	0	0	0	0	0	0	1465	0.78	0.28
Decapoda_large	2.4	2	0	7	0	0	0	0	23	0	0	**410**	0	0	0	1	0	0	0	2	0	445	0.92	0.12
Egg-like	1.9	12	26	3	7	4	0	12	3	0	0	0	**259**	0	1	0	0	8	0	0	15	350	0.74	0.20
Fibers	6.4	10	0	84	0	0	21	0	7	14	0	0	0	**1034**	0	0	0	0	12	28	0	1210	0.85	0.13
Medusae	0.7	16	0	0	2	0	0	0	2	0	0	0	2	0	**100**	18	0	0	0	0	0	140	0.71	0.19
Nectophores	6.1	26	0	21	32	0	1	3	5	0	0	0	1	1	10	**1029**	20	0	1	0	5	1155	0.89	0.19
Thalicea	1.4	8	0	5	6	0	2	0	0	0	0	0	0	0	3	108	**123**	0	0	0	0	255	0.48	0.17
Limacina	4.5	2	267	0	0	8	0	9	0	0	0	0	4	0	0	0	0	**558**	0	0	7	855	0.65	0.28
Scratch	1.6	2	0	0	5	0	0	0	1	0	0	0	0	6	0	2	4	0	**285**	0	0	305	0.93	0.14
Pteropoda_other	2.3	9	0	2	1	0	0	3	55	23	23	20	0	11	0	0	0	0	0	**291**	2	440	0.66	0.15
Radiolaria	5.1	23	45	11	8	0	0	45	12	1	3	0	1	0	0	3	0	3	0	0	**800**	955	0.84	0.18
Total	100	1182	1120	1546	1373	837	402	1219	2124	1513	1594	468	324	1185	123	1268	149	774	330	341	978	**18 850**	**0.78**	**0.19**

Corresponding correct identifications (in bold) are in the diagonal.

Table III: Final categories used for classifying the 2007–2008 time series in the Bay of Villefranche, after initial machine classification, followed by manual validation, then manual subdivision into additional categories

Categories used for classifying the Villefranche time series	
Aggregates	Decapoda_large
Aggregates_dark	Decapoda_other
Algae	Echinodermata
Amphipoda	Egg like
Annelids	Fiber
Appendicularia	Fish
Bad focus	Heteropoda
Bivalves	Medusae_ephyrae
Bubbles	Medusae_other
Chaetognatha	Multiple
Cladocera	Nauplii
Copepoda_*Acartia*	Ostracoda
Copepoda_*Centropages*	Other
Copepoda_*Euterpina*	Pteropoda_*Limacina*
Copepoda_Harpacticoida	Pteropoda_other
Copepoda_*Oithona*	Radiolaria
Copepoda_Poecilostomatoida	Scratch
Copepoda_*Temora*	Siphonophora_eudoxid
Copepoda_other	Siphonophora_nectophores
Copepoda_other_multiples	Siphonophora_other
Copepoda_other_small	Thaliacea

The CM (Table II) of this learning set shows that most of the groups have a recall (rate of true positives) of about 80% and a contamination rate (false positives) smaller than 20%. Thus, this classifier performed moderately well, however not sufficiently accurately for ecological studies. The classifications of vignettes were then manually validated by sorting all vignettes into appropriate categories, a process which was facilitated by the prior machine classification. Many users may wish to use the classifications obtained at this point (i.e. in this case, to 14 categories).

Analysis of seasonal variations

For the present study, while manually validating image assignments to appropriate categories we chose to create additional categories beyond those in Table II. Several additional taxonomic categories could be reliably distinguished manually, although not by the machine classifiers. The result was a total of 42 categories, 33 of them zooplankton (see Table III), and all verified with essentially 100% accuracy. The use of the automated classifier greatly facilitated manual validation; it was simple to subsequently drag and drop their vignettes into the correct categories. Representative vignettes of some of the identified taxa may be seen in Fig. 8.

Our semi-automated analysis of an annual cycle of zooplankton variation in the Bay of Villefranche revealed pronounced seasonal variation in abundance, with substantial changes in the composition of the mesozooplankton (Fig. 9). Calanoid copepods were the numerically dominant organisms at all times of year, increasing from 75% before the peak of the bloom to a maximum of 95% at the peak and declining to 55% afterwards, as cladocerans, appendicularians and other taxa increased in relative importance (Figs 9 and 10). Poecilostome and oithonid copepods were abundant prior to the peak (16 and 8%, respectively). The community appears to be more diverse in summer.

In Fig. 10, we compare time series of individual major taxa both before and after manual validation of the sorted vignettes. While automated classification ("unvalidated") shows very good agreement with the manually validated time series for total copepods, this was not the case for other categories of organisms. For 4 of the 5 other groups of organisms in Fig. 10 (i.e. Appendicularia, chaetognaths, Cladocera, *Oithona*), the typical error was an overestimate, with moderate to high contamination with other organisms (false positives). For the sixth group (Decapoda), the usual error was underestimation (i.e. false negatives). This result underscores the importance of manual validation, even for classifiers that seem to have an overall acceptable error rate.

Sensitive ZooScan size measurements make it possible to readily reconstruct size spectra of organisms. For example, Fig. 11 illustrates the overall size spectrum for all copepods combined, as well as spectra for some of the dominant genera, including the smaller-bodied *Oithona*, intermediate-sized *Acartia* and larger-bodied *Centropages*.

DISCUSSION

The ZooScan–ZooProcess–PkID system is an end-to-end approach for digital imaging of preserved zooplankton, segmentation and feature extraction, and design and application of machine learning classifiers. The results from ZooScan analyses lead readily to numerical abundances as well as construction of size and biomass spectra. The calibration of each digital scan with reference to an OD standard makes it possible to directly compare images from ZooScans used in different laboratories.

Many existing zooplankton sampling programs have archived large numbers of plankton samples that have yet to be fully analyzed. Analysis of such samples is recognized as a high priority (Perry *et al.*, 2004), but this is an expensive task when carried out by trained microscopists. Complete analysis has awaited the development of machine learning or automated molecular methods.

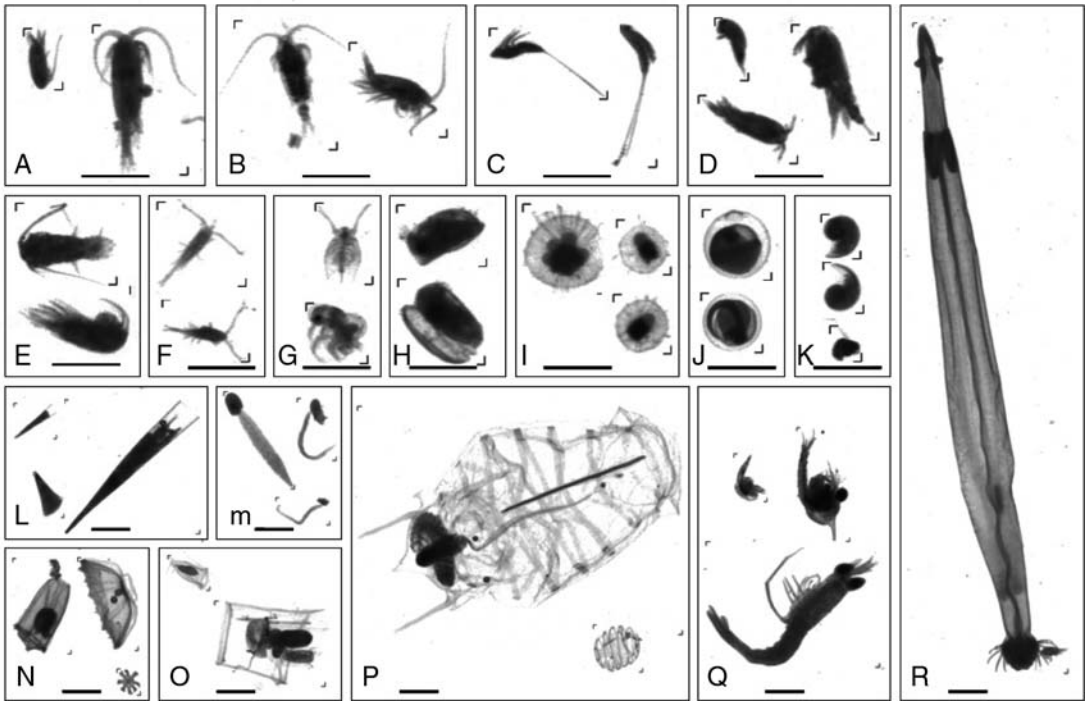


Fig. 8. Examples of vignettes of organisms from ZooScan analysis of the 2007–2008 time series in the Bay of Villefranche-sur-mer (scale bar = 1 mm). (**A**) Copepods, (**B**) *Centropages*, (**C**) Harpacticoida, (**D**) Poecilostomatoida, (**E**) *Temora*, (**F**) *Oithona*, (**G**) Cladocera, (**H**) Ostracoda, (**I**) Radiolaria, (**J**) eggs, (**K**) *Limacina*, (**L**) Pteropoda, (**M**) Appendicularia, (**N**) medusae, (**O**) Siphonophora, (**P**) Thaliacea, (**Q**) Decapoda, (**R**) Chaetognatha.

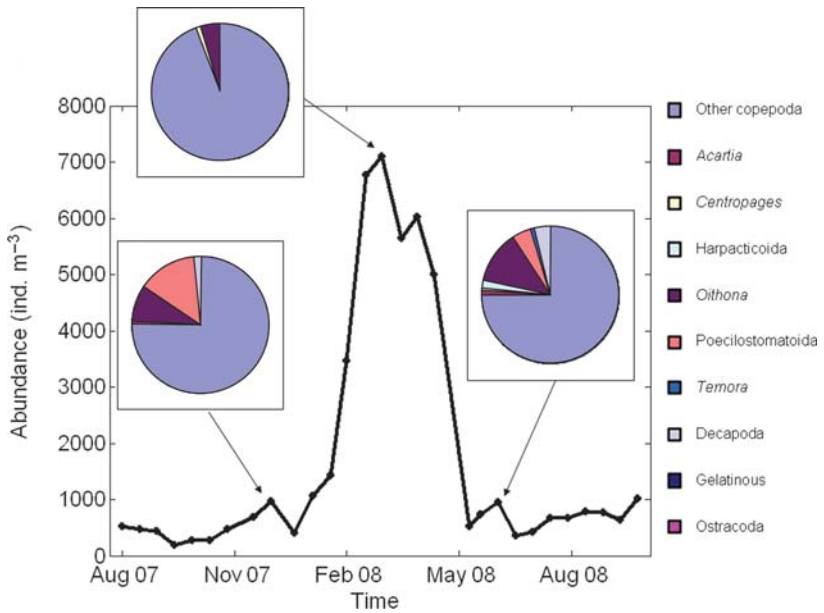


Fig. 9. Total abundance of mesozooplankton from 2007 to 2008, and the proportion of primary mesozooplankton categories (inset pie diagrams) before, during and after the 2008 spring bloom in the Bay of Villefranche. All classifications were validated manually.

Equally important for such sample collections is the archiving of digital representations of the samples, to facilitate permanent records of their contents as a complement to the conservation of the physical samples themselves. Such digital images permit automatic or semiautomatic image analysis, rapid measurement of

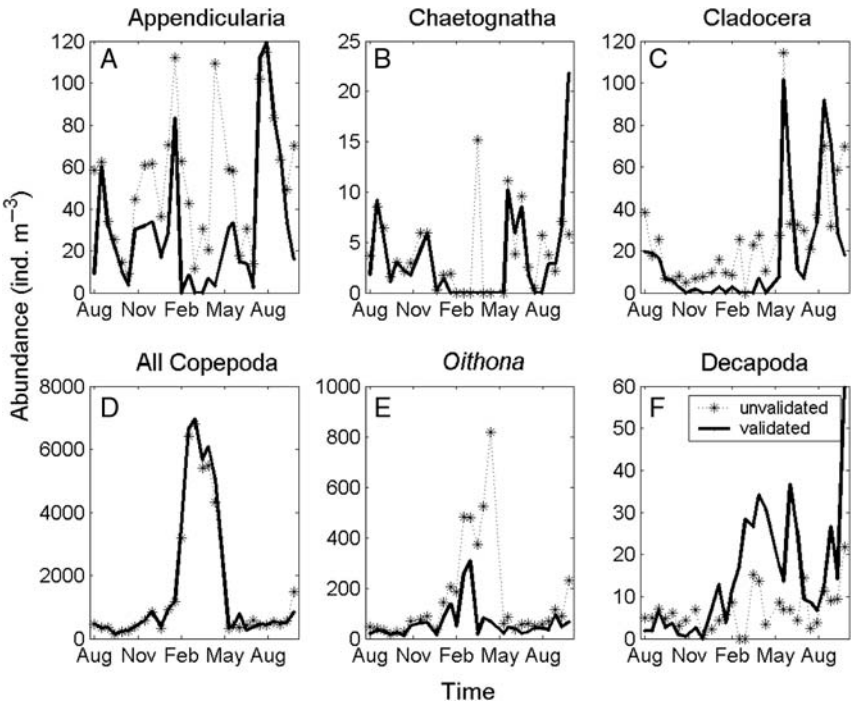


Fig. 10. Abundance of six major groups of mesozooplankton from 2007 to 2008 in the Bay of Villefranche. Time series of each category are illustrated as classified automatically by the Random Forest algorithm without manual validation (dotted line) and after manual validation (solid line).

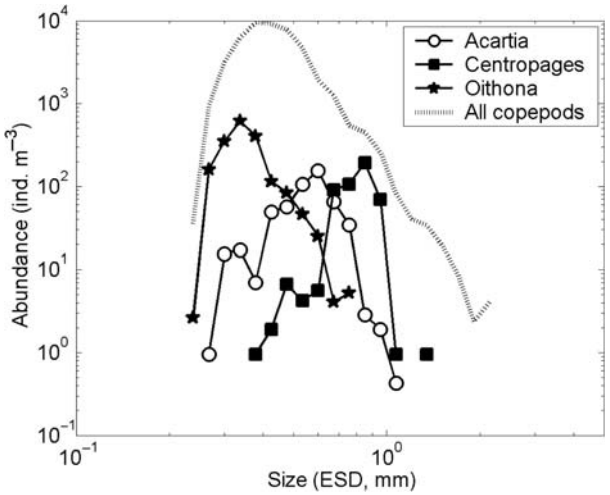


Fig. 11. Size spectrum of total copepods and three different copepod genera (*Acartia*, *Centropages* and *Oithona*), from all sampling dates in the Bay of Villefranche. All classifications were validated manually.

organisms and a permanent record of the sample contents that can be revisited in the future. The ZooScan system fulfils many of these objectives. It permits relatively rapid analysis of zooplankton samples combining automated classification and manual validation, digital archiving of images [for example the Villefranche time series ZooScan images are stored in PANGAEA®-Publishing Network for Geoscientific and Environmental Data] in databases accessible to the scientific community, standardization of images from different ZooScans allowing the construction of combined learning sets, non-destructive analysis so the samples can be used for other purposes and safe laboratory operation with aqueous samples.

Several classification algorithms have already been tested in the plankton recognition literature (Culverhouse *et al.*, 1996; Grosjean *et al.*, 2004; Blaschko *et al.*, 2005; Hu and Davis, 2005). The Random Forest algorithm seems to be one of the most promising (Grosjean *et al.*, 2004; Bell and Hopcroft, 2008). However, care is needed in design and testing of learning sets. Bell and Hopcroft (Bell and Hopcroft, 2008) built a learning set of 63 categories, but reduced this to two categories and correctly identified copepods 67.8% of the time and euphausiid eggs with even lower accuracy. Following automated classification in Irigoien *et al.* (Irigoien *et al.*, 2009), only four categories (two size categories of copepods, the euphausiids and mysids category, and chaetognaths) out of 17 had an acceptable error level. Hu and Davis (Hu and Davis, 2006) proposed use of a sequential dual classifier, using first a shape-based feature set and a neural network classifier, followed by a texture-based feature set and a support vector machine classifier.

Here we endorse a practical semi-automated method that may help biologists obtain taxonomically more detailed data sets with sufficient accuracy. Comparison between machine predicted and manually validated classifications showed that for dominant taxa such as copepods, automatic recognition was sufficiently accurate. However, for less abundant taxa such as appendicularians and chaetognaths, automatic recognition generally overestimated true abundances (but underestimated the abundance of decapods). Fully automated classification would have resulted in inaccurate descriptions of seasonal cycles of key zooplankton taxa and produced biased size spectra. Such biases result from contamination by other abundant groups, especially in winter/spring in the present study when copepods strongly dominate. There are no simple conversion factors that could be used here because the error is not constant through the seasonal cycle. The total time required for classification with manual validation is only slightly longer than with a fully automated classifier, because there is no need to construct a detailed learning set. Moreover, the results are significantly improved over an automated method alone. Proper design of the initial classifier makes the subsequent manual validation step proceed relatively quickly. The initial classifier will then facilitate subsequent subdivision into categories that are easily classified manually.

It is important to keep in mind when classifying the sample automatically that all types of objects that are encountered in a sample, including artifacts, must have a corresponding category in the learning set. If not, they will systematically contaminate other categories, leading to lower recognition performance.

Our results are encouraging for the estimation of zooplankton size and biomass spectra from ZooScan analyses. Many ecological traits (including metabolic rates, population abundance, growth rate and productivity, spatial habitat, trophic relationships) are correlated with body size (e.g. Gillooly *et al.*, 2002; Brown *et al.*, 2004). Hence, because body size captures so many aspects of ecosystem function, it can be used to synthesize a suite of co-varying traits into a single dimension (Woodward *et al.*, 2005). However, with some automated measurement methods for reconstructing size spectra from *in situ* measurements, all the *in situ* objects are treated as living plankton, though it has been shown that a significant proportion of objects can be marine snow (Heath *et al.*, 1999; González-Quirós and Checkley, 2006; Checkley *et al.*, 2008). The ZooScan imaging system provides an efficient means to reconstruct plankton size spectra from taxonomically well-characterized zooplankton samples. In addition, automated measurements of either linear or areal dimensions of digitized organisms can be related to their biomass, applied on a taxon-specific basis.

The classification method proposed here allows a relatively detailed taxonomic characterization of zooplankton samples and provides a practical compromise between the fully automatic but less accurate and the accurate manual classification of zooplankton. Useful size and biomass estimations may be rapidly obtained for ecologically oriented studies. Results from different ZooScan data sets can be combined using PANGAEA®'s data warehouse, thus encouraging cooperative, networked studies over broad geographic scales.

ACKNOWLEDGEMENTS

We thank Todd Langland and Corinne Desnos for assistance with measurements.

FUNDING

The ZooScan development was funded by CNRS/INSU ZOOPNEC program, by the UPMC and the EU FP6 programs SESAME and EUR-OCEANS under contracts GOCE-2006-036949 and 511106, respectively. C.G.-C. was supported by an EUR-OCEANS PhD fellowship. This work was stimulated by SCOR WG130 and was supported by the Mediterranean Scientific Commission (CIESM) Zooplankton Indicator program, and by the US National Science Foundation via the California Current Ecosystem LTER program.

REFERENCES

Abramoff, M. D., Magelhaes, P. J. and Ram, S. J. (2004) Image Processing with ImageJ. *Biophotonics Int.*, **11**, 36–42.

Bell, J. L. and Hopcroft, R. R. (2008) Assessment of ZooImage as a tool for the classification of zooplankton. *J. Plankton Res.*, **30**, 1351–1367.

Benfield *et al.* (2007) RAPID Research on Automated Plankton Identification. *Oceanography*, **20**, 172–187.

Berman, M. S. (1990) Enhanced zooplankton processing with image analysis technology. *Int. Counc. Explor. Sea Comm. Meet*, **1990/L:20**, 5.

Berube, D. and Jebrak, M. (1999) High precision boundary fractal analysis for shape characterization. *Comp. Geosci.*, **25**, 1059–1071.

Blaschko, M. B., Holness, G., Mattar, M. A. *et al.* (2005) Automatic in situ identification of plankton. *Seventh IEEE Workshops on Application of Computer Vision (WACV/MOTION'05)*, **1**, 79–86.

Bollmann, J., Quinn, P. S., Vela, M. *et al.* (2004) Automated particle analysis: calcareous microfossils. In Francus, P. (ed.), *Image Analysis, Sediments and Paleoenvironments*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 229–252.

Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Brown, J. H., Gillooly, J. F., Allen, A. P. *et al.* (2004) Toward a metabolic theory of ecology. *Ecology*, **85**, 1771–1789.

Chang, C.-C. and Lin, J. (2001) Training nu-support vector classifiers: theory and algorithms. *Neural. Comp.*, **13**, 2119–2147.

Checkley, D. M., Jr, Davis, R. E., Herman, A. W. *et al.* (2008) Assessing plankton and other particles in situ with the SOLOPC. *Limnol. Oceanogr.*, **53**, 2123–2126.

Culverhouse, P. F., Williams, R., Reguera, B. *et al.* (1996) Automatic categorisation of 23 species of Dinoflagellate by artificial neural network. *Mar. Ecol. Prog. Ser.*, **139**, 281–287.

Culverhouse, P. F., Williams, R., Reguera, B. *et al.* (2003) Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Mar. Ecol. Prog. Ser.*, **247**, 17–25.

Dundar, M., Fung, G., Bogoni, L. *et al.* (2004) A methodology for training and validating a CAD system and potential pitfalls. *CARS*, **July**, pp. 1010–1014.

Fernandes, J. A., Irigoien, X., Boyra, G. *et al.* (2009) Optimizing the number of classes in automated zooplankton classification. *J. Plankton Res.*, **31**, 19–29.

Gasparini, S. (2007) PLANKTON IDENTIFIER: a software for automatic recognition of planktonic organisms., http://www.obs-vlfr.fr/~gaspari/Plankton_Identifier/index.php.

Gillooly, J. F., Charnov, E. L., West, G. B. *et al.* (2002) Effects of size and temperature on developmental time. *Nature*, **417**, 70–73.

González-Quirós, R. and Checkley, D. M., Jr (2006) Occurrence of fragile particles inferred from optical plankton counters used in situ and to analyze net samples collected simultaneously. *J. Geophys. Res.*, **111**, C05S06. doi:10.1029/2005JC003084.

Gorsky, G., Guilbert, P. and Valenta, E. (1989) The autonomous image analyzer: enumeration, measurement and identification of marine phytoplankton. *Mar. Ecol. Prog. Ser.*, **58**, 133–142.

Grosjean, P., Picheral, M., Warembourg, C. *et al.* (2004) Enumeration, measurement, and identification of net zooplankton samples using the ZOOSCAN digital imaging system. *ICES J. Mar. Sci.*, **61**, 518–525.

Guyon, I. and Elisseeff, A. (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157–1182.

Heath, M. R., Dunn, J., Fraser, J. G. *et al.* (1999) Field calibration of the optical plankton counter with respect to *Calanus finmarchicus*. *Fish. Oceanogr.*, **8**(Suppl. 1), 13–24.

Herman, A. W. (1992) Design and calibration of a new optical plankton counter capable of sizing small zooplankton. *Deep-Sea Res.*, **39**, 395–415.

Hu, Q. and Davis, C. (2005) Automatic plankton image recognition with co-occurrence matrices and support vector machine. *Mar. Ecol. Prog. Ser.*, **295**, 21–31.

Hu, Q. and Davis, C. (2006) Accurate automatic quantification of taxa-specific plankton abundance using dual classification with correction. *Mar. Ecol. Prog. Ser.*, **306**, 51–61.

Irigoien, X., Fernandes, J. A., Grosjean, P. h. *et al.* (2009) Spring zooplankton distribution in the Bay of Biscay from 1998 to 2006 in relation with anchovy recruitment. *J. Plankton Res.*, **31**, 1–17.

Jeffries, H. P., Sherman, K., Maurer, R. *et al.* (1980) Computer processing of zooplankton samples. In Kennedy, V. (ed.), *Estuarine Perspectives*. Academic Press, New York, pp. 303–316.

Jefferies, H. P., Berman, M. S., Poularikas, A. D. *et al.* (1984) Automated sizing, counting and identification of zooplankton by pattern recognition. *Mar. Biol.*, **78**, 329–334.

Kennett, J. P. (1968) *Globorotalia truncatulinoides* as a paleo-oceanographic index. *Science*, **159**, 1461–1463.

Ortner, P. B., Cummings, S. R., Aftring, R. P. *et al.* (1979) Silhouette photography of oceanic zooplankton. *Nature*, **277**, 50–51.

Perry, R. I., Batchelder, H. P., Mackas, D. L. *et al.* (2004) Identifying global synchronies in marine zooplankton populations: Issues and opportunities. *ICES J. Mar. Sci.*, **61**, 445–456.

Peters, A., Hothorn, T. and Lausen, B. (2002) Ipred: improved predictors. *R News*, **2**, 33–36.

Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufman, San Francisco.

Rakotomalala, R. (2005) TANAGRA: une plate-forme d'expérimentation pour la fouille de données. *MODULAD*, **32**, 71–85.

Rasband, W. S. (2005) *ImageJ, U. S. National Institutes of Health*. Bethesda, MD, USA, http://rsb.info.nih.gov/ij/.

Rolke, M. and Lenz, J. (1984) Size structure analysis of zooplankton samples by means of an automated image analyzing system. *J. Plankton Res.*, **6**, 637–645.

Santos Filho, E., Sun, X., Picheral, M. *et al.* Implementation and evaluation of new features for zooplankton identification: Zooscan Case Study. *J. Plankton Res.*, (submitted for publication).

Simpson, R., Williams, R., Ellis, R. *et al.* (1992) Biological pattern recognition by neural networks. *Mar. Ecol. Prog. Ser.*, **79**, 303–308.

Sternberg, S. R. (1983) Biomedical image processing. *Computer*, **16**, 22–34.

Wiebe, P. H., Gallager, S. M., Davis, C. S. *et al.* (2004) Using a high-powered strobe light to increase the catch of Antarctic krill. *Mar. Biol.*, **144**, 493–502.

Woodward, G., Ebenman, B., Emmerson, M. *et al.* (2005) Body-size in ecological networks. *Trends Ecol. Evol.*, **20**, 402–409.

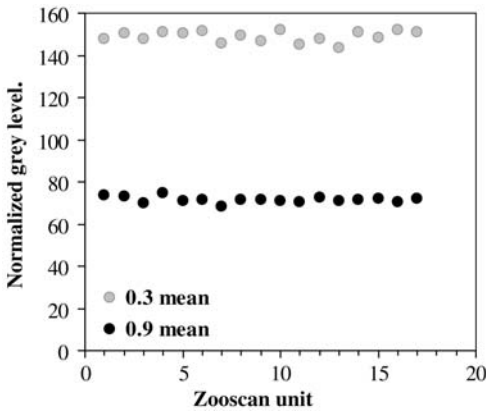
APPENDIX 1. GLOSSARY OF TERMINOLOGY USED

Accuracy	The proportion of the total number of classified objects that is correctly classified

Category	A taxon or group of taxa used in the learning set and confusion matrix
Classifier	A supervised learning algorithm applied to automated classification of objects. Classifiers are developed from a suite of characteristics extracted from each object
Confusion matrix (CM)	A matrix illustrating both predicted (from the classifier) and true classifications of all object categories
Contamination	See false positive rate
dat1.txt file	The PID file completed with the predicted and validated categories, if a validation has been performed
Error rate	Proportion of mispredicted organisms (to be manually corrected in order to obtain a fully validated data set)
False positive rate	The proportion of objects that is incorrectly classified as belonging to a category of interest; also called *contamination*
Learning set	A set of vignettes of organisms sorted in categories by an expert and used in a supervised learning model; also called training set
Log file	A text file containing details concerning the analyzed sample image
.pid file	A data file resulting from image analysis by ZooProcess. Includes the LOG file above the data section. Each identified object occupies one row, with all the variables extracted from that object in columns
Plankton Identifier	Software for automatic recognition of plankton
Precision	The proportion of predicted positive objects that was correctly assigned
Recall	see true positive rate
True positive rate	The proportion of objects that is correctly classified as belonging to a category of interest; also called *recall*
Validation	Manual sorting of vignettes to the correct category, following initial automated classification of vignettes
Variable	Attributes extracted from every detected object (see the list of extracted variables in Appendix 3)
Vignette	An image of a single detected object; also called *ROI* (region of interest)
ZooProcess	Software for image acquisition, treatment and analysis built for the ZooScan system

APPENDIX 2.

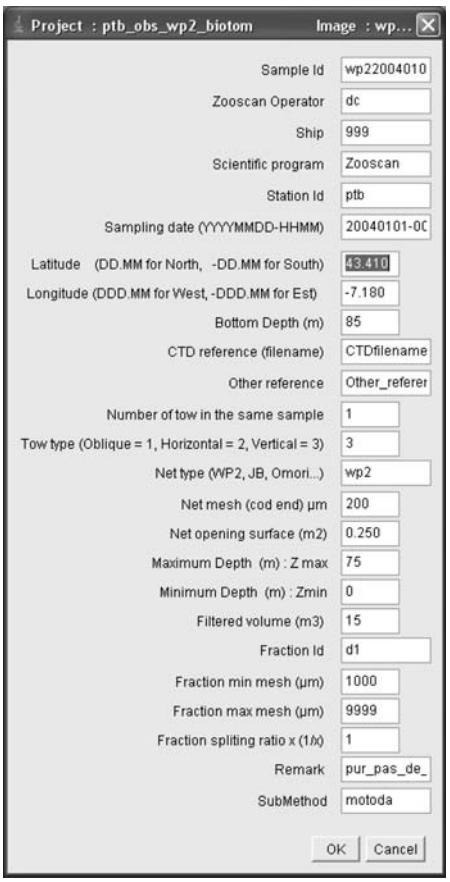
Grey level control of 17 ZooScan units using two different optical density calibration disks (OD 0.3 and 0.9). The variability among different instruments is lower than the variability within the same category of objects in one image (not shown).



APPENDIX 3. THE METADATA WINDOW IN ZOOPROCESS WITH THE DETAILS RECORDED WITH EACH SCAN

Sample Id	Incorporating sample date and time in filename assists with subsequent file retrieval
ZooScan operator	
Ship	
Scientific program	
Station Id	Name of sampling location
Sampling date	Date and time of sample collection
Latitude	Coordinates of the station, degrees·minutes
Longitude	Coordinates of the station, degrees·minutes
Bottom depth (m)	Bottom depth of the station
CTD reference filename	(Permits CTD data to be associated with plankton results)
Other reference	(Can be used to record the name of the collector)
Number of tows in the same sample	(Useful where samples are pooled)
Tow type	
Net type	
Net mesh (cod end)	Cod-end mesh size (if different from the net mesh size, this information should be recorded elsewhere in the remarks field)

Net opening surface (m^2)	Area of net mouth
Maximum depth (m)	Maximum depth reached by the net when collecting the sample
Minimum depth (m)	Minimum depth reached by the net when collecting the sample
Filtered volume (m^3)	Flowmeter readout (alternatively, derived from mouth area and tow length)
Fraction id	Identifies the fraction name when the sample has been sieved in different size categories (e.g. D1 for fraction >1 mm and D2 for fraction between 200 μm and 1 mm)
Fraction min mesh (μm)	Lower mesh size for sieving the sample
Fraction max mesh (μm)	Upper mesh size for sieving the sample
Fraction splitting ratio	Ratio of total sample volume to volume of aliquot scanned
Remarks	Free text field
Submethod	Method used to subsample the original sample



APPENDIX 4. LIST OF VARIABLES RECORDED IN THE DATA SECTION OF THE PID FILES

Standard ImageJ variables

Angle	Angle between the primary axis and a line parallel to the *x*-axis of the image
BX	X coordinate of the top left point of the smallest rectangle enclosing the object
BY	Y coordinate of the top left point of the smallest rectangle enclosing the object
Height	Height of the smallest rectangle enclosing the object
Width	Width of the smallest rectangle enclosing the object
X	X position of the center of gravity of the object
XM	X position of the center of gravity of the object's grey level
XMg5	X position of the center of gravity of the object, using a gamma value of 51
XStart	X coordinate of the top left point of the image
Y	Y position of the center of gravity of the object
YM	Y position of the center of gravity of the object's grey level
YMg5	Y position of the center of gravity of the object, using a gamma value of 51
YStart	Y coordinate of the top left point of the image

Other variables

Area	Surface area of the object in square pixels
Mean	Average grey value within the object; sum of the grey values of all pixels in the object divided by the number of pixels
StdDev	Standard deviation of the grey value used to generate the mean grey value
Mode	Modal grey value within the object
Min	Minimum grey value within the object (0 = black)
Max	Maximum grey value within the object (255 = white)
Slope	Slope of the grey level normalized cumulative histogram
Histcum1	grey level value at 25% of the normalized cumulative histogram of grey levels
Histcum2	grey level value at 50% of the normalized cumulative histogram of grey levels
Histcum3	grey level value at 75% of the normalized cumulative histogram of grey levels

Perim	The length of the outside boundary of the object
Major	Primary axis of the best fitting ellipse for the object
Minor	Secondary axis of the best fitting ellipse for the object
Circ	Circularity = (4 * Pi * Area) / $Perim^2$; a value of 1 indicates a perfect circle, a value approaching 0 indicates an increasingly elongated polygon
Feret	Maximum feret diameter, i.e. the longest distance between any two points along the object boundary
IntDen	Integrated density. The sum of the grey values of the pixels in the object (i.e. = Area*Mean)
Median	Median grey value within the object
Skew	Skewness of the histogram of grey level values
Kurt	Kurtosis of the histogram of grey level values
%area	Percentage of object's surface area that is comprised of holes, defined as the background grey level
Area_exc	Surface area of the object excluding holes, in square pixels (=Area*(1−(%area/100))
Mean_exc	Average grey value excluding holes within the object (= IntDen /Area_exc)
Fractal	Fractal dimension of object boundary (Berube and Jebrak, 1999)
Skelarea	Surface area of skeleton in pixels. In a binary image, the skeleton is obtained by repeatedly removing pixels from the edges of objects until they are reduced to the width of a single pixel

APPENDIX 5.

Examples of extracted vignettes and measurements. Vignettes of (a) an appendicularian, (b and c) copepods with antennules in different orientations and (c) a chaetognath. Feret diameter (grey line), major and minor elliptical axes (black lines) and the smallest rectangle enclosing the object are delineated on the leftmost image. Silhouettes illustrate the surface area of each organism when the contiguous regions of background pixels are excluded ("area excluded," center image) and the total surface area (rightmost image). Scale bar = 1 mm.