ICES Journal of Marine Science, 61: 518–525 (2004)
doi:10.1016/j.icesjms.2004.03.012

Enumeration, measurement, and identification of net zooplankton samples using the ZOOSCAN digital imaging system

Philippe Grosjean, Marc Picheral, Caroline Warembourg, and Gabriel Gorsky

Grosjean, P., Picheral, M., Warembourg, C., and Gorsky, G. 2004. Enumeration, measurement, and identification of net zooplankton samples using the ZOOSCAN digital imaging system. – ICES Journal of Marine Science, 61: 518–525.

Identifying and counting zooplankton are labour-intensive and time-consuming processes that are still performed manually. However, a new system, known as ZOOSCAN, has been designed for counting zooplankton net samples. We describe image-processing and the results of (semi)-automatic identification of taxa with various machine-learning methods. Each scan contains between 1500 and 2000 individuals <0.5 mm. We used two training sets of about 1000 objects each divided into 8 (simplified) and 29 groups (detailed), respectively. The new discriminant vector forest algorithm, which is one of the most efficient methods, discriminates between the organisms in the detailed training set with an accuracy of 75% at a speed of 2000 items per second. A supplementary algorithm tags objects that the method classified with low accuracy (suspect items), such that they could be checked by taxonomists. This complementary and interactive semi-automatic process combines both computer speed and the ability to detect variations in proportions and grey levels with the human skills to discriminate animals on the basis of small details, such as presence/absence or number of appendages. After this checking process, total accuracy increases to between 80% and 85%. We discuss the potential of the system as a standard for identification, enumeration, and size frequency distribution of net-collected zooplankton.

Keywords: image analysis, long-term series, machine-learning, net samples, pattern recognition, size spectrum, zooplankton.

P. Grosjean: Laboratoire d'écologie numérique, Université de Mons-Hainaut, Avenue Maistriau, 19, B-7000 Mons, Belgium. M. Picheral, C. Warembourg, and G. Gorsky: Laboratoire d'Océanographie de Villefranche (UMR 7093), Station Zoologique, Observatoire Océanologique, BP 28, F-06234 Villefranche sur mer Cedex, France. Correspondence to P. Grosjean: tel: +32 65 37 34 97; fax: +32 65 37 33 12; e-mail: philippe.grosjean@umh.ac.be.

Introduction

Zooplankton play a central role in aquatic ecosystems relative to phytoplankton and higher trophic levels (Banse, 1995). Yet understanding the influence of physical forcing on zooplankton population dynamics is still a gap in our knowledge. Likewise, the steady increase of demographic pressure and industrial activity, the overfishing of commercial resources, and the destruction of natural habitats all continue to stress marine and freshwater environments, thus raising the question: how do changes in the global environment affect the abundance, diversity, and production of plankton and nekton?

In the oceans, the role of zooplankton in the transformation and flux of organic matter is not fully understood, especially the production of, and interaction with, marine snow. Furthermore, most marine metazoans spend at least part of their lifetime (larval stages) in planktonic form – a stage that is ecologically critical for species survival and dispersion.

Zooplankton is sensitive and reactive to external perturbations (Lenz in Harris *et al.*, 2000) and is, consequently, an indicator of environmental change, i.e. of the possible impacts of phenomena such as global warming (Beaugrand *et al.*, 2002) or a rapid and exponential increase in CO_2 partial pressure in the atmosphere (Siegenthaler and Sarmiento, 1993). Various other perturbations (including anthropogenic: fisheries, chemical and organic pollutions, …) also influence the composition and structure of different trophic levels, but their effects are not known completely (Planque and Ibanez, 1997).

Despite these observations, the study of zooplankton populations is not a major priority in some large-scale

1054-3139/$30.00

programmes and international projects. The reasons are partly related to the difficulty in collecting data, leading to fragmented information that is hard to interpret. Sampling, manual identification, and counting of zooplankton are labour-intensive and time-consuming and limit the number of net tows that can be processed. Moreover, net-sampling always integrates spatial information. Acoustics and optical counters give insight into spatial distribution, but at the cost of taxonomic identification (for reviews of the various methods, see Foote and Stanton and Foote in Harris *et al.* (2000) and Wiebe and Benfield (2003)).

In such a context, the development of new technologies that can provide rapid, unbiased, and quantitative data about zooplankton is likely to significantly advance our knowledge. One aspect that deserves attention is the retrospective analysis of historical samples, particularly the study of time-series. More detailed analyses of such series are likely to provide a better interpretation of long-term changes in ecosystems. Indeed, we should focus on series as old as possible to avoid the "syndrome of baseline shift" formulated by Pauly (1995) (but see also Myers, 2000). This implies dealing with historical samples, and consequently new technologies developed must remain compatible with previous sampling techniques.

Image analysis has been considered a potential alternative to traditional manual treatment of plankton samples (Jefferies *et al.*, 1984; Rolke and Lenz, 1984; Gorsky *et al.*, 1989; Steidinger *et al.*, 1990; Tang *et al.*, 1998), but a marriage between optical systems and software development has never matured. One reason is the difficulty in identifying and measuring flexible objects with highly variable shapes. Zooplankton can rest in lateral or ventral positions; extensions (e.g. spines, antennae, appendages) from the body are often in different planes; individuals can overlap; or individuals may be damaged. In recent years, the steady increase in the power of computers, the development of faster and more accurate digital acquisition hardware, and the progress made in machine-learning techniques used to analyse such data enable us to reconsider the problem today.

We have designed a new system called ZOOSCAN. Here we describe the zooplankton image-processing and the (semi)-automatic recognition system using various machine-learning methods. We discuss the system's potential as a standard for obtaining identification, enumeration, and size frequency distribution of net-collected zooplankton samples.

Digitizing zooplankton samples

ZOOSCAN (Gorsky and Grosjean, 2003; but see also http://www.zooscan.com) permits rapid and complete analysis of preserved zooplankton samples and stores the data in digital form (allowing easy sharing and retrieval of the information; Grassle, 2000). The sample, or the subsample, is poured directly into the scanning cell. Any overlapping organisms are manually separated before the sample is digitized, and both introduction and recovery of the sample are simple and rapid. This process takes about 15 min. The instrument is not illustrated here because of an ongoing patenting process.

The samples used in this study come from a 50-year series sampled weekly at the permanent station off Villefranche sur mer (point B, see http://www.obs-vlfr.fr/Rade) and are collected with vertical tows from 60 m to the surface with a WP-2 net (200-μm mesh size), (UNESCO, 1968).

Image quality and image-processing

Samples are digitized with 2400-dpi resolution and the resulting images are 17 500 × 7000 pixels in size. The resulting quality (Figure 1) is suitable for taking morphometric measurements and for classifying species, genera, or families.

Pixel size is measured as 10.58 μm, with a standard deviation/mean of 0.28%. Thus, distortions and variations are negligible. Such a resolution is appropriate for mesozooplankton analysis. A standard image with 16-bit grey level requires approximately 250 Mb and can be handled by recently acquired PCs. Figures 2–4 illustrate the various steps in processing the picture.

Sample size

Both sampling and analysis processes should be optimized in relation to the accuracy of the results. Not all species appear with the same probability in a plankton community. If rare species have to be considered, the sample size must be large enough to include at least a few tens of individuals of each taxon. Individual size-spectra by taxa can be computed with ZOOSCAN because each individual is measured. Hence, a sample of 2000 individuals enables ca. 100 individuals to be measured for taxa that occur in as little as 5% of the whole population. This is a suitable starting point for standard treatment of zooplankton samples.

If a whole sample is considered, large animals can be present at very low density, compared to more abundant, small animals (in an equivalent biomass in the same water volume). Consequently, obtaining a sufficient number of larger animals can result in digitization of an unnecessarily large number of small ones. We have found that the best strategy is to divide the sample by gently sieving into two fractions, one containing animals < 0.5 mm, the other of animals > 0.5 mm. We have experimentally determined that around 1500–2000 individuals for the small fraction is a reliable number to allow separation of specimens in the scanning cell of a 15×10 cm area. For the large fraction, the number of individuals to scan ranges between 500 and 800. Here we deal only with the small fraction. ZOOSCAN performs well with the large fraction, but in this article we focus on the smaller size classes because they are digitized with a lower resolution (and thus are more difficult to identify).

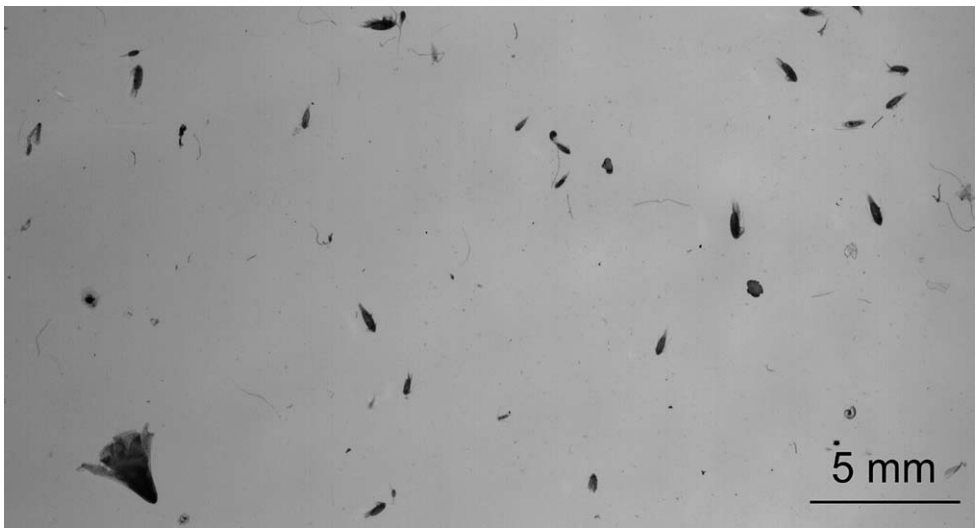

Figure 1. Part of a raw digitized image at 2400 dpi (resolution of the displayed portion is 3000×1600 pixels). The whole scanned area contains between 1000 and 2000 individuals.

Subsampling

What is the optimal number of replicates or aliquots containing around 2000 individuals of the small fraction that can be scanned with ZOOSCAN for accuracy of the measurements to be increased? To answer this question, we used a Motoda splitter to subdivide a sample into 16 aliquots containing about 2000 individuals each. The 16 aliquots were scanned and analysed using ZOOSCAN. The data from these replicates were pooled for analysis. No attempt was made to identify animals. All images pooled were considered to originate from a single unique population (the zooplankton community).

As shown in Figure 5, the total number of blobs (separate regions in the picture, detected as objects by the image analysis) identified in the subsamples is not constant. The total number ranges from 1457 to 3259, which is a twofold increase between the two extremes. The mean number of blobs identified in the 16 replicates is 2183. Obviously, a single subsample does not allow accurate estimation of the total number of individuals, and this conclusion will probably apply to separate taxa, too.

To determine the 95% confidence interval on the total number of blobs identified, we made a bootstrap analysis of these replicates, pooled by 1, 2, 3, …, 16. The results are

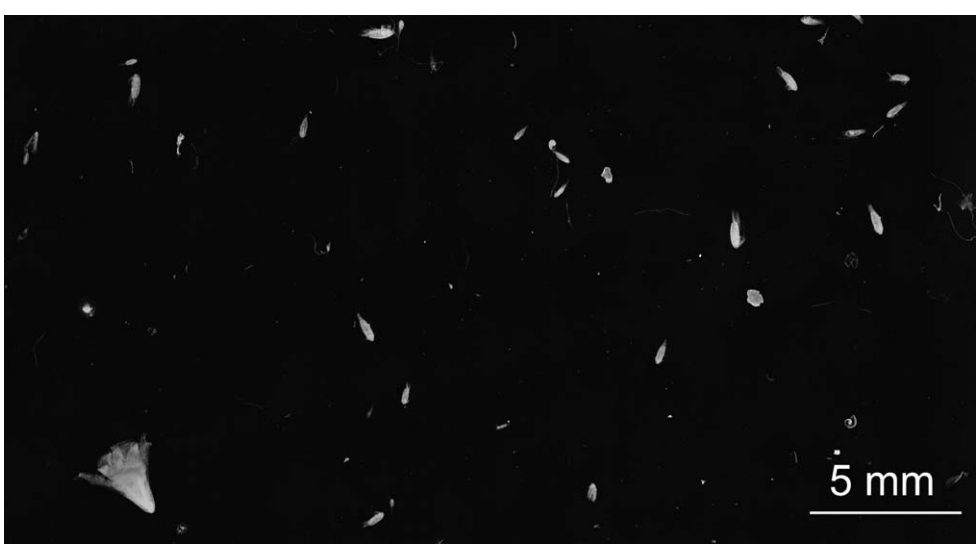

Figure 2. Same as Figure 1, but negative after background elimination and image enhancement. This picture is used for image analysis.

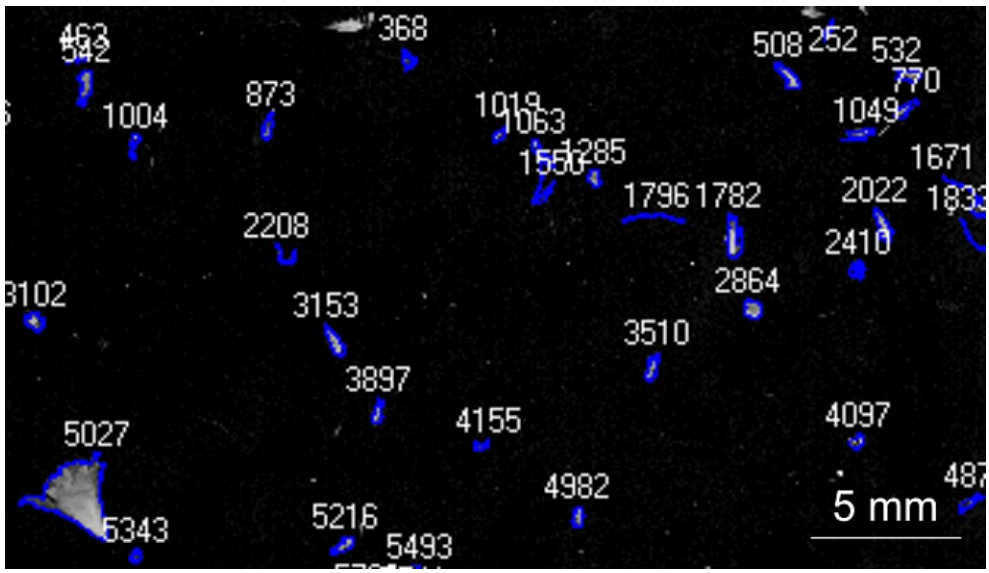

Figure 3. The same area as Figures 1 and 2, but objects are detected, contoured, and labelled by the image analysis.

presented in Figure 6. The accuracy of the mean number of blobs is improved from a single subsample to duplicates (from 2200±980 to 2200±690), but the gain decreases exponentially with the number of replicates. Three or four replicates (2200±570 and 2200±480, respectively) appear to be a good trade-off between accuracy of the mean number of blobs and sampling effort. Duplicates are probably also acceptable when the number of samples is high (high frequency time-series). It should be noted, however, that these conclusions are drawn for samples divided with the Motoda splitter. They could be different with other splitting devices (Youngbluth, 1980). Longhurst and Seibert (1967) found that use of the Folsom plankton splitter is dependent on the skill of the operator, and this may also be the case for the Motoda splitter.

Object recognition

In assessing the potential of discriminating various taxa using automatic analysis of the images, we used a training set of about 1000 objects from 14 different scans (various samples in different seasons and years taken at a permanent station with a vertical tow from 60 to 0 m using a WP-2

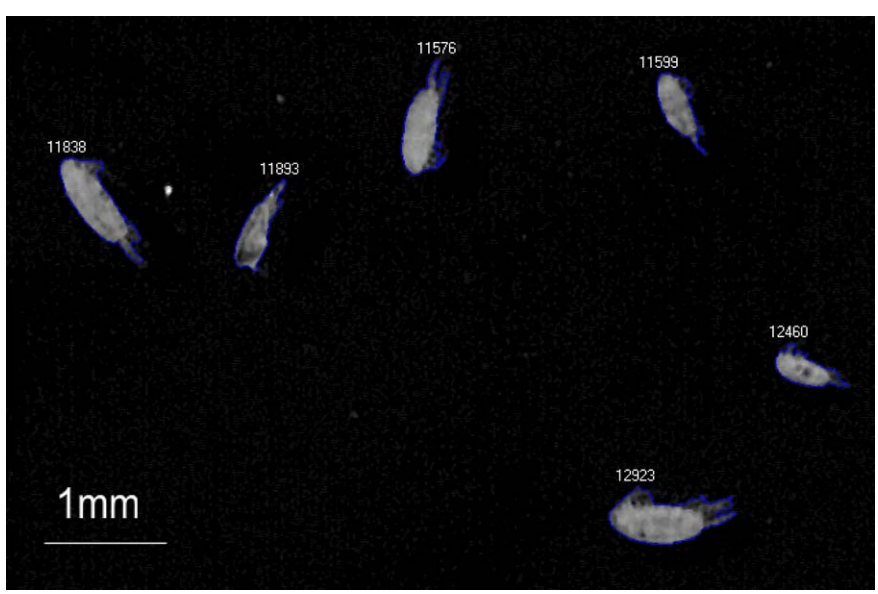

Figure 4. Examples, at actual resolution, of contoured objects (blobs) during image analysis: copepods.

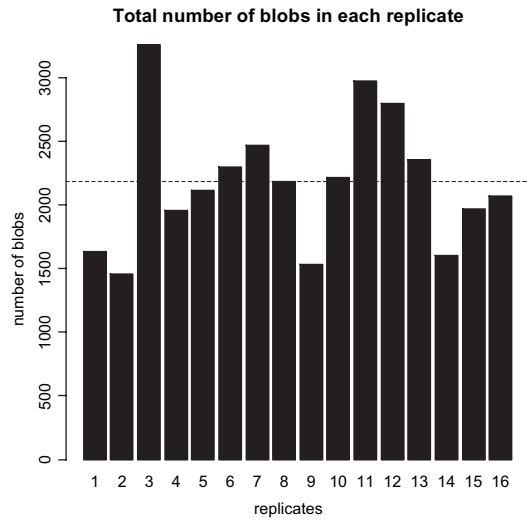

Figure 5. Total number of blobs identified in each replicate. The average number of blobs (dotted line) is: 2183±515 (mean ± standard deviation).

net). The objects, selected so that the largest diversity in the training set could be obtained, were manually classified into eight groups (see Figure 7), and then into a more detailed training set of 29 groups (Figure 8. In this more detailed training set, we used a slightly larger number of individuals (1127 instead of 1035 items in the 8-group set). With the 92 additional individuals, each category contains enough individuals in the training set, that is, at least 8−10 items per taxon.

Different classification algorithms were tested with both training sets [linear, quadratic, mixture, and flexible discriminant analysis (Hastie *et al.*, 1994); k-nearest neighbours; learning vector quantization (Tang *et al.*, 1998); tree and recursive partitioning methods, including ensembles of bootstrapped tree, such as bagging and

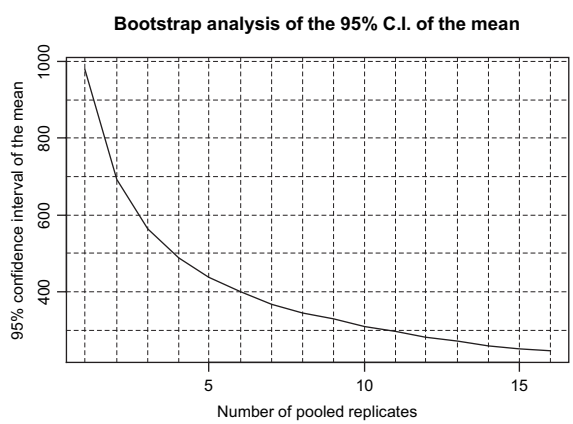

Figure 6. Confidence interval of the mean number of images identified in function of the number of replicated scans (10 000 bootstraps on the 16 samples for each number of replicates).

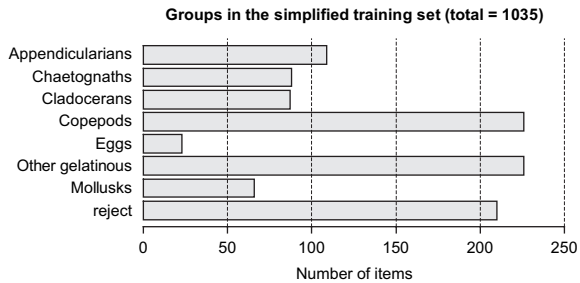

Figure 7. Composition of the simplified training set and number of individuals in each group. The "reject" group contains all objects that do not belong to the seven others (marine snow, phytoplankton, and some other taxa of zooplankton that are present in low proportions, that is, less than 0.5% in the series).

random forest (Breiman, 2001); support vector machine (Meyer, 2001); feed-forward, single hidden layer neural network (Simpson *et al.*, 1992)]. We also tested methods in which two or more different algorithms are combined, such as double bagging with linear discriminant analysis or k-nearest neighbours (Peters *et al.*, 2002), and also discriminant vector forest, a new combined method that we have set up specifically for analysing ZOOSCAN data and that mixes linear discriminant analysis, learning vector quantization, and random forest (see Table 1). It would be inappropriate here to detail the algorithms of all these methods. Readers should consult the references cited.

Basically, all these techniques search for rules for predicting the class of an object based on all the measurements made. These are computed using the training set where the class of these items is known (because it was manually identified by the operator during the training stage). At the end of this training stage, these various methods are capable, with varying degrees of accuracy, to predict the class of unknown objects, providing measurements only, and this is applied on the whole digitized series. It is this degree of accuracy that quantifies the overall quality of a given algorithm in a particular application. This is used as a criterion for deciding which method is best suited for identifying automatically digitized zooplankton measured with ZOOSCAN. Each object is described by 27 parameters: size (length, width, …), shape (elongation, compactness, …), moments (first and second order), and grey-levels distributions (minimal, maximal, mean grey values, …).

It appears that the automatic recognition of zooplankton is a difficult task for all of these methods, because the intragroup variability is large and because the training set is probably contaminated by errors made during manual recognition of these objects by experts (Culverhouse *et al.*, 2003).

We have also observed limitations of several methods (some discriminant analyses, as well as the neural network) when the number of taxa increases and/or the number of items in each taxon decreases (detailed training set). In

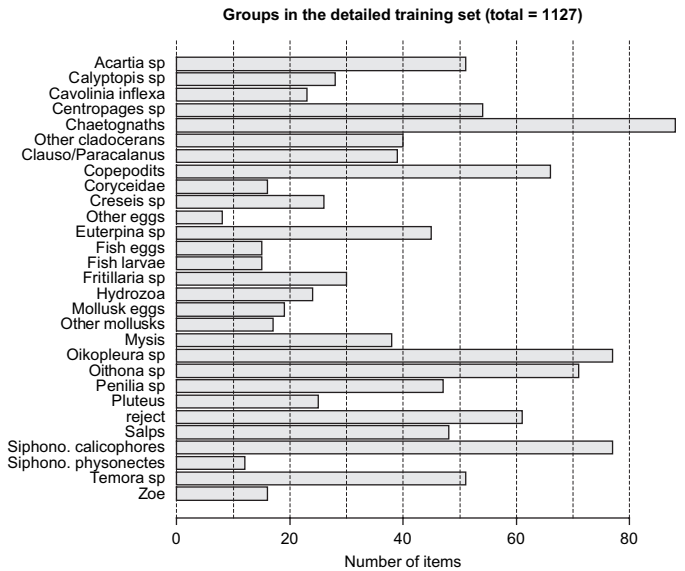

Figure 8. Composition of the detailed training set and number of items in each group. The "reject" group is the same as in Figure 7.

these cases, the learning phase either required an extremely large amount of memory (the computer used was a Pentium IV 1.6 Ghz with 1-Gb RAM memory) or took too long (the process was stopped after 3 h of calculation). Other methods appear much more robust to the number of taxa simultaneously recognized (random forest and discriminant vector forest performed almost equally well with the simplified and detailed training sets, in terms of both accuracy and speed). They are among the best methods in each case. It is probably possible to develop even more detailed training sets with such methods.

Combined methods appear more efficient in this context, particularly double bagging with linear discriminant analysis and the new discriminant vector forest method. With the latter, we reached an accuracy level of almost 75% with the detailed training set.

A supplementary algorithm tags objects that are classified with low accuracy by the discriminant vector forest method

Table 1. Comparison of various recognition methods with both the simplified and the detailed training set (100 replicates with random 2/3 training set and 1/3 test set). Accuracy is the mean total recognition success as evaluated on the 100 replicates of the test set only. Speed is the time required to perform one whole cycle (training + test). The symbol "−" means that the method did not succeed in making the training set: either the 1-Gb RAM memory was exhausted or the operation took more than 3 h.

	Simplified (8 groups)		Detailed (29 groups)	
Method	Accuracy (%)	Speed (s)	Accuracy (%)	Speed (s)
Linear discriminant analysis	76.8	0.1	70.6	0.2
Quadratic discriminant analysis	82.9	0.2	−	−
Mixture discriminant analysis	81.4	2.4	−	−
Flexible discriminant analysis	77.6	1.8	72.7	6.0
k-nearest neighbour analysis	77.2	0.1	60.4	0.1
Learning vector quantization	76.6	0.3	60.0	0.4
Tree method	72.0	0.5	55.1	2.3
Recursive partitioning	72.8	1.2	57.7	3.1
Bagging (bootstrap on trees)	81.7	3.6	69.8	8.0
Double bagging with LDA	85.0	10.3	74.6	25.5
Double bagging with k-n.n.	81.9	8.9	70.1	13.8
Random forest	83.9	1.7	73.4	2.5
Support vector machine	68.5	1.2	47.8	1.9
Neural network	73.9	25.8	−	−
Discriminant vector forest	83.6	2.7	74.4	4.0

Table 2. Percentage of suspect items with the detailed training set (29 groups) and the discriminant vector forest method as a function of severity. The accuracy level increases after re-identification of these suspect items by the specialist.

Severity parameter	% Suspect items	Accuracy after re-identification of suspect items (%)
0.00	0.0	74.4
0.25	7.0	79.8
0.5	13.0	83.8
0.75	16.8	85.6
1.00	22.1	87.2
1.25	27.1	89.6
1.50	30.6	89.9
1.75	35.4	90.9
2.00	40.7	94.1

(suspect items). Using this algorithm, we designed a complementary semi-automatic and interactive analysis where specialists can check and modify the choice made by the computer for these suspect items. It is possible to adjust the severity with which items are tagged as suspect. With the severity parameter value between 0.25 and 0.75, up to 7−17% of items are tagged (see Table 2). After the checking process, total accuracy increases to 80−85%.

Using a 1.6-Ghz Pentium IV computer with 1-Gb RAM memory under Windows XP, our algorithm is capable of recognizing 10 000 items in less than 5 s. This rate is fast enough for a routine processing of large series containing several million items requiring identification. The speed criterion is one key aspect of such a system, although it was rarely mentioned in early studies of this nature. There is not much advantage in using a computer-based recognition system if it is so slow that it does not speed up sample treatment significantly.

Discussion

The study of zooplankton is traditionally conducted on preserved samples from net tows. Various taxa (order, family, genera, or species) are enumerated in each sample. The degree of accuracy in the identification is experimenter-dependent. Hence, systematic bias is introduced when different specialists measure different fractions of the same series. Moreover, the whole analysis must be done again when a fine separation of taxa is required. Finally, the size of the organisms is rarely recorded simultaneously with identification. Consequently, the sample treatment is usually not optimal and represents only a small fraction of the information contained in the sample.

The capability of ZOOSCAN in digitizing and (semi)-automatically classifying zooplankton to taxa was explored using images from a multi-annual series of WP-2 net tows in the northwest Mediterranean Sea (data and analysis of the whole series will be published elsewhere). A training set was developed and manually identified by experts.

1. It appears that the quality of the images is high enough to discriminate among at least 29 different taxa, by examining the organisms on screen and also by using custom-made software, with a reasonable level of accuracy (about 75% for the automated computer method).
2. It appears that the complementary semi-automatic method that combines both computer and human skills increases the recognition level to 85% or even more.
3. It appears that gain in speed (one scan is acquired and treated in less than 20 min), ease of use (digital images are easier to analyse than biological material under the microscope, and they can be shared, possibly through the Internet), and quantity of information (both individual size and nature of the particles is determined) mean that use of computerized systems based on image analysis is more advantageous in processing net zooplankton samples in comparison to manual processing.

The object classification algorithm must be able to discriminate between a fairly large number of taxa (typically, a few tens) with an accuracy level between 75% and 90%. The computerized system must also be much faster than manual handling of the same samples. Those two aspects, accuracy of recognition level with large numbers of taxa (which require high-resolution imaging of the organisms) and speed, were previously the major impediments to the development of automated methods. The current version of the ZOOSCAN system matches these requirements, thanks to a new combined algorithm known as discriminant vector forest.

A complete automated system is not, and should not be, the solution for zooplankton samples that require identification to the species level in species-rich collections. A certain level of control by the biologist is necessary. In this way, one can combine human skill to discriminate animals on basis of small details (presence/absence or number of appendages, for instance) with the computer potentials to better analyse volumes and grey-levels distribution. The automated approach simplifies and speeds up the classification process by computer recognition. The complementary semi-automated approach that we propose here uses both the computer and human classification for the 10−15% most difficult specimens in the samples in order to increase the overall recognition accuracy up to 85%.

Acknowledgements

We are grateful to Dr M. Youngbluth for constructive comments on an earlier version of the manuscript. We thank the local taxonomists J. C. Braconnot, C. Carré, and

J.-C. Molinero for help in identifying zooplankton of the training set.

References

Banse, K. 1995. Zooplankton: pivotal role in the control of ocean production. ICES Journal of Marine Science, 52: 265–277.

Beaugrand, G., Reid, P. C., Ibanez, F., Lindley, J. A., and Edwards, M. 2002. Reorganization of North Atlantic marine copepod biodiversity and climate. Science, 296: 1692–1694.

Breiman, L. 2001. Random forests. Machine Learning, 45: 5–32.

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

Gorsky, G., and Grosjean, Ph. 2003. Qualitative and quantitative assessment of zooplankton samples. GLOBEC International Newsletter, 9: 5.

Gorsky, G., Guilbert, P., and Valenta, E. 1989. The autonomous image analyzer: enumeration, measurement and identification of marine phytoplankton. Marine Ecology Progress Series, 58: 133–142.

Grassle, J. F. 2000. The Ocean Biogeographic Information System (OBIS): an on-line, worldwide atlas for accessing, modelling and mapping marine biological data in a multidimensional geographic context. Oceanography, 13: 5–7.

Harris, R. P., Wiebe, P. H., Lenz, J., Skjoldal, H. R., and Huntley, M. 2000. ICES Zooplankton Methodology Manual. Academic Press, San Diego, CA.

Hastie, T., Tibshirani, R., and Buja, A. 1994. Flexible discriminant analysis by optimal scoring. Journal of the American Statistical Association, 89: 1255–1270.

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

Longhurst, A. R., and Seibert, D. L. R. 1967. Skill in the use of Folsom's plankton sample splitter. Limnology and Oceanography, 12: 334–335.

Meyer, D. 2001. Support vector machine. R News, 1: 23–26.

Myers, R. A. 2000. The synthesis of dynamic and historical data on marine populations and communities, putting dynamics into the Ocean Biogeographical Information System (OBIS). Oceanography, 13: 56–59.

Pauly, D. 1995. Anecdotes and the shifting baseline syndrome of fisheries. Trends in Ecology and Evolution, 10.

Planque, B., and Ibanez, F. 1997. Long-term time series in *Calanus finmarchicus* abundance: a question of space? Oceanologica Acta, 20: 159–164.

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

Rolke, M., and Lenz, J. 1984. Size structure analysis of zooplankton samples by means of an automated image analyzing system. Journal of Plankton Research, 6: 637–645.

Siegenthaler, U., and Sarmiento, J. L. 1993. Atmospheric carbon dioxide and the ocean. Nature, 365: 119–125.

Simpson, R., Williams, R., Ellis, R., and Culverhouse, P. F. 1992. Biological pattern recognition by neural networks. Marine Ecological Progress Series, 79: 303–308.

Steidinger, K. A., Chase, C., Garrett, J., Mahmoudi, B., Roberts, B., Thomas, C., and Truby, E. 1990. The use of optical pattern recognition in dinoflagellate taxonomy. *In* Toxic Marine Phytoplankton. Ed. by E. Graneli, B. Sunstrom, L. Edler, and D. M. Anderson. Elsevier, Amsterdam.

Tang, X., Steward, W. K., Vincent, L., Huang, H., Marra, M., Gallager, S. M., and Davis, C. S. 1998. Automatic plankton image recognition. Artificial Intelligence Review, 12: 177–199.

UNESCO, 1968. Zooplankton Sampling. Paris. 174 pp.

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

Youngbluth, M. J. 1980. Daily, seasonal, and annual fluctuations among zooplankton populations in an unpolluted tropical embayment. Estuarine and Coastal Marine Science, 10: 265–287.