Nimis P. L., Vignes Lebbe R. (eds.)
Tools for Identifying Biodiversity: Progress and Problems – pp. 231-236.
ISBN 978-88-8303-295-0. EUT, 2010.

VeSTIS: A Versatile Semi-Automatic Taxon Identification System from Digital Images

Nikos Nikolaou, Pantelis Sampaziotis, Marilena Aplikioti,
Andreas Drakos, Ioannis Kirmitzoglou, Marina Argyrou,
Nikos Papamarkos, Vasilis J. Promponas

Abstract — In this work we present a flexible Open Source software platform for training classifiers capable of identifying the taxonomy of a specimen from digital images. We demonstrate the performance of our system in a pilot study, building a feed-forward artificial neural network to effectively classify five different species of marine annelid worms of the class Polychaeta. We also discuss on the extensibility of the system, and its potential uses either as a research tool or in assisting routine taxon identification procedures.

Index Terms — digital image analysis, open source, semi-automatic taxon identification.

1 Introduction

Automated taxon identification (ATI) can be defined as the process of automating the routine identification of specimens [1] through the exploitation of modern computer science technologies and domain knowledge. ATI methods are based on mathematical descriptors of morphological [1], [2], [3], behavioural [4] or genetic [5] characters. These data are used as input into pre-processing and analysis pipelines, which are most often based on statistical or machine learning methods. ATI procedures are quickly becoming a necessity in the effort to understand and monitor global biodiversity.

So far, several research efforts to deal with ATI from digital images have been

N. Nikolaou, I. Kirmitzoglou, V.J. Promponas are with the Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus. E-mail: nkleopas@gmail.com, bip6ki2@ucy.ac.cy, vprobon@ucy.ac.cy.
M. Aplikioti, A. Drakos, M. Argyrou are with the Department of Fisheries and Marine Research, 101 Vithleem Street, 1416 Nicosia, Cyprus. E-mail: maplikioti@dfmr.moa.gov.cy, andreas_drakos@hotmail.com, margyrou@dfmr.moa.gov.cy.
P. Sampaziotis, N. Papamarkos are with the Department of Electrical and Computer Engineering, Democritus University of Thrace, 67100 Xanthi, Greece. E-mail: psampaz@gmail.com, papamark@ee.duth.gr.

reported [2], with three major ones focusing on the implementation of semi-automatic species identification systems; (i) DAISY [3], (ii) SPIDA web [6], and (iii) ABIS [7]. Important drawbacks of such systems are that they are either suitable for a relatively narrow taxonomic range (e.g., SPIDA, ABIS) or unavailable for public use (e.g. DAISY, ABIS). Nevertheless, both these shortcomings could be eliminated in a community-based approach with the availability of suitable extensible platforms open for further development. Extensibility can be achieved in a dual manner: (i) at the software component level (e.g. by an Open Source modular software), and (ii) at the data level, with a flexible scheme to permit incorporation of novel data types regarding the taxonomic range accepted by the system, or data and feature types utilized in the ATI task.

In this work, we present our progress in designing and implementing such an Open Source computer system, VeSTIS. We demonstrate VeSTIS in the systematic identification of 5 species of the Class Polychaeta (Phylum Annelida), a marine macroinvertebrate group well known for the identification difficulties it presents.

2 Materials and Methods

2.1 Description and Key Features of the System

VeSTIS is intended to be a generic user-friendly platform capable of virtually identifying any taxonomic unit. It currently embeds a large number of state-of-the-art digital image analysis, enhancement and pattern recognition algorithms making it independent from the use of commercial software. Moreover, VeSTIS incorporates an SQL-based database schema and client-server technology to allow multiple users working simultaneously. The database schema was specifically designed to aid easy storage and retrieval of meta-data and to allow publishing of its contents on the Internet. This adds to the extensibility of the platform by facilitating the development of web-modules, such as a national biodiversity portal or a web-application for remotely identifying specimens through the users' browser. Finally, a very important characteristic is the ability to train VeSTIS with user-selected features in order to optimize the ATI process.

2.2 Species Selection, Sample Collection and Image Acquisition

In order to test the functionality of the system, five Polychaete species were used: *Nematonereis unicornis* (Smarda, 1861), *Marphysa bellii* (Audouin & Milne-Edwards, 1833), *Polyophthalmus pictus* (Dujardin, 1839), *Armandia polyophthalma* (Kükenthal, 1887) and *Terebellides stroemi* (Sars, 1835). These species were selected due to: (i) their high abundance in the coastal waters of Cyprus, and (ii) the relatively few problems in their identification compared to other Polychaete species.

Samples were collected with a Van Veen grab from a number of coastal sampling stations at depths of 25-35m in soft substrates. They were then sieved with a 0.5mm sieve, fixed and properly preserved. Finally, all Polychaete specimens were identified to species level with the use of stereoscopes and microscopes.

Prior to finalising the exact photo-shooting conditions, we evaluated a series of factors directly related to the quality of the shots; i.e. various magnifications, background colour, lighting source and homogeneity, specimen body parts and their orientation, as well as specimen fixation. The best results were obtained by fixing the specimens between slides against a uniformly black background. For illuminating the system we used two Leica CLS150X cold light sources with the optic fibres oriented in a way that minimized shadows. For this demonstration, we focused on the frontal body part of the animals and specifically on the head and the first 10 segments.

All images used for training and validating VeSTIS were acquired using a Leica DFC290 camera mounted on a Leica MZ7.5 stereo-microscope. Photos were taken under specimen-size dependent magnifications (in the 12.6x-32x range) with the maximum resolution supported by the camera (3.2 MP) through the Leica Application Suite (LAS) software. Image pre-processing was carried out within VeSTIS.

2.3 Image Pre-processing and Feature Extraction

Object (specimen) orientation correction: Image orientation is corrected, for the specimen to lay in a horizontal direction (Fig. 1A and 1B). This is important for object contour representation (see below).

Image segmentation and object isolation: In order to isolate the object in the image, we used the Otsu binarization method [8]. This is a segmentation process which automatically creates a black (object) and white (background) image (Fig. 1C) based on the image histogram. Using connected component analysis, VeSTIS locates and isolates the object.

Object contour representation & feature vector generation: Upper/lower object profile features are computed by recording the distance of the lower boundary of the bounding box to the furthest/closest object pixel for each image column.

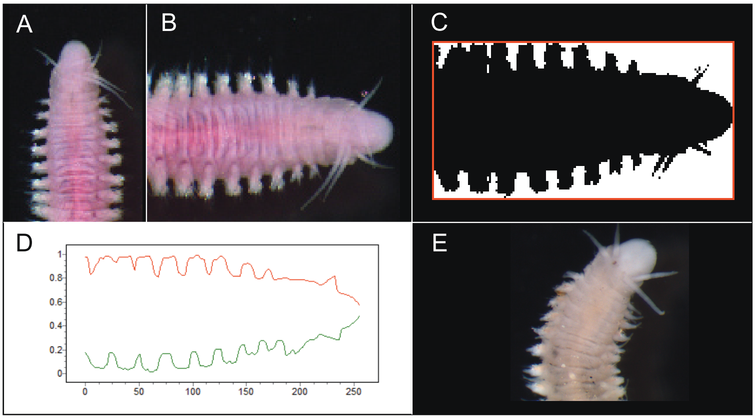

Fig. 1 – (A) Original, and (B) Corrected specimen orientation. (C) Segmentation and object isolation. (D) Contour representation. (E) An image classified as *bad* due to curvature.

All values vary between 0 and 1, since they are normalized by the height of the object (Fig. 1D). These two profiles can be considered to form a closed curve, allowing the use of Fourier descriptors [9] to mathematically describe the object's contour. Fourier descriptors allow bringing the power of Fourier theory to shape parameterisation by characterising a contour with a set of numbers that represent the frequency content of a whole shape. They are invariant to rotation, scale, and translation and are used as the input vector for the feed-forward artificial neural network (FFANN).

2.4 Generation of training and validation data sets

For generating training/validation sets, we manually classified all images based on species, specimen, orientation and condition. For four of the species in question, orientation was either dorsal or ventral. For *T. stroemi* only lateral-view photos were taken, mainly because of the species' morphology. Images were classified as *good* (G) whenever the specimen was in a good condition or *bad* (B) if the specimen was curved or moderately destroyed (Fig. 1E). We then created 3 training sets based on specimens' orientation, using only images flagged as *good*. Following a similar procedure we generated 9 different sets for evaluation purposes using both *good* & *bad* images (Tab. 1). We only included *good* images in training sets to reduce noise and test the ability of our approach to correctly classify problematic images/specimens. Multiple images were acquired for each specimen. However, a single image of each individual was included in either the training or the validation sets, in order to (i) avoid over-fitting during training, and (ii) minimize any bias on the estimation of the performance of the classifier. Thus, any pair of training-validation sets was strictly disjoint.

Training sets			Validation sets								
TS1	**TS2**	**TS3**	**VS1**	**VS2**	**VS3**	**VS4**	**VS5**	**VS6**	**VS7**	**VS8**	**VS9**
DLG	VLG	DVLG	DLG	VLG	DVLG	DLB	VLB	DVLB	DLGB	VLGB	DVLGB

Tab. 1 – Image types included in different training and validation data sets. D, V, L = Dorsal, Ventral, Lateral view; G, B = Good, Bad image classes.

3 Results and Discussion

Two FFANNs were trained for each training set in batch mode with the resilient back-propagation learning algorithm [10], each initialized with different random weights. A fully connected architecture, with a single hidden layer of 30 neurons and a sigmoid activation function, proved to be good choices after experimentation. Five output units served for classifying each specimen to the respective species using a 'winner-take-all' output encoding scheme. FFANNs were trained for 2000 epochs, and in all cases the mean squared error of desired versus predicted outcomes of the networks converged to very small values. The performance of each FFANN was evaluated with the independent validation sets (Tab. 2). We also observed the performance of a simple ensemble average of independent classifiers trained with different types of data. In several cases the performance was drastically improved (Tab. 2).

	FFANN-TS1	FFANN-TS2	FFANN-TS3	ENSEMBLE
VS1	**0.702 (0.017)**	0.667 (0.034)	0.702 (0.051)	0.738
VS2	0.693 (0.016)	**0.727 (0.000)**	0.727 (0.000)	0.705
VS3	0.698 (0.000)	0.698 (0.016)	**0.715 (0.025)**	0.709

Tab. 2 – Evaluation of identical FFANNs trained with different training sets on independent validation data sets. For each FFANN the performance reported corresponds to the average overall performance (and standard deviation) of two independently trained networks initialized with different random weights. A simple ensemble averaging approach often seems to outperform individual classifiers. Data sets are described in Tab. 1. Specimen orientation seems to be an important factor affecting classification accuracy. As expected, results obtained with *bad* images were clearly inferior (data not shown).

Species identification is a painstaking and time-consuming task, which requires highly skilled and adequately trained scientific personnel. Although the design and implementation of reliable and accurate ATI methods is a challenging problem, it will definitely give rise to more experimentation and thus to the growth and evolution of systematics. It is anticipated that Open Source solutions will boost development, applicability and usage of ATI methods similar to what has been experienced in the field of computational molecular biology.

We are currently working on adding more software components to VeSTIS (feature extractors, classifiers, etc.). We specifically plan to address the feature selection task, since classification quality is expected to depend mainly on the features rather than the classifier. This is also an attempt to cover a gap in the literature that mainly deals with the effectiveness of classifiers.

VeSTIS, although currently in alpha phase, is being actively developed. We expect to release the first beta binaries and source code at the url http://troodos.biol.ucy.ac.cy/BRL/ within late 2010.

Acknowledgement

This work was co-funded by the Republic of Cyprus and the EU European Regional Development Fund (ERDF) through a grant from the Cyprus Research Promotion Foundation (AEIFORIA/FISI/0308(BIE)/10).

References

[1] K. J. Gaston and M. A. O'Neill, "Automated Species Identification: Why Not?", *Philos. Trans. R. Soc. Lond. B. Biol. Sci.,* vol. 359, pp. 655-67, 2004.

[2] N. MacLeod, *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*. Boca Raton, FL: CRC Press, 2008.

[3] A. T. Watson, M. A. O'Neill and I. J. Kitching, "Automated Identification of Live Moths

(Macrolepidoptera) Using Digital Automated Identification System (Daisy)", *Systematics and Biodiversity,* vol. 1, pp. 287-300, 2004.

[4] J. Tanttu, J. Turunen, A. Selin and M. Ojanen, "Automatic Feature Extraction and Classification of Crossbill (*Loxia* spp.) Flight Calls", *Bioacoustics,* vol. 15, p. 251, 2006.

[5] A. Valentini, F. Pompanon and P. Taberlet, "DNA Barcoding for Ecologists", *Trends in Ecology & Evolution,* vol. 24, pp. 110-117, 2009.

[6] K. N. Russell, M. T. Do, J. C. Huff and N. I. Platnick, "Introducing Spida-Web: Wavelets, Neural Networks and Internet Accessibility in an Image-Based Automated Identification System". In: N. MacLeod (ed.), *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*, Boca Raton, FL: CRC Press, pp. 131-152, 2008.

[7] T. Arbuckle, S. Schroder, V. Steinhage and D. Wittmann, "Biodiversity Informatics in Action: Identification and Monitoring of Bee Species Using Abis". In: *15th International Symposium Informatics for Environmental Protection*, Zurich, pp. 425-430, 2001.

[8] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms", *IEEE Transactions on Systems, Man and Cybernetics,* vol. 9, pp. 62-66, 1979.

[9] O. Petkovic and J. Krapac, *Shape Description with Fourier Descriptors*, Technical Report*,* 2002.

[10] M. Riedmiller and H. Braun, "A Direct Adaptive Method for Faster Backpropagation Learning: The Rprop Algorithm". In: *IEEE International Conference on Neural Networks,* San Francisco, 1993.