Feature Extraction and Interactive Visualization to Assist Green Algae Taxonomic Classification

Vinicius R. P. Borges, Maria Cristina F. de Oliveira
University of São Paulo
Institute of Mathematics and Computer Science
São Carlos, São Paulo, Brazil
viniciusrpb@icmc.usp.br, cristina@icmc.usp.br

Thais Garcia Ferreira, Armando Augusto H. Vieira
Federal University of São Carlos
Bothanical Department
São Carlos, São Paulo, Brazil
thais.garcia.bio@gmail.com, ahvieira@ufscar.br

***Abstract*—This paper describes ongoing work on the use of visualization, mining and image analysis techniques to support biologists conducting taxonomic classification of freshwater green microalgae. The classification of such organisms is highly problematic because traditional taxonomy is inconsistent and biologists must carry out complex and meticulous procedures that demand considerable expert knowledge. We are working with biologists to define a visual exploration process characterized by user interaction with visualizations based on similarity trees, that attempt to provide a hierarchical representation of the relationships among green algae families. This requires obtaining representative feature vectors and developing automatic feature extractors from green algae images. Preliminary experiments indicate that tree-based visualizations coupled with a visual exploration strategy and an automatic extractor for computing algae morphological features can assist biologists and potentially improve the taxonomic classification process.**

***Keywords*-Visual exploration, Similarity trees, Classification, Feature extraction, Green algae**

I. Introduction

Green algae have an important role in nature and human life. These organisms are effective indicators of freshwater quality, affecting water properties such as colour, odour and taste, and interacting with chemical compounds that are potentially hazardous to human and/or animal health. They are highly sensitive to changes in their environment and therefore can signal the deterioration of ecological conditions [1].

Taxonomic classification of green algae is an important topic in Ficology. DNA-Barcode is a powerful tool for classifying algae at specie level, but its computation is not integrated into traditional taxonomical practices. In this work the interest is on a particular family of algae, the *Selenestraceae*. Classifying species from this family is particularly problematic due to inconsistencies introduced because characteristics that identify a particular species are also applicable to others. Such limitations motivate additional efforts from biologists into deriving a precise classification of the species.

An additional difficulty faced by biologists in this scenario is the time consuming tasks of observing and computing morphological features from microscopy green algae images. The manual classification procedure starts by sampling an algae culture, which has cells from an unique specie. The culture is observed on a microscope, when the taxonomists analyse some behaviors such as group formation, shape characteristics, pirenoid and presence of mucilage. A specific software is employed to obtain a digital image from which measures such as width, height and curvature from the algae shapes are computed. The classification is conducted by selecting a subset of morphological features and analysing them according to an identification key. Obtaining, interpreting and selecting the best features are hard tasks, since they require identifying and handling manually multiple types of information as well the image collection is constantly increasing.

Hence, it is interesting to investigate computational techniques to support biologists in this task. Information Visualization can contribute providing interactive visual representations that can facilitate the identification and preliminary analysis of interesting patterns and similarity relations among data. Image Analysis provides methods to describe relevant physical properties of an image and represent it in an appropriate structure for further processing.

In this paper we outline a visual exploration process to support taxonomic classification of green algae, aiming to improve classification correctness and reduce human effort, while retaining user control over the classification steps. In this process, users can interact with the visualizations to identify relevant relationships amongst images and help adjusting classifier parameters based on his domain knowledge. The visual classification system for image collections proposed by Paiva et al. [2] provides a starting point for our studies. Similarly to that work, we explore tree-based visualizations [3], [4], since hierarchical similarity relations among different species of green algae can be showed.

Furthermore, we discuss our attempts for the automatic extraction of morphological features from microscope images, as obtaining the relevant features is paramount to the success of the proposed strategy. Shape-based features are the focus since traditional green algae taxonomy considers it an important visual property for discrimining species.

This paper is organized as follows: Section II presents green algae feature extraction and the required preprocessing steps, as image filtering and segmentation. Section III briefly describes a similarity tree-based technique for visualizing image collections that can be helpful to handle green algae classification domain. Section IV outlines a process to support biologists in the task of green algae taxonomic classification based on visual exploration. Section V reports some preliminary experiments performed in visualization and automatic classification scenarios. Finally, Section VI describes some conclusions and the current stage of this work.

II. Methodology for Feature Extraction

The objective of feature extraction is to describe relevant visual properties of algae images as feature vectors, which will be the input for the automatic classifiers and for creating the visualizations. Traditional taxonomy considers shape as a relevant feature for green algae classification and the goal is to model an extractor capable of computing discriminant features. The next sections present the required steps that are important for successful computation of representative green algae morphological features.

A. Image Preprocessing

Generally, green algae images are noisy and have low contrast as a consequence of the image acquisition process due to high resolution and the imersion fluid that is inserted in microscope slide. Original RGB digital images are transformed into grayscale intensity levels, as shown in Figure 1(a), since color information is not relevant for the task. Moreover, the grayscale image is subsampled in 40% from the original dimension $(4,164 \times 3,120)$ in order to reduce execution times of further image analysis steps.

A constrast stretching is applied into grayscale image to improve contrast and correct illumination. The obtained enhanced image, shown in Figure 1(b), is used for noise removal and contour enhancement using the partial differential equation (PDE) based filter known as nonlinear annisotropic diffusion [5], given by Eq(1):

$$u_t = g|\nabla u| div \left(\frac{\nabla u}{|\nabla u|} \right) - \lambda(1-g)(u-I) \qquad (1)$$

in which $g = g(s) = \frac{1}{1+|\nabla s|^2}$ is a smooth non-increasing function, div is the divergent operator, $I = I(x,y)$ is the image to be filtered and $u = u(x,y,t)$ is a smooth version of I at time step $t > 0$, knowing that $u(x,y,0) = I(x,y)$. The parameter λ balances the smoothness during region boundaries. Eq. (1) is numerically solved using Euler-Lagrange equations associated with a gradient descent scheme. This PDE-based filter is useful for smoothing image regions with uniform intensities while preserving region contours. The resulting filtered image is presented in Figure 1(c).

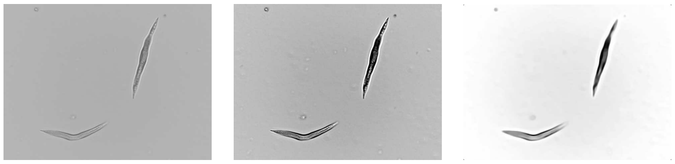

(a) (b) (c)

Figure 1. Preprocessing steps: (a) Original grayscale image; (b) Enhanced image by contrast stretching; (c) Nonlinear diffusion filtering technique [5].

B. Segmentation

The goal of segmentation is to subdivide an image in green algae regions (foreground) and background. The technique for green algae segmentation is based on threshold for image binarization and histogram analysis to compute a threshold value. Before histogram analysis step, a histogram equalization, shown in Figure 2(a), is applied to the filtered image, reducing the quantity of intensities representation to 128.

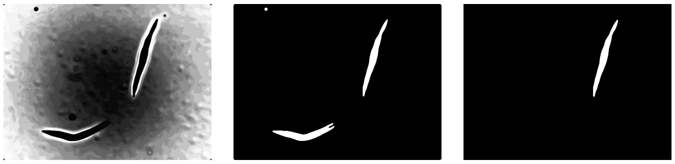
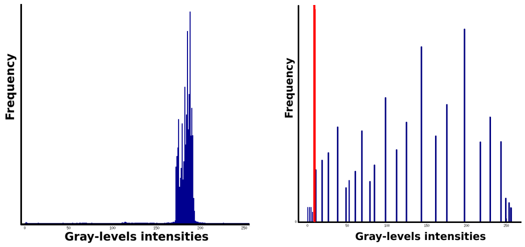

(a) (b) (c)

Figure 2. Green algae image segmentation: (a) Equalization on image from *Figure 1(c)*; (b) Segmentation carried out by threshold; (c) The final image containing one algae cell.

The histogram equalization is presented in Figure 3. Figures 3(a) and 3(b) show respectively the filtered and equalized images histograms relative to Figure 1(c), in which the last one presents a better intensities compacting than the first. In the equalized histogram, the gray-level intensities related with green algae regions are represented by the leftmost bars, since these correspond to the lower intensities in the respective image. In this sense, an optimal threshold value should be close to the leftmost group of bars in the histogram, as indicated by the red line in Figure 3(b).



(a) (b)

Figure 3. Histogram analysis procedure: (a) Filtered image histogram; (b) Equalized image histogram.

Assume that h_i is the histogram of an image I, for $i = 0,...,255$ denoting the gray-levels intensities. As aforementioned, we observed that, in most cases, green algae group intensities in equalized histogram are alternately separated by zero intensities frequencies. Thus, starting from gray-level zero, the idea is to find the first difference between consecutive non-zero frequencies intensities that is greater than two. **Algorithm 1** shows the required steps for computing threshold value t^*, in which $zeroCount$ variable counts consecutive non-zero intensities frequencies.

The equalized image is thresholded using t^*, obtaining a binary image as shown in Figure 2(b). The next step is to select one of the regions that must represent an algae cell, in

Algorithm 1: Determining an optimal threshold.

```
Input: $h$ is the histogram function and $I_{EQ}$ is
the equalized image
Output: $t^*$ threshold value
$zeroCounter \leftarrow 0$;
for $i \leftarrow 0$ to $255$ do
    if $h_i(I_{EQ}) = 0$
        then $zeroCounter \leftarrow zeroCounter + 1$;
        else $zeroCounter \leftarrow 0$;
    if $zeroCounter > 2$
        then $t^* \leftarrow i$;
            break;
end
```

which the criteria is the biggest white area in binary image. Figure 2(c) presents the final segmentation result.

C. Shape Representation and Feature Extraction

In this step, shape contour is extracted from binary segmented image using an adapted Chain Code [6] algorithm, aiming to represent contour points as a set of cartesian coordinate pairs. A discrete shape contour C is a set of points $\{p_1, ..., p_N\}$, in which $p_i = (x_i, y_i)$, for $i = 1, ..., N$. From this representation, some geometrical basic descriptors are computed [7]: *convexity*, *circularity*, *circle variance*, *retangularity*, *area* and *solidity*.

Traditional green algae taxonomy considers curvature as a relevant algae property for specie identification and it has been successfully applied for diatom recognition [8], [9], [10]. The curvature function ϕ_i at a given point $p_i = (x_i, y_i)$ of a discrete contour C can be computed according to Eq. (2):

$$\phi_i = \frac{x'_i y''_i - x''_i y'_i}{((x'_i)^2 + (y'_i)^2)^{3/2}}, \tag{2}$$

in which $x'_i = x_{i+1} - x_i$ and $y'_i = y_{i+1} - y_i$ are x and y first-order derivatives approximations, respectively. Similarly, $x''_i = x_{i+1} - 2x_i + x_{i-1}$ and $y''_i = y_{i+1} - 2y_i + y_{i-1}$ are x and y second-order derivatives approximations. However, curvature values can be corrupted by abnormal high values or noisy points [11]. Thus, a partial differential equation given by Eq. (3) is applied to curvature ϕ to smooth the contour shape without losing relevant general curvature information:

$$\beta_i^{t+1} = \beta_i^t + \Delta t((g(\beta^+)\beta^+) - (g(\beta^-)\beta^-)) \quad t = 0, ..., \tau. \tag{3}$$

in which Δt is the time step parameter, τ is the iterations number, set as $\tau = 50$, $\beta_i^0 = \phi_i$, $\beta^+ = \beta_{i+1}^t - \beta_i^t$ and $\beta^- = \beta_i^t - \beta_{i-1}^t$. To ensure scale invariance, curvature needs to be normalized by the mean absolute curvature, as Eq. (4) shows:

$$\kappa_i = \frac{1}{N} \frac{\beta_i}{\sum_{i=1}^{N} |\beta_i|} \tag{4}$$

Finally, average bending energy and maximum curvature value are computed from points curvature κ_i. Further studies will rely on shape matching using curvature points and another approaches as scale space analysis [12], [8].

A major challenge that still needs to be considered is the temporal nature of the algae morphological features, as their shapes evolve along their life cycle, an aspect that has not been considered hitherto.

III. Tree-based Visualization

We propose applying a tree-based visualization technique to generate intuitive layouts that express relevant relationships between images from different species of green algae. The technique requires computing dissimilarities from the images, which can be approximated by a distance function computed over the given feature space. As long as the feature vectors capture the relevant shape (and possibly others) characteristics for discriminating the different species, the tree-based visualization will convey the relevant information.

The Neighbor-Joining [3] is the most popular algorithm to construct similarity trees. A rootless tree is constructed, in which each graphical mark represents an image, and pirwise image distances, mapped to the edge lengths, reflect dissimilarity. The node colors denote the image classes, if available. As a result, the tree depth reflects the dissimilarity levels in the green algae image collection, in which algae from the same species should be placed in close branches.

Although the NJ tree representation avoids clutter and preserves local relationships, its computation as proposed by *Valdivia et. al* [4] demands high computational cost and creates many virtual nodes. Improved NJ-based tree algorithms have been proposed to reduce execution times and to ensure a better usage of visual space, as the *Promoting Neighbor-Joining* [2]. That technique relies on graph theory to reorganize leaf nodes and reduce the number of virtual nodes. Moreover, the original NJ algorithms ignore the temporal nature of green algae morphological features when constructing the phylogenetic tree, an aspect that needs further investigations.

IV. Visual exploration to Support Classification

Visual exploration is an interactive process that employs visualization techniques to support data analysis tasks, in scenarios where the relevant relations between data objects are assumed to be unknown. Dealing with classification tasks, visual exploration allows users to be immediately exposed to the results of a classification, including false positives, false negatives, mismatches and outliers, and also to input information into the process.

Up to now we have been studying how to perform automatic feature extraction from green algae images, since features with high capability of discriminating between species are required for accurate and effective visualization. Nevertheless, the ultimate goal is to define an interactive visual exploration framework to support the classification procedure, that allows users to gradually refine and improve the classification results. Considering that biologists are the potential users of the proposed system, their domain knowledge about taxonomy, ancestrality relations between species and relevant morphological features can render this process more accurate and effective than conventional or fully automated classification.

Paiva et al. [2] introduced a visual classification methodology of image collections with user support, and their work will provide the starting point for our studies in visual exploration. The proposed framework should contemplate the following steps, which are repeated until satisfactory results are obtained:

1) Identify and extract the relevant features;
2) Derive an initial classification and the corresponding NJ tree visualizations;
3) Using appropriate interaction techniques, biologists dinamically handle the visualization and assess the classification/visualization results;
4) Using their knowledge and previous results, biologists can update classifier settings and/or modify the feature space for classification.

V. Experiments and Preliminary Results

A. Tree-based Visualization

We generated NJ tree visualizations of a collection of 120 green algae images, composed by an uniform sampling of 4 genres: *Ankistrodesmus*, *Nephrocytium*, *Monoraphidium* and *Kirchneriella*. The feature vectors were computed with the approach described in Section II and the distance matrix that serves as input of the NJ algorithm was obtained with the Euclidean distance. Figure 4 shows the resulting tree *layout*.

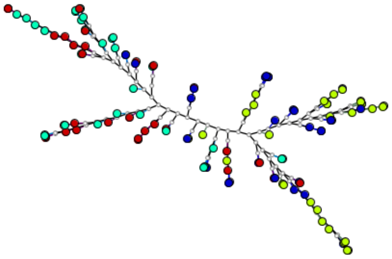


Figure 4. NJ tree obtained using the *VisPipeline* software [2].

The tree shown in Figure 4 has low and average depth branches, and poor segregation of green algae genres, leading to the conclusion that computed features are not discriminative enough. Notice that the temporal aspect of algae morphological features has not been handled.

B. Automatic Classification

We used the previous described algae image collection as input to three automatic classifiers implemented in Weka library: K-Nearest Neighbors (KNN), Naive-Bayes and Support Vector Machine (SVM). We tested KNN with values of K in the range $[3, ..., 19]$, choosing $K = 9$ as it produced the best accuracy rate. The chosen kernel for SVM classifier was the linear function. The leave-one-out cross validation strategy was adopted to evaluate the classification performance.

Table I
Classification results

Classifier	ROC Area	Pr	Rc
K-NN	0.831	0.591	0.583
Naive-Bayes	0.829	0.553	0.558
SVM	0.788	0.569	0.558

Table I shows the measures ROC Area, Precision (Pr) and Recall (Rc), and the achieved classifiers performances. It can be seen that KNN had the better classification results compared with SVM and Naive-Bayes. Considering the overall Precision and Recall results, true positives rates need to be improved, an indication that extracted features are not powerfully discriminative among algae species.

VI. Final Considerations

In this paper we introduced the problem of green algae taxonomical classification and outlined a visual exploration process to support biologists in handling this task. Preliminary studies have shown that tree-based visualizations can provide an intuitive approach to show the similarity relations between different algae species by means of a hierarchical structure. Building this structure requires handling the feature extraction problem. We are currently studying alternative strategies to obtain more discriminative features, since preliminary results on automatic classification are still not satisfactory.

Acknowledgment

The authors acknowledge the financial support of the State of São Paulo Research Funding Agency (FAPESP) and the Brazilian Federal Funding Agency (CNPq).

References

[1] P. McCormick and J. Cairns, “Algae as indicators of environmental change,” *Journal of Applied Phycology*, vol. 6, no. 5, pp. 509–526, 1994.

[2] J. G. Paiva, L. Florian, H. Pedrini, G. P. Telles, and R. Minghim, “Improved similarity trees and their application to visual data classification,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2459–2468, 2011.

[3] N. Saitou and M. Nei, “The neighbor joining method: a new method for reconstructing phylogenetic trees,” *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[4] A. M. C. Valdivia, F. V. Paulovich, R. Minghim, and G. P. Telles, “Point placement by phylogenetic trees and its application to visual analysis of document collections,” *IEEE Symposium on Visual Analytics Science and Technology*, vol. 1, pp. 99–106, 2007.

[5] C. A. Z. Barcelos and V. B. Pires, “An intelligent method for edge detection based on nonlinear diffusion,” *3rd International Conference on Artificial Intelligence in Theory and Practice*, pp. 329–338, 2008.

[6] L. da Fontoura Costa and R. M. Cesar Jr, *Shape analysis and classification: theory and practice*. CRC press, 2010.

[7] Y. Mingqiang, K. Kidiyo, and R. Joseph, “A survey of shape feature extraction techniques,” *Pattern Recognition*, vol. 1, pp. 1–38, 2008.

[8] A. C. Jalba, M. H. Wilkinson, J. B. Roerdink, M. M. Bayer, and S. Juggins, “Automatic diatom identification using contour analysis by morphological curvature scale spaces,” *Machine Vision and Applications*, vol. 16, no. 4, pp. 217–228, 2005.

[9] R. Loke, M. Bayer, D. Mann, and J. du Buf, “Diatom recognition by convex and concave contour curvature,” *OCEANS ’02 MTS/IEEE*, vol. 4, pp. 2457–2465, 2002.

[10] M. H. Wilkinson, J. B. Roerdink, S. Droop, and M. Bayer, “Diatom contour analysis using morphological curvature scale spaces,” *International Conference on Pattern Recognition*, vol. 3, pp. 652–655, 2000.

[11] G. V. Pedrosa and C. A. Z. Barcelos, “Anisotropic diffusion for effective shape corner point detection,” *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1658 – 1664, 2010.

[12] R. D. S. Torres and A. X. Falcao, “Contour salience descriptors for effective image retrieval and analysis,” *Image and Vision Computing*, vol. 25, no. 1, pp. 161–185, 2006.